Improving Transparency: Extracting, Visualizing, and Analyzing Corporate Relationships from SEC 10-K Documents

Gabriel Lucas, Michael Gebbie, Kim Norlen and John Chuang
School of Information Management and Systems
University of California at Berkeley

Abstract. We present a system to extract, visualize, and analyze inter-corporation relationships disclosed by public companies in their annual reports to the U.S. Securities and Exchange Commission (SEC). In improving the transparency of these disclosures, we allow policy makers, analysts, investors, and the general public to analyze these relationships at both the firm level and the industry level. Using probabilistic information retrieval and extraction techniques, we automatically extract a dataset of 45,000 relationships between 26,000 companies from over 15 gigabytes of SEC 10-K documents. These relationships range from ownerships, agreements, and personal connections to competition and legal disagreements. Information visualization and social network analytic techniques can then be applied to explore and analyze the dataset.

7th International Conference on Technology Policy and Innovation, Monterrey Mexico, June 2003.

1 Introduction

With the spectacular failures of multiple large companies in the past two years, there is an increasing realization that corporate transparency is critical to the robustness of a highly interconnected global economy. On the one hand, higher standards for corporate transparency have prompted firms to more fully disclose their relationships with partners, customers, and competitors. On the other hand, corporate transparency can also shed light on the extent of interdependence between firms and how a single failure can have ripple effects throughout the rest of the economy.

Enacting more stringent laws and regulations for public disclosure is a first step toward greater corporate transparency. However, regulatory filings are usually submitted in unstructured free-text, with key information embedded in legalistic jargon or technical footnotes. Armies of research analysts must be recruited to read and extract the relevant information. This process is tedious, time consuming, and prone to omission and error.

We believe that new information technologies can improve the process of transforming public disclosure as it exists today into corporate transparency. In this paper, we demonstrate the use of probabilistic information extraction techniques to automatically extract corporate relationships in 10-K documents (annual reports) filed with the SEC. The output of this process is a graph of nodes and edges that shows different types of reported relationships between firms across industries. We then apply information visualization techniques and social network analysis to this dataset to explore and analyze these relationships, at both the firm and industry levels.

Since 1993, the SEC has hosted EDGAR, an online repository for annual, quarterly and other corporate filings by public U.S. firms. As of the first quarter of 2002, there were 45,012 10-K documents submitted by 26,372 companies in the archive. This corpus constitutes approximately 15GB of text. Searching the corpus by hand for relationships would be exceedingly laborious. We estimate that it would take a person over 100 weeks to identify and categorize the relationships in this corpus. We reduce the time to extract and categorize relationships to four weeks – including all programming, data processing, and debugging – using commercial off-the-shelf hardware and software.

From 10-K documents we collect relationships that appear to describe a connection between two companies. We then use a probabilistic ranking heuristic to determine the likelihood that these relationships exist, discarding potential relationships that are less likely to be valid. At the same time, we categorize the relationships into one of five types: *Ownership*, *Agreement, People*, *Competition*, and *Legal Disagreement*. The result is a network of 44,570 relationships that has a precision of 92%. We find that 14,801 of the 26,372 companies listed in the online SEC database have at least one relationship. We develop a visualization tool to allow exploration of individual segments of the relationship network.

Applying social network analytic techniques to our dataset, we can gain insight into the nature of corporate inter-relationships at both the industry level and the firm level. Our key findings include:

- The economy is highly connected – 97% of companies with at least one relationship are connected to one another in a single component, either directly or indirectly.
- There is a highly skewed distribution of relationships reported by the companies – the top two companies each reported over 1000 relationships, the top 11 companies each reported over 100 relationships, while 90% of companies reported fewer than 10 relationships.
- Only a very small fraction of relationships (2%) are reported by both companies.

- The network of *People* relationships is the most expansive and least centralized of all individual relationship networks.
- The most active companies primarily have large numbers of *Agreements*.
- Financial companies dominate many of the top 10 ranked lists based on a variety of network analysis metrics, such as degree and centrality

This paper is structured as follows: Section 2 covers related work; Section 3 defines company sectors and relationship categories; Section 4 documents the probabilistic extraction methodology; Section 5 describes the visualization tool; and Section 6 discusses the results of our network analysis.

2 Related Work

Building upon our previous work in (Norlen et al., 2002), our system applies techniques and theory from three fields: information retrieval and extraction, visualization of large graphs, and social network analysis. At the core of our extraction heuristic is a probabilistic term weighting formula pioneered by (Robertson and Sparck-Jones, 1976) and advanced by several information retrieval systems at the Text REtrieval Conference (TREC). In the commercial sphere, Edgar Online provides similar extraction results in an online interface (via the Lycos Finance portal) that shows the *People* relationships pulled from SEC online documents.[1] The visualization applet for this dataset is similar to those from Orgnet[2] and TheyRule.[3] Comparable commercial products include Strategic Landscapes' Goldridge[4] and the Centre for Global Corporate Positioning's Alliance Mapping System.[5] Other researchers (Berkowitz et al., 1978 and Burt, 1983) apply social network analysis to corporate ownership networks. We use network analysis metrics described in Wasserman and Faust (1994) and calculate these metrics using both UCINET (Borgatti, 1999) and our own customized scripts.

3 Company Sectors, Relationship Categories

Our system automatically groups companies by standard industry sectors and relationship categories. The system collects U.S. Standard Industrial Classification (SIC) codes from 10-K documents and uses those classifications to categorize companies into nine top-level categories.[6] Table 1 lists these nine sector groupings. If a company's 10-K lists a relationship, we define that company as an "author." Table 1 shows the count of relationships authored.[7]

---

[1] http://www.edgar-online.com/lycos/quotecom/people/companypeople.asp?cik=789019

[2] http://www.orgnet.com/inetindustry.html

[3] http://www.theyrule.net/

[4] http://www.goldridge.net/

[5] http://www.cgcpmaps.com/demo.php

[6] The SIC hierarchical structure, including the top-level categories, can be found at http://www.osha.gov/cgi-bin/sic/sicser5. NAICS superceded SIC in 1997, but has not yet been adopted by the SEC.

[7] The number of relationships authored, in graph theory and social network terminology, reflects a directed network.

Table 1. Companies and relationships by SIC sector group.

SIC Sector Group	2-Digit SIC Codes	# of Companies	# of Relationships Authored
Agriculture, Forestry, and Fisheries	01 – 09	47 (0.18%)	106 (0.23%)
Mineral Industries	10 – 14	597 (2.26%)	1307 (2.87%)
Construction Industries	15 – 17	119 (0.45%)	328 (0.72%)
Manufacturing Industries	20 – 39	4311 (16.35%)	15432 (33.90%)
Communications, Transportation, and Utilities	40 – 49	1444 (5.48%)	6417 (14.10%)
Wholesale Trade	50 – 51	477 (1.81%)	1494 (3.28%)
Retail Trade	52 – 59	731 (2.77%)	2614 (5.74%)
Finance, Insurance, and Real Estate	60 – 67	5544 (21.02%)	7752 (17.03%)
Service Industries	70 – 88	2760 (10.47%)	9408 (20.67%)
Unclassifiable or Missing	9999	10342 (39.22%)	666 (1.46%)
Total	-	26372 (100.00%)	45524 (100.00%)

Table 2. Relationship category counts.

Relationship Category	Relationships
Ownership	5378 (12.1%)
Agreement	9628 (21.6%)
People	27390 (61.5%)
Competition	2047 (4.6%)
Legal Disagreement	127 (2.8%)
Total	44570 (100.0%)

The top four sectors – Manufacturing, Finance, Service, and Communications – contain 53% of all companies and 85% of all authored relationships. We note that nearly 40% of companies in the SEC database did not file any 10-K documents, or are categorized as Unclassifiable (SIC code 9999).

Our system automatically categorizes relationships into one of five types (Table 2):

- *Ownership*: Any type of ownership of stock, property, assets, or other items via merger, acquisition, purchase or other means.
- *Agreement*: Any customer, partner, legal, purchase, financial, or other agreement between two companies.
- *People*: Current or former employees or board members connecting two companies.
- *Competition*: Any kind of competitive relationship between two companies.
- *Legal Disagreement*: Current, past, or possible lawsuits involving two companies either as plaintiffs, defendants, or other parties in legal actions other than *Agreements.*

Our system places the majority of relationships into the *People* category. *Competition* and *Legal Disagreement* relationships, on the other hand, were rare and accounted for only 7.4% of all relationships. Only a very small fraction of relationships (2%) are reported by both parties.

4 Probabilistic Extraction of Relationships from Free-text

The goal of automatic relationship extraction and categorization is to find all of the relationships between publicly traded U.S. companies that file 10-Ks. Our process emphasizes precise and accurate results at the expense of recall. Therefore, although we found 44,570 relationships between 14,801 companies, our results may omit many relationships. We analyze the relationships among the 26,372 companies that have submitted 10-K documents electronically to the SEC between the years 1993 to the 1st quarter of 2002. To do so, we downloaded all of the 45,012 10-K documents available on the SEC's Edgar site. This corpus constitutes approximately 15GB of text.

Searching the corpus by hand for relationships would be exceedingly laborious. We estimate that it would take a person working 40 hours a week processing 10 documents per hour over 100 weeks to identify and categorize the relationships in our corpus. We reduce the time to extract and categorize relationships to under one month – including all programming, data processing, and debugging – using low-cost hardware and off-the-shelf software tools.

4.1 Extraction and Categorization Heuristic

Our basic approach is to locate potential relationships in documents and examine the surrounding text for evidence of their validity and type. The system first extracts 700-character "paragraphs" bordering instances of company names (excluding those of the document author), then classifies these paragraphs into pre-defined relationship categories based on category confidence scores. Finally, it reduces the number of erroneous relationships by using probabilistic confidence cutoff thresholds.

Paragraph Extraction. We use company names as keywords to find paragraphs that contain relationships. However, prior to paragraph extraction, we first prune all names in order to get accurate matches. For instance, "Capital Realty Investors IV Limited Partnership" may be too long to extract all paragraphs concerning this company, since the company is unlikely to be referenced using its full name. On the other hand, "Capital Realty" may be too short because another company name may contain this string. We shorten the company names to a length that is as short as possible without being ambiguous (in this case, "Capital Realty Investors IV").

Optimizing short company names is difficult in some cases and our dictionary contains some less than perfect short company names. For example, the company "Capital Trust" co-exists with 44 other companies that contain "capital trust," such as "Capital Trust, Inc" and "Enron Capital Trust I." In cases like these, our approach may count relationships twice. To dampen this effect, when there are two or more possible company names in the same paragraph, we keep relationships for the longer company names ("Enron Capital Trust I") and eliminate the relationships for the shorter names ("Capital Trust"). In addition to the short name ambiguity problem, our dictionary is missing variants of company names, such as "IBM" as an acronym for "International Business Machines, Inc."

Once we have a dictionary of pruned company names, the system searches all documents for matches. For each instance of a name we extract the 350 characters before and after the company name (rounded off to the nearest complete word) and write the paragraph to a database. For our dataset, we extracted 502,293 candidate paragraphs containing possible relationships.

Training. Some candidate paragraphs don't actually contain valid relationships because of ambiguous company names. Other paragraphs are not categorizable due to insufficient information in the surrounding words. To eliminate

relationships that are likely to be false positives, we probabilistically train a weighting index and regression equation to create confidence scores for each relationship.

We use two steps to train. First, we bootstrap the weighting index by manually examining documents and selecting around 100 paragraphs that represent commonly occurring categories. By indexing the words in the bootstrap paragraphs and calculating the Robertson Spark-Jones (RSJ)[8] weight for each of words, we create a weighting index for each category to judge whether or not other paragraphs belong to a category. We reuse each paragraph as both positive and negative examples of categories. For example, we use a paragraph placed in the "Competition" category as a positive example for the Competition category as well as a negative example in all other categories. To judge and categorize a new paragraph, we simply add the RSJ weights for each word in the paragraph to produce confidence scores for each category.

It is possible to rely on the bootstrap weighting index alone to evaluate all candidate paragraphs. However, there is a risk that the manually selected paragraphs used to bootstrap may not represent all the candidate paragraphs extracted from the corpus. We mitigate this risk in the second training step by randomly sampling 30 paragraphs from each category (as determined by the bootstrap index confidence scores) for manual evaluation.

As a part of the second training step, we create a new RSJ weighting index that again reuses paragraphs as both positive and negative examples. In addition to the RSJ category score for each paragraph, we aggregate paragraph scores to derive relationship scores for the count of paragraphs, the maximum weighting score, the sum of all weighting scores, and the standard deviation of weighting scores for each relationship by category. We use linear regression to combine these scores into an overall confidence score (with an average R-squared value over 0.5 across all categories).

We use the confidence scores to determine whether a relationship inside a category is good or bad. To reduce the most frequent kinds of errors (invalid company names and improper categorization), we manually evaluate a second random sample. We select our final cutoff scores so that we achieve at least 90% precision (i.e., false positives occur less than 10% of the time).

4.2 Extraction Results and Evaluation

Our extraction and categorization process yields a dataset of 44,570 relationships between 14,801 companies within 45,012 10-K documents (15GB of text). The remaining 11,571 companies are isolates that do not participate in any relationships with other companies in our database. To measure the precision of our dataset, we randomly select 30 relationships from each category (150 samples total) and weight these according to the distribution of relationships across categories. We find an overall precision of 92%. We estimate our absolute recall to be between 20% and 30%, based on a thorough examination of 10 SEC documents. Most missed relationships are between companies not included in our dictionary of 26,372 companies. We estimate recall within our dictionary to be between 45% and 55%.

---

[8] RSJ is a probabilistic term weighting formula pioneered by Robertson and Sparck-Jones (Robertson and Sparck-Jones, 1976). We also tested the widely used OKAPI-BM25 weighting formula designed to normalize document lengths and found that it performed the same as RSJ in our domain.

5 Graphical Interface

To aid the consumption of our dataset, we present a browser-based visual tool prototype. Users may search for companies, generate relationship graphs, and read original source documents.

The main elements of the tool are the search feature, the graph display panel, and the metadata panel (Figure 1). Users generate graphs by searching for specific companies and adding companies' relationship networks to the display panel. Searches can be limited to companies involved in certain types of relationships by selecting the category checkboxes in the search window. The display panel is a Java applet implementing a version of the spring-embedder graphing algorithm described in Eades (1984). This algorithm draws a graph that pushes networked nodes apart, arranging itself into a legible configuration. The metadata panel on the left lists the selected company's details and includes links to the source documents.

When a company is added to the display from the search panel, the company's "authors" (companies mentioning it in a 10-K) appear as well. This feature shows how local clusters fit into the overall network and identifies previously hidden relationships. Edges between company nodes denote relationship types by color.

Relationship paths are browsable (Gram and Cockton, 1996) and allow users to visually explore "subjects" (companies mentioned in a 10-K) within limited screen space. Users navigate networks by double clicking on nodes with thick borders to display all the subjects of the author company. Conversely, double clicking on a company node whose subjects are already showing hides that company's subjects. Figure 2 shows a company node expanded to reveal a network of People and Agreement relationships.

6 Network Analysis and Discussion

The extraction process collects relationships disclosed in annual reports. The resulting dataset is a network of relationships between companies. In this section, we discuss the results of our analysis of this network.

With the exception of the Legal Disagreement network, we find that each network has a single large group of companies connected to each other. The term *component* refers to a set of companies that are connected to each other (either directly or indirectly), but are isolated from the remaining companies in the network. Table 3 lists the sizes of the largest components for each relationship type.

In the network of all relationships, we find that 97% of non-isolate companies are related to each other, albeit distantly in many cases. We note nearly the same percentage in the *People* network. We verify that the topology of the overall network is primarily affected by the topology of the *People* network.

Despite the presence of these large components, the overall relationship network is sparsely connected, with a density of only 0.013%. We calculate density as the number of extracted relationships (44,570) divided by the $(n)(n\text{-}1)/2$ total possible number of relationships among n companies (where n = 26,372). One reason for the sparseness of our network is that over 40% of the companies in our network do not have any relationships at all. Table 4 lists the number of companies with zero, one, and more than one relationship for each type of relationship.

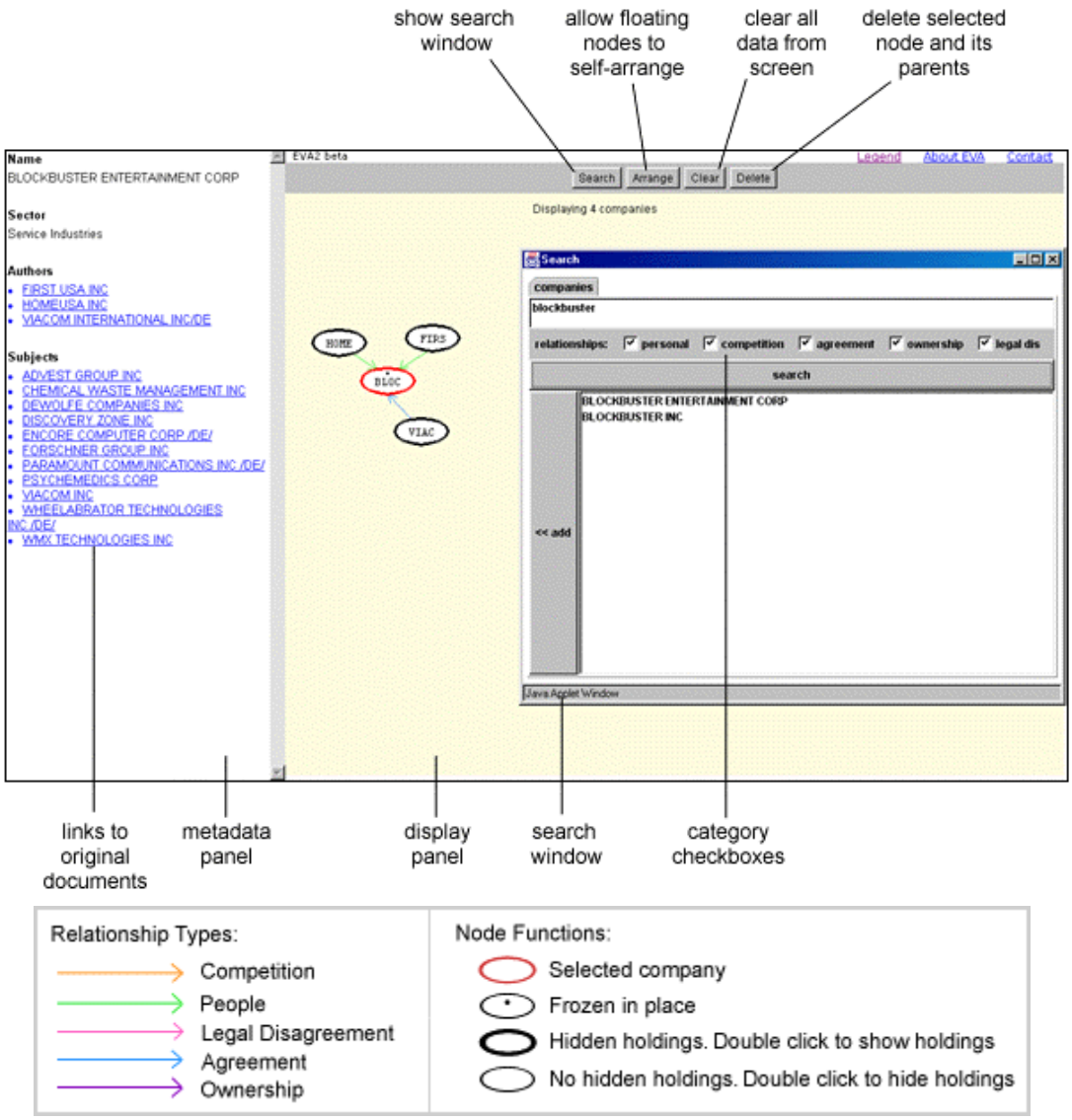

Figure 1. Visual interface and legend.

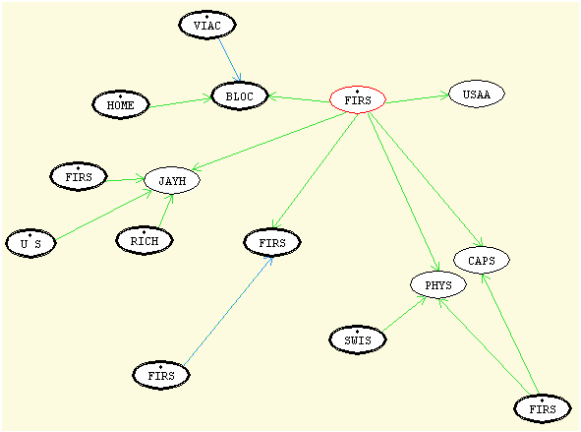

Figure 2. Sample relationship network. The node representing First USA Inc. is expanded to show relationships extracted from its annual report.

Table 3. Largest components for each relationship type.

Relationship	Number of companies in largest component	Total number of companies participating in this network	Percent of participating companies in largest component
All	14,377	14,801	97%
Ownership	2,648	5,738	46%
Agreement	5,971	7,210	83%
People	11,857	12,341	96%
Competition	1,098	2,147	51%
Legal Disagreement	8	210	4%

Table 4. Company relationship count distribution by relationship type.

Relationship	Companies		
	2 or more relationships	1 relationship	0 relationships (isolates)
All	43.48%	12.65%	43.88%
Ownership	7.99%	13.76%	78.24%
Agreement	12.99%	14.35%	72.66%
People	33.44%	13.35%	53.20%
Competition	3.15%	4.99%	91.86%
Legal Disagreement	0.11%	0.69%	99.20%

We observe that the Competition and Legal Disagreement networks have extremely high percentages of isolate companies. We conclude that these companies either do not have such relationships, or did not describe such relationships in their 10-K's. In contrast, we find a much higher percentage of companies active in the *Ownership*, *Agreement*, and *People* networks. In particular, we find that the largest network, in terms of number of companies participating, is the *People* network. Our finding suggests that the relationships people have with current and former employers may be as important as the formal ties companies form with each other.

Within these large components, we find groups of companies that are strongly connected. The term *bi-component* defines a group of companies that are so strongly connected that the removal of a single company would not disconnect the bi-component. In the *People* network, we find 8,242 companies that are strongly connected; in the *Agreement* network, we find 2,473 companies that are strongly connected. For other networks, the numbers are much smaller.

Finally, we examine the likelihood of a relationship between two companies given the presence of another type of relationship between those same companies. Overall, given any type of relationship between two companies, the likelihood that those two companies have a second relationship is approximately 5%. However, given a legal disagreement between two companies, the likelihood of an *Ownership* relationship is 33%; the likelihood of an *Agreement* is 50%; and that of a *People* relationship is 33%. This finding validates a common social network phenomenon: that disagreements rarely happen between two otherwise unrelated parties. We also find relatively higher conditional probabilities given an ownership relationship: the likelihood of a *People* relationship is 11%, and that of an *Agreement* is 15%. This observation also seems logical, as company purchases and partnerships are often initiated and cemented with personal relationships.

6.1 Company Prominence

Degree. Node degree indicates how many relationships a company is involved in. Degree can be interpreted as a proxy for a company's level of "activity" in a network. Table 5 lists average degree by relationship type.

Companies have on average slightly more than 3 relationships, but the variance is quite high. In particular, a few companies participate in a large number of relationships, while the vast majority of companies participate in two or fewer relationships. We also observe that average node degrees for *People* and *Agreement* relationships are much higher than the average node degree for *Ownerships*. From this finding, it appears that *Agreement* and *People* relationships are critical ties that bring together large groups of companies. Finally, we see lower average node degrees for *Competitive* and *Legal Disagreement* relationships. It can be difficult to determine whether companies are more eager to disclose "constructive" relationships, or whether a typical company simply doesn't have as many competitors or disagreements with other companies.

We rank companies in descending order by degree for each type of relationship and plot the resulting degree graphs on a log-log scale (Figure 3). We find that node degree approximately follows power law distribution for all but the *People* relationship type. This is consistent with findings showing power law degree distributions in natural, engineered and social networks (Barabasi and Albert, 1999, Faloutsos et al., 1999). We also observe a knee at a company rank value of 10. For rank 1 through 10, we observe that degree for all relationships closely mirrors degree for *Agreement* relationships. For ranks greater than 10, we observe that degree for all relationships parallels degree for *People* relationships. We conclude that companies most active in the network are primarily doing so through *Agreements*, but for the vast majority of companies, their overall activity level corresponds to how active they are in the *People* relationship network.

When we analyze average degree by sector (Table 6), we see that the average Communications company has the most relationships while the average Finance company has the fewest. We rank companies in descending order for each sector by their overall degrees and plot the resulting degree graphs on a log-log scale (Figure 4). We observe that all sectors have similar degree distributions, with the exception of the beginning of the Finance sector, where a few companies have large degrees that distinguish them from all other companies in the sector. These two findings suggest that the Finance sector consists of two types of companies: financial institutions that are extremely active in forming relationships with other companies, and those with much smaller relationship networks.

Table 5. Node degree by relationship type.

Relationship	Node Degree	
	Average	Standard Deviation
All	3.213	12.550
Ownership	0.408	1.709
Agreement	0.730	9.899
People	2.077	4.312
Competition	0.155	0.796
Legal Disagreement	0.010	0.122

Table 6. Node degree by sector.

Sector	Average Node Degree
Agriculture	3.872
Mineral	3.389
Construction	4.050
Manufacturing	5.929
Communications	6.737
Wholesale trade	5.205
Retail trade	5.761
Finance	3.059
Service	5.206
Unclassified	3.213

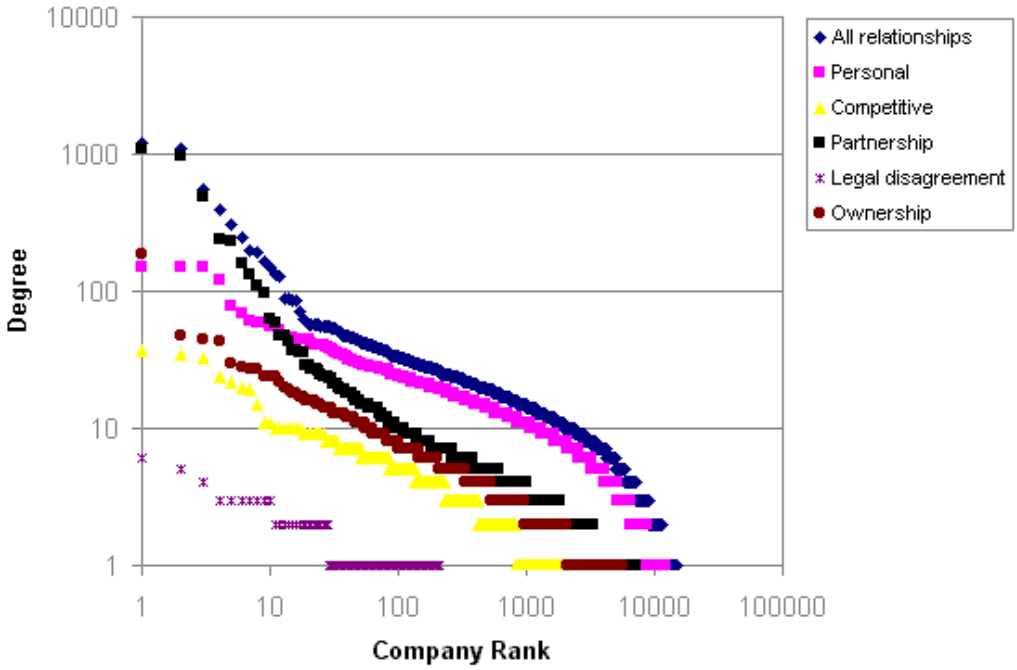

Figure 3. Degree distribution by relationship type

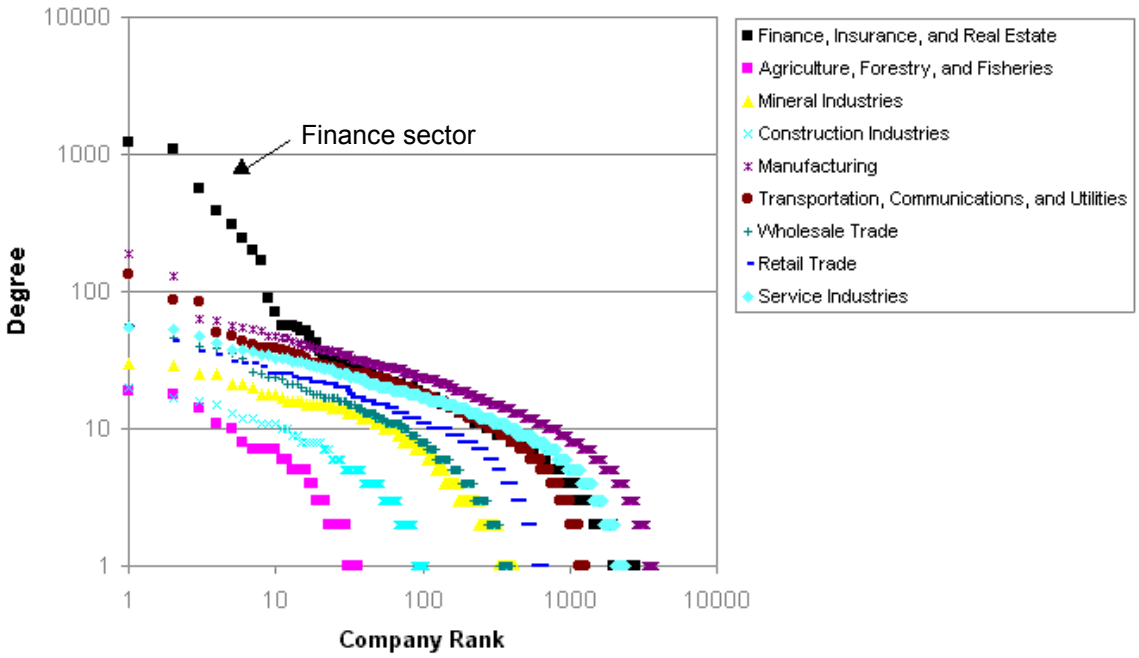

Figure 4. Degree distribution by sector.

Cutpoints. Companies that join otherwise disconnected portions of the relationship network are important to the stability of the network. These companies are called *cutpoints*, because if they were removed from the network, they would "cut" the network into two or more components. We analyze the companies whose removal from the largest component of each network would create the most disconnected, non-isolate components, and list the top ten in Table 7. We find that they are all part of either the *Ownership* or *Agreement* networks. Given the small size of the *Legal Disagreement* and competitive networks, we would not expect any companies from these networks to make the list. However, since no company from the *People* network made the list, we conclude that this network is more stable than the *Ownership* and *Agreement* networks.

When we analyze which sectors account for the most cutpoints for each relationship network, we find that 50% of all Finance companies in the largest component of the competition network are cutpoints. Furthermore, we find that 51% of all Retail Trade companies in the largest component of the *Ownership* network are cutpoints. For the *People* and *Agreement* networks, we find that the distribution of cutpoints is relatively even across all sectors. However, we do observe that in the *Agreement* network, the removal of Enron in particular would create more disconnected non-isolate components (4) than would any other company in the wholesale trade sector.

Betweenness. *Betweenness* measures how often a node appears on the shortest path between all pairs of nodes in the network. Companies with high Betweenness values are important because they are key links lying between many pairs of companies. We examine which companies are most often between other pairs of companies and present the top ten companies in Table 7. We observe that a few companies are between a high percentage of companies for the *Ownership*, *Agreement*, and *Competition* networks. However, like the cutpoint metric, we again observe that no companies from the *People* network made the top ten list. We conclude that the *People* network is less centralized and

more spread out than the other networks. In the *People* network, no companies have positioned themselves as the central players in the network. Given the constant movement of employees and directors among companies, we are not surprised by this result.

Standout Companies. We calculate which companies have the highest degree, Betweenness, and cutpoint impact (Table 7). We observe that the Finance industry contains: 8 of 10 companies with highest degree; the top three "Between" companies, and all ten companies with the greatest cutpoint impact. We conclude that companies from the Finance industry are central companies in our network. Other industries with central companies include Manufacturing, Communications, and Service.

Table 7. Top 10 companies based on degree, Freeman Betweenness, and impact as a cutpoint

(a) Node Degree		*(b) Freeman Betweenness*		**(c) Impact as a cutpoint**	
Company (Sector)	**Node Degree**	***Company (Sector)***	***Betweenness* (Relationship Type)***	***Company (Sector)***	***Number of components** (Relationship Type)***
Chase Manhattan Bank (Finance)	*1215*	Capital Trust* *** (Finance)	57.32% (Ownership)	*Bank Of America (Finance)*	73 (Agreement)
Bank Of America (Finance)	*1097*	Chase Manhattan Bank (Finance)	45.53% (Agreement)	*Chase Manhattan Bank (Finance)*	64 (Agreement)
Chemical Bank (Finance)	*559*	Bank Of America (Finance)	35.76% (Agreement)	*Capital Trust*** *(Finance)*	39 (Ownership)
Citicorp (Finance)	*391*	AT&T Corp (Communications)	30.00% (Competition)	*Chemical Bank (Finance)*	18 (Agreement)
Bank Of Boston Corp (Finance)	*306*	Lucent Technologies (Communications)	20.03% (Competition)	*Bank Of Boston Corp (Finance)*	17 (Agreement)
Capital Trust (Finance)*	*245*	CompUSA Inc (Retail)	19.77% (Competition)	*Bank Of America (Finance)*	10 (Ownership)
Credit Suisse First (Finance)	*201*	AT&T Corp (Communications)	16.96% (Ownership)	*Citicorp (Finance)*	10 (Agreement)
Johnson & Johnson (Manufacturing)	*191*	Microsoft Corp (Service)	16.44% (Competition)	*Credit Suisse First (Finance)*	9 (Ownership)
Bank Of America NTSA (Finance)	*166*	Broadvision Inc (Service)	15.82% (Competition)	*Chase Manhattan Bank (Finance)*	9 (Ownership)
First Interstate Bank (Finance)	*150*	Fresenius Medical Care Hldg Inc (Manufacturing)	14.17% (Competition)	*Credit Suisse First (Finance)*	9 (Agreement)

* Within all largest components of each relationship network, percentage of shortest paths between two companies that include this company.

** Number of non-isolate components that would be created by removing this company from the network.

*** As stated earlier, the names of several companies contained such common words that our dataset may contain false positive relationships for these companies. Capital Trust is one such company; the results reported here for Capital Trust may not be completely accurate. Also, it is possible that the network measurements for Capital Trust and other companies with name difficulties are slightly overstated.

7 Conclusion

We present a system to extract, visualize, and analyze inter-corporation relationships disclosed by public companies in their annual 10-K reports to the SEC. In improving the transparency of these disclosures, we allow policy makers, analysts, investors, and the general public to analyze these relationships at both the firm and industry levels. Using probabilistic information retrieval and extraction techniques, we extract 45,000 relationships between 26,000 companies from over 15 gigabytes of SEC 10-K documents. We estimate that a manual extraction and categorization process would take 25 times longer than the four weeks our system required for programming, training, and processing. Our categorization and evaluation algorithm has an overall precision of 92 percent. While our dictionary of company names

contains some ambiguous company names, the network we have constructed is robust, given its nonrandom characteristics. We also present a visual interface for users to explore the relationships dataset.

Applying social network analysis techniques to our extracted dataset, we gain insight into the nature of corporate inter-relationships. Among our key findings are: (1) 97% of companies involved in at least one relationship are connected to each other, either directly or indirectly; (2) there is a highly skewed distribution of relationships reported by the companies – the top two companies each reported over 1000 relationships, the top 11 companies each reported over 100 relationships, while 90% of companies reported fewer than 10 relationships; (3) only a very small fraction of relationships (2%) are reported by both companies; (4) the network of *People* relationships is the most expansive and least centralized of all individual relationship networks; (5) the most active companies in the network primarily have large numbers of *Agreements*; and (6) Finance industry companies dominate many of the top 10 ranked lists of central companies. We believe this dataset can be used to answer other research questions of interest to regulators, investors, or academics.

In light of recent high profile corporate collapses in the energy and telecommunications industries, companies are now motivated to meet the demands of regulators, investors, and the general public in providing full disclosure of all aspects of their business activities and interests. Against this backdrop, we believe regulatory agencies such as the SEC should adopt a corporate reporting mechanism built upon a standardized schema like XML, so that future extractions are both easier and more complete. Such a schema would dramatically improve the efficiency and accuracy of a system like ours and bring greater transparency of corporate inter-relationships to public discourse.

Acknowledgment

This work is supported by the U. S. National Science Foundation under Cooperative Agreement Number ITR-0085879.

References

Barabasi, A. and Albert, R. 1999. Emergence of scaling in random networks. *Science* 286:509-512.

Berkowitz, S.D., P.J. Carrington, Y. Kotowitz, and L. Waverman. 1979. The determination of enterprise groupings through combined ownership and directorship ties. *Social Networks* 1: 391-413

Borgatti, S.P., M.G. Everett, and L.C. Freeman. 1999. UCINET 5.0 Version 1.00. Analytic Technologies.

Burt, R.S. 1983. *Corporate Profits and Cooptation: networks of market constraints and directorate ties in the American economy*. New York: Academic Press.

Eades, P. 1984. A heuristic for graph drawing. *Congressus Numerantium* 42, pp. 149 – 160.

Faloutsos, M., P. Faloutsos, and C. Faloutsos. 1999. On Power Law Relationships of the Internet Topology. *Proceedings of ACM SIGCOMM'99.*

Gram, C. and G. Cockton. 1996. *Design Principles for Interactive Software*. New York: Chapman & Hall.

Norlen, K., M. Gebbie, G. Lucas, and J. Chuang. EVA: Extraction, Visualization and Analysis of the Telecommunications and Media Ownership Network. Proceedings of International Telecommunications Society 14th Biennial Conference (ITS2002), Seoul Korea, August 2002.

Robertson, S. E. and K. Sparck-Jones. 1976. Relevance weighting of search terms. *Journal of the American Society for Information Science.* 27:129-146.

U.S. Securities and Exchange Commission. 2002. EDGAR Form Pick Search. Available from <http://www.sec.gov/edgar/searchedgar/formpick.htm>. [March 20, 2002].

Wasserman, S. and K. Faust. 1994. *Social Network Analysis.* New York: Cambridge University Press.