Decision Support Systems 47 (2009) 408–414

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Discovering company revenue relations from news: A network approach

Zhongming Ma [a,*], Olivia R.L. Sheng [b], Gautam Pant [b]

[a] Computer Information Systems Department, California State Polytechnic University, Pomona, United States
[b] Department of Operations and Information Systems, The University of Utah, United States

ARTICLE INFO

Article history:
Received 31 December 2007
Received in revised form 6 April 2009
Accepted 8 April 2009
Available online 16 April 2009

Keywords:
Web mining
Revenue comparison
Social network analysis
Business news
Intercompany network

ABSTRACT

Large volumes of online business news provide an opportunity to explore various aspects of companies. A news story pertaining to a company often cites other companies. Using such company citations we construct an intercompany network, employ social network analysis techniques to identify a set of attributes from the network structure, and feed the attributes to machine learning methods to predict the company revenue relation (CRR) that is based on two companies' relative quantitative financial data. Hence, we seek to understand the power of network structural attributes in predicting CRRs that are not described in the news or known at the time the news was published. The network attributes produce close to 80% precision, recall, and accuracy for all 87,340 company pairs in the network. This approach is scalable and can be extended to private and foreign companies for which financial data is unavailable or hard to procure.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

Business news contains rich and current information about companies. Investment and business analysts often need to spend significant amounts of time scanning business news to compare a pair of companies (possibly competitors or partners) or to identify business relationships on the basis of revenues, sales, debts, or other financial or operating metrics. However, the huge volume of news stories makes discovering interesting information for a large number of companies nontrivial and nonscalable. Content providers like Yahoo! Finance [35] typically organize online business news by company. A news story belonging to a company often mentions several other companies. The company and any of the mentioned companies may have a relation, such as in a partnership, which is covered by the news. Alternatively, the companies may simply cooccur in the same piece of news and have no relation at all. In this paper we identify company citations from large number of news stories, construct an intercompany network from the company citations, and examine whether such a network can be used to infer some meaningful relations. To explore the suggested methodology, we experiment with a company revenue relation (CRR) between two companies. For a directed company pair (i.e., source to target), their CRR is positive if the target company's revenue measure is not lower than the source's and negative otherwise. Therefore, CRR is a binary value simply indicating which company in the pair is more "powerful" in terms of their revenues. Because revenue-based comparisons of companies are common to investment and business analysis, we choose to study this paired revenue-based measurement of CRR as an example of business relationships to test our methodology.

Using news we build the intercompany network in which each node is a company and a link between two companies indicates that a news story pertaining to one company cites/mentions the other. The intercompany network is viewed as a social network [33,28] whose structure can be quantified through graph-theoretic attributes. We employ and extend a set of graph-based measurements from social network analysis (SNA) literature, report their distributions, and measure how well CRR between two companies can be predicted by those graph-based measurements.

Our approach is based on prior findings about graph-based attributes. Literature in different domains (e.g., sociology and computer science) finds that graph-based attributes reflect certain properties of nodes in the network. For example, outdegree is a simple measure of centrality [33] and indegree represents a prestige [33] or authority measure [20]. Hence an intuition is that when company A is mentioned many times in news stories pertaining to other companies, A is likely to be powerful (e.g., high revenue). Even though we expect a lot of noise in the company citations due to meaningless cooccurrence, we hope that by deriving data from large number of news stories over a certain time and for thousands of companies, the effect of noise may be diminished. So the novelty of this research is in the use of structural attributes of networks derived from seemingly irrelevant data (company citations) to discover knowledge (i.e., CRR) given the fact

* Corresponding author.
E-mail addresses: zma@csupomona.edu (Z. Ma), olivia.sheng@business.utah.edu (O.R.L. Sheng), gautam.pant@business.utah.edu (G. Pant).

0167-9236/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2009.04.007

that news stories were not necessarily written to describe CRR (and thus our approach does not employ Natural Language Processing or NLP techniques).

In this study the news is collected from a time period before the company revenue information, which is used for determining CRRs, is available. In practice, prediction for a relationship, such as CRR, would likely be derived from previous earnings data. Forecast models (e.g., [22,2]) predict business performance measurements, such as future return on equity, but require previous financial and/or operating information as input. The performance metrics for companies can be purchased from data providers who compile data from various analysts following the companies and producing such results. So the availability of forecasts depends on resources (e.g., manpower, accurate financial and operational data), and is possibly available only for some (mostly large public) companies. In addition there may be issues of timeliness in the availability of data that can be used for predictions (i.e., data may not be available when it is needed). Our automatic approach predicts CRRs for a great number (e.g., over 6000 in this paper) of large and small companies without using any of these potentially costly resources. However, our approach is by no means a replacement for the informative earnings forecast produced by financial predictive models or analysts. Rather, it complements these traditional approaches. Moreover, since we use SNA-based graph-theoretic attributes, after constructing the intercompany network our approach is language neutral for CRR prediction and can be applied to news written in languages other than English. Hence, the approach can be used to predict CRRs among foreign companies for which reliable and timely revenue data may be hard to procure. We have validated our approach on public companies (since data is available for them), and we expect that it can be of potential value to private companies. However we could not test our approach for private companies since we do not have access to the necessary financial data that is needed for measuring CRR. Nevertheless, we realize that such data for private firms can be obtained through methods such as surveys and interviews.

Our prediction models show good predictive performance but they are less conducive to explaining the relative significance of various attributes in predicting CRR. Hence, we perform logistic regression to identify a subset of attributes (independent variables, IVs) that significantly discriminate positive and negative CRRs. Our approach is generalizable with respect to other types of business relationships, network attributes, and prediction analysis techniques. Therefore, it provides a foundation for broad applied research and decision support applications of knowledge discovery on the Web.

2. Literature review

Many researchers in areas such as organizational behavior and sociology have investigated the nature and implications of social networks created by business relationships. For example, Walker et al. [32] examine an interfirm network on the basis of cooperative relationships from a commercial directory of biotechnology firms. It demonstrates that network structure strongly influences the choices of a biotechnology startup in terms of establishing new relationships (licensing, joint venture, and R&D partnership) with other companies. Uzzi [30] investigates how social relationships and networks affect a firm's acquisition and cost of capital. Gulati and Gargiulo [17] demonstrate that an existing interorganizational network structure affects the formation of new alliances which eventually modifies the existing network. Ganley and Lampe [14] test the relationship between user reputation level and structural holes of a social network. Valck et al. [31] examine how virtual communities as social networks affect consumer decision-making processes. A major difference between those prior studies and ours is that prior works construct a social network using explicit relationships from given data sources, whereas our network links are company citations identified from various kinds of business news which does not describe CRRs and very often the company citations merely reflect the fact that those companies cooccur in the same piece of news.

Research in information retrieval and bibliometrics has employed SNA and graph-theoretic techniques on a network of documents. They consider implicit signals, such as URL links, email communications, or article citations, as links between nodes (i.e., documents). They use the resulting network of documents to study problems such as measuring the importance of individual documents (e.g., [7,20]), discovering communities on the Web (e.g., [19,16]), and measuring the impact of published articles and journals (e.g., [15]). However, they do not focus on discovering business relationships between companies.

The economic signals contained in news and identified by human readers have been well explored. Researchers have studied news of macro events, such as earnings announcements and volatility (e.g., [11,9]). In studying exchange-rate movements, Dominguez and Panthaki [10] include not only the macro announcements, but also non-scheduled news. By examining the daily response of stock prices to economic news, Pearce and Roley [26] demonstrate empirical results that support the efficient markets hypothesis. Key differences between these studies and ours are that (1) we do not manually read a large volume of news stories to label events as positive or negative, or identify any business relationships described in the news; (2) we automatically extract company citations that can represent certain business relationships or just cooccurrence in news.

After analyzing the text content of online Chinese news and extracting phrases, Newsmap [24] generates a hierarchical knowledge map as a tool for exploring business intelligence from news, where knowledge is represented as phrases. Bernstein et al. [4] apply ClearForest, a commercial text analytics system, to extract company entities from Yahoo! business news and posit that all companies are linked to each other if they appear in the same piece of news (cooccurrence approach). They construct an undirected and unweighted (binary weight) network with 315 companies and 1047 links, count how many other companies are connected with each company, rank all companies by the counts, and report that some of the 30 top-ranked companies in the computer industry are also *Fortune* 1000 companies. Their work is somewhat similar to our study, in that they use online business news to construct an intercompany network. The difference between their work and ours is that, firstly, we study a different problem (i.e., predicting CRR) and, secondly, we qualify links in the constructed network by both direction and weights. Furthermore, different from all past related research we employ various graph-based metrics to predict the CRR between any pair of companies linked in a network that contains tens of thousands of such company pairs.

3. Problem analysis

3.1. News-driven SNA-based CRR prediction

In our approach, nodes in an intercompany network consist of companies mentioned in news stories. When determining a link between two nodes, unlike traditional SNA that uses explicitly given social relationships (e.g., [21,31]), we assume a directed link from company A to company B if a news story pertaining to the company A mentions (cites) company B. Moreover, a link from company A to company B carries a weight that equals the total number of citations for company B in a set of news stories belonging to company A. The direction and weight should provide additional information about the flow and strength of business relationships in the constructed network. The weights in our network reflect the accumulated citations between a pair of companies and enable us to quantitatively identify a relationship between two companies over time. Hence, our approach is more comprehensive than prior related literature on several dimensions, including a richer network (with weights and direction), larger data sets, and various analyses related to CRR prediction.

Before we present our research questions in detail, we describe how we measure CRR, and then introduce notation for this study. Hereafter, we use the following pairs of terms interchangeably: network and graph, node and company, link and company pair or pair of companies.

3.2. Measurements for CRR

As we mentioned in the Introduction, a positive or negative revenue relation exists between a pair of companies. However, when the two companies come from different sectors, their (absolute) revenue values may not be comparable. Therefore, besides a direct comparison of revenues in dollars, we derive the following three metrics to determine a positive or negative CRR by taking the size of a sector into consideration:

- Revenue rank, or the rank of the company's revenue in its sector, namely, revenue rank$(n_i) \in [1, |\text{sector}(n_i)|]$, where revenue rank$(n_i)$ is company n_i's rank in its sector by revenue and $|\text{sector}(n_i)|$ is the total number of companies in the sector to which company n_i belongs;
- Normalized revenue rank$(n_i) = \frac{\text{revenue rank}(n_i)}{|\text{sector}(n_i)|}$; and
- Revenue share$(n_i) = \frac{\text{revenue}(n_i)}{\sum_{n_j \in \text{sector}(n_i)} \text{revenue}(n_j)}$,

where revenue(n_i) is company n_i's revenue value (in dollars).

In Section 6 we report the detailed results measured by normalized revenue ranks. The results measured by the other three metrics are similar and therefore are not included in the paper.

3.3. Network terminology

In this section, we first introduce relevant notation in directed graphs, followed by notation in directed, weighted graphs.

3.3.1. Notation in directed graphs

Fig. 1 presents a directed graph (digraph) that consists of four nodes joined by eight directed links. More formally, a digraph $G_d(N, L)$ consists of a set of nodes N and a set of links L [33], where

$$N = \{n_1, n_2, \ldots, n_m\} \text{ and}$$

$$L = \{l_1, l_2, \ldots, l_k\}, \text{ where link } l_i = \left(n_{\text{source}}, n_{\text{target}}\right).$$

The node indegree, NID(n_i), in a digraph is the number of nodes linked to n_i; the node outdegree, NOD(n_i), is the number of nodes linked from n_i [33]. Node indegree, or a metric based on it, has been used often to represent trustworthiness, authority, and prestige in many prior works (e.g., [29,6,20]). In this figure NID(n_1) and NOD(n_1) are 3 and 2.

3.3.2. Notation in weighted, directed graphs

Fig. 2 depicts a digraph in which each link carries a weight. This is a small portion of the intercompany network and it consists of four nodes/companies and 12 links. More formally, a weighted digraph G_{wd} (N, L, W) includes N, L, and W is a sequence of weights associated with the set of links, where $W = (w_1, w_2, \ldots, w_k)$.

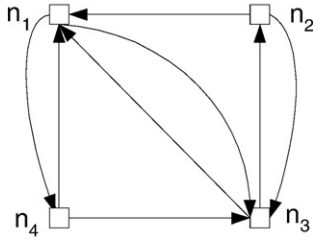

Fig. 1. Directed graph.

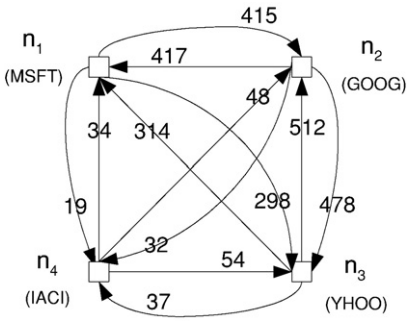

Fig. 2. Weighted, directed graph. MSFT: Microsoft Corp., GOOG: Google Inc., YHOO: Yahoo! Inc., IACI: IAC/InterActive Corp.

The degrees described in Section 3.3.1 consider only the number of neighbor nodes and ignore weights of the links. We introduce two degree-based concepts, weight of node indegree (WNID(n_i)) and outdegree (WNOD(n_i)), by accumulating the weights of neighbors that the node is linked to or from. For example, in Fig. 2 WNID(n_1) and WNOD(n_1) are 765 and 732.

Each of these degree-based attributes measure the connectivity at the node level by considering all (directly connected) neighbor nodes. Thus, we call them node degree-based attributes. However, since CRR is about just two companies, we are also interested in measurements for just one pair of nodes or dyad. For a directed dyad (n_i, n_j), we define the following equivalent dyad degree-based terms:

- Weight of dyad indegree (WDID), WDID(n_i, n_j), is the weight of the link from n_j to n_i;
- Weight of dyad outdegree (WDOD), WDOD(n_i, n_j), is the weight of the link from n_i to n_j;
- Net weight of dyad (NWD), NWD(n_i, n_j) = WDOD(n_i, n_j) − WDID(n_i, n_j).

For instance, for pair (n_3, n_2) or (YHOO, GOOG) in Fig. 2, its WDID and WDOD, and NWD are 478 and 512, and 34 respectively.

In addition to these various degree-based measurements, we also use a network analysis package, JUNG, [23] to compute scores on the basis of three different centrality/importance measuring schemas: pagerank [7], HITS [20], and betweenness centrality [5]. These schemas extend beyond immediate neighbors to compute the importance or centrality of a given node in the whole network. The pagerank algorithm computes a popularity score for each Web page on the basis of the probability that a random surfer will visit the page [7]. The HITS algorithm in JUNG [23] generates a node authority score for each node. Both HITS and pagerank compute principal eigenvectors of matrices derived from graph representations of the Web [20], so our use of them for a graph whose nodes are companies differs from their original use. As a node centrality measurement, betweenness measures the extent to which a node lies between the shortest paths of other nodes in the graph [13] and it can indicate the power of a node [6]. Finally we divide the various attributes into three groups (see Table 1) on the basis of the range of the network covered for computing the attributes.

3.4. Research questions

We want to explore the broad hypothesis that structural attributes derived from a network that is constructed from news stories can infer CRR. Therefore, we identify attributes that capture the pairwise/local relationships between companies (dyad degree-based) or estimate the global importance of each company (node degree-based and node

centrality-based). On the basis of these structural attributes, we ask the following specific research questions:

1. How well can the attributes derived purely from network structure, as shown in Table 1, predict CRRs for company pairs in the network?
2. How does CRR prediction performance differ among the three groups of attributes, which represent different amount of network covered?
3. Which of the network structure-based attributes (when combined linearly) are significant in distinguishing positive and negative CRRs?

4. Data

We now describe the source and nature of our raw data (news stories) and the process by which we constructed the intercompany network from them. To provide statistical insights into the data, we briefly report distributions of the various attributes identified in the previous section.

4.1. Raw data

The first raw data set consists of eight months (July 2005–February 2006) of business news for all companies on Yahoo! Finance [35]. We include all companies across all nine sectors in Yahoo! Finance whose annual revenue records appeared in the company statistics section in Yahoo! Finance as of early April 2006. The revenue values represent total revenues in the previous four quarters. So we predict revenue relations using news collected before the revenue records become available. In addition, we use three months' (October–December 2005) news stories from the first data set as a second data set to validate the major results we obtain from the first, but with the second data set we study CRRs on the basis of quarterly revenues.

4.2. Preliminary data processing

The news stories on Yahoo! Finance are not limited to those available from yahoo.com but also include those from other news sources, such as forbes.com, thestreet.com, and businessweek.com. In other words, URL links corresponding to news titles that have been organized under a company in Yahoo! Finance may point to Web pages located on different sites. Yahoo! Finance organizes the business news stories by company and date. Taking advantage of this organizing mechanism, we programmatically fetch news stories for each company during the eight-month period. We observe that very often Yahoo! organizes the same piece of news under different companies if the news mentions those companies (e.g., contains stock tickers for those companies); we treat such a news story as belonging to each of the companies that Yahoo! identifies for the story. For example, provided that a news article mentions companies A (identified by its ticker) twice and company B once and it is organized under each of the two companies, using just this news article we derive WDID and WDOD as 2 and 1 for the company pair (A, B).

Table 1
Three groups of attributes.

Attribute	Example	Range of network covered
Dyad degree-based	WDID, WDOD, NWD	A given node and only one directly connected node
Node degree-based	WNID, WNOD	A given node and all directly connected nodes
Node centrality-based	Pagerank, HITS, betweenness	Whole network

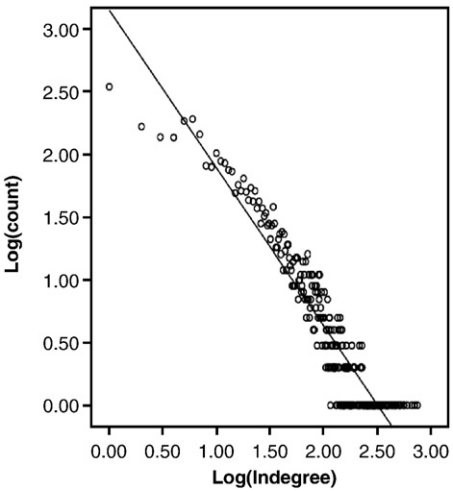

Fig. 3. Node indegree (NID) distribution.

4.3. Node and link identification

A news story identifies a company according to its stock ticker on NYSE, NASDAQ or AMEX. If a piece of news is organized under a company n_i and mentions another company n_j, we consider the news to belong to n_i and there is a directed link from n_i to n_j, denoted as (n_i, n_j). If company n_j is cited several times in the same piece of news, each citation adds to the accumulated weight for the directed link. We aggregate citation frequency across all news stories in a data set. Furthermore, we do not count self-references; therefore, we ignore citations to company n_i if they appear in a news story belonging to n_i. To illustrate, if a news story pertaining to company n_1 mentions the companies in the sequence $(n_2, n_1, n_3, n_4, n_4, n_2, n_5)$, we derive the set of links and weight sequence as $\{(n_1, n_2), (n_1, n_3), (n_1, n_4), (n_1, n_5)\}$ and (2, 1, 2, 1), respectively. We filter out news stories that do not mention any other company.

After we collected the annual revenues and news stories for all companies across all nine sectors in Yahoo! Finance, we emerged with a total of 6428 companies and 60,532 news stories for the first data set and 6246 companies and 36,781 news stories for the second data set. For the first data set, we note that the early months (i.e., July–September 2005) included fewer news stories than later months, because Yahoo! does not archive as many historical news stories as recent ones.

4.4. Attribute distributions

Several variables derived from social phenomena and networks, such as distribution of wealth and the frequency of word usage in the English language [1], follow a power law distribution. Recent research shows several aspects of digital networks such as the Internet follow power law distributions as well. For example, the rank and frequency of the outdegrees of Internet domains [12] and the indegree and outdegree of Web page links [3,8,21] reflect the power law distribution. With the directed, weighted intercompany network, we observe a similar distribution for various node degree measurements (NID, NOD, WNID, and WNOD) and link weight. For example Fig. 3 shows a log–log plot for NID indicating a power law distribution where X axis is the (log transformed) NID value and Y axis is the (log transformed) count (i.e., number of nodes having such a NID value).

5. Research methods

With Fig. 4 we introduce the specific procedures and methods we use to address our research questions. In our prediction of CRR for company pairs, we use NWD to identify the source and target and ensure each pair is selected only once: If (n_i, n_j) is identified as a pair,

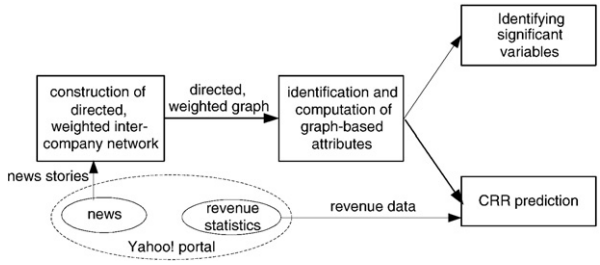

Fig. 4. Diagram of methodology and analysis approaches.

(n_j, n_i) will not be selected for CRR prediction because their CRR values are just opposite of each other. We sort all the links by their NWD values in descending order and consider only those links whose NWD values are greater than or equal to 0. For any link (n_i, n_j) in the network with a NWD value of 0, we ignore the opposite link (n_j, n_i). We identify 87,340 company pairs from the first data set and 46,725 pairs from the second one and use them to predict CRRs.

5.1. Classification methods

Using Weka [34] as a data analysis tool, we employ two classification methods to evaluate the CRR prediction performance for company pairs. For our classification methods, we select logistic regression and C4.5 [27] decision tree (i.e., J48 classifier in Weka). Logistic regression is frequently used in business research for problems with a binary class label (as for our CRR prediction problem); decision tree is one of the commonly used classifiers in data mining, because it is highly accurate for binary classification problems, does not impose assumptions about the distribution of data, and its results are well suited for human interpretation [25]. We use two different methods so that we can compare their performances for our applications. For each of the classification methods we report results on the basis of 10-fold cross-validation. In line with standard metrics used in information retrieval, we report precision and recall for positives and negatives, and accuracy to evaluate the performance of the predictive models:

$$precision = \frac{number\ of\ correctly\ predicted\ positive\ (negative)\ instances}{number\ of\ predicted\ positive\ (negative)\ instances},$$

$$recall = \frac{number\ of\ correctly\ predicted\ positive\ (negative)\ instances}{number\ of\ actual\ positive\ (negative)\ instances},$$

$$accuracy = \frac{number\ of\ correctly\ predicted\ instances}{number\ of\ instances}.$$

5.2. Identifying significant variate

The main purpose of this paper is to explore the power of structural attributes in predicting CRRs. However, we would also like to investigate the significance (if any) of individual IVs in discriminating between positive and negative CRRs, and we use logistic regression to perform this task. The linear nature in which attributes are combined in logistic regression allows for a simplistic understanding of their individual significance. In particular, from the 87,340 pairs in the first data set we randomly select 1000 pairs such that each company in the chosen pairs is distinct. As a result, there are 2000 unique companies in the 1000 pairs and hence these 1000 pairs are considered independent. The independence of each pair is required for conducting this analysis. With 12 IVs (NWD and WDOD for the pair, WNID, WNOD, pagerank, HITS and betweenness scores for source and target) and CRR as the dependent variable (DV), following procedures illustrated in [18], we employ binary logistic regression in SPSS (version 12.0) to find the significant variables. In particular, we start with a base model that uses the mean of the DVs and does not include any IVs. Then from a list of candidate IVs which have statistically significant differences between the two DV groups, we add an additional IV at each step by choosing the IV having the highest and significant score statistic (method "Forward: LR" in SPSS) until the stepwise estimation procedure stops (i.e., no remaining IV is significant).

Table 3
Classification results of CRR using dyad degree-based attributes (NWD and WDOD).

Classification method	CRR	Precision	Recall	Accuracy
Logistic regression	0	52.6%	99.2%	52.6%
	1	54.5%	1.1%	
Decision tree	0	52.6%	97.1%	52.5%
	1	49.1%	3.1%	

Notes: Attributes are NWD, WDOD, and source and target WNID, WNOD, pagerank, HITS, and betweenness.

6. Results and analyses

Using the first data set, first we report how well the various attributes derived from network structure predict CRRs for company pairs. To tease out the effects of the three different groups of attributes — dyad degree-based, node degree-based, and node centrality-based — we repeat the prediction experiment with each set of attributes separately. Using logistic regression we report what IVs are significant in distinguishing CRRs. For the second data set, we briefly report results similar to those obtained for the first data set. In particular, we provide prediction performance of CRR on the basis of Q4 2005. The class label therefore is a binary number whose values correspond to positive (1) and negative (0) CRRs.

6.1. Predicting CRR with annual revenues

6.1.1. All three groups of attributes

To predict the CRR for each pair of companies, we use a total of 12 attributes (2 dyad degree-based, 4 node degree-based, and 6 node centrality-based). For the node degree-based and node centrality-based measures, we employ a pair of attributes for the source and target companies of each link. Of the dyad degree-based attributes, we do not use WDID because it can be derived directly from NWD and WDOD. Table 2 shows the results of the two classification methods for the first data set (87,340 company pairs).

From Table 2 we observe that using attributes derived from network structure without resorting to any information about a company's financial or operational data, we achieve precision, recall, and accuracy

Table 2
Classification results of CRR with 12 attributes (first data set).

Classification method	Class label (CRR)	Number (percentage) of pairs	Precision	Recall	Accuracy
Logistic regression	0	45,907 (52.6%)	74.8%	77.1%	74.3%
	1	41,433 (47.4%)	73.7%	71.2%	
Decision tree	0	45,907 (52.6%)	80.5%	81.1%	79.7%
	1	41,433 (47.4%)	78.9%	78.2%	

Table 4
Classification results of CRR using node degree-based attributes (WNID and WNOD for both source and target).

Classification Method	CRR	Precision	Recall	Accuracy
Logistic regression	0	71.3%	84.1%	73.8%
	1	78.0%	62.4%	
Decision tree	0	80.1%	80.9%	79.4%
	1	78.6%	77.7%	

Table 5
Classification results of CRR using centrality-based attributes (pagerank, HITS, betweenness scores for both source and target).

Classification method	CRR	Precision	Recall	Accuracy
Logistic regression	0	74.6%	77.6%	74.3%
	1	74.0%	70.7%	
Decision tree	0	80.2%	80.0%	79.1%
	1	77.9%	78.1%	

of approximately 70–80% in predicting the CRR between companies, given that our data set consists of an almost equal number of positive and negative CRR instances (see the third column in the table). In addition we divide the 87,340 pairs into two subsets: (1) all pairs in which both companies in a pair belong to the same sector and (2) the remaining pairs (different sectors). We examine the prediction performance for each subset separately, and again, the precision, recall, and accuracy fall around the 70–80% range, similar to those in Table 2.

6.1.2. Each individual group of attributes

We are also interested in comparing the performances with individual groups of attributes separately; in Tables 3, 4, and 5, we provide the associated results for the first data set.

The two dyad degree-based attributes, NWD and WDOD, fail to predict revenue relations well, whereas the four node degree-based and six node centrality-based attributes produce results nearly as good as those from using all 12 attributes together.

The poor performance of dyad degree-based attributes may be due to their reliance on the local (pairwise) flow of citations between the two companies. This localized property of the dyadic attributes may fail to capture the relative importance of the two companies, which is formed by all the citations they receive from or provide to many other nodes in the network. Thus the more global node degree- and node centrality-based measures can better predict CRR.

As described in Section 2, Bernstein et al. [4] find that large degree values (in an undirected and unweighted graph) indicate large computer companies. Following their approach we convert our graph into an undirected and unweighted one, compute the degree values for all the nodes, and further derive CRR for pairs using the degree values. We consider this a baseline approach and it produces a single fixed accuracy value at 71.9%. To compare the baseline with our approach we conduct one sample *t*-test to see if the mean of the results by our method is significantly different from the result by the baseline. We ran our approach (logistic regression and decision tree with two node degree-based attributes as shown in Table 4) 20 times to produce 20 different values of accuracy and found that the average accuracy of our approach is significantly better than that of the baseline ($p<0.001$). The average accuracy for logistic regression is 73.7% and 79.7% for decision tree.

6.1.3. Significant variates

At the first step of regression with the 1000 pairs (2000 unique companies), before adding the first IV into the model, we find that ten IVs (four node degree-based and six centrality-based) are significant in score statistic (with significance equal to or less than 0.05) and the (two) dyad degree-based IVs are not. The result for dyad degree-based IVs is consistent with what we see in Table 3. In other words dyad-based IVs produce very poor prediction results. The first IV included in the regression model is source HITS score as it has the largest score statistics.

Table 6
Prediction results for 1000 pairs by logistic regression with two significant IVs.

Classification method	CRR	Precision	Recall	Accuracy
Logistic regression	0	69.4%	54.9%	66.8%
	1	64.2%	68.3%	

Table 7
Classification results of CRR with 12 attributes (second data set).

Classification method	CRR	Precision	Recall	Accuracy
Logistic regression	0	75.0%	80.1%	75.5%
	1	76.1%	70.4%	
Decision tree	0	76.4%	76.2%	75.4%
	1	74.3%	74.6%	

After including source HITS and repeating the evaluation procedures, the second IV to be added is target HITS score. At this step, all the eight IVs that were significant before including the first IV become insignificant due to a high multicollinearity among the IVs (i.e. HITS, pagerank, betweenness, NWD and WNOD). The high multicollinearity among those IVs explains the similar performance by different sets of IVs in Tables 4 and 5. The coefficient β for source HITS is negative (-1863.7) and for target HITS is positive (1627.5), which indicates that an increase in source HITS decreases the likelihood of positive CRR; and increase in target HITS increases the likelihood of positive CRR. In other words, global centrality of source or target company is indicative of its higher revenues. Hence, the global centrality-based HITS metrics for the source and the target company form the significant variates. The prediction results of the 1000 pairs using the logistic regression (with a constant and the two IVs — source HITS and target HITS) are shown in Table 6.

6.2. Predicting CRR with quarterly revenues

With the second data set we report the CRR prediction performance on the basis of quarterly revenues. We present the CRR prediction results in Table 7 and the CRRs are determined by revenues of Q4 2005. The prediction performance is very similar to those in Table 2 that are generated on the basis of annual revenues.

7. Conclusions

We propose a news-driven, SNA-based business relationship discovery approach to harvesting the predictive value of business news in discerning revenue relations between companies. Our approach uses company citations in news to understand the direction and strength of the relative importance between a pair of companies. In our intercompany network, nodes are companies, and links are directed and weighted on the basis of the direction and frequency of citations in news stories. We identify and quantify various attributes of the network using standard network analysis metrics. We then use these attributes to predict the (future) relative revenue relation between a pair of companies as an example of business relationships the approach might predict. We process and employ two sets of multi-month data from the online business news available at Yahoo! Finance. Both data sets reaffirm the robustness of our findings on the basis of annual and quarterly revenues. By applying logistic regression we are able to identify a smaller set of significant IVs. The identified significant IVs are consistent with the performance results of predictive models which indicate that the global measures of node importance are better at discriminating between positive and negative CRR. Once companies are identified our approach is intrinsically language independent and can be extended to news in various languages. Hence, it can be easily extended to private and/or foreign firms where accurate financial data is scarce. Another desirable property of our approach is that it does not use any financial or operational data for prediction of CRR.

Similar to many other networks constructed from the Internet, we find that various attributes of our network, such as NID, NOD, WNID, WNOD, and link weight, follow the power law distribution. We study the CRR prediction problem by using three groups of attributes together, as well as individual groups separately. Different groups of attributes vary in the range of the network covered for their computations. More global

measures, such as node degree- and node centrality-based attributes, are better predictors of CRR than are the dyad degree-based attributes that concentrate only on pairwise relationships and ignore the rest of the network. In terms of CRR prediction performance, the precision, recall, and accuracy are in the range of 70–80%.

We take advantage of two features of Yahoo! Finance [35] when constructing our intercompany network: (1) Yahoo! Finance organizes news by company so we did not collect news from company websites or categorize news under different companies. (2) Companies in news under Yahoo! Finance are identified by their tickers and therefore we did not apply NLP techniques to identify company entities. In case a news source does not provide tickers within news stories one could use either existing NLP tools or custom-developed programs to identify company names. For example, Bernstein et al. [4] employ a commercial tool to extract companies from news. We note that we predict a binary CRR value for a pair of companies instead of estimating their revenues in dollar values or ranking a set of (more than two) companies by revenues, which limits our approach to only provide a high level forecast.

We plan to further validate our approach with a variety of business relationships that can be based on quantitative (e.g., CRR) or qualitative (e.g., competitors) data. In a separate study we have applied our methodology to discovering competitor relationships from news and achieved interesting results. In addition it would be interesting to validate the usefulness of CRR or apply our approach to news from different languages (and countries), various types of companies (e.g., private versus public), and over time. Further research might also attempt to derive and evaluate additional graph attributes that synthesize the global and dyadic measures that potentially provide more effective predictors of business relationships between a pair of companies. Another direction is to compare our approach with other types of methods, for instance comparing our approach with earning forecast models on the basis of prediction performance.

References

[1] Adamic, L.A. Zipf, Power-laws, and Pareto — a Ranking Tutorial. http://ginger.hpl.hp. com/shl/papers/ranking/ranking.html, 2002.
[2] R.D. Banker, L.C. Chen, Predicting earnings using a model based on cost variability and cost stickiness, Accounting Review 81 (2) (2006) 285–307.
[3] A.L. Barábasi, R. Albert, H. Jeong, Scale-free characteristics of random networks the topology of the world wide web, Physica A 281 (2000) 69–77.
[4] A. Bernstein, S. Clearwater, S. Hill, F. Provost, Discovering knowledge from relational data extracted from business news, Proc. of the KDD 2002 Workshop on Multi-relational Data Mining, Edmonton, Alberta, Canada, 2002.
[5] U. Brandes, A faster algorithm for betweenness centrality, Journal of Mathematical Sociology 25 (2) (2001) 163–177.
[6] D.J. Brass, Being in the right place: a structural analysis of individual influence in an organization, Administrative Science Quarterly 29 (1984) 518–539.
[7] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Computer Networks and ISDN Systems 30 (1–7) (1998) 107–117.
[8] A.Z. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J.L. Wiener, Graph structure in the web, Proc. of the 9th World Wide Web Conference, 2000, pp. 309–320.
[9] J. Conrad, B. Cornell, W.R. Landsman, When is bad news really bad news? Journal of Finance 57 (6) (2002) 2507–2532.
[10] K. Dominguez, F. Panthaki, What defines 'news' in foreign exchange markets? Journal of International Money and Finance 25 (2006) 168–198.
[11] R.F. Engle, V.K. Ng, Measuring and testing the impact of news on volatility, Journal of Finance 48 (5) (1993) 1749–1778.
[12] M. Faloutsos, P. Faloutsos, C. Faloutsos, On power-law relationships of the Internet topology, Proc. ACM SIGCOMM, 1999, pp. 251–262.
[13] L.C. Freeman, Centrality in social networks: conceptual clarification, Social Networks 1 (1979) 215–239.
[14] D. Ganley, C. Lampe, The ties that bind: social network principles in online communities, Decision Support Systems 47 (3) (2009) 266–274.
[15] E. Garfield, Citation Indexing: Its Theory and Application in Science, Technology, and Humanities, Wiley, New York, 1979.
[16] D. Gibson, J. Kleinberg, P. Raghavan, Inferring web communities from link topology, Proc. of 9th ACM Conference on Hypertext and Hypermedia, Pittsburgh, PA, USA, 1998, pp. 225–234.
[17] R. Gulati, M. Gargiulo, Where do interorganizational networks come from? American Journal of Sociology 104 (5) (1999) 1439–1493.
[18] J.F. Hair, W.C. Black, B.J. Babin, R.E. Anderson, R.L. Tatham, Multivariate Data Analysis, 6th editionPrentice Hall, 2006.
[19] Kautz, H., Selman, B., and Shah, M. The Hidden Web. AI Magazine, 18, 2, 27–36, 1997.
[20] J. Kleinberg, Authoritative sources in a hyperlinked environment, Journal of ACM 46 (5) (1999) 604–632.
[21] R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, Trawling the web for emerging cyber-communities, Computer Networks 31 (11–16) (1999) 1481–1493.
[22] R. Lipe, The information contained in the components of earnings, Journal of Accounting Research 24 (1986) 37–64.
[23] O'Madadhain, J., Fisher, D., White, S., and Boey, Y.B. JUNG: The Java Universal Network/Graph Framework (ver. 1.7.4). http://jung.sourceforge.net, 2006.
[24] T.H. Ong, H. Chen, W.K. Sung, B. Zhu, Newsmap: a knowledge map for online news, Decision Support Systems 39 (2005) 583–597.
[25] B. Padmanabhan, Z. Zheng, S. Kimbrough, An empirical analysis of the value of complete information for eCRM models, MIS Quarterly 30 (2) (2006) 247–267.
[26] D.K. Pearce, V.V. Roley, Stock prices and economic news, Journal of Business 58 (1) (1985) 49–67.
[27] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufman, San Mateo, CA, 1993.
[28] J. Scott, Social Network Analysis: a Handbook, 2nd ed.Sage Publications, London, 2000.
[29] W. Tsai, Social capital, strategic relatedness and the formation of intraorganizational linkages, Strategic Management Journal 21 (2000) 925–939.
[30] B. Uzzi, Embeddedness in the making of financial capital: how social relations and networks benefit firms seeking financing, American Sociological Review 64 (1999) 481–505.
[31] K. Valck, G.H. Bruggen, B. Wierenga, Virtual communities: a marketing perspective, Decision Support Systems 47 (3) (2009) 185–203.
[32] G. Walker, B. Kogut, W. Shan, Social capital, structural holes and the formation of an industry network, Organization Science 8 (2) (1997) 109–125.
[33] S. Wasserman, K. Faust, Social Network Analysis: Methods and Applications, Cambridge University Press, Cambridge, UK, 1994.
[34] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed.Morgan Kaufmann, San Francisco, 2005.
[35] Yahoo, http://finance.yahoo.com.

Zhongming Ma is an Assistant Professor in the Computer Information Systems Department of California State Polytechnic University, Pomona. His main research interest is Web Mining for Knowledge Discovery and he has published in ACM Transactions on Information Systems. He received his PhD from the University of Utah in 2007.

Olivia R. Liu Sheng is Presidential Professor and Emma Eccles Jones Presidential Chair of Information Systems at the David Eccles School of Business, University of Utah. She also directs the Global Knowledge Management Center (http://gkmc.utah.edu) to seek research and education extension of data driven business optimization. Her research focuses on data mining and optimization techniques for ebusiness management, behavior targeting, personalization, recommendation, fraud/intrusion detection, bio-medical, digital government, telemedicine, telework and distributed learning applications. Her research has received funding from various Utah State agencies, Wasatch Advisors, Overstock, Optatio, U.S. Army, NSF, IBM, Tivoli, Toshiba Corp., Sun Microsystems, Hong Kong Research Grants Council, Asia Productivity Organization, SAP University Alliance, and Bureau of Land Management. She has published over 50 papers in such journals as Management Science, ACM Trans. on Information Systems, ACM Trans. on Internet Technology, Information Systems Research, INFORMS Journal on Computing, Communications of ACM, IEEE Trans. on Man, Machine and Cybernetics, IEEE Trans. on Biomedical Computing, and IEEE Trans. on Engineering Management. She is on the editorial board for various journals.

Gautam Pant is an Assistant Professor in the Department of Operations and Information Systems at the University of Utah. His current research is focused on online visibility, web mining, and business intelligence. His research has been published in IEEE Transactions on Knowledge and Data Engineering, ACM Transactions on Information Systems, and ACM Transactions on Internet Technology. He received his PhD from the University of Iowa in 2004.