# Extracting a Corporate Social Network from Text

Bill McDowell[*] Lingpeng Kong[*] Bryan R. Routledge[†] Noah A. Smith[*]
[*]School of Computer Science and [†]Tepper School of Business, Carnegie Mellon University
{wmcdowel,lingpenk,nasmith}@cs.cmu.edu, routledge@cmu.edu

February 2014

**Abstract**

This paper describes the construction of a computational representation of a social network of corporations in the United States. Applying tools from natural language processing and machine learning to a corpus of publicly available press releases, we have extracted a network of more than 1.7 million connections. We describe the methods used, preliminary analysis, and a visualization tool enabling further exploration.

## 1 Introduction

We have carried out a project to construct a social network representation of corporations in the United States and the relationships between them. The network has been extracted from a corpus of more than 200,000 press release documents published by more than 13,000 unique companies (discussed in Section 2).

The relationships in the network are labeled according to a novel taxonomy of corporate relationships (depicted in Figure 1 and discussed in Section 3). Extracting a network of this scale from a largely unstructured text corpus is a new technical challenge. We describe our extraction method, which uses basic text analysis tools from natural language processing, along with a statistical model that classifies a mention of a company in another company's press release into our taxonomy.

While our network is neither exhaustive nor perfectly clean, it provides a view of the corporate ecosystem that, to our knowledge, is unprecedented. We describe the network in Section 4, including a visualization interface that enables exploration of the network and how it varies over time.

## 2 Text Dataset: Press Releases

Our text data is 229,420 press release documents of 13,190 unique companies filed during the period 1994–2012. The press releases are obtained from mandatory corporate filings to the Securities Exchange Commission (SEC) 8-K filings. The 8-K forms are filed by companies to inform investors of "material events." As such, they cover a wide variety of corporate topics ranging from changes in management, major legal agreements, mergers, and announcements about financial results. These filings are unscheduled (in contrast to quarterly 10-Q or annual 10-K reports). The forms are similar to press releases and, conveniently for our purposes, we key on the press release document that is attached to many 8-K filings. The SEC filings are publicly available at `http://www.sec.gov/edgar.shtml`.

Starting with all 8-K documents filed between 1994 (start of SEC's on-line data) and 2012, we select those 8-K reports that are from a publicly-listed corporation as identified by cross-referencing to the COMPUSTAT financial database. This removes 8-K filings of "special purpose vehicles" that bundle other financial contracts like mortgages (e.g., mortgage backed securities). From the 8-K we extract, using a simple keyword-based algorithm, the company press release attached to the 8-K. We omit 8-K filings that do not contain a press release. Document counts and an overview of the data are discussed in more detail in Section 4 along with characteristics of the network we extract.

## 3 Network Extraction

To extract a network of organizations from our text corpus we develop a pipeline process that consists of three main steps.

1. **Preprocessing** to clean out non-text from the documents, segment the cleaned text into sentences, and enrich the resulting sentences with structure that could be useful in the construction of the network. Most importantly, the preprocessing runs named entity recognition (NER) to identify mentioned organizations which eventually become the nodes in the network.
2. Author-to-mention **relationship classification**, to characterize the social relationships between the corporate document authors and the organizations that they mention.
3. **Merging**, in which the organization mentions scattered across many documents are merged into entities, and the classified relationships between authors and mentions are consolidated into relationships between these entities.

The resulting network consists of nodes derived from document authors and mentioned organizations, connected by typed edges derived from relationship-type distributions between the authors and the organizations they mention. We present each step of this pipeline in more detail. In Section 4 we discuss the properties and preliminary results of the extracted network.

### 3.1 Text Preprocessing

The text preprocessor cleans unnecessary material from each document, and enriches the text with structure output by various NLP tools. The cleaning part of this step removes tabular data and garbled text from the input, which prevents the rest of the pipeline from attempting to infer structure from non-textual content. Sentences are segmented and tokenized using the Stanford CoreNLP library.[1] We applied handwritten rules that, for example, remove sentences not containing alphabetical characters, or that start with multiple hyphens or underscores. A word frequency filter removes lines starting with ten or more words not in a list of the most frequent 2,000 English words.

The CoreNLP library's named entity recognizer (Finkel et al., 2005) was used to augment the text with annotations for people, organizations, locations, and other entity types. The library also includes modules for part-of-speech tagging, syntactic parsing, and coreference resolution, all of which we believe might be useful in later stages of the pipeline. However, these additional tools are computationally expensive to run, and so we opted to run NER alone to ensure that data from the entire corpus of 229,420 documents could be included in the final network.

### 3.2 Relationship Classification

The relationship classifier produces representations of the relationships between corporate document authors and the organizations that they mention (as identified by NER). The representations of the relationships are given by probability distributions over possible relationship types, and the possible relationship types are defined by a fixed, manually constructed taxonomy.

[1] Available at `http://nlp.stanford.edu/software/corenlp.shtml`

The model for the classifier was constructed using a data-driven, supervised learning approach. For training, the supervised model requires a set of mentions annotated with gold-standard types for relationships between the authoring corporation and mentioned organization. The model provides a probability distribution over types for a mention's relationship to the mention's author, given features of the text surrounding the mention. The classification step of the pipeline outputs this distribution for every mention in the corpus.

### 3.2.1 Relationship Type Taxonomy

One of the main reasons we are interested in extracting a network from textual data is the ability to infer the nature of the relationship between the author and the entity being mentioned. To guide our modeling we developed a hierarchical taxonomy of relationships and hand annotated a sample of our data. Since there is no pre-existing and well-established taxonomy, we developed our categories in conjunction with the labeling of the data. Our taxonomy is hierarchical because relationships, even to well-informed expert annotators, are not unambiguous (e.g., customer-supplier vs. strategic partners). However, the higher levels in the hierarchy are less ambiguous. More generally, some ambiguity in categorizing relationships is to be expected given the longstanding questions about the legal versus economic about the nature of a corporation and the "boundaries of the firm" (Jensen and Meckling, 1976). The taxonomy we settled on is in Figure 1.

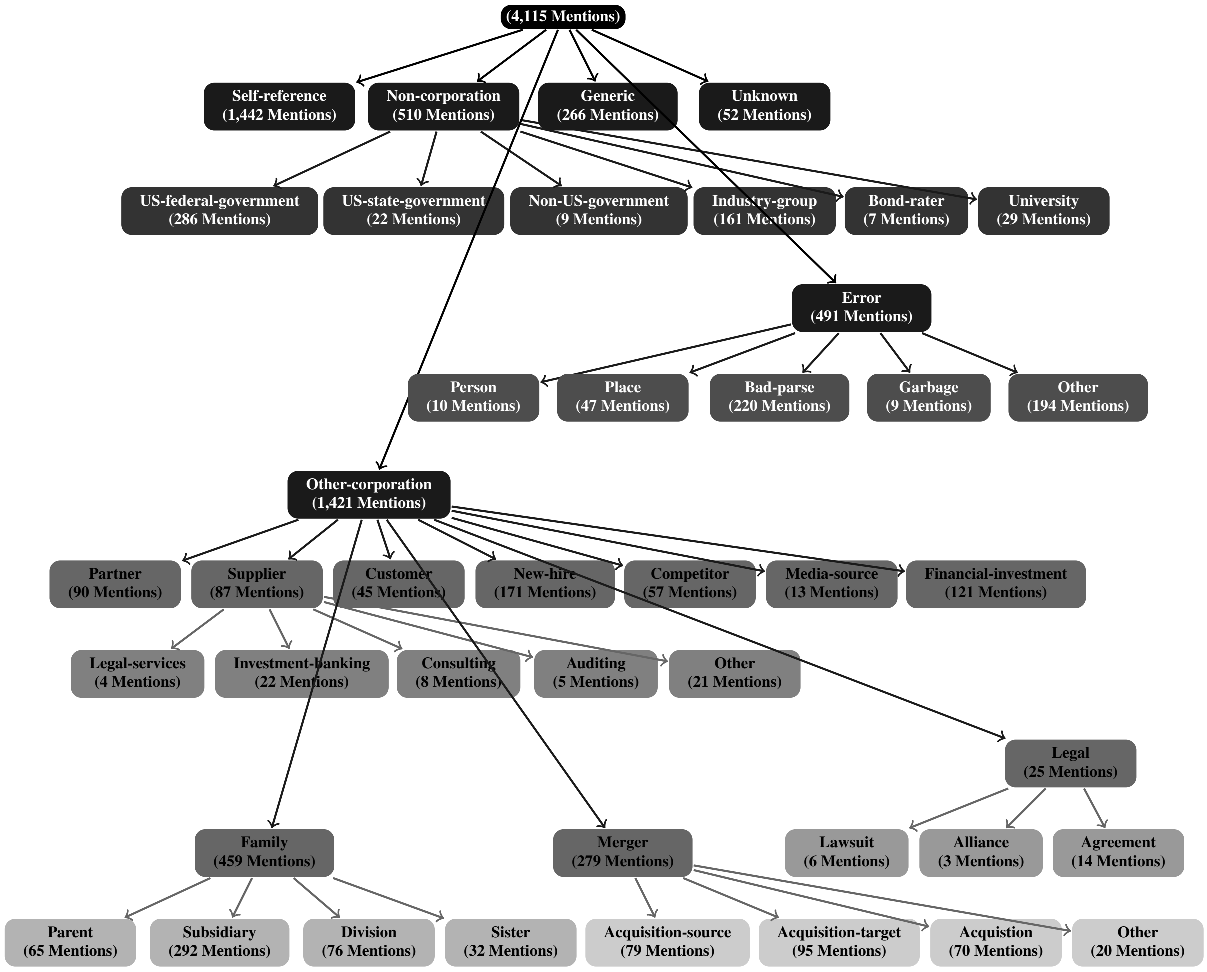
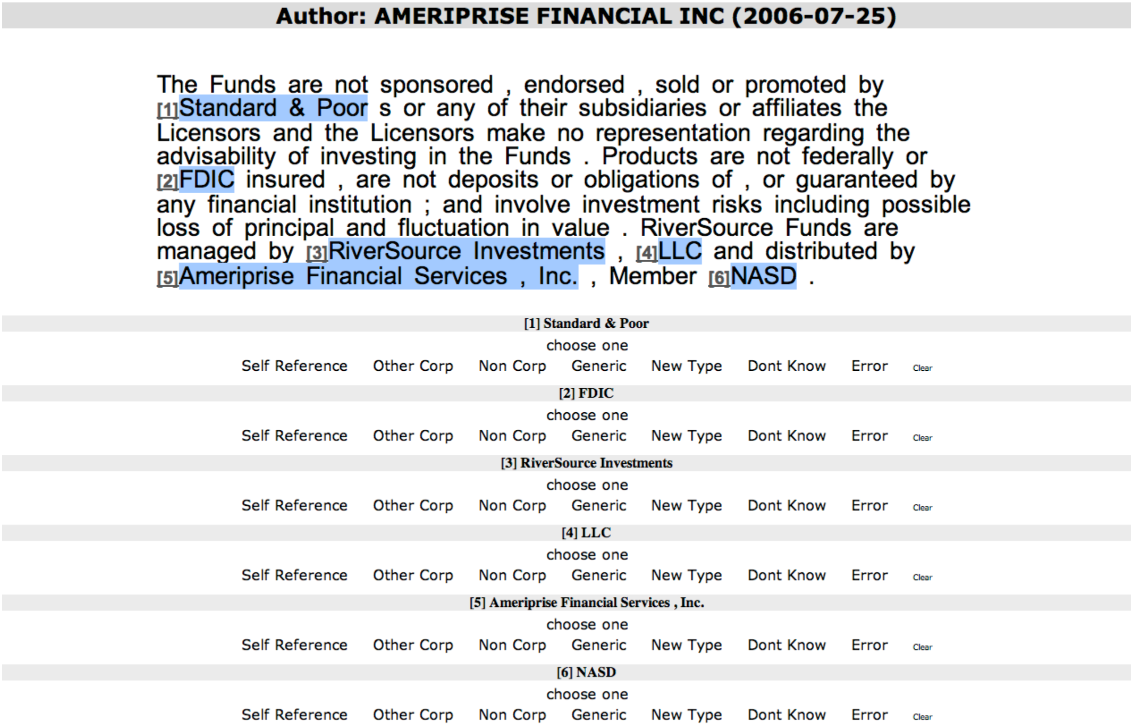


Figure 1: Relationship type taxonomy. Definitions are provided in Table 14 in Appendix A. Each node in the figure shows a label followed by the number of mention examples annotated with that label.



Author: AMERIPRISE FINANCIAL INC (2006-07-25)

The Funds are not sponsored , endorsed , sold or promoted by [1]Standard & Poor s or any of their subsidiaries or affiliates the Licensors and the Licensors make no representation regarding the advisability of investing in the Funds . Products are not federally or [2]FDIC insured , are not deposits or obligations of , or guaranteed by any financial institution ; and involve investment risks including possible loss of principal and fluctuation in value . RiverSource Funds are managed by [3]RiverSource Investments , [4]LLC and distributed by [5]Ameriprise Financial Services , Inc. , Member [6]NASD .

[1] Standard & Poor

choose one

Self Reference Other Corp Non Corp Generic New Type Dont Know Error Clear

[2] FDIC

choose one

Self Reference Other Corp Non Corp Generic New Type Dont Know Error Clear

[3] RiverSource Investments

choose one

Self Reference Other Corp Non Corp Generic New Type Dont Know Error Clear

[4] LLC

choose one

Self Reference Other Corp Non Corp Generic New Type Dont Know Error Clear

[5] Ameriprise Financial Services , Inc.

choose one

Self Reference Other Corp Non Corp Generic New Type Dont Know Error Clear

[6] NASD

choose one

Self Reference Other Corp Non Corp Generic New Type Dont Know Error Clear

Figure 2: A screenshot of the annotation web-interface.

Each path starting at the root of the taxonomy represents a type of relationship, with longer paths representing more fine-grained types. So, paths of length 1 to the first level of the taxonomy represent coarse relationship types, paths of length 2 to the second level represent more fine-grained types, and paths of length 3 represent the most fine-grained types. Table 14 in Appendix A contains definitions for the types at each level of granularity.

### 3.2.2 Annotation for Training

In order to provide our supervised relationship classifier with training data, we recruited a group of 10 business school graduate students, and worked with them to annotate a small set of organization mentions from the press release corpus. The annotation was collaborative (we were in the same room) and we used the process to develop the taxonomy as we annotated data.

The annotators used a custom web-interface (see Figure 2) to establish the most appropriate relationship type for each organization mention within a random sample of three-sentence texts taken from the corpus. The sample to annotate was stratified across dates. We selected a random sentences that contained a named-entity (identified by our pre-processing stage). To provide context, we showed the annotators the three sentences surrounding the mention (i.e., the sentence prior, the sentence with a named-entity, and the sentence following). Finally, we took advantage of the fact that named entities tend to cluster and so annotated all named entities that appeared in the three sentence set. In addition to the text, the annotators were given the (formal) name of the author and the date. For each mentioned organization within the texts, the annotators selected the relationship type between the mention and the author corporation given the information provided by the three surrounding sentences. Since the relationship types were defined by a taxonomy with varying levels of granularity, the annotators were instructed to only select relationship types of granularity they could resolve.

| Annotator 1 | Annotator 2 | Shared Mentions | Cohen's $\kappa$ (Full) | Cohen's $\kappa$ (Limited) |
|---|---|---:|---:|---:|
| *1* | *5* | 54 | **0.73 | **0.75 |
| *2* | *5* | 48 | *0.58 | **0.65 |
| *1* | *2* | 43 | **0.66 | **0.69 |
| *6* | *1* | 28 | **0.61 | **0.78 |
| *6* | *2* | 27 | **0.65 | **0.64 |
| *3* | *5* | 26 | *0.55 | **0.68 |
| *6* | *5* | 26 | **0.77 | **0.77 |
| *6* | *3* | 26 | *0.55 | **0.59 |
| *5* | *7* | 25 | 0.39 | *0.43 |
| *1* | *4* | 23 | **0.75 | ***0.90 |
| *1* | *3* | 21 | *0.55 | ***0.94 |

Table 2: Cohen's $\kappa$ statistics for pairs of annotators who annotated at least 20 of the same mentions. The second to last column ("Full") shows agreement computed over the full taxonomy. The last column ("Limited") shows agreement computed only over direct descendants of *Other-corporation* and *Non-corporation* in the taxonomy. We mark scores conventionally interpreted as "fair" (0.41–0.60) with *, "substantial" (0.61–0.80) with **, and "almost perfect" (0.81–1.0) with ***.

The annotators annotated a total of 2,761 three-sentence chunks containing 4,115 organization mentions. The sentences assigned to each annotator were randomly selected with replacement, and so some sentences were assigned to more than one annotator, but no annotator annotated all of the sentences. Table 1 shows the number of mentions that were annotated by a given number of annotators, and Figure 1 shows the number of mentions annotated with each label in the taxonomy.

| # Annotators | Mentions |
|:---:|---:|
| 1 | 3,794 |
| 2 | 271 |
| 3 | 43 |
| 4 | 7 |

Table 1: Number of mentions annotated by a given number of annotators.

We computed Cohen's (1960) $\kappa$ to measure the agreement between pairs of annotators who annotated at least 20 of the same mentions.[2] Since we limited the final version of the relationship classification model to only include labels directly under the *Other-corporation* and *Non-corporation* nodes in the taxonomy, we also computed agreement scores limited to these labels. Table 2 shows the agreement computed over both the full and limited sets of labels for each pair of annotators who share at least 20 mentions. The table marks scores according to conventional interpretations, but we caution against taking these as more than an indication that annotators were more or less on the same page. A more careful annotation effort, with training and periodic agreement checks among annotators, would likely produce higher quality data and better agreement. Now that we have a working taxonomy and an appreciation of the characteristics of the resulting network, this more careful annotation is now feasible.

[2] Cohen's $\kappa$ measures the agreement between two annotators by $\frac{P(a)-P(e)}{1-P(e)}$ where $P(a)$ estimates the probability two annotators agree based on the observed agreement, and $P(e)$ estimates of the probability that they agree due to chance. We computed $P(a)$ only over mentions annotated by both annotators, and $P(e)$ using marginal distributions for each annotator over all mentions annotated by the annotator.

### 3.2.3 Model Definition

Using the annotated data, we trained a supervised probabilistic model to classify the relationship types between author corporations and the organizations that they mention. Our hierarchical arrangement of relationship types (Figure 1) motivated us to use a hierarchy of multinomial logistic regression models, each associated with a non-leaf category in the taxonomy. Each of these multinomial logistic regressions defines a distribution over the more fine-grained labels associated with the label's children in the taxonomy. For a given organization mention, the probability of a label corresponds to the product of multinomial logistic regression probabilities along the path from the root node to that label. The path with the highest score corresponds to the classifier's most probable hypothesized label for the mention's relationship type.

More formally, let $\mathcal{T}$ be the set of labels in the taxonomy. For $t \in \mathcal{T}$, let $\pi(t) \in \mathcal{T}$ be $t$'s parent, and let $d(t)$ be the depth of $t$ in the taxonomy. Assuming $t_0$ is the empty label at the root, each $t \in \mathcal{T}$ corresponds to a (unique) path of the form $\langle t_0, t_1, \ldots, t_{d(t)} = t\rangle$ where for $i \geq 1$, $t_{i-1} = \pi(t_i)$. Then, given organization mention $m$, the distribution over labels for $m$ has the following factorized form:

$$p(t \mid m) = \prod_{i=1}^{d(t)} p(t_i \mid t_{i-1}, m) \tag{1}$$

Each of the factors takes the following parametric form (multinomial logistic regression):

$$p(t_i \mid t_{i-1}, m) = \frac{\exp\left(\mathbf{f}_{t_{i-1}}(t_i, m)^\top \mathbf{w}_{t_{i-1}}\right)}{\sum_{t \in C(t_{i-1})} \exp\left(\mathbf{f}_{t_{i-1}}(t, m)^\top \mathbf{w}_{t_{i-1}}\right)} \tag{2}$$

where $\mathbf{f}_{t_{i-1}}$ is a vector of features derived from the text characterizing the context of $m$, and $\mathbf{w}_{t_{i-1}}$ is a vector of parameters estimated by numerical optimization of the training data log-likelihood.[3] Note that each parent label $t_{i-1}$ has its own feature representation and its own parameters.

Once the model is estimated, $p(t \mid m)$ can be computed recursively for a given $m$ using the above formulas.

### 3.2.4 Model Development

Extensive research in natural language processing has found that the most important aspect of a statistical model for text is the choice of features, and so we experimented with several variations of the model involving different sets of textual and metadata features. In these experiments, we focused on developing each taxonomy label's multinomial logistic regression model separately before joining them together. This piecewise development allowed us to improve the predictions for labels toward to the top of the taxonomy before working with the fine-grained labels at the bottom. Since the lack of training examples at the fine-grained labels limits the potential reliability of statisical estimates, we prioritized the development of the models at the more coarse-grained labels. As a result of this focused development, the current working version of the model only outputs relationship types from the sub-taxonomy that terminates at the immediate children of the root, *Other-corporation*, and *Non-corporation* labels.

Recall that we annotated 4,130 organization mentions in context. When developing the model, we randomly split these mentions into a training set $D$ (90%) and a test set (10%). The model at label $t$, which classifies amongst its children, was only trained on data that falls beneath its descendants within the taxon-

[3] We optimize parameters using creg, available at `https://github.com/redpony/creg`

omy according to the annotations.[4] More precisely, the model at $t$ was trained only on data:

$$D_t = \{m \mid \exists t' \in \mathcal{T} \text{ s.t. } \pi(t') = t \wedge t' \in \langle t_0, \ldots, t_m \rangle\} \subseteq D \tag{3}$$

where $t_m$ is the annotated label for mention $m$. So, for example, the *Other-corporation* model was only trained using examples $m$ such that $m$'s annotated label's taxonomy path includes one of *Other-corporation*'s children (e.g. *Partner*, *Supplier*, etc.) When experimenting with the logistic regression at $t$ on $D_t$, we used 10-fold cross validation on $D_t$, computing the accuracy on each fold and the average accuracy across all folds. For the subset of data $D_t^i$ in fold $i$, this accuracy was computed for the model at $t$ as:

$$\frac{\sum_{m \in D_t^i} \mathbf{1}(c_t(m) = t_m)}{|D_t^i|} \tag{4}$$

where

$$c_t(m) = \operatorname*{argmax}_{t' : t = \pi(t')} p(t' \mid t, m) \tag{5}$$

is the classification for $m$ computed by the logistic regression at $t$.

*Regularization* refers to the addition of a term in the log-likelihood objective function that penalizes large magnitudes for $\mathbf{w}_t$. Regularization is important in avoiding overfitting, but it complicates parameter estimation because it is itself parameterized (i.e., it introduces "hyperparameters" whose values must be selected). In this work, we use an elastic net regularizer (Zou and Hastie, 2005). Elastic nets have two hyperparameters, one for the strength of a penalty on the $\ell_1$ norm of $\mathbf{w}_t$ and one for the strength of a penalty on the squared $\ell_2$ norm of $\mathbf{w}_t$. These are denoted by $\lambda_1$ and $\lambda_2$, respectively. Within each iteration of the cross-validation, we used a grid-search to select hyperparameter values.

### 3.2.5 Model Features

Table 3 lists the feature types we explored while developing the model. Several of these features have extra parameters whose values we varied in our development experiments. Below, we summarize the feature variations with which we experimented, and in Section 3.2.6, we evaluate the effectiveness of the features that we included in the final model.

[4] We also excluded all data labeled with the *Unknown* label, and we did not include this label in any version of our models. This is also true for evaluation results in Section 3.2.6. Additionally, we used a single, arbitrarily chosen annotation in cases where more than one annotator annotated a mention. In future work, more sophisticated methods might take advantage of the remaining annotations as well.

| Feature | Description |
|---|---|
| *Gazetteer-Contains* | $\mathbf{1}(O(m) \in G)$ |
| *Gazetteer-Edit Distance* | $\min_{g \in G} E(O(m), g)$ where $E$ is normalized edit-distance |
| *Gazetteer-Initialism* | $\max_{g \in G} \mathbf{1}(O(m)$ is an initialism for $g)$ |
| *Gazetteer-Prefix-Tokens* | $\max_{g \in G} \mathbf{1}(O(m)$ and $g$ share at least $k$ prefix tokens) |
| *Self-Edit-Distance* | $E(O(m), A(m))$ where $E$ is normalized edit-distance |
| *Self-Equality* | $\mathbf{1}(O(m) = A(m))$ |
| *Self-Initialism* | $\mathbf{1}(O(m)$ is an initialism for $A(m)$ or $A(m)$ is an initialism for $O(m))$ |
| *Self-Prefix-Tokens* | $\mathbf{1}(O(m)$ and $A(m)$ share at least $k$ prefix tokens) |
| *Self-Share-Gazetteer-ID* | $A(m)$ and $O(m)$ share the same identifier in gazetteer $G$ |
| *N-gram-Context* | $\forall w \in V_n$, $\mathbf{1}(w$ is at most $k$ tokens away from $O(m)$ in $s(m))$ |
| *N-gram-Dependency* | $\forall w \in V_n$, $\mathbf{1}(w$ is related to $O(m)$ in the dependency parse for $s(m))$ |
| *N-gram-Sentence* | $\forall w \in V_n$, $\mathbf{1}(w \in s(m))$ |
| *Metadata-Attribute* | $\forall w, w' \in V(a)$, $\mathbf{1}(a(M(o)) = w)$, $\mathbf{1}(a(A(o)) = w)$, $\mathbf{1}(a(M(o)) = w \wedge a(A(o)) = w')$ |
| *LDA* | Topic distribution for ’document’ $A(m).O(m).s(m)$ computed by LDA |

Table 3: Model features. Each of the above features returns a vector of values in $[0, 1]$ for a given organization mention. The description column contains some extra notation: $m$ is an organization mention, $s(m)$ is the sentence containing $m$, $A(m)$ is the name of the author corporation for the document containing $m$, and $O(m)$ is the name of the mentioned organization. For the gazetteer features, $G$ is a gazetteer supplied to the feature. $\mathbf{1}(p)$ is an indicator function which returns 1 if $p$ is true and 0 otherwise. $k$ is an extra integer parameter. $V_n$ is a vocabulary of normalized $n$-gram tokens from the annotated documents, and $V(a)$ is the vocabulary for corporation metadata attribute $a$.

***Gazetteer* Feature Variations**: We experimented with the *Gazetteer* features in the root logistic regression model.[5] These features check to see whether a given gazetteer contains a mention—or some close variation on the mention—in order to help sort out corporations from non-corporations and generic names. The variations with which we experimented used the gazetteers listed in Table 4. For the *Gazetteer-Prefix-Tokens* feature, we varied the minimum number of shared prefix tokens from 1 to 2.

| Feature | Description |
|---|---|
| *Compustat Corporations* | Corporation names from Compustat metadata |
| *Bloomberg Corporations* | Corporation names from Bloomberg metadata |
| *Scraped Non-corporations* | Non-corporate organization names from wikipedia |
| *Scraped Corporations* | Corporation names SEC master file |
| *Ticker* | Corporation ticker symbols mapped to corporation names for publicly traded companies (from Bloomberg and Compustat) |
| *Non-corporation Initialisms* | Non-corporate initialisms mapped to non-corporation names |
| *Stop Words* | Corporation stop-words (e.g. ”Corp”, ”Inc”, ”Company”, etc) |

Table 4: Gazetteers. The *Compustat Corporations*, *Bloomberg Corporations*, and *Bloomberg Ticker* gazetteers were constructed from the Compustat and Bloomberg metadata. These metadata include—but are not limited to—the fields listed in Table 5 for several corporations. The *Non-corporation Initialisms* gazetteer was constructed from the same data as the *Scraped Non-corporations* gazetteer. The *Stop Words* gazetteer was manually constructed from uninformative terms that frequently occur in corporation names.

***Self* Feature Variations**: We used the *Self* features in the root logistic regression to check for sim-

[5] A “gazetteer” is a list of named entities from an outside source. We used several sources. The SEC text data also comes with meta data identifiers of Central Index Key or CIK and company name. COMPUSTAT also has formal corporate names and industry codes. We also scraped data from Bloomberg’s (`http://www.bloomberg.com/markets/companies`) list of 65,000 company names, industries, and ticker symbols. For a list of government agency names we combined from wikipedia and other sources: `http://en.wikipedia.org/wiki/List_of_United_States_federal_agencies`, `http://www.wikinvest.com/concept/Government_Regulatory_Agencies`, `http://academics.smcvt.edu/cbauer-ramazani/BU113/fed_agencies.htm`.

ilarities between the author and mention organization names, with the intention of sorting *Self-Reference* from other mention types. We tried variations of the *Self-Prefix-Tokens* feature requiring either 1 or 2 shared prefix tokens, mirroring our experiments with the *Gazetteer-Prefix-Tokens* feature. We used the *Self-Share-Gazetteer-ID* feature with the *Bloomberg Ticker* gazetteer to determine if a mention acts as the ticker of the author corporation.

| **Attribute** | **Description** |
|---|---|
| *CIK* | Central Central Index Key is a unique company ID |
| *Name* | Corporation name (formal company name) |
| *Ticker* | Ticker symbol for listing on stock echange |
| *Country* | Country in which the corporation is located (head office location) |
| *Type* | Corporation type (e.g., common stock) |
| *Industry* | Corporation industry (e.g. Gold Mining, Airline, etc.) via Bloomberg based and Standard Industrial classification (SIC code) |

Table 5: Metadata Attributes.

***N-gram* Feature Variations**: We tried several variations on the *N-gram* features in the root, *Other Corporation*, and *Non-corporation* logistic regressions, with the aim of classifying the relationships based on varying sized chunks of the surrounding text. In each classifier, we tried unigram, bigram, and trigram versions of the *Dependency*, *Sentence*, and *Context* features. Unigrams generally gave performance improvements, whereas bigrams only helped in the other corporation model, and trigrams did not help in any experiment. This is unsurprising; bigram and trigram features are more sparse and typically parameter estimation for models including them requires a great deal more data than we have.

The *N-gram Dependency* features generally did not provide any performance improvement, possibly because any inference that could be gained from dependencies could be made from *N-gram Sentence* features alone. With the *N-gram Context* feature, we varied the window-size from 0 to 3. The window-size 0 in the root classifier was the most useful *Unigram Context* variation—providing a way to distinguish between corporations and non-corporations by the terms contained within their names.

Another variation on the unigram features mapped each unigram to a Brown cluster (Brown et al., 1992).[6] This variation was motivated by the effectiveness of Brown clustered features in other supervised learning tasks (Owoputi et al., 2013), but the clustering only gave a minor improvement in the performance of the *Non-corporation* model, and it had no effect on the accuracy of the other models at the other labels.

***Metadata-Attribute* Feature Variations**: In development of the *Other-corporation* logistic regression, we tried several versions of the *Metadata-Attribute* feature using the metadata attributes listed in Table 5, with the hope that metadata describing the author and mentioned corporations might provide some information relevant to their relationship. Except for *SIC*, none of the metadata attributes improved the performance of the model, and *SIC* only resulted in a minor improvement.

***LDA* Feature Variations**: We experimented with features derived from latent Dirichlet allocation (LDA; Blei et al., 2003) topic-distributions in the root, *Other-corporation*, and *Non-corporation* classifiers.[7] To compute these features, we first trained LDA on the entire corporate press release corpus to provide topic distributions for single-sentence documents. Then, the trained LDA model produced topic distribution for sentence containing mentions, and these distributions served as features for the mentions. We hoped that the topic distributions might provide the classifiers with useful information scattered across the full press release corpus, but these features generally gave no performance improvements for the logistic regressions.

**Other Feature Variations**: In our experiments with each of the feature types described above, we

[6] We used the Brown clustering implementation available at `https://github.com/percyliang/brown-cluster` applied to the corpus of press release documents.

[7] We used the LDA implementation provided in the Mallet library at `http://mallet.cs.umass.edu/`

generally supplied the features with the exact strings for the organization names and the surrounding sentence text. We tried a few variations on this. Most notably, one variation removed organization stop words supplied by the *Stop Words* gazetteer, whereas other versions simply removed non- alpha-numeric characters and extra whitespace. In the end, we settled on a cleaning function that performs the same operations as the hashing function $H$ described in Section 3.3.1, except that it does not remove stop-words or map initialisms and tickers to names.

### 3.2.6 Model Evaluation

After performing several experiments in development—trying many sets of features from Section 3.2.5 at each label's model, and comparing the resulting accuracies—we finalized the feature sets for each logistic regression, and evaluated the resulting models. In general, we included features in the final models if their use in experiments during development resulted in improved accuracy in the model.

The final sets of feature types for each label's logistic regression are shown in Table 6 along with the results of the ablation studies we performed to evaluate the usefulness of each feature type within the final model. In the ablation studies, we computed the average cross-validation accuracies of the root, *Non-corporation*, and *Other-corporation* classifiers with their final feature sets on the appropriate sets of data $D_t$ (defined in Section 3.2.4). Then, for each classifier, we repeatedly recomputed the accuracy with one of the final features removed. Comparing the accuracies of the final classifiers to the classifiers with features ablated is a standard way to measure their role in the final model.

From each of the ablation study results in Table 6, we see that there was no feature set whose removal caused a huge drop in accuracy. This suggests that for a given mention example, there are several features which are good for classifying its relationship, so removing any single feature does not result in a misclassification, and there may be a large amount of redundant information provided across the features in the final model. For example, Table 6 shows that the root classifier accuracy was not affected by removing any single *Scraped Corporations* or *Scraped Non-corporations* gazetteer feature set, but removing all of them drops the accuracy from 79% to 78%,[8] suggesting that each gazetteer feature provided roughly the same information to the model.

[8] We expected the gazetteer features to have a greater effect on performance based on our observations during development, but we had overlooked the fact that the *N-Gram Context* feature explains a large number of the same mention examples.

| **Root Model** | **Mean Accuracy** |
|---|---:|
| Final | 0.79 |
| Without Gazetteer | 0.75 |
| Without Gazetteer (*Scraped Corporations*, *Scraped Non-corporations*) | 0.78 |
| Without *Gazetteer-Contains* (*Scraped Corporations*) | 0.79 |
| Without *Gazetteer-Contains* (*Scraped Non-corporations*) | 0.79 |
| Without *Gazetteer-Contains* (*Stop Words*) | 0.77 |
| Without *Gazetteer-Edit Distance* (*Scraped Corporations*) | 0.79 |
| Without *Gazetteer-Edit Distance* (*Scraped Non-corporations*) | 0.79 |
| Without *Gazetteer-Initialism* (*Scraped Corporations*) | 0.79 |
| Without *Gazetteer-Initialism* (*Scraped Non-corporations*) | 0.79 |
| Without *Gazetteer-Prefix Tokens* (*Scraped Corporations*) | 0.79 |
| Without *Gazetteer-Prefix Tokens* (*Scraped Non-corporations*) | 0.79 |
| Without *Self-Edit Distance* | 0.77 |
| Without *Self-Initialism* | 0.77 |
| Without *Self-Prefix Tokens* | 0.76 |
| Without *Self-Share Gazetteer-ID* | 0.79 |
| Without *Unigram-Context* (Window-size=0) | 0.73 |
| ***Other-corporation*** **Model** | **Mean Accuracy** |
| Final | 0.69 |
| Without *Unigram-Context* (Window-size=1) | 0.69 |
| Without *Unigram-Sentence* | 0.64 |
| Without *Bigram-Sentence* | 0.67 |
| Without *Metadata-Attribute* (*SIC*) | 0.68 |
| ***Non-corporation*** **Model** | **Mean Accuracy** |
| Final | 0.82 |
| Without *Unigram-Context* (Window-size=0) | 0.80 |
| Without *Unigram-Context* (Window-size=0, Brown Clustered) | 0.80 |
| Without *Unigram-Sentence* | 0.82 |
| Without *Unigram-Sentence* (Brown Clustered) | 0.83 |

Table 6: Ablation studies for logistic regression models at the root, *Other-corporation*, and *Non-corporation* labels in the taxonomy. Each line shows the accuracy averaged across cross validation folds (Equation 4) for a final model, or a final model with some of its features removed. The final root model was run with hyper-parameters $\lambda_1 = 0$ and $\lambda_2 = 1$, the final *Other-corporation* model was run with $\lambda_1 = 0$ and $\lambda_2 = 3$, and the final *Non-corporation* model was run with $\lambda_1 = 0.5$ and $\lambda_2 = 1.0$.

Table 6 also shows that removing the *Unigram-Sentence* feature with Brown clustering actually increases the accuracy from 82% to 83%. This feature—and the other features that are seemingly redundant—was included in the model because its addition had resulted in accuracy improvements during some of the experiments in development. Some amount of variation between development and the final test is expected, due to the differences in data, hyperparameter values, and imperfect convergence of numerical algorithms. It is always possible to continue seeking improved feature sets, and these are expected pay off more as more data comes available.

We also evaluated the final model that combined the root, *Other-corporation*, and *Non-corporation* classifiers. The combined model only labeled mentions with the leaves $L$ of the sub-taxonomy that includes the root, *Other-corporation*, and *Non-corporation* labels and their children (using a rule of the form given in Equation 5). Since the output labels were limited to $L$, the combined classifier was only trained using mentions in:

$$D_L = \{m \in D \mid \exists t \in L \text{ s.t. } t \in \langle t_0, \ldots, t_m \rangle\} \tag{6}$$

The average 10-fold cross validation accuracy on $D_L$ for the combined classifier was 70%, and the average accuracy on the test set was slightly lower at 62%. These scores are both far above the *Self-reference* majority baselines of 36% on $D_L$ and 35% on the test set, which suggests that the time we spent developing our model paid off to some extent. Whereas these *Self-reference* baselines give a lower bound for the desired performance, the inter-annotator agreement gives an upper bound on achievable accuracy, since we do not expect a computer program to agree with human annotators more closely than they agree with each other. The maximum agreement between any pair of annotators was 95%, which sets a very high standard. Among other annotator pairs that shared more than 20 examples, one agreement score was 91%, three were between 80% and 85%, four were between 70% and 75%, and the remaining two were 52% and 65%. So, our model performs near the lower end of the range of annotator agreement scores, far below the upper end of the range. The agreement scores were computed using only the few examples shared between annotators, and the annotators carried out their work in a collaborative environment, so we cannot draw a definite conclusion from this comparison. However, it seems promising that our model's accuracy falls in the range of agreements, rather than below it.[9]

The confusion matrix for the model's performance on $D_L$ is given in Table 15 in Appendix A. The matrix shows that the model performed with less than 50% accuracy on the 11 labels for which there was only a small amount of training data (fewer than 108 examples for each). Except for *Error*, the model performed with greater than 50% accuracy on the remaining 8 labels.[10] This suggests that the model might show significant improvement if supplied with more data for the labels that currently have few training examples.[11]

### 3.2.7 Running the Model

As part of the network construction pipeline, the final trained version of the root, *Other-corporation*, *Non-corporation* combined relationship classification model takes in the set of organization mentions $M$ from the full preprocessed press release corpus, and outputs a distribution over relationship types $p(t \mid m)$ for each $m \in M$. To save computation time when running the model over the full corpus, the running classifier treats all instances of a single NER-identified organization name within one document as a single mention, and computes the model features only over the first sentence in which the name occurs.[12] Given the output distributions for the mentions, the final merging step of the pipeline can construct the nodes and edges of the network.

## 3.3 Merging

The final step of the network construction pipeline merges the output of the relationship classifier into a network of organization entities connected by typed business relationships. For each mention $m \in M$, the merging step takes in a distribution $p(t \mid m)$ output from the relationship classifier. This distribution ranges over relationship types $t \in \mathcal{T}$ for the relation between the author $A(m)$ of the document containing $m$ and the organization $O(m)$ represented by $m$. The merging step resolves $A(m)$ and $O(m)$ to normalized entities

[9] The inter-annotator agreement used in these comparisons is not the same as the Cohen's $\kappa$ score in Table 2. The $\kappa$ score takes chance agreements into account, whereas the agreement score in this section is the estimated probability that the two annotators give the same label. This estimated probability can be thought of as one annotator's accuracy with respect to the other, which is comparable to model accuracy—unlike $\kappa$.

[10] The poor performance on *Error* examples is not surprising. These examples were incorrectly identified as organizations by NER. Assuming the NER tool was well-developed, it's unlikely that our system would be able to identify its errors without a concentrated effort on our part. Otherwise, we would have accidentally developed an improvement to NER.

[11] We included the matrix for $D_L$, but left out the matrix for the test set because the smaller size of the test set made its confusion matrix much more difficult to interpret.

[12] Organization names $x_1$ and $x_2$ are treated as the same if $H'(x_1) = H'(x_2)$ where $H'$ performs the same operations as cleaning function $H$ described in Section 3.3.1, except that it lacks the gazetteer checks.

$H(A(m))$ and $H(O(m))$ using a hash function $H$, and computes relations between entities by aggregating mentions and their type distributions as:

$$S_{n_1 \to n_2} = \sum_{m \in M} \mathbf{1}(H(A(m)) = n_1 \wedge H(O(m)) = n_2) \tag{7}$$

$$P_{n_1 \to n_2, t} = \sum_{m \in M} \mathbf{1}(H(A(m)) = n_1 \wedge H(O(m)) = n_2) \cdot p(t \mid m) \tag{8}$$

$S_{n_1 \to n_2}$ represents the number of mentions of entity $n_2$ by entity $n_1$, and $P_{n_1 \to n_2, t}$ represents the expected number of mentions in which $n_1$ holds a relationship of type $t$ to $n_2$.[13] The final network $(N, E)$ is a directed graph of normalized entities with:

$$N = \{n \mid \exists m \in M \text{ s.t. } n = H(A(m)) \vee n = H(O(m))\} \tag{9}$$

$$E = \{(n_1 \to n_2) \mid S_{n_1 \to n_2} > 0\} \tag{10}$$

Additionally, we want to view changes in relationships over time, so the merging step also produces sub-networks for each year. These sub-networks are easily computed using the above definitions, except with $M$ restricted to the set of mentions $M_y$ in documents from year $y$.

### 3.3.1 Entity Resolution Function

Our current implementation of the entity resolution function $H$ applies the sequence of operations listed in Algorithm 1, producing a common form for many coreferring names.

**Algorithm 1** Our implementation of the entity hash function $H$. Note that **Stop Words** is the gazetteer containing terms like “Company,” “Corp,” “Inc,” etc.

```
begin $H(x)$:
$x \leftarrow x$ without terms that start with non-alphanumeric characters
$x \leftarrow x$ without non-alphanumeric characters
$x \leftarrow Lowercase(x)$
if $x$ without corporate Stop Words is non-empty then
    $x \leftarrow x$ without corporate Stop Words
end if
if $\exists$ unique $c \in$ Bloomberg Ticker Gazetteer s.t. $x = H(\text{Ticker}(c))$ then
    return $H(c)$
else if $\exists$ unique $o \in$ Non-corporation Initialism Gazetteer s.t. $x = H(\text{Initialism}(o))$ then
    return $H(o)$
else
    return $x$ with whitespace character sequences replaced by underscores
end if
end $H$
```

Optimally, two organization names $o_1$ and $o_2$ refer to the same organization if and only if $H(o_1) = H(o_2)$. We evaluated our implementation of $H$ against this standard using our relationship type annotated data by comparing the values of $H(A(m))$ and $H(O(m))$ with the relationship type annotation for $m$. In particular, assuming the following function definitions for mention $m$ (where $A(m)$, $O(m)$, and $T(m)$ are the author, mentioned organization, and true relationship type):

[13] Because of the way in which we grouped all instances of the same name within a single document into one mention, $S_{n_1 \to n_2}$ counts documents between $n_1$ and $n_2$, assuming that $n_1$ refers to $n_2$ by a single name within a given document.

$$EH(m) = \begin{cases} 1 & H(A(m)) = H(O(m)) \\ 0 & \text{Otherwise} \end{cases}$$

$$ET(m) = \begin{cases} 1 & T(m) = \textit{Self-reference} \\ 0 & \text{Otherwise} \end{cases}$$

We considered $H$ to be correct for mention $m$ if $EH(m) = ET(m)$. Applying this criteria across all annotated mentions, we estimate that the current implementation of $H$ has accuracy 83%, precision 99%, and recall 55%, and $F_1$ score 71%. Thus it is highly conservative, opting to avoid over-merging of mentions. This evaluation is limited to testing $H$ against author names paired with their mentioned organizations, and it leaves out several important cases—for example, when $o_1$ and $o_2$ refer to the same non-corporation (since the author must be a corporation). If a more sophisticated implementation of $H$ is developed in the future, more sophisticated evaluations are advised.

## 4 Exploring the Corporate Network

To summarize so far, we have generate the corporate network by applying our classification empirical model to the complete data set of 229,420 press release documents (from from SEC filed 8-K documents) over the period 1994 to 2012. These documents were written by 13,190 unique companies (defined by the SEC "central index key" or CIK identifier). Here we present some preliminary exploration of the network.

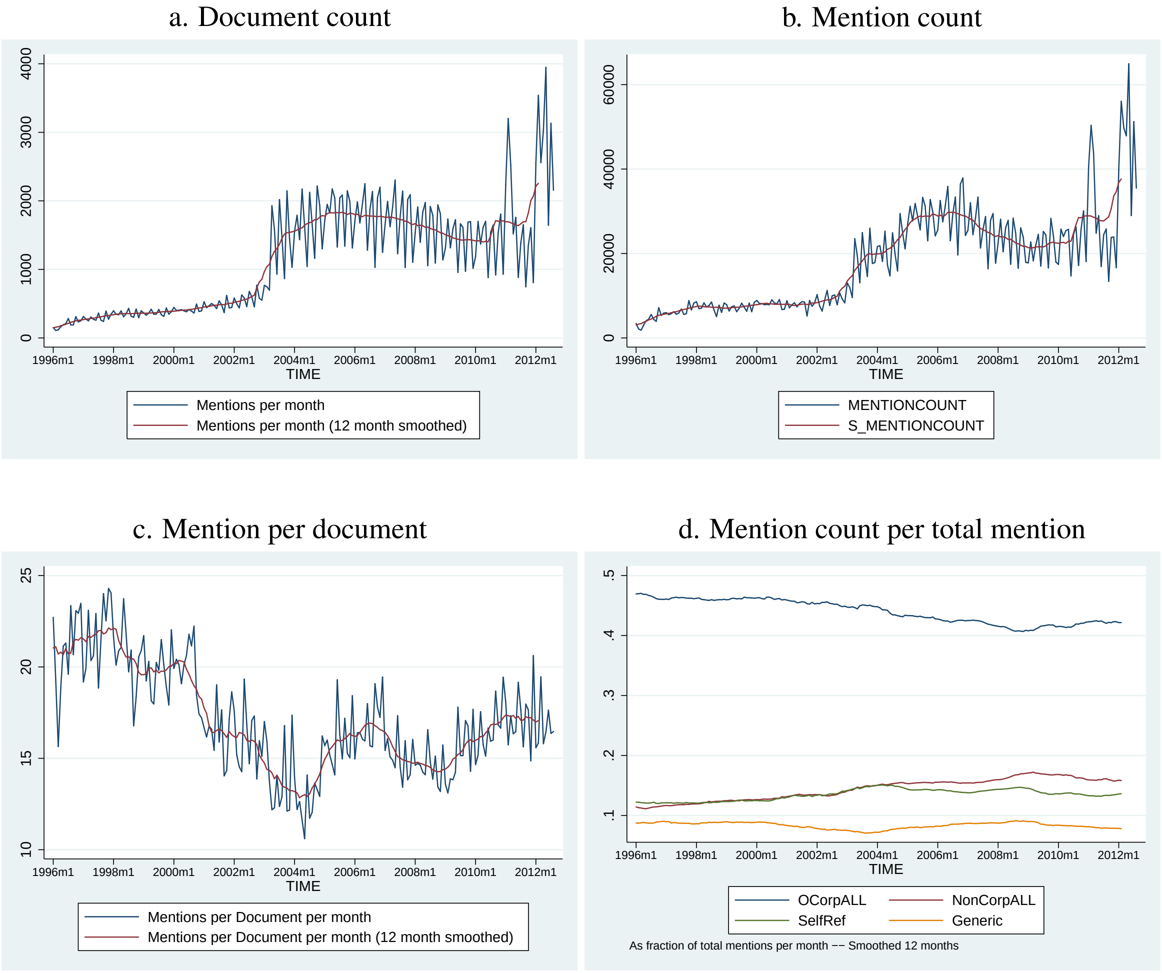
### 4.1 Description of the Extracted Corporate Network

Over the period of our data, the number of filings increased. Figure 3(a) shows the time series of documents. The increase is due primarily to the adoption of electronic filing in the early years of our sample. The number of 8-K filings also sharply increased following "Regulation FD" (August 21, 2000) that required firms to disseminate information more widely and at the same time resulting in more 8-K documents. The Sarbanes-Oxley Act ("SOX" of July 31, 2002) also increased disclosure requirements and increased the volume of 8-K filings. The sharp increase in 8-K filings in 2011 is a bit of a mystery that warrants more investigation; the overall number of filings was relatively unchanged over this period, but there happen to be more with attached press releases.

As mentioned previously, we identified named entities in the documents ("mentions"), which correspond to the nodes of network. As shown in Figure 3(b) the number of these mentions tracks similarly to the document counts. The mentions per document drops slightly over this period, shown in Figure 3(c), reaching its low around the 2000 "dot-com" era. Overall, is mostly in the range of 10 to 15 mentions per document. Across the whole dataset there are 618,067 unique mentions. Of those, there are 30,611 unique mentions that occur more than 15 times in the data. As expected, the mention data has a long tail.

Our model categorized each mention into a type. The four main categories are mentions of other companies (*Other-corporation*), mentions of non-corporation entities, like the Securities Exchange Commission (a *Non-corporation*), self-references where the author company mentions itself (*Self-reference*), and generic entities like The Board of Directors (*Generic*). We also have an type to catch possible errors of items that are not organizations or have been poorly parsed (*Error*). Examples of each of these are discussed below. Figure 3(d) shows that each of these main categories. Specifically, each mention is categorized by our classifier. Here we are counting the type according to what the model states as most likely. The plot is normalized by the total mentions each month. Surprisingly, there is, overall, little variation across the sample period.

Table 7 shows the most frequently mentioned items in our data. Note that many of these mention-nodes are not "authors" of the data and so have only "in" connections. Not surprising at least with hindsight, the


a. Document count

b. Mention count

c. Mention per document

d. Mention count per total mention

Figure 3: Various counts over time.

Securities Exchange Commission (SEC) is the most common organization mentioned. The SEC is the main financial regulatory body and is the agency requiring the 8-K reports be filed. The SEC is an example of a "non-corporate" or *NonCorp* mention-node in our data. These are organizations like the Food and Drug Administration (FDA) and the New York Stock Exchange (NYSE) which play a large and important roll in the corporate economy but are not themselves corporations.

In the table, "EBITDA" is identified as a named entity mention. Of course, this is not an organization but an acronym for "earnings before interest, taxes, depreciation,and amortization." This is an example of an *Error*. Also listed in the table are "Insetco" and "Lorraine Copper," both of which are corporations but misclassified as *Error*. "Enron" is an example of a mention-node that is a corporate organization. The category *OCorp-Family* indicates the mention is part of the cooperate structure of the Enron Corporation (a parent, subsidiary, and similarly related). Enron went bankrupt (spectacularly) in the fall of 2001 in part related to accounting fraud that used a complex structure of subsidiaries and related-party partnerships.

The general taxonomy of our labels, in Figure 1, is hierarchical. Tables 8–11 show the most frequent entities appearing by category and in some of the more common subcategories. Specifically, the tables sort on total mentions and displays according to the most frequent assigned category.[14]

It is interesting how frequent *Non-corporation* mentions are, accounting for about 15% of the mentions. Table 8 lists the most frequent. However, there are 1,746 nodes mentioned more than 15 times and labeled *Non-corporation* most frequently. In general, it is not easy to list and rank importance of governmental and industry organizations, and the network from the set of mentions gives and interesting perspective. By far the most frequently cited organizations are security regulators (SEC), tax authorities (IRS), banking and commerce entities (FDIC, Federal Reserve, Treasury), as well as more industry-specific regulatory organizations (FDA, FIRC). Industry organizations include stock markets.[15] Another example of note is the bond-rating service of Standard and Poors. As noted previously, the more specific the taxonomy, the less accurate our classification. Some federal organizations like the Labor Relations Board and The Pension Guarantee Trust are classified under *Non-corporation*, *Industry-group*. Other interesting examples of state institutions and universities are listed in Tables 10 and 11.

Table 12 has examples of organizations mentioned generically, *Generic*, such as "the board of directors". Here, we have not made any attempt to resolve this generic term to a specific organization. An interesting problem for future research might be to incorporate entities into the network that are *parts* of organizations. The companies that are listed most frequently as refereeing to themselves, *Self-reference*, is largely correlated with companies that have more and longer filings.

### 4.2 Industry Dynamics by Mention Type

It is interesting to look at mentions (and the mention network) by industry. To define industries we used the Standard Industrial Classification (SIC) four-digit code and grouped companies (authors, in our data) into 49 industry groups. The mapping to these industry groups is commonly used in financial economics research.[16]

Figure 4 shows the proportion of mentions of *Other-corporation* by industry (smoothed with a 12 month centered moving average) for several industries. Interestingly, the mention of *Other-corporation* increased for banks around the time of the financial crisis.[17] Presumably, this a reflection of the heavy involvement

[14] The SEC (and variations of its name) is a very frequent mention and is classified (correctly) as a *Non-corporation*, *US*, 95% of the time. However, since it is so common, even its misclassifications are common and skew lists of less common categories like *University*.

[15] In hand annotating the data we labeled stock markets consistently as *Non-corporation*, *Industry-group*, even though some, like NASDAQ-OMX are actually stand-alone corporate organizations.

[16] The definitions are standard in financial empirical analysis and are listed at Ken French's data library: `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`

[17] The most striking manifestation of the financial crisis was the bankruptcy of Lehman Brothers on September 14, 2008. Earlier indications of trouble in the financial sector appeared in August of 2007 (an event call the "Quant Meltdown") and the Federal

|  |  | Mentions |  | Most Frequent Category |  |  |
|---|---|---|---|---|---|---|
| Node | Total | In | Out | In | Out | Comments |
| securities exchange commission | 192,526 | 192,526 | 0 | *US-fed* |  |  |
| company | 77,086 | 77,086 | 0 | *Generic* |  |  |
| new york stock exchange | 61,309 | 61,309 | 0 | *NonCorp-Ind* |  |  |
| insetco | 53,473 | 53,473 | 0 | *Error* |  | an investment fund |
| gaap | 44,315 | 44,315 | 0 | *US-fed* |  | acronym: "generally accepted accounting principles" |
| board directors | 41,257 | 41,257 | 0 | *Generic* |  |  |
| lorraine copper | 21,184 | 21,184 | 0 | *Error* |  | a company |
| internal revenue service | 20,519 | 20,519 | 0 | *US-fed* |  |  |
| bank | 12,841 | 11,892 | 949 | *Generic* | *OCorp-Family* |  |
| us securities exchange commission | 11,403 | 11,403 | 0 | *US-fed* |  |  |
| food drug administration | 11,162 | 11,162 | 0 | *US-fed* |  |  |
| treasury | 10,985 | 10,985 | 0 | *Generic* |  |  |
| ebitda | 10,643 | 10,643 | 0 | *Error* |  | acronym for earnings |
| boardroom | 10,399 | 10,399 | 0 | *Generic* |  |  |
| commission | 9,531 | 9,531 | 0 | *Generic* |  |  |
| d | 8,818 | 8,818 | 0 | *Error* |  |  |
| erisa | 7,625 | 7,625 | 0 | *Error* |  | acronym for "Employee Retirement Income Security Act" |
| common stock | 7,268 | 7,268 | 0 | *Generic* |  |  |
| enron | 7,261 | 461 | 6,800 | *OCorp-Family* | *OCorp-Family* | a company notable for its complex inter-company structure related to its accounting fraud and bankruptcy in 2001 |

Table 7: Most frequent mentions.

| Node | Total | In | Out | Comments |
|---|---|---|---|---|
| securities exchange commission | 192,526 | 192,526 | 0 | |
| gaap | 44,315 | 44,315 | 0 | |
| internal revenue service | 20,519 | 20,519 | 0 | |
| us securities exchange commission | 11,403 | 11,403 | 0 | |
| food drug administration | 11,162 | 11,162 | 0 | |
| united states securities exchange commission | 7,190 | 7,190 | 0 | |
| fdic | 5,753 | 5,753 | 0 | |
| federal trade commission | 4,531 | 4,531 | 0 | |
| us food drug administration | 4,343 | 4,343 | 0 | |
| federal reserve system | 4,104 | 4,104 | 0 | |
| premier beverage | 3,648 | 3,648 | 0 | an error |
| federal deposit insurance | 3,364 | 3,364 | 0 | |
| department justice | 3,215 | 3,215 | 0 | |
| federal energy regulatory commission | 2,889 | 2,889 | 0 | |
| us treasury | 2,518 | 2,518 | 0 | |

Table 8: Most frequent *Non-corporation* mentions.

| Node | Total | In | Out | Comments |
|---|---|---|---|---|
| new york stock exchange | 61,309 | 61,309 | 0 | |
| american stock exchange | 5,170 | 5,170 | 0 | |
| pension benefit guaranty | 4,905 | 4,905 | 0 | |
| nasdaq | 4,738 | 4,738 | 0 | |
| nasdaq stock market | 4,167 | 4,167 | 0 | |
| nasdaq market | 3,888 | 3,888 | 0 | |
| labor relations board | 2,514 | 2,514 | 0 | |
| nasdaq capital market | 2,345 | 2,345 | 0 | |
| fannie mae | 1,952 | 1,031 | 921 | |
| european union | 1,865 | 1,865 | 0 | |
| nyse euronext | 1,369 | 151 | 1,218 | |
| standard poor ratings services | 1,286 | 1,286 | 0 | |
| nyse amex | 1,286 | 1,286 | 0 | |
| toronto stock exchange | 1,188 | 1,188 | 0 | |

Table 9: Most frequent *Non-corporation*, *Industry-group* mentions.

| Node | Total | In | Out | Comments |
|---|---|---|---|---|
| california superior court | 96 | 96 | 0 | |
| ontario superior court justice | 87 | 87 | 0 | |
| public utilities commission nevada | 83 | 83 | 0 | |
| supreme court british columbia | 56 | 56 | 0 | |
| connecticut department public utility control | 53 | 53 | 0 | |
| minnesota public utilities commission | 49 | 49 | 0 | |
| california department insurance | 48 | 48 | 0 | |
| pennsylvania insurance department | 42 | 42 | 0 | |
| kentucky public service commission | 40 | 40 | 0 | |
| tennessee valley authority | 37 | 19 | 18 | |
| pennsylvania public utility commission | 33 | 33 | 0 | |
| london court arbitration | 33 | 33 | 0 | |
| north atlantic treaty organization | 31 | 31 | 0 | |
| massachusetts department public utilities | 30 | 30 | 0 | |
| maine public utilities commission | 30 | 30 | 0 | |

Table 10: Most frequent *Non-corporation*, *US-state-government* mentions.

| Node | Total | In | Out | Comments |
|---|---:|---:|---:|---|
| education realty | 1,172 | 0 | 1,172 | |
| american campus communities | 1,058 | 6 | 1,052 | |
| university california | 774 | 774 | 0 | |
| stanford university | 483 | 483 | 0 | |
| harvard university | 482 | 482 | 0 | |
| university pennsylvania | 449 | 449 | 0 | |
| northwestern university | 385 | 385 | 0 | |
| university texas | 364 | 364 | 0 | |
| university | 334 | 334 | 0 | |
| university chicago | 332 | 332 | 0 | |
| columbia university | 294 | 294 | 0 | |
| duke university | 282 | 282 | 0 | |
| university michigan | 260 | 260 | 0 | |
| university illinois | 240 | 240 | 0 | |
| university southern california | 239 | 239 | 0 | |

Table 11: Most frequent *Non-corporation*, *University* mentions.

| Node | Total | In | Out | Comments |
|---|---:|---:|---:|---|
| *Corporate self-reference* | | | | |
| surviving | 4,159 | 4,159 | 0 | commonly used term in a merger the “surviving company” |
| standard poor | 3,525 | 3,525 | 0 | |
| el paso | 3,365 | 464 | 2,901 | |
| exelon | 3,094 | 617 | 2,477 | |
| covanta | 2,971 | 24 | 2,947 | |
| chesapeake energy | 2,637 | 109 | 2,528 | |
| wintrust financial | 2,462 | 11 | 2,451 | |
| csc | 2,368 | 268 | 2,100 | |
| general law | 2,181 | 2,181 | 0 | |
| harris | 2,135 | 120 | 2,015 | |
| *Generic self-reference* | | | | |
| company | 77,086 | 77,086 | 0 | |
| board directors | 41,257 | 41,257 | 0 | |
| bank | 12,841 | 11,892 | 949 | |
| treasury | 10,985 | 10,985 | 0 | |
| boardroom | 10,399 | 10,399 | 0 | |
| commission | 9,531 | 9,531 | 0 | |
| common stock | 7,268 | 7,268 | 0 | |
| indemnified party | 7,069 | 7,069 | 0 | |
| securities | 4,294 | 4,294 | 0 | |
| indemnifying party | 4,166 | 4,166 | 0 | |

Table 12: Most frequent self mentions.

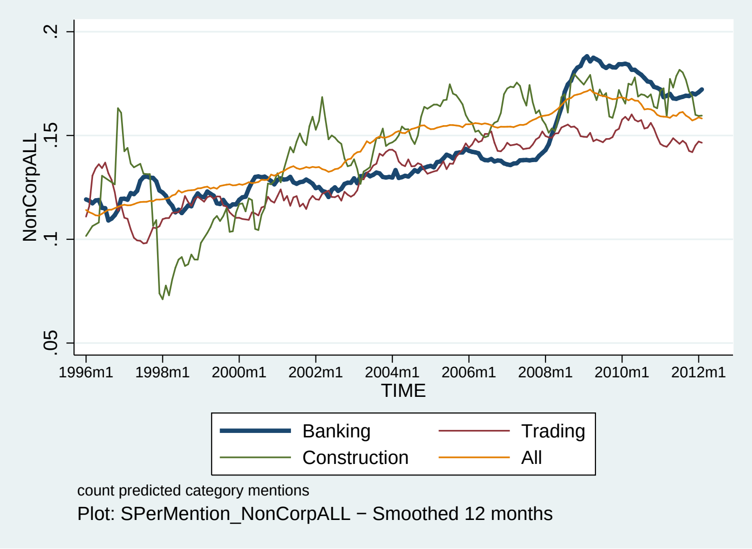


Figure 4: Mention type by industry.

of government agencies through the period. It is interesting that similar changes are not seen other related industries like financial trading companies. This example points to an interesting perspective on the role non-corporate entities play in the economy. The specific example we discuss later, in the context of our visualizer, picks up this theme by illustrating how these organizations can connect coporations.

### 4.3 Network: Preliminary Analysis

Putting all the 618,067 nodes together as defines a network of 1,710,767 edges for the full dataset of 1994-2012. The size of the network broken down by nodes mentioned within a specific year largely follows year-counts of our documents. This is shown in Figure 5. The network has several characteristics that future research can explore. The network is "directed" because we have the author company making the link to the mention. Since we have categorized each link (with noise) the network is also "multimodal." Since edges may be defined by multiple documents, the network can be weighted by the frequency of the mention. Finally, all the root documents are time-stamped allowing a rich temporal dimension along which to explore the network.

The main structure of our network stems from the fact we have a small number (13,190) of authors (out-link generators), linking to the 618,067 mentioned nodes. Recall that of those, there are 30,611 unique mentions that occur more than 15 times in the data. One measure of the connectedness of the network is the number of directed edges that are reciprocated. This measure is called the return. Across the 13,190 author-nodes, there are 3,610 pairs that are connected with such a reciprocal link (i.e., author Company A mentions Company B and author Company B mentions Company A). Looking at this in Figure 5, the dashed line, we see that the occurrences of these return links is proportional to the edge count. The plot counts the number of return links where the reciprocal mentions both happen in the same year. The overall count of 3,610 does not restrict the mentions to occur contemporaneously. Looking at the set of documents that define a return edge (e.g., reciprocal mentions) and selecting the pair of documents with the least time between filings gives an indication of the importance of the reciprocal mention. Of the 3,610 mentions, 11% are reciprocally mentioned within one day. That is company A mentions B and company B mentions A in filings that occur within one day or less. 16% are occur within a week, 23% within one month, and 55% are within one year.

Reserve bank of New York orchestration of the acquisition of Bear Sterns by J.P. Morgan in March of 2008.

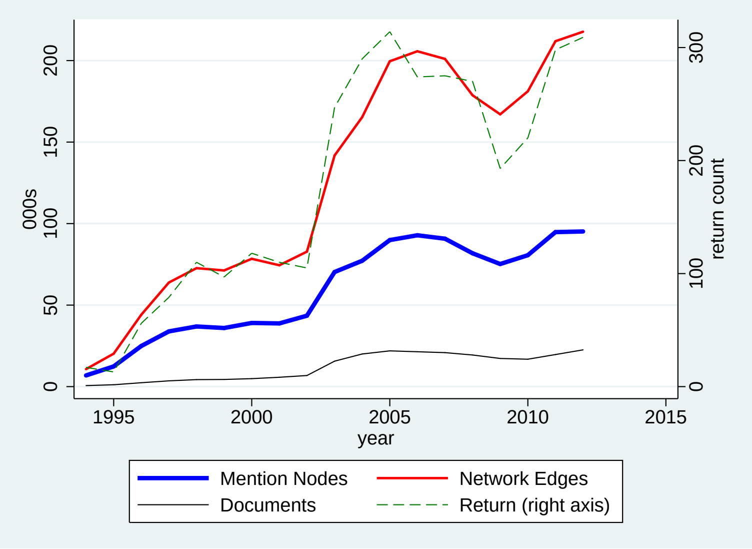


Figure 5: Network size by year.

| Query Type | Description |
|---|---|
| *Organization Name* | For nodes whose normalized names contain the given terms |
| *Metadata CIK* | For nodes that have a given CIK |
| *Metadata Country* | For nodes that have a given country |
| *Metadata Industry* | For nodes that have a given industry |
| *Metadata SIC* | For nodes that have a given SIC |
| *Metadata Ticker* | For nodes that have a given ticker |
| *Metadata Type* | For nodes that have a given organization type |
| *In Mention Count* | For nodes that are mentioned by others more than a given number of times |
| *Out Mention Count* | For nodes that mention others more than a given number of times |
| *Self Mention Count* | For nodes that mention themselves more than a given number of times |
| *In Maximum Posterior Type* | For nodes $n$ that, for a given $t$ have $t = \text{argmax}_{t'} \Sigma_{n' \neq n} P_{n' \rightarrow n, t'}$ |
| *Out Maximum Posterior Type* | For nodes $n$ that, for a given $t$ have $t = \text{argmax}_{t'} \Sigma_{n' \neq n} P_{n \rightarrow n', t'}$ |
| *Self Maximum Posteror Type* | For nodes $n$ that, for a given $t$ have $t = \text{argmax}_{t'} P_{n \rightarrow n, t'}$ |

Table 13: Search query types supported by the network visualization tool.

### 4.4 Visualization

There are many interesting topics and research questions to explore using our inferred network structure. As a start, and as an exploratory tool others can use, we built a web interface to visualize parts of the network. The visualization tool can be accessed at `http://demo.ark.cs.cmu.edu/cre/`, which provides searchable graphical representations of subsets of the network.[18]

The visualization tool takes a search query of a type listed in Table 13, and responds with a list of entities in the network. From these returned entities, we can select a subset for which to display a graph. The resulting graph will show the selected nodes and their neighbors, along with all the relationships between them. Given this graph, we can click on any of the neighbors of the selected nodes to select them as well, traversing the network outward from the initially selected set.

The visualization also displays information about each node and relationship in the selected sub-network. For each node $n$, the visualization shows the sums over posteriors and mention counts for its relationships

[18] The tool was developed from a existing open-source library of code previously written by one of the authors of this report.

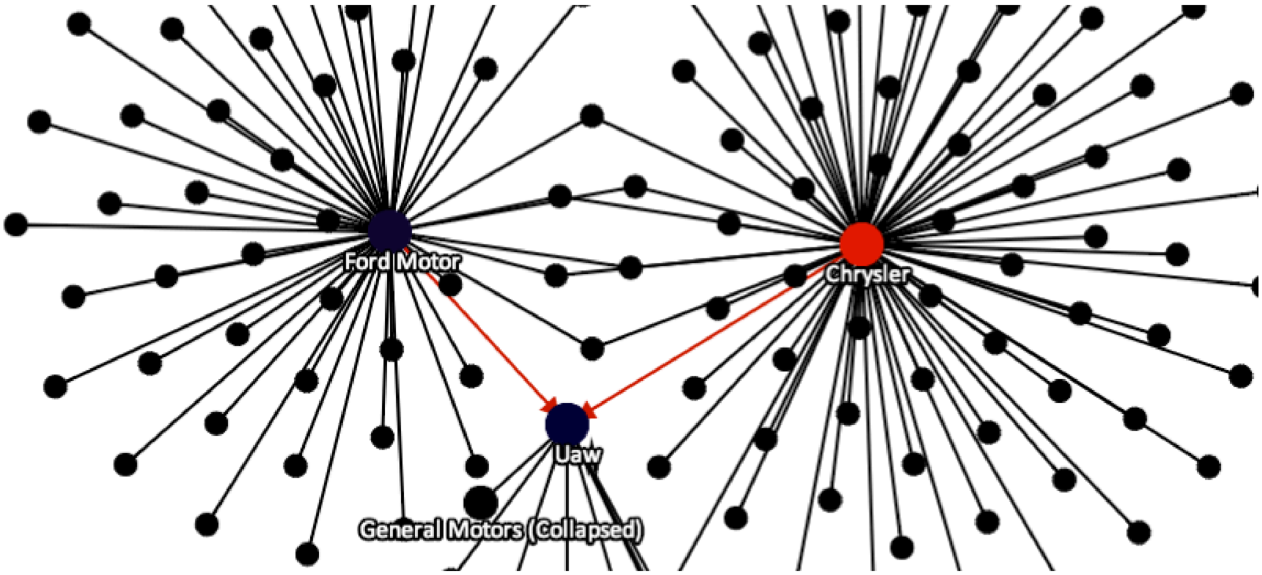


Figure 6: Screen shot of network visualizer showing Ford and Chrysler connected by their common union, the UAW. See `http://bit.ly/fordchrysleruaw` to recreate the example.

outgoing, incoming, and self-directed relationships:

$$\Sigma_{n' \neq n} P_{n \to n',t} \text{ and } \Sigma_{n' \neq n} S_{n \to n'}$$
$$\Sigma_{n' \neq n} P_{n' \to n,t} \text{ and } \Sigma_{n' \neq n} S_{n' \to n}$$
$$P_{n \to n,t} \text{ and } S_{n \to n}$$

For a relationship from $n_1$ to $n_2$, the visualization shows the expected number of mentions $P_{n_1 \to n_2,t}$ for each type $t$, the relationship type with the highest expected number of mentions $\text{argmax}_t\, P_{n_1 \to n_2,t}$, the number of mentions from which the relationship was merged $S_{n_1 \to n_2}$, and the sentence texts from which the relationship was was derived.

The tool also allows viewing slices of the network that are aggregated from data filtered down to a single year (see Section 3.3). We can limit the queries to search nodes of these yearly networks. If we select these nodes, and then change the filtering year, the tool will display the network for the same selected set of nodes at a different time. This allows us to see changes in organizations' relationships over time.

### 4.5 Visualization Example

As an example usage of our visualization tool, Figure 6 shows the network around the Ford Motor Company and Chrysler Group, LLC in the year 2011.[19] Interestingly, despite the fact that both companies are large components of the automobile industry, neither company mentions the other in the year shown, or in the other years of our dataset. However, both companies mention the United Auto Workers, since the union plays a large roll in both companies. The text from the company press releases that generate this connection both speak of the outcome of labor contract negotiations From Chrysler's release of 10/27/2011: "New four-year national labor agreement with the UAW ratified on October 26 , 2011." Similarly, from Ford on 10/04/2011 "Ford Motor Company and the United Auto Workers union (UAW) have reached a tentative agreement on a new four-year labor contract ..." Note that the United Auto Workers also appear in our

[19] The URL to recreate this picture is `http://bit.ly/fordchrysleruaw`.

data as node `united_auto_workers` highlighting that resolving nodes to mentions more accurately is an important remaining task, and that the network clues might be helpful. The companies are also linked by the the node of Generally Accepted Accounting Principles (`gaap`). The connection through this node is less interesting since this node is mentioned frequently by almost every company in our data (see Table 7). The UAW is an example of a node in our network that is not a corporation *Non-corporation*. As mentioned above, understanding the role of these entities in the economy is an unexplored research direction.

## 5 Products

As discussed above, a tool for visualizing our extracted network is available publicly at: `http://demo.ark.cs.cmu.edu/cre/`. A database containing the entire network will be made publicly available in the near future at the project website, `http://www.ark.cs.cmu.edu/CorporateNetwork`.

## 6 Conclusion and Future Work

We have successfully constructed a large network of the American corporate ecosystem, and developed software to visualize it. We briefly note several directions for continued research.

First, there are many ways to improve the network itself. We believe that a greater investment in clean annotations of relationship types of mentions will go a long way toward more accurate categorization of links among entities. We also believe that improved models, for example using latent variables and semi-supervised learning, will lead to improved accuracy by taking advantage of the large corpus of press releases. A more usable version of the network might also be constructed with more aggressive filtering, and with the direct incorporation of temporal information within the model. The annual groupings presented here are likely not the most appropriate for the data. Finally, much can be done to more accurately resolve organization mentions and authors to entities, and to evaluate the accuracy of such solutions.

Second, further questions about our approach are motivated by substantive research questions. One important question not considered yet is how this network compares to methods of constructing a social network of entities from other sources. For example, how closely does our text-defined network map to industry definitions or those that use macroeconomic data from input-output tables (Acemoglu et al., 2012). Similarly, comparing our network to those created by looking at the co-movement in (Kolar et al., 2010). More generally, we believe the network can serve as a tool for exploration and hypothesis generation in analysis of industries, changes over time, and the role of various non-corporate institutions and entities in the corporate world.

## Acknowledgments

This project was supported by a grant from the Alfred P. Sloan Foundation. The authors thank members of the Noah's ARK research group at Carnegie Mellon University for helpful advice and graduate students at the Tepper School of Business at Carnegie Mellon for help and advice with the annotation.

## References

Daron Acemoglu, Vasco M. Carvalho, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. The network origins of aggregate fluctuations. *Econometrica*, 80(5):1977–2016, 2012. ISSN 1468-0262. doi: 10.3982/ECTA9623. URL `http://dx.doi.org/10.3982/ECTA9623`.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jennifer C. Lai. Class-based $n$-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.

J. A. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association of Computational Linguistics*, pages 363–370, Ann Arbor, Michigan, USA, June 2005.

Michael C. Jensen and William H. Meckling. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, 3(4):305–360, 1976.

Mladen Kolar, Le Song, Amr Ahmed, and Eric P. Xing. Estimating time-varying networks. *The Annals of Applied Statistics*, 4(1):94–123, 03 2010. doi: 10.1214/09-AOAS308. URL `http://dx.doi.org/10.1214/09-AOAS308`.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, Atlanta, GA, June 2013.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

# A Appendix

| **Coarse Relationship Types (first level of taxonomy)** | | |
|---|---|---|
| *Self-reference* | (*Sel*) | Mention is the author corporation |
| *Other-corporation* | (*OCorp*) | Mention is an organization different corporation from the author |
| *Non-corporation* | (*NonCorp*) | Mention is an organization that is not a corporation |
| *Generic* | (*Gen*) | Mention is a generic organization (e.g. ”the company”) |
| *Error* | (*Err*) | Mention is not an organization (Named entity classification mistake) |
| *Unknown* | | Mention cannot be related to author by information given in the text |
| **Other Corporation Relationship Types (second level of taxonomy)** | | |
| *Family* | (*Fam*) | Corporations are in the same corporate family (e.g., parent, subsidiary) |
| *Merger* | (*Mer*) | Mention is discussing a corporate merger or takeover or change in corporate control |
| *Legal* | (*Leg*) | Corporations are legally involved with each other |
| *Partner* | (*Par*) | Corporations are partners |
| *New-hire* | (*Hir*) | Corporation is hiring a new employee (typically an executive) and mentioned corporation is related to the new employee (past place of employment, on the board of) |
| *Customer* | (*Cus*) | Mentioned corporation is a customer of the author |
| *Supplier* | (*Sup*) | Mentioned corporation is a supplier for the author |
| *Competitor* | (*Com*) | Corporations compete with each other |
| *Media-source* | (*Med*) | Mentioned corporation is a media source (e.g. New York Times) |
| *Financial-investment* | (*Fin*) | One corporation is financially investing in the other |
| **Non-corporation Relationship Types (second level of taxonomy)** | | |
| *US-federal-government* | (*US-fed*) | Mention is part of the U.S. federal government |
| *US-state-government* | (*US-state*) | Mention is part of a U.S. state government |
| *Non-US-government* | (*Non-US*) | Mention is part of a non-U.S. government |
| *Industry-group* | (*Ind*) | Mention is an industry group |
| *Bond-rater* | (*Rat*) | Mention is a bond rater |
| *University* | (*Uni*) | Mention is a university |
| **Error Relationship Types (second level of taxonomy)** | | |
| *Person* | | Mention is a person |
| *Place* | | Mention is a place |
| *Bad-parse* | | Mention contains incorrectly parsed text (Stanford NLP error) |
| *Garbage* | | Mention contains garbage text (not properly cleaned in preprocessing) |
| *Other* | | Mention is an error in some other way |
| **Family (Other Corporation) Relationship Types (third level of taxonomy)** | | |
| *Parent* | | Mention is the author’s parent |
| *Subsidiary* | | Mention is the author’s subsidiary |
| *Division* | | Mention is a division of the author |
| *Sister* | | Mention shares a parent with the author |
| **Merger (Other Corporation) Relationship Types (third level of taxonomy)** | | |
| *Aquisition-source* | | Author is acquiring mention |
| *Aquisition-target* | | Mention is acquiring author |
| *Aquisition* | | One corporation is acquiring the other, but direction is not specified by the text |
| *Other* | | One corporation is merging with the other through a non-acquisition |
| **Legal (Other Corporation) Relationship Types (third level of taxonomy)** | | |
| *Lawsuit* | | Corporations are involved in a lawsuit |
| *Alliance* | | Corporations are forming a strategic alliance |
| *Agreement* | | Corporations are participating in a legal agreement |
| **Supplier (Other Corporation) Relationship Types (third level of taxonomy)** | | |
| *Legal-services* | | Mention provides legal services to author |
| *Investment-banking* | | Mention is the author’s investment banker |
| *Consulting* | | Mention consults for the author |
| *Auditing* | | Mention audits for the author |
| *Other* | | Mention provides some other service to author |

Table 14: Definitions of labels in the taxonomy in Figure 1. Some labels also have abbreviations which are used in the confusion matrix in Table 15.

| $t_a/t_p$ | *Sel* | *Gen* | *Err* | *Fam* | *Mer* | *Leg* | *Par* | *Hir* | *Cus* | *Sup* | *Com* | *Med* | *Fin* | *US-fed* | *US-state* | *Ind* | *Rat* | *Uni* | ***Total*** | ***Incorrect*** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Sel* | 1166 | 25 | 24 | 54 | 23 | 0 | 3 | 3 | 0 | 1 | 1 | 2 | 5 | 1 | 0 | 1 | 0 | 0 | 1309 | 11% |
| *Gen* | 12 | 131 | 37 | 18 | 6 | 1 | 1 | 2 | 0 | 0 | 0 | 1 | 4 | 8 | 0 | 5 | 0 | 0 | 226 | 42% |
| *Err* | 41 | 48 | 186 | 101 | 25 | 0 | 1 | 12 | 0 | 3 | 1 | 0 | 3 | 5 | 0 | 3 | 0 | 0 | 429 | 57% |
| *Fam* | 41 | 0 | 26 | 299 | 25 | 1 | 1 | 8 | 2 | 1 | 0 | 1 | 3 | 1 | 0 | 0 | 0 | 1 | 410 | 27% |
| *Mer* | 24 | 0 | 18 | 44 | 140 | 0 | 2 | 1 | 0 | 1 | 0 | 2 | 4 | 1 | 0 | 2 | 0 | 0 | 239 | 41% |
| *Leg* | 2 | 0 | 2 | 9 | 4 | 6 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 24 | 75% |
| *Par* | 8 | 0 | 14 | 19 | 11 | 0 | 22 | 2 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 79 | 72% |
| *Hir* | 9 | 4 | 8 | 11 | 0 | 2 | 0 | 104 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 4 | 0 | 4 | 148 | 30% |
| *Cus* | 3 | 0 | 4 | 9 | 3 | 0 | 0 | 0 | 12 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 32 | 63% |
| *Sup* | 4 | 0 | 10 | 16 | 4 | 0 | 0 | 0 | 0 | 39 | 1 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 79 | 51% |
| *Com* | 3 | 0 | 3 | 17 | 6 | 0 | 0 | 0 | 0 | 1 | 15 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 46 | 67% |
| *Med* | 0 | 0 | 2 | 4 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 11 | 100% |
| *Fin* | 9 | 1 | 9 | 4 | 10 | 0 | 0 | 5 | 0 | 1 | 0 | 0 | 44 | 1 | 0 | 4 | 0 | 0 | 108 | 59% |
| *US-fed* | 1 | 1 | 5 | 9 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 214 | 0 | 15 | 2 | 3 | 252 | 15% |
| *US-state* | 1 | 0 | 7 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 5 | 3 | 0 | 0 | 0 | 19 | 84% |
| *Ind* | 3 | 1 | 11 | 9 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 15 | 0 | 96 | 0 | 0 | 139 | 31% |
| *Rat* | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 7 | 71% |
| *Uni* | 1 | 0 | 4 | 1 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 17 | 27 | 37% |
| *Non-US* | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 7 | 100% |
| ***Total*** | *1328* | *211* | *373* | *649* | *264* | *10* | *30* | *140* | *14* | *48* | *21* | *6* | *76* | *253* | *3* | *134* | *4* | *27* | | |
| ***Incorrect*** | *12%* | *38%* | *50%* | *54%* | *47%* | *40%* | *27%* | *26%* | *14%* | *19%* | *29%* | *100%* | *42%* | *15%* | *0%* | *28%* | *50%* | *37%* | | |

Table 15: 10-fold cross validation confusion matrix for the current root, *Other-corporation*, *Non-corporation* combined model. A cell in the matrix shows the number of mentions annotated with label $t_a$ that the model predicted to have label $t_p$. Columns represent predicted labels, and rows represent annotated labels. The labels are shown through their abbreviations specified in Table 14. There were no predictions for the *Non-US-government* label, so that column was removed from the matrix.