

EVA: Extraction, Visualization and Analysis of the Telecommunications and Media Ownership Network

Kim Norlen, Gabriel Lucas, Michael Gebbie, and John Chuang¹
School of Information Management and Systems, University of California Berkeley

ABSTRACT

We present EVA, a prototype system for Extracting, Visualizing, and Analyzing corporate ownership information as a social network. Our extraction methodology uses probabilistic information retrieval techniques to gather relationships from heterogeneous sources of online text. We store these relationships in a database that can reflect industry changes over time. The browser-based visualization tool allows users to query the database and explore large networks of companies. We demonstrate this system with data from the telecommunications and media industries, and our analysis identifies influential companies and power law distributions in the network. We believe this system can aid government regulators, policy researchers, and the general public to interpret complex corporate ownership structures.

1 Introduction

Ownership is a fundamental element of analysis in economics and public policy. An ownership relationship may indicate the flow of capital, information, and influence between two firms. It may also have broader implications for industrial organization, competition, and antitrust. In an era of industry convergence, ownership relationships among companies have become so complex that they resemble directed social networks rather than simple hierarchies. Yet, tracking and analyzing ownership networks is a prerequisite to informed public debate on proposed mergers or government regulation. While federal regulations dictate full ownership disclosure for public firms, such data are often decentralized and unstructured, making systematic documentation and analysis of ownership very difficult. Researchers must often sift through large volumes of free-text, a process that is time-consuming, tedious, and non-scalable.

¹ The authors are listed in reverse alphabetical order. We thank Hal Varian, Michael Buckland, and Clifford Lynch for helpful discussions on this work. Supported by the National Science Foundation through ITR grant ANI-0085879.

The EVA project has three objectives. The first is to efficiently gather and consolidate from multiple, heterogeneous sources large amounts of ownership data about the telecommunications and media industries. We define ownership as one company's possession of equity in another company. The second is to allow the public to explore this information through a simple, intuitive interface.² The third is to analyze the data as a social network, so that at a macro level we can understand the topology of the network and at a micro level identify the prominent companies in the network.

Although EVA could be applied to any industry, we concentrate on the telecommunications and media industries for three reasons. First, these industries represent more than 3% of the GDP (U.S. Bureau of Economic Analysis, 2002). Second, companies in these industries control information content and delivery, publishing, broadcasting, and global networking—all essential ingredients to public speech. Third, these industries are in a state of flux. In particular, the Telecommunications Act of 1996 opened up possibilities for new constellations of ownership by lifting regulatory barriers between media and telecommunications companies. The ability to track these changes will remain an important element of telecommunications policy for years to come (Compaine and Gomery, 2000).

Our network contains 6,726 relationships among 7,253 companies; an additional 1,090 companies have no relationships. Relationships were valid at least some point between January 1, 1998, and December 31, 2001. However, because we have not yet searched for sales and dispositions, we caution that some relationships may no longer be valid. While the data set is by no means complete, one can observe interesting trends. We find that two metrics—number of companies clustered together (component size), and number of companies to which a company is connected (company degree)—are characterized by power law distributions. We find that over half of the companies in our database are connected to each other, and that 234 companies are so strongly connected that the removal of any one of those 234 would not disconnect the other

² The EVA visualization tool is located at <<http://denali.berkeley.edu/eva/>>.

233. Finally, we find that ten companies are the owners in over 24% of all relationships that EVA has gathered.

This paper has five additional sections. Section 2 gives an overview of the EVA system. Section 3 describes the process for extracting and storing ownership relationships. Section 4 explains the visualization interface. Section 5 contains our network analysis and discusses its significance. Finally, Section 6 offers concluding thoughts and suggestions for improving not only EVA, but also the way ownership data are published and tracked.

2 System Overview

EVA comprises three components: an extraction engine, a database, and a visualization interface. Figure 1 shows how these components interact:

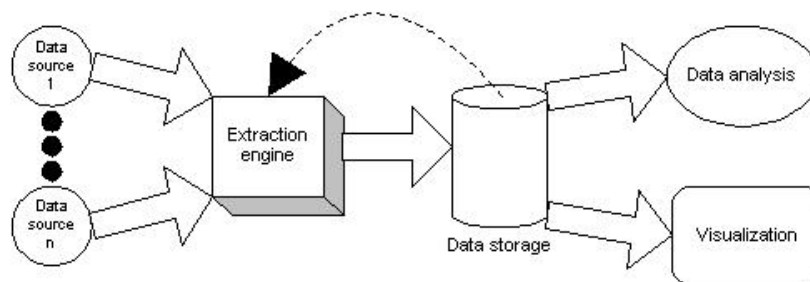


Figure 1. EVA data flow diagram.

The extraction engine is both a primary and secondary research tool for gathering ownership data. As a primary research tool, the extraction engine identifies ownership data buried within lengthy free-text documents; as a secondary research tool, it gathers ownership data summarized in documents published by organizations that did their own prior research. The extraction engine can search any number of heterogeneous data sources. If the data from a source are well-formatted, EVA can gather such data automatically. On the other hand, if the data are not well-formatted, EVA either probabilistically ranks likely ownership relationships for human review, or else offers an interface for humans to manually enter the data. The visual interface displays subsets of the network and lets users explore different paths among companies. The interface is browser-based and connects live to our centralized database over the Internet, thereby

guaranteeing that users will always see the most recently compiled data. The database stores the information as a directed network, enabling calculation of overall connectedness and identification of prominent companies. Finally, the data analysis primarily uses UCINET (Borgatti et al., 1999), a well-respected network analysis software package. However, we have written some custom scripts for additional metrics not supported by UCINET.

Our current sources include three online document collections:

- *Columbia Journalism Review's Who Owns What* (Moore, 2001)
- *The Industry Standard's online Deal Tracker Database* (Industry Standard, 2001)
- Public company 10-K annual reports (U.S. Securities and Exchange Commission, 2001)

Columbia Journalism Review (CJR) and the Industry Standard are both secondary sources because the ownership data contained in these documents were gathered, compiled, and verified by other researchers. The 10-K annual reports (10-Ks) are primary sources, for EVA had to gather, compile, and verify the ownership data from these documents. Figure 2a shows the number of companies for which each source found at least one relationship, and Figure 2b shows the number of relationships each source found.

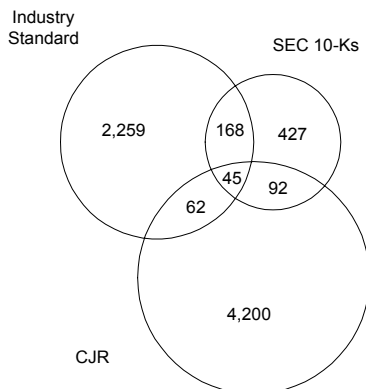


Figure 2a. Number of companies for which each source found at least one relationship (Total = 7,253)

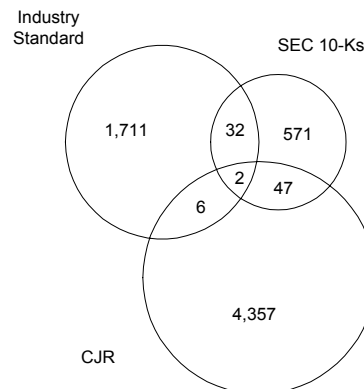


Figure 2b. Number of relationships found by each source (Total = 6,726)

Because our analysis depends on the quality of our sources and on the extraction process, we note some limitations of our data set:

- Omissions are possible due to inherent limitations of information extraction techniques

- Omissions are possible due to a lack of coverage in source documents
- Companies do not necessarily disclose all holdings, or if they do, that information is sometimes buried in confusing documents.³

The lack of overlap in relationships found among our three sources suggests that we are only seeing part of the whole picture, and that coverage may be further improved with additional data sources. For example, Thomson Financial and Mergerstat are two leading sources of mergers and acquisitions data, commercially compiled by teams of full-time research staff. Discussions with one vendor reveal a data set for the telecommunications and media industries that contains slightly fewer relationships than our data set.

3 Probabilistic Extraction of Relationships from Free-text

As we discussed in Section 2, EVA can gather data from secondary sources. However, this method is dependent on time-consuming and expensive primary research of other individuals. This method may also require manual entry if the data are not well-formatted. Thus, we designed EVA to identify and present to human reviewers likely relationships from primary sources, such as free-text 10-K documents.

The main goal of the extraction engine is to minimize the human effort required to conduct primary research. Below is a breakdown of how long it took to design and run the free-text extraction engine in EVA:

- 160 hours to code and test
- 20 hours to download and process documents (3,374 10-K documents, 1.5GB of text)
- 30 hours to manually train the system and evaluate the top 3,249 relationship paragraphs

Since the extraction code can be reused, EVA would require significantly less time to process additional sources of data.⁴

³ For example, Tyco spent about \$8 billion in its past three fiscal years on more than 700 acquisitions that were never announced to the public (Maremont et al., 2002).

⁴ We conducted a rough experiment to approximate how much time EVA can save researchers looking for acquisition data in 10-K documents. We timed ourselves reading sampled 10-K documents and manually recorded each acquisition that we found. We calculated that we would have needed about 293 hours to process the entire 1.5GB of text. This is about six times more than the 50 hours needed by the EVA extraction engine to process an additional 1.5GB of data.

3.1 Extraction Heuristic

EVA uses keywords to find relevant parts (“paragraphs”) of documents and then probabilistically ranks those paragraphs on the likelihood that they include valid acquisition data. The steps of the heuristic are:⁵

1. Start with a seed list of company names
2. Extend the list of company names using both probabilistic extraction and manual review
3. Use keywords (like “acquisition” and “merger”) to identify candidate paragraphs containing at least one company name
4. Use simple noise filter rules (like deleting duplicates) to eliminate paragraphs not likely to be useful
5. Rank the remaining paragraphs using a probabilistically trained weighting index and regression-based weighting formula
6. Present the highest ranked paragraphs to humans who:
 - a. Eliminate bad relationships
 - b. Identify relationships that are missed by the extraction engine

3.1.1 Probabilistic Weighting

At the core of our heuristic is a probabilistic term weighting formula pioneered by Robertson Sparck-Jones (Robertson and Sparck-Jones, 1976) and advanced by several information retrieval experimental systems at the Text REtrieval Conference (TREC, <http://trec.nist.gov/>).⁶ To probabilistically train the weighting index, we manually rate randomly sampled paragraphs as “good” or “bad.” Good paragraphs contain valid acquisition data between a parent and child company. We use SQL to convert rated paragraphs into a weighting table in a database. By summing the word weights for each paragraph, we compute a confidence score that ranks the likelihood that future paragraphs contain valid relationships.

Using linear regression, we find that paragraph word weights alone explain 45.7% of the variance in the paragraphs. We improve this result by adding two more variables to the formula: the probabilistic weight of the words in the sentence containing the keyword and the proximity of the keyword to the acquired company’s name. Using linear regression a second time, we add together the paragraph weights, sentence weights, and proximity measures after multiplying them

⁵ A detailed account of the extraction heuristic is in our technical report (<http://denali.berkeley.edu/eva/tech-report>).

⁶ We also tested the widely used OKAPI-BM25 weighting formula designed to normalize document lengths. It performed slightly worse than RSJ in terms of precision and recall because all our paragraphs are the same length (600 characters).

by their regression coefficients.⁷ The combined weighting formula explains 52.4% of the variance and is used to produce a ranking confidence value for each paragraph.

3.1.2 Probabilistic Weighting Performance: Precision and Recall

The EVA extraction engine is primarily designed to save time, but we also measure precision and recall.⁸ Examining the 35% of paragraphs with the highest probabilistic rankings, we find that the extraction module has a precision of 55.4% and a recall of 50.0%. This performance compares favorably with the performance of DARPA-sponsored Message Understanding Conferences (MUC) systems,⁹ where good performance in less complex event extraction domains translated to precision and recall measurements between 50% and 70% (Grishman, 1997). We present two examples of false positives that underscore the difficulty in automatic extraction of acquisition events that result in changes in equity holdings.¹⁰ In many cases, the language in 10-K documents is so ambiguous that even humans are confused.

The first example comes from a 10-K filed by Aether Systems and contains an acquisition that does not meet our definition of equity ownership. At one point, text in the document appears to describe Aether's acquisition of Motient:

... In connection with the acquisitions of Cerulean, Sinope, RTS and Motient, the Company [Aether Systems] has accrued \$29,800 as of December 31, 2000 for the remaining portion of the purchase price. Such amount has been allocated to the fair value of the assets purchased and the liabilities assumed... (SEC, 2001a)

Subsequent text, however, clarifies that Aether has simply acquired one of Motient's business units, rather than any equity in Motient itself:

... On November 30, 2000, we [Aether Systems] acquired Motient's retail transportation business unit for \$49.2 million in cash... (SEC, 2001a)

⁷ We use a linear regression method similar to the logistic regression technique by Cooper, Gey and Dabney (1992) to estimate weighting formula coefficients.

⁸ Precision is the number of good records returned from a search, divided by the number of records returned from a search. Thus, if four of ten records returned by a search engine were valid, then precision would be 40% (4 / 10). Recall is the number of good records returned from a search, divided by the total number of good records in all documents searched. Thus, if there are eight total good records, but the search engine only returned four, then recall would be 50% (4 / 8). For these calculations, we need to know the number of good relationships that EVA finds for a given set of documents, as well as the total number of good relationships in those documents. The first metric simply requires that we track the number of good acquisitions found during our reviews. The second metric is more difficult; without reading all documents thoroughly, we do not know how many acquisitions are contained in all the documents. However, while reviewing paragraphs we search for additional acquisition data to manually enter, so as an approximation we use the total number of acquisitions found during the review phase.

⁹ MUC evaluations helped researchers set standards to evaluate the performance of tasks such as named entity recognition (NER), entity attribute recognition, entity relationship fact-finding, and entity event finding. Soderland (1999) compares several successful information extraction systems that MUC participants have created.

¹⁰ Freitag points out additional difficulties when extracting information within the "acquisition" domain (Freitag, 1998).

Companies routinely report acquisitions of assets in their 10-K filings, but by definition these events do not give rise to equity ownerships. However, these acquisitions are sometimes financed by the equity of the acquirer, so a transfer of equity ownership can actually occur in the reverse direction. In this second example, the excerpt from the 10-K filing of Nextel Communications is tagged as containing an acquisition event that results in Nextel becoming an equity owner of Motorola:

... we [Nextel] acquired all of Motorola's 800 MHz SMR licenses in the continental United States in exchange for 41.7 million shares of class A common stock and 17.8 million shares of nonvoting class B common stock ... (SEC, 2001b)

However, human review reveals that Nextel acquired some licenses from Motorola using its own stock, and therefore Motorola emerges as an equity owner of Nextel as a result of this transaction.

To help improve the automatic extraction process we have two proposals. First, we propose an XML schema in Appendix 1 for companies to label ownership relationships and transactions in their 10-Ks. Such a schema would greatly reduce the problem of ambiguity and generally lead to a more transparent understanding of major acquisition events.

A second way to improve precision and recall is to use category filters. As the examples indicate, we want the extraction engine to exclude acquisition events such as asset acquisitions, proposed, pending and future acquisitions, and acquisitions of warrants. We create category filters that probabilistically identify the type of acquisition contained in each paragraph. Paragraphs with a high likelihood of being in a category other than equity ownership are eliminated and excluded from manual review. Preliminary experiments on our data set show that, given a fixed number of paragraphs, category filters can increase precision up to 6% (Figure 3). Although category filters do not eliminate the need for human review, they do allow reviewers to examine fewer paragraphs yet still receive a higher yield.

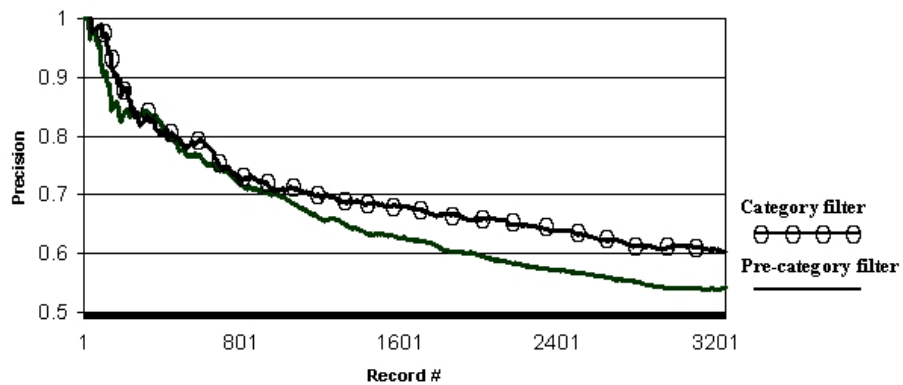


Figure 3. Automatic extraction precision by top-ranked paragraph.

3.1.3 Manual Evaluation

The final step in the free-text extraction process is to manually review the top 35% ranked paragraphs. Our system allows people to evaluate as many paragraphs as they have time for. Ranking the paragraphs allows people to focus first on the paragraphs that are most likely to be useful. Our interface allows reviewers to quickly accept, reject, reverse and add relationships. Altogether, manual evaluation process takes approximately 22 seconds per paragraph, or 1.6 minutes per acquisition.

4 Information Visualization

Understanding networks can be difficult without a visual explanation. Graphs have long been the primary method of representing social networks (Brandes et al., 2001). Because EVA treats corporate ownership as a social network, it is logical to expect a graphical component as an attachment to our work. Graphical representation reveals a macroscopic view of an industry, plus sub-structures that would otherwise remain hidden. Our goal is to present this information without overwhelming the user or cluttering the display. Tufte (1990) outlines principles for displaying information visually; we attempt to follow these principles whenever possible.

We present a browser-based prototype display tool for visual exploration of the EVA database. Users may search for companies, generate ownership graphs on the fly, and read original

information sources. Given the appropriate data, the tool can also visualize changes to ownership networks over time.

4.1 Related Visualization Work

Visualizing ownership networks is an interdisciplinary endeavor drawing from the fields of social network analysis, business intelligence, media criticism, and information design. Specific works related to the EVA visualization tool therefore include graphics and software from several sources. Rosenwein (2000) contains a good example of a network graphic explaining media mergers described in a news article. Krackplot (Krackhardt et al., 1994) and Graphviz (Gansner et al., 1993) are examples of software for generating images of social networks from given data sets. Some applets available on the Internet, such as those from Orgnet (<http://www.orgnet.com/inetindustry.html>) and They Rule (<http://theyrule.orgo.org/>), are similar to the EVA display in spirit, if not in content.

Fewer visualizations deal specifically with corporate ownership relationships. Those we are aware of tend to be directed toward the business intelligence market. Strategic Landscapes, an online tool from Goldridge¹¹ (<http://www.goldridge.net/>), generates textual reports and graphical maps of companies related by many factors, including ownership. The Centre for Global Corporate Positioning (<http://www.cgcpmaps.com/demo.php>) offers a similar service.

4.2 Graphical Interface

The interface prototype (Figure 4) provides an interactive way to explore relationships in the EVA database. Major elements include the search feature, the graph display panel, and the metadata panel. Users generate graphs by searching for specific companies and adding the ownership networks of those companies to the display panel. The display panel is a Java applet implementing a version of the spring embedder graphing algorithm described in Eades (1984). Changing the panel's date range alters the displayed graph to reflect a different ownership

¹¹ Goldridge is now owned by Vizigence, Ltd.

network for the new time period. The metadata panel on the left lists details about the selected company and includes links to source documents substantiating its ownership relationships.

Figure 5 shows the graph's legend with definitions for colors, arrows, node sizes, and borders.

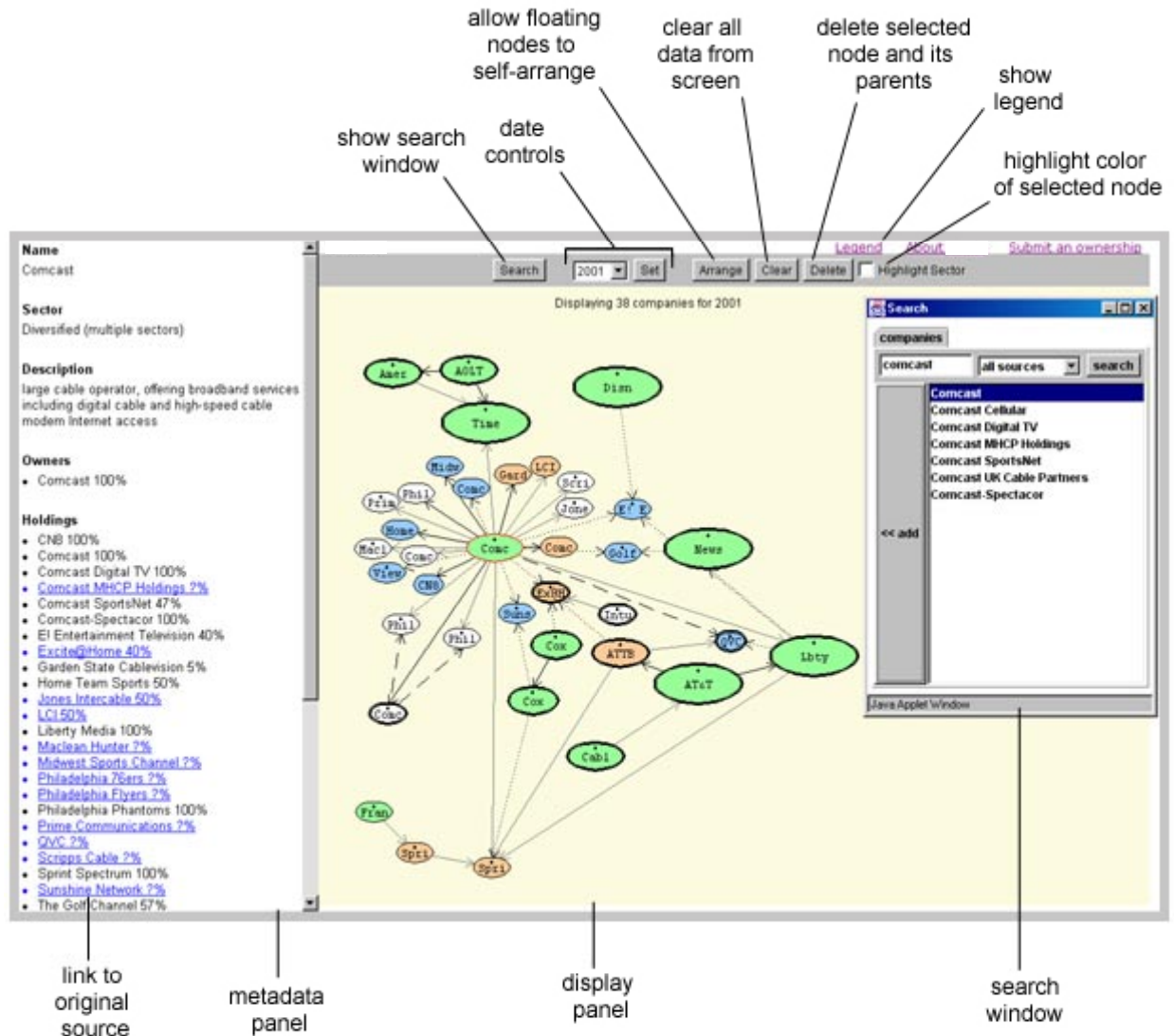


Figure 4. EVA display.

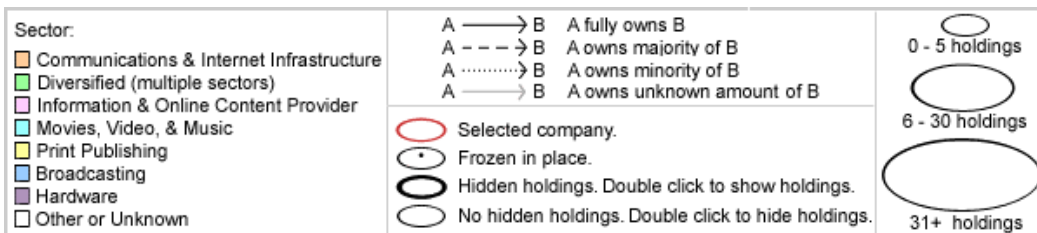


Figure 5. Legend for EVA display.

For every company on the screen, all parents of that company are also always on the screen. This principle helps users to quickly identify top-level companies, see how local clusters fit into the overall network, and identify previously hidden relationships. As an example, a recent user of our interface selected *Sunset Magazine*, and to his surprise found that AOL was an indirect owner (Figure 6). However, this discovery provided the user with an explanation for why he found an AOL trial disk enclosed between the pages of the latest issue of *Sunset Magazine*.¹²

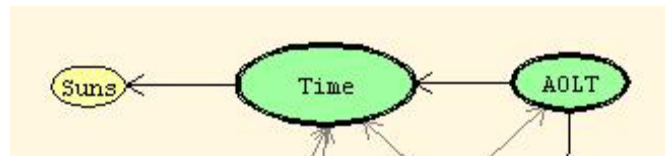


Figure 6. All parents are automatically added to the display along with the selected company. Here, *Sunset Magazine* is shown with its parents Time Warner and AOL-Time Warner.

To allow visual exploration of subsidiaries in limited screen space, ownership paths are browsable (Gram and Cockton, 1996). Users navigate networks by double clicking on nodes with thick borders. This action displays all the holdings (children) of that company. Conversely, double clicking on a company whose children are already showing hides that company's children. As with the persistent display of parents, browsing children can also lead to the unintended discovery of pathways between companies. Figure 7 shows how a user can find a path between Bertelsmann and AT&T simply by displaying the children of AOL-Time Warner.

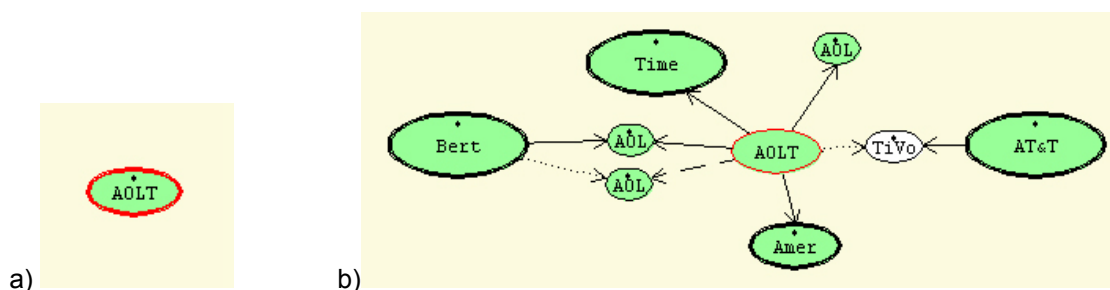


Figure 7. a) Parent companies appear automatically when a company displays in the EVA interface. Here, AOL-Time Warner has been added to the screen. b) Double clicking the node AOL-Time Warner reveals its direct subsidiaries their other owners.

Future work on the EVA display would include several improvements. First, the browsability principle should be extended to the search feature by adding an alphabetical index of company

¹² Thanks to Lincoln Cushing for this example.

names, so finding them is easy even if an exact spelling is unknown. Second, the ability to compare two display panels with different date ranges would simplify the task of comparing changes to ownership networks over time. Finally, the graph layout algorithm could be altered to use the vertical or horizontal dimensions of the display to denote metrics such as prestige or degree (see Section 5), as suggested in Brandes et al. (2001).

5 Network Analysis

In this section, we apply several network analysis metrics to our data set to illustrate the potential this technique has for revealing influential telecommunications and media companies. We use the standard network analysis software package UCINET together with custom PERL scripts to make these calculations.

To summarize our findings:

- Two metrics follow power law distributions: the number of relationships each company has (node degree), and the sizes of connected clusters (component size)
- The largest cluster contains 53.6% of companies
- Ten companies are the parents for over 24% of all relationships
- 87% of companies are involved in at least one ownership relationship. However, only 10% of companies are involved in more than one ownership relationship
- The greatest outdegree (number of children) for a company is 552; the greatest indegree (number of parents) for a company is six
- Removing a random relationship is likely to increase the number of components in our network, while removing a random company is not

Table 1 (see next page) is a summary of the companies with the largest values for eleven network analysis measurements that we performed. We refer to this table throughout this section.

(a) Cutpoints (total components)		(b) Cutpoints (total non-isolates)		(c) Outdegree		(d) Indegree		(e) Overall degree / Ego network size		(f) Betweenness ¹³	
Company	Components created	Company	Non-isolates created	Company	Outdegree	Company	Indegree	Company	Overall degree	Company	Normalized Betweenness
Clear Channel Communications	550	Time Warner	16	Clear Channel Communications	552	United Video Satellite Group	6	Clear Channel Communications	552	Liberty Media	18.35%
Liberty Group Publishing	288	Viacom	15	Liberty Group Publishing	288	Sprint Spectrum	5	Liberty Group Publishing	288	Time Warner	10.12%
CNHI	209	News Corp	12	CNHI	209	Excite@Home	5	CNHI	209	Clear Channel Communications	8.87%
News Corp	164	Microsoft	9	News Corp	177	Open Market	5	News Corp	178	AT&T	7.18%
CBS Radio	146	Bertelsmann	8	CBS Radio	147	Go2Net	5	CBS Radio	148	News Corp	6.76%
Lee Enterprises	146	Advance Publications	7	Lee Enterprises	146	OmniSky	5	Lee Enterprises	146	Emmis Communications	6.25%
Gannett	134	Cox Enterprises	7	Gannett	134	Thirteen other companies have an indegree of 4	4	Gannett	134	Viacom	5.68%
Disney	125	Hollinger	7	Disney	130			Disney	130	WALC-FM	5.55%
PRIMEDIA	124	Liberty Media	7	PRIMEDIA	125			PRIMEDIA	127	Disney	3.62%
Time Warner	100	PRIMEDIA	7	Time Warner	110			Time Warner	114	Open Market	3.42%

(h) Cliques		(i) Ego networks with the largest number of relationships between alters				(j) Ego networks that can reach the largest number of non-isolate companies within two hops		(k) Ego networks with the highest density		
Company	Cliques	Company	Number of relationships between alters	Size of ego network	Density of ego network	Company	Percent of non-isolates within two hops of ego network	Company	Density	Size of ego network
Liberty Media	23	Liberty Media	23	85	0.32%	Liberty Media	10.28%	CommTouch	100%	2
AT&T	12	Comcast	12	50	0.49%	WALC-FM	8.27%			
Comcast	8	AT&T	8	30	0.92%	Clear Channel Communications	7.69%			
UnitedGlobalCom	6	Ticketmaster-Online CitySearch	7	10	7.78%	Radio One	7.65%	39 companies have an ego network density of 50%	50%	2
United Video Satellite Group	5	USA Networks	6	29	0.74%	American Tower	7.64%			
Go2net	5	UnitedGlobalCom	6	25	1%	Hispanic Broadcasting	7.64%	TCI Ventures	33.33%	3
Six companies are involved in 4 cliques	4	Ticketmaster	6	19	1.75%	SFX Entertainment	7.63%	NetStream	33.33%	3
		Go2net	5	19	11.90%	547 other companies can reach 7.61% non-isolates within two hops	7.61%	MuchMusic	33.33%	3
		United Video Satellite Group	5	7	1.46%			United Pan-Europe Com.	33.33%	3
		USA Information and Services	5	7	11.90%			Wabash	20%	5

Table 1. Network Analysis: Metric summaries

¹³ We ignore relationship directionality for calculating Betweenness, a common practice according to Wasserman and Faust (1994). When directionality is preserved, Liberty Media still ranks highest and appears in 0.01% of all geodesics.

5.1 Related Network Analysis Work

Network analysis has been applied to many different fields, including: engineered systems such as the Internet (Faloutsos et al., 1999), the WWW (Broder et al., 2000, Kleinberg and Lawrence, 2001), and electric power grids; biological systems such as the neural network of the *Caenorhabditis elegans* (Watts and Strogatz, 1998); and social networks such as movie actor collaboration (Albert and Barabasi, 2002) and terrorist networks (Picarelli, 1998). Social network analysis has also been applied to corporate ownership networks (Berkowitz et al., 1978, Burt, 1983). For example, Stark and Vedres (2000) and Vedres (2000) recently investigate changes to the ownership networks of Hungarian companies during the 1990's and show how these networks dissolved after a period of industrial privatization.

We base our network analysis methods on the principles outlined in Wasserman and Faust (1994), particularly those used to determine network prominence. We use UCINET 5 (Borgatti et al., 1999) to calculate various prominence measurements, including degree and Freeman Betweenness. We also use it to identify cliques, cutpoints, and bridges, as well as to determine the number of components in the network.

5.2 Network Topology

5.2.1 Component Distribution

A component is a maximal connected sub-graph, or, a group of nodes connected only to each other. The largest component in our data set contains 4,475 companies, or 53.6% of the entire network, while the next largest component contains only 3.5% of the network. We plot component sizes in descending rank order (Figure 8) and observe that component size follows a power law distribution with a slope $\alpha = -0.56$ ($r^2 = 0.87$). In other words, the i -th largest component has a size of $c \cdot i^\alpha$, where c is a constant. The fit improves further if we consider the top ten components ($\alpha = -2.23$, $r^2 = 0.96$) and the remaining components ($\alpha = -0.48$, $r^2 = 0.89$) separately.

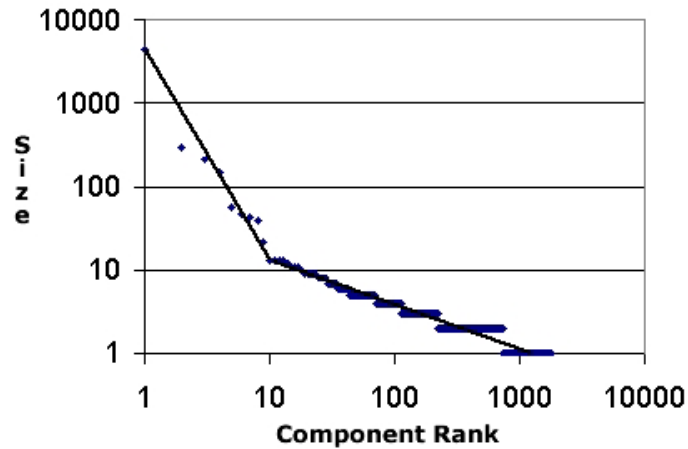


Figure 8. Population distribution of all components, rank-ordered by size.

5.2.2 Density

Density is defined as the number of relationships in the network, divided by the maximum possible number of relationships among all companies in the network. Given a network of size N , the maximum number of relationships is $(N)(N-1)/2$. A dense network would indicate overall strong connectivity among the companies in our network, while a sparse network would indicate overall weak connectivity. The maximum possible number of relationships among the 8,343 companies in our database is 34,798,653. However, we have identified only 6,726 relationships, or fewer than 0.02% of the maximum possible. Thus, we conclude that the network is sparsely connected. However, as Wasserman and Faust (1994) points out, a large network with a low density measurement can still have prominent actors. Indeed, we find several companies whose prominence measurements are significantly higher than those for most other companies in the network (see Section 5.5). Table 2 lists different types of nodes. The population distribution of these node types explains why the network density is so low: only 10% of companies participate in more than one relationship.

Ownership role	Role	Subtype	Number of children	Number of parents	Population distribution
Parent only	Transmitter	Root	1	0	8%
		Star	> 1	0	3%
Child only	Receiver	Leaf	0	1	69%
		Magnet	0	> 1	2%
Parent and child	Carrier	Link	1	1	1%
		Hub	>1	1	4%
			1	>1	
None	Isolate		0	0	13%

Table 2. Population distribution by ownership role

5.2.3 Depth and Diameter

Depth is defined as the longest shortest path (longest geodesic) between two nodes, considering the direction of each relationship. Diameter is the longest geodesic, regardless of directionality. The depth of the EVA network is 12 relationships long, and the diameter is 23; both occur in the largest component. If the largest component were excluded, the depth would be three and the diameter four. Thus, the largest component is four times “deeper” and over five times “wider” than the next highest and widest components.

Among all companies in the largest component, Comcast has the shortest radius and the greatest depth. That is, ignoring relationship directionality, every other company has at least one longer shortest path to the edge of the component than does Comcast. Yet, preserving relationship directionality, no other company has a longer shortest path to the edge of the component than does Comcast. Thus, Comcast is in the center of the component when relationships are treated as bidirectional links, yet at its top when relationship directionality is preserved.

5.2.4 Source Comparison

We compare the topologies of the ownership networks that arise from each of the three independent data sources to the combination of all three. We find significant differences in the number of components and component sizes across the data sources (Table 3). The CJR data contains fewer but larger components. On the other hand, the Industry Standard and SEC 10-K sources contain more components, only a few of which are large. The combination of all three sources connects many otherwise disconnected components because each data source contains

relationships that the other two lack. As a result, the size of the largest component is even greater than the sum of the largest components of each individual source.

Source	Total components	Components of size ≥ 100	Size of largest component
<i>CJR</i>	18	7	2,563
<i>Industry Standard</i>	789	1	354
SEC 10-Ks	127	1	203
All three	723	4	4,475

Table 3. Population and component distribution of network based on data source

5.3 Sensitivity Analysis

Our sensitivity analysis examines the likelihood that a removal of a random relationship or company would increase the number of components in the network and thereby change the overall network topology.

5.3.1 Bridges

A bridge is a relationship that, when removed, increases the number of components in the network. Relationships that are bridges are important because they would disconnect two or more companies from each other and alter the overall topology of the network. 89% of the relationships in our data set are bridges. This number is high because so many companies participate in only one relationship (see Table 2). Relationships that connect to these companies are bridges since their removal would isolate those companies from the rest of the network. Another 2% of relationships are bridges whose removal would disconnect a component into two components, each of which contained two or more companies. These bridges are not always obvious because both of the companies connected by the bridge are involved in at least one other relationship.

5.3.2 Cutpoints

Similar to a bridge, a cutpoint is a node that, when removed, increases the number of components in the network. A company that is a cutpoint is important because it is the only company connecting two or more otherwise disconnected companies. We found 742 total cutpoints; however, only 273 of these would leave behind two or more non-isolate components.

Table 1(a) lists the cutpoints whose removal would most increase the number of components, and Table 1(b) lists the cutpoints whose removal would most increase the number of non-isolate components.

5.4 Prominence

Identifying which nodes are the most important, or prominent, in a network is a common task in social network analysis. A company ownership network presents an interesting twist since certain well-known companies are often assumed to be prominent merely because of their reputation or name recognition. However, the most prominent companies may turn out to be lesser-known but well-positioned companies in the network. We examine two measurements of prominence: degree and betweenness.

5.4.1 Degree

Degree measures the number of ownership relationships in which a company is engaged, either as a parent (outdegree), as a child (indegree), or both (overall degree). In social network analysis, outdegree indicates expansiveness, indegree indicates popularity, and overall degree indicates activity (Wasserman and Faust, 1994). These characterizations are likewise valid for our network; companies with high outdegrees may have expanded their operations, companies with high indegrees may have attracted much attention from other companies, and companies with high overall degrees are likely very involved in the network. Figure 9 shows plots of the three degree distributions in descending rank order. Like the plot for component size (Figure 8), these plots resemble power law functions. We use linear regression to compute $\alpha = -0.96$ ($r^2 = 0.94$) for outdegree, $\alpha = -0.13$ ($r^2 = 0.51$) for indegree, and $\alpha = -0.89$ ($r^2 = 0.97$) for overall degree. From these calculations, we conclude that indegree does not fit very well to a power law function. A power law function for degree means that $D = c \cdot i^\alpha$. That is, the i -th largest company degree = $c \cdot i^\alpha$, where c is a constant. A power law distribution for company degree is consistent with findings showing power law degree distributions in other naturally occurring, social, and engineered networks as reported in (Barabasi and Albert, 1999, Faloutsos et al., 1999).

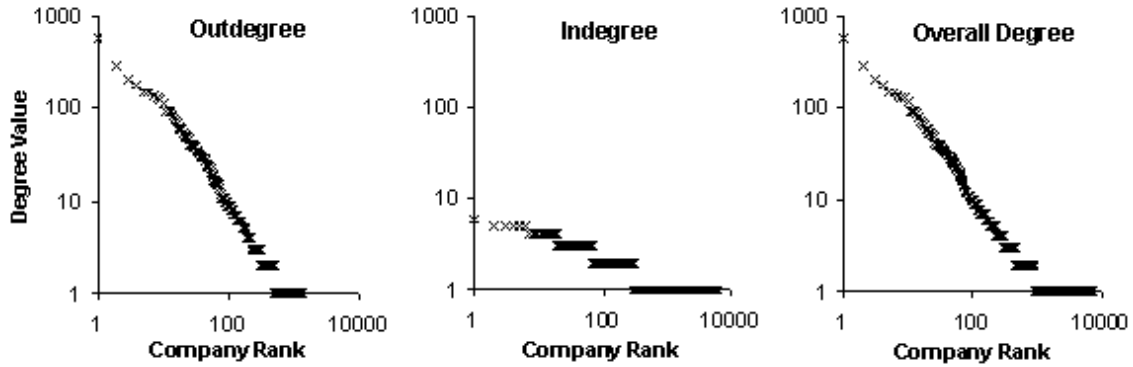


Figure 9. Distribution of node degrees

Only 16% of companies have an outdegree of one or more, while 76% have an indegree of one or more. In other words, most companies are not owners themselves but rather owned by another company. Table 1(c-e) indicates the companies with the highest outdegrees, indegrees, and overall degrees. Of the ten companies with the greatest outdegrees, six are not owned by any other company. Furthermore, 24% of all relationships involve one of these ten companies as a parent. Thus, ownership is concentrated among a few companies that are located primarily at the top of the network. However, the converse does not hold true for indegree; most companies with high indegrees are not at the bottom of the network. Rather, 17 of the top 20 companies are also owners themselves. Finally, because the maximum outdegree is 552 yet the maximum indegree is just six, overall degree is nearly identical to outdegree in most cases. In other words, if a company is active in the network, it is most likely doing so as a parent, rather than as a child. This finding is consistent with the similar values for α as computed above in the linear regressions.

Table 4 provides summary statistics for the three degree measurements.

	OUTDEGREE			INDEGREE			OVERALL DEGREE		
	Mean	Median	Std Dev	Mean	Median	Std Dev	Mean	Median	Std Dev
All companies	0.93	0	9.67	0.93	1	0.46	1.85	1	9.65
Companies with outdegree > 0	4.97	1	21.94	0.43	1	0.79	5.41	2	21.99
Companies with indegree > 0	0.42	0	4.53	1.06	1	0.32	1.48	1	4.58

Table 4. Degree summary statistics

5.4.2 Freeman Betweenness

Betweenness measures how often a node appears in the shortest path between all other node pairs, regardless of relationship directionality. Nodes with high betweenness scores are like hubs in an airport system, linking together more distant outliers. Companies with a high betweenness are important because they are well-positioned between many other pairs of companies in the network. In our network, only 841 companies (10% of the network) have a betweenness greater than zero, meaning that all other companies do not lie on the shortest path between even one pair of companies. Of those 841 companies, several are between many pairs of companies. In particular, Liberty Media lies on over 18% of all shortest paths. In other words, when two companies are indirectly connected through one or more intermediate companies, 18% of the time Liberty Media is one of those intermediates. Table 1(f) lists companies with the top betweenness measurements.

5.5 Analysis of Important Subsets of the Network

5.5.1 Bi-Components

A bi-component is a group of nodes that could not become disconnected by the removal of just one node. In a bi-component there is at the very least a circle path that loops through all companies; often, though, there are additional relationships within the bi-component. In our network, bi-components are significant because they indicate strong connections among groups of companies. We found 28 bi-components containing three or more companies. The largest bi-component contains 234 companies and includes AOL-Time Warner, AT&T, Bertelsmann, British Telecom, CBS, Cisco, Comcast, Deutsche Telecom, Disney, Intel, MCI WorldCom, Microsoft, NBC, Sony, and Yahoo!. Its density is 1.4%, which is 72 times greater than the density of the overall network. Notably, many of the companies in this bi-component are competitors with certain other companies in the bi-component.

5.5.2 Cliques

A clique is a group of nodes that all have a direct relationship to each other. In our case, cliques are significant because they indicate the densest concentrations of ownership within a network that is otherwise quite sparse. We found 136 companies participating in 75 ownership cliques. Only one of these cliques has four members; all others have three. The companies involved in the most cliques are listed in Table 1(h).

5.5.3 Ego Networks

Every node has an ego network, consisting of itself (the ego) and all nodes to which it is immediately connected (its alters). Note that the number of alters equals the sum of a node's indegree and outdegree. Thus, the overall degree values in Table 1(e) also quantify the sizes of the largest ego networks. An ego network is significant because it indicates the size and composition of a company's close circle of other companies. In the EVA network, most ego networks are small, a fact that is not surprising given that most companies are only involved in one relationship. Furthermore, most ego networks are not dense. Only 130 ego networks contain additional relationships between alters, while all other ego networks contain relationships only between the ego and its alters. In other words, inbreeding is uncommon: new ownership relationships between two companies rarely lead to new ownership relationships between one of those two companies and the other's alters.

Table 1(i) lists the companies whose ego networks contain the largest number of relationships between alters. For these ego networks, we have also provided the size and density of those networks. Liberty Media, for example, has an ego network with 85 alters and 23 relationships among those alters. Table 1(j) lists the companies whose ego networks with the greatest "reachability," which measures what percent of non-isolate companies can be reached within two hops of any company in the ego network. For example, over ten percent of non-isolate companies can be reached within two hops of Liberty Media's ego network. Finally, Table 1(k)

lists the companies with the densest ego networks. The density of CommTouch's ego network is 100%, for example, although the ego network itself is quite small, with only two alters.

6 Conclusion

EVA is a prototype research tool for extracting, visualizing, and analyzing corporate ownership relationships. Applying EVA to the telecommunications and media industries, we find that over half the companies—including some competitors—are connected to one another in a single, large component. We find power law distributions in the number of relationships for each company (node degree) and in the number of companies in each connected cluster (component size). We identify a single component with over 4,400 connected companies, including forty-four Fortune 500 companies. Certain companies appear repeatedly at the top of prominence measurements, suggesting that they are among the most prominent companies in the network.

There are several directions for future work. First, we can process additional data sources and add to the company name dictionary to expand coverage. Second, we can extend EVA to gather information on additional events, such as divestitures and spin-offs, to more accurately reflect the current network topology. Third, we can develop additional features for the visualization interface, such as multiple display panels with different data ranges to facilitate the comparison of ownership network changes over time.

We have demonstrated that information retrieval, extraction, storage, and visualization techniques can be used to build cost-effective systems to gather and present corporate ownership information from multiple data sources. More fundamentally, we believe regulatory agencies such as the SEC should adopt a corporate reporting mechanism built upon standardized XML schema, such as the one described in Section 3 and Appendix 1. As companies are increasingly intertwined with one another in this era of convergence, they have also become more vulnerable to collapses of other companies in the network. As a result, 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. An XML based reporting mechanism can dramatically improve the efficiency and accuracy of a system like EVA and bring greater transparency of corporate ownerships to public discourse.

REFERENCES

- Albert, R. and A.-L. Barabasi, 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics* 74(47). Available at <<http://www.nd.edu/~networks/PDF/rmp.pdf>>. [March 20, 2002].
- Barabasi, A. and Albert, R. 1999. Emergence of scaling in random networks. *Science* 286:509-512.
- Borgatti, S.P., M.G. Everett, and L.C. Freeman. 1999. UCINET 5.0 Version 1.00. Natick: Analytic Technologies.
- Brandes, U., T. Raab, and D. Wagner. 2001. Exploratory Network Visualization: Simultaneous Display of Actor Status and Connections. *Journal of Social Structure* 2(4).
- Brill, E. 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. *Computational Linguistics*. 21(4):543-566.
- Broder, A., R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. 2000. Graph Structure in the Web: experiments and models. In *Proceedings of the 9th International World Wide Web Conference, Amsterdam, The Netherlands, May 15-19*. Available from <<http://www.www9.org/w9cdrom/160/160.html>>. [March 20, 2002].
- 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.
- Burt, R.S. 1983. *Corporate Profits and Cooptation: networks of market constraints and directorate ties in the American economy*. New York: Academic Press.
- Cali, M.E. and R.J. Mooney. 1997. Relational learning of pattern-match rules for information extraction. In *Working Papers of ACL-97 Workshop in Natural Language Learning, Madrid, Spain, July 11*. Available from <<http://lcg-www.uia.ac.be/conll97/proceedings.html>>. [March 20, 2002].
- Compaine, B.M. and D. Gomery. 2000. *Who Owns the Media?: competition and concentration in the mass media industry*, 3rd ed., Lawrence Erlbaum Associates: New Jersey.
- Cooper, W. S., F.C. Gey, and D.P. Dabney. 1992. Probabilistic Retrieval Based on Staged Logistic Regression. In *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, June 21-24*, N.J. Belkin, P. Ingwersen, A.M. Pejtersen, eds. Available from <<http://dblp.uni-trier.de/db/conf/sigir/CooperGD92.html>>. [March 20, 2002]
- Faloutsos, M., P. Faloutsos, and C. Faloutsos. 1999. On Power Law Relationships of the Internet Topology. *Proceedings of ACM SIGCOMM'99*. Available from <<http://www.acm.org/sigcomm/sigcomm99/papers/session7-2.html>>. [March 20, 2002].
- Fellbaum, C., ed. 1994. *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press.
- Freitag, D. 1998. Machine Learning for Information Extraction in Informal Domains. Ph.D. diss., Carnegie Mellon University. Available from <<http://reports-archive.adm.cs.cmu.edu/anon/1999/CMU-CS-99-104.pdf>>. [March 20, 2002].
- Gansner, E. R., E. Koutsofios, S. C. North and K.-P Vo. 1993. A Technique for Drawing Directed Graphs. *IEEE Trans. of Software Engineering*. 19(3):214-230.
- Gebbie, M. and Y. Zhang. 2001. Competitors+. Available from <<http://www.temika.com>>. [March 20, 2002].
- Gram, C. and G. Cockton. 1996. *Design Principles for Interactive Software*. New York: Chapman & Hall.
- Grishman, R. 1997. Information Extraction: Techniques and Challenges. *Information Extraction (International Summer School SCIE-97)*, M.T. Paziienza, ed. New York: Springer-Verlag.

- Huffman, S. 1996. Learning Information Extraction Patterns from Examples. In *Symbolic, Connectionist, and Statistical Approaches to Learning for Natural Language Processing*, S. Wernter, E. Riloff, and G. Scheller, eds. 246-260. New York: Springer-Verlag.
- Industry Standard. Deal Tracker. 2001. Available from <<http://www.thestandard.com/search/deals>>. [June 30, 2001].
- Kleinberg, J. and S. Lawrence. 2001. The Structure of the Web. *Science*, 294, 1849-1850.
- Krackhardt, D., J. Blythe, and C. McGrath. 1994. KrackPlot 3.0: An Improved Network Drawing Program. *Connections* 17(2): 53-55.
- Maremont, M., J. Hechinger and G. Zuckerman. 2002. Tyco to Tap Backup Credit Lines, Shares Drop 19% Following News. *The Wall Street Journal*, Feb 5.
- Moore, A., ed. Who Owns What. *Columbia Journalism Review* web site. Available from <<http://www.cjr.org/owners/>>. [October 1, 2000].
- Mueller, E.T. Making news understandable to computers. Available from <<http://www.signiform.com/newsextract/newsund.htm>>. [March 31, 2000].
- Norlen, K., Lanard, V., and Lucas, G. 2001. Media Map. Available at <<http://dream.sims.berkeley.edu/media-map/>>. [March 20, 2002].
- Picarelli, J.T. 1998. Transnational threat indications and warning: The utility of network analysis. In *AAAI Fall Symposium on Artificial Intelligence and Link Analysis Technical Report, October 23-25, Orlando, Florida*. Available at <<http://www-eksl.cs.umass.edu/aila/picarelli.pdf>>. [March 20, 2002].
- Riloff E. 1993. Automatically constructing a dictionary for information extraction tasks. In *Eleventh National Conference on Artificial Intelligence, Washington, D.C., August 18-20*. 811-816. Menlo Park: AAAI Press.
- Robertson, S. E. and K. Sparck-Jones. 1976. Relevance weighting of search terms. *Journal of the American Society for Information Science*. 27:129-146.
- Robertson, S. E., S. Walker, and M. Beaulieu. 1998. Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive. In *Proceedings of the 7th Text REtrieval Conference*. 253-264.
- Soderland, S. 1999. Learning Information Extraction Rules for Semi-structured and Free Text. *Machine Learning*. 34(1/3):233-272.
- Stark, D. and B. Vedres. 2001. Pathways of property transformation: Enterprise network careers in Hungary, 1989-2000. Unpublished manuscript. Available from <<http://www.santafe.edu/sfi/publications/01wplist.html>>. [March 20, 2002].
- Telecommunications Act of 1996, Pub. LA. No. 104-104, 110 Stat. 56.
- Tufte, E.R. 1990. *Envisioning Information*. Cheshire: Graphics Press.
- U.S. Bureau of Economic Analysis Industry Accounts Data. 2002. Gross domestic product by industry. Available from <<http://www.bea.doc.gov/bea/dn2/gpoc.htm#1994-2000>>. [March 20, 2002].
- U.S. Securities and Exchange Commission. 2002. EDGAR Form Pick Search. Available from <<http://www.sec.gov/edgar/searchedgar/formpick.htm>>. [March 20, 2002].
- U.S. Securities and Exchange Commission. 2000. Aether Systems, Inc. Annual report on form 10-K for the year ended December 31, 2000. Available at <<http://www.sec.gov/Archives/edgar/data/1093434/000095013301500380/w47071e10-k.txt>>. [March 4, 2002].
- U.S. Securities and Exchange Commission. 2000. Nextel Communications, Inc. Annual report on form 10-K for the year ended December 31, 1998. Available from <<http://www.sec.gov/Archives/edgar/data/824169/0000950133-99-001031.txt>>. [March 4, 2002].
- Vedres, B. 2000. A Tulajdonosi Hálózatok Felbomlása (The Dissolution of Ownership Networks). *Közgazdasági Szemle (Hungarian Review of Economics)* 47 (in Hungarian) English version available at <<http://www.columbia.edu/~bv2002/pages/papers/pdf/discons.pdf>>. [March 4, 2002].
- Wasserman, S. and K. Faust. 1994. *Social Network Analysis*. New York: Cambridge University Press.
- Watts, D. and S. Strogatz. 1998. Collective dynamics of 'small world' networks. *Nature*. 393:202-204.

Appendix 1. Proposed Corporate Ownership DTD

We suggest two XML document types for describing corporate ownership information. The first represents a state, i.e., a relationship with a beginning and end date. The second documents an event, a transaction of a specific type occurring at a specific time. These DTDs could be designated as "open" for use with other namespaces. Mueller (2000) suggests a similar approach for adding topical information to news articles. (This solution does not address the problem of parsing and resolving company names to unique ids in the database.)

```
-----
<!--This is an XML schema for corporate ownership relationships.-->

<?xml version="1.0"?>
<!DOCTYPE ownership-relationship [
<!ELEMENT ownership-relationship
      (start-date?,end-date?,parent,child,
       stake?,source*,author?,(#PCDATA))>
<!ELEMENT start-date (year,month?,day?)>
<!ELEMENT end-date (year,month?,day?)>
<!ELEMENT parent (#PCDATA)>
<!ELEMENT child (#PCDATA)>
<!ELEMENT stake (#PCDATA)>
<!ELEMENT source (#PCDATA)>
<!ELEMENT author (author-name, author-date, author-email*)>
<!ELEMENT author-name (#PCDATA)>
<!ELEMENT author-date (year,month?,day?)>
<!ELEMENT author-email (#PCDATA)>
<!ATTLIST source confidence (low | medium | high) "low" #IMPLIED]>]

-----

<!--This is an XML schema for representing corporate ownership transactions.
(Transactions imply preceding and proceeding relationships.) -->

<?xml version="1.0"?>
<!DOCTYPE ownership-transaction [
<!ELEMENT ownership-transaction
      (date,parent,child,stake?,source*,author?,(#PCDATA))>
<!ELEMENT date (year,month?,day?)>
<!ELEMENT parent (#PCDATA)>
<!ELEMENT child (#PCDATA)>
<!ELEMENT stake (#PCDATA)>
<!ELEMENT source (#PCDATA)>
<!ELEMENT author (author-name, author-date, author-email*)>
<!ELEMENT author-name (#PCDATA)>
<!ELEMENT author-date (year,month?,day?)>
<!ELEMENT author-email (#PCDATA)>

<!ATTLIST ownership-transaction type
      (acquisition | sale | spin-off | combination | rename ) #REQUIRED>
<!ATTLIST source confidence (low | medium | high) "low" #IMPLIED]>]

-----

<!--This is an example ownership-transaction record. -->

<?xml version="1.0"?>
<!DOCTYPE ownership-relationship
      SYSTEM "http://denali.berkeley.edu/eva/ownership_relationship.dtd">
<ownership-relationship>
  <start-date>
    <year> 1999 </year>
    <month> 02 </month>
    <day> 18 </day>
  </start-date>
  <parent>AT&T</parent>
  <child>TCI</child>
  <stake>100</stake>

```

```

<source confidence = "high">
  http://www.fcc.gov/ccb/Mergers/ATT_TCI/
</source>
<author>
  <author-name> EVA Group (KN) </author-name>
  <author-date>
    <year> 2001 </year>
    <month> 12 </month>
    <day> 30 </day>
  </author-date>
  <author-email> knorlen@sims.berkeley.edu </author-email>
</author>
</ownership-relationship>

```

<!--This is an example ownership-transaction record. -->

```

<?xml version="1.0"?>
<!DOCTYPE ownership-transaction
  SYSTEM "http://denali.berkeley.edu/eva/ownership_transaction.dtd">
<ownership-transaction type = "acquisition">
  <date>
    <year> 1999 </year>
    <month> 02 </month>
    <day> 18 </day>
  </date>
  <parent> AT&T </parent>
  <child> TCI </child>
  <stake> 100 </stake>
  <source confidence = "high">
    http://www.fcc.gov/ccb/Mergers/ATT_TCI/
  </source>
  <author>
    <author-name> EVA Group (KN) </author-name>
    <author-date>
      <year> 2001 </year>
      <month> 12 </month>
      <day> 30 </day>
    </author-date>
    <author-email> knorlen@sims.berkeley.edu </author-email>
  </author>
</ownership-transaction>

```