Death by Committee? An Analysis of Corporate Board (Sub-) Committees

ONLINE APPENDIX

Renée B. Adams[*]
University of Oxford and ECGI

Vanitha Ragunathan
University of Queensland

Robert Tumarkin
UNSW Australia

October 28, 2020

[*] Corresponding Author. Saïd Business School, University of Oxford, Park End Street, Oxford, OX1 1HP, U.K. Telephone: +(44) 01865 288824. E-mail: renee.adams@sbs.ox.ac.uk

Contents

1 Grammatical natural language processing **2**

1.1 Data collection . . . 2

1.2 Accuracy . . . 3

1.3 Methodological benefits . . . 4

2 Textual information gathering and decision-making data **5**

2.1 Sentence partitioning . . . 6

2.2 Topic modeling . . . 15

2.3 Sentence assignment . . . 20

2.4 Algorithm discussion . . . 21

3 Computational notes **23**

3.1 Grammatical dependency graph search . . . 23

3.2 doc2vec and latent Dirichlet allocation . . . 24

3.3 *k*-nearest neighbors . . . 24

4 Variable definitions **24**

4.1 Board characteristics . . . 24

4.2 Director characteristics . . . 25

4.3 Firm characteristics . . . 26

4.4 Trade characteristics . . . 26

4.5 Acquisition characteristics . . . 27

Figures

1 Sentence partitioning hyperparameter cross-validation: Univariate results . . . 31

Tables

1 Sentence partitioning hyperparameter cross-validation: Grid search results . . . 32

2 Latent dirichlet allocation: Topic coherence . . . 34

3 Latent dirichlet allocation: Descriptive topic data . . . 35

Board and committee LDA-based responsibility and meeting frequency data used in the main paper is available from the authors. At the time of this publication, data may be downloaded from www.tumarkin.net.

1. Grammatical natural language processing

1.1. Data collection

Mathematically, the grammatical structure of a sentence is most easily represented by a labelled directed graph, which consists of nodes connected by edges that indicate direction and type. For a given sentence, the nodes of the graph are the words of the sentence. The edges of the graph link the words from governing word to dependent word and are labelled by the type of grammatical relationship. The tree structure presented in Figure 1 of the main paper is a sub-type of the labelled directed graph structure, which permits more general cyclical relationships.

After representing sentences as directed graphs, we identify key grammatical relationships between words in sentences describing either board or committee meetings. The words and relationships mirror those described in the examples of Section 3.4.1 of the main paper. We allow for several synonyms that describe board or committee meetings and several additional grammatical relationships. However, to ensure that we do not generate false data, we allow for only a modest number of alternatives to each.

For example, in the active voice grammatical pattern, we look for a nominative subject of either "board" or "committee." The root verb may be "met", "held", or "conducted". The object is "meeting", "time", or "occasion", or pluralized forms of these words. The object may appear as either a direct object, a clausal complement (*ccomp*), a temporal modifier (*tmod*), or a preposition of the word "on" (*prep on*).

Once we identify the appropriate grammatical sub-graph, we can extract the number of committee meetings. This generally shows up as a numeric modifier of the object. However, when a committee meets infrequently (e.g., "the committee met once"), an object will not be present. Instead the number of meetings is found in the words "once" or "twice," which appear as adverbial modifiers (*amod*) or indirect objects (*iobj*) of the verb. We assign zero meetings to any committee that matches one of our alternative sub-graph patterns for a non-meeting committee in a fiscal year.

We identify the committee type by looking at words that are dependent noun compound modifiers or adjectival modifier (*amod*) of the governing word "committee." In the case of a committee with multiple roles, all descriptive words link directly to the governing "committee." So, for a Nominating and Governance Committee, "nominating" and "governance" will both be direct dependents of "committee." Companies often name committees using the preposition "on," as in a Committee on Governance. In these cases, the type of committee is marked with the prepositional-on relationship (*prep on*) by CoreNLP. Finally, in some cases, a firm may spend a paragraph discussing the composition, charter, and meetings of a committee. As such, the sentence discussing meetings may refer to the entity generically as "the committee." In such cases, the algorithm scans the paragraph to find the most recently mentioned committee type.

1.2. Accuracy

We evaluate the performance of our grammatical data extraction technique by comparing the results to a hand-collected dataset. We collect board and committee meeting data from 200 proxy statements containing approximately 1,000 total board and committee meetings. Observations on which the algorithmic and hand-collected data agree are considered correct. We recheck the source proxy statement whenever the two methods disagree, identify the correct information, and consider the observation a data collection error for the method that produced the inconsistency.

Based on this comparison, we are confident that our approach is accurate. It correctly records the number of meetings in 98.5% of observations. Errors were primarily due to proxy statements providing multiple data points, such as when a proxy statement lists the typical number of scheduled annual board meetings and the actual number that were held. Interestingly, the hand-collected data had more errors than the algorithmic data despite our best efforts to ensure that the hand-collected data was completely accurate. The hand-collected error rate of 3.5% was more than twice that of the algorithmic error rate of 1.5%. Nearly all hand-collected errors were due to data entry mistakes.

1.3. Methodological benefits

There are three key benefits of our approach. First, grammatical analysis identifies commonalities in written information. While there are myriad ways of verbalizing information about the meetings of boards and committees, there are, in effect, only a few underlying grammatical structures linking an entity with the number of meetings it holds. Whereas other natural language processing (NLP) techniques may need to adapt to a multitude of possible formulations of word orders in key sentences, grammatical techniques need only focus on a few key structures. This helps makes grammar-based parsing robust and accurate.

Second, this approach does not suffer from issues that plague other NLP techniques. Grammatical analysis considers words and the context in which they are used. Commonly used dictionary methods look at words without considering their context. To overcome this intrinsic shortcoming, such techniques often incorporate ad hoc rules. For example, researchers may place an upper bound on the number of words between keywords. Such bounds are designed to allow for flexibility in the way sentences are structured but can create false positives. Intervening clauses can make keyword searches not viable, as relaxing word distance bounds may lead the algorithm to pick up a large amount of irrelevant information. However, grammatical parsing still identifies grammatical relationships between governing words and dependents regardless of intervening clauses.

For example, Bruker Corporation's proxy statement for the 2008 fiscal year included the following statement: "The Audit Committee of the board of directors, which is currently comprised of Brenda J. Furlong, Collin J. D'Silva and Richard A. Packer, each of whom satisfy the applicable independence requirements of the SEC rules and regulations and NASDAQ Marketplace Rules, met six times during the 2008 fiscal year." The grammatical parsing technique identifies the relationship between the subject "committee" and the root verb "met" despite 38 intervening words. It distills the meeting-related content in this complex sentence to "The Audit Committee met six times."

Third, grammatical data extraction can be logically parsimonious. In our application, only four target grammatical relationships are needed to collect meeting data. These include refined versions of the active voice and passive-voice grammatical patterns described in Section 3 of the main paper. Two other grammar patterns, one active voice and one passive voice, account for cases when a committee did not meet during a fiscal year.

Finally, grammatical analysis easily allows for sentences that contain multiple pieces of information. Firms often disclose all the meetings of its committees in a single sentence. One such example is "During the last fiscal year, the Audit Committee met five times, the Compensation Committee met three times, and the Nominating and Governance committee met once." Grammatical parsing of this sentence creates a nested structure, with each clause discussing a committee and its meetings recognized individually. This makes it straightforward to extract the meeting data.

2. Textual information gathering and decision-making data

This appendix provides an overview of the machine learning and natural language processing (NLP) algorithms used in the paper to examine how firms describe the allocation of information gathering and decision-making responsibilities across corporate boards and sub-committees. The discussion emphasizes practical implementation issues of these tools particular to our study of corporate boards. We assume the reader has a basic familiarity with neural networks and Bayesian statistical methods; a complete technical discussion is beyond the scope of this document.

The methodology consists of three principal steps:

1. *Sentence partitioning:* We build a collection (corpus) of all 5.9 million sentences in our sample of proxy statements that mention either a board or a committee.[1] We then represent each sentence as a semantic vector that describes its content using the *doc2vec* neural network algorithm (Le and Mikolov 2014). Finally, we remove sentences that are unlikely to contain information on information gathering and decision making using a *k*-Nearest neighbors (*k*-NN)

[1] Sentence boundaries in documents are identified with the Stanford CoreNLP stemmer, which uses a deterministic rule-based algorithm to identify sentence endings. Sentences that mention boards and committees are identified by using a case-insensitive search for "board" or "committee". We do not require that these patterns are surrounded by word boundaries and, thereby, capture sentences with pluralized forms.

machine learning algorithm.[2]

2. *Topic modeling:* Sentences remaining after partitioning are modeled for topics using Latent Dirichlet Allocation (LDA). We review topics manually to determine if they describe information gathering and decision-making responsibilities.

3. *Sentence attribution and measure calculation*: Sentences are attributed to their underlying board and committees using the Stanford Natural Language Processing Group's CoreNLP dependency parser (Chen and Manning 2014) to identify entity names. We then calculate a unique firm-year measure of how these information gathering/decision-making topics are allocated across boards and committees. This measure is discussed in the main paper.

2.1. Sentence partitioning

We first build a sample of sentences that are expected to describe corporate boards information gathering and decision-making processes. To do so, we actually focus on the related problem: removing sentences that do not describe these responsibilities.

2.1.1. k-*Nearest neighbors*

We use *k*-Nearest neighbors, a standard non-parametric machine learning algorithm, as the core of our partitioning approach. *k*-NN begins with the researcher manually categorizing a training sample of observations. An uncategorized observation is then compared to each observation in the training sample using a distance measure (e.g. Euclidean distance) and categorized using a decision rule. The majority decision rule assigns an uncategorized observation to the group that is most common among the *k* closest training observations. The supermajority rule only assigns a category to an observation when a supermajority fraction of the *k* closest training observations are of the same category. For example, a 0.60 supermajority fraction requires than at least six out of the ten nearest observations are of the same category. Supermajority rules can result in

[2] *k*-Nearest neighbors is a well-known machine learning algorithm with a history that arguably spans a thousand years (Pelillo 2014).

indeterminacy; an observation that does not have a supermajority of *k* closest training observations is left uncategorized.

We select 10,000 sentences randomly from our corpus and categorize each as either *relevant* (i.e. describing information gathering or decision-making responsibilities) or *irrelevant*. These are translated into a semantic vector representation (i.e. a vector in $\mathbb{R}^n$) as discussed in the next subsection. This vector representation allows us to calculate simple Euclidean distances between sentences. An uncategorized sentence is then classified using a *k*-NN majority or supermajority decision rule.

The number of nearest neighbors *k* and the decision rule are hyperparameters that suit the classification environment. For example, if $k = 25$ and a majority decision rule is used, then a sentence that has at least 13 irrelevant sentences close to it will be categorized as irrelevant. If $k = 101$ and a supermajority of 0.7 is required, then a sentences that has at least 71 irrelevant sentences near it will be categorized as irrelevant.[3] We return to the determination of *k* and the decision rule in Section 2.1.3.

2.1.2. Semantic sentence vectorization

Semantic sentence vectorization algorithms improve on "bag-of-words" models often used in natural language processing. In a bag-of-words model, each sentence (or document) in a corpus is broken into constituent tokens. Tokens may be the words themselves or normalized forms (e.g. by eliminating conjugation differences). These constituents tokens form an unordered set (bag); the source document token-order is lost. "One-hot encoding" assigns each token a value of 1 in one dimension and 0 in all others. That is, each token is encoded as a unique standard base vector. A corpus of documents containing *m* unique tokens corresponds to a *m*-dimensional vector space. A sentence or document is typically the set (or a linear combination) of the relevant token vectors.[4]

[3] We round decision rules that necessitate a fractional plurality upwards to the nearest integer.

[4] Bag-of-words techniques may optionally include processing of words into a root form via stemming (for example Porter 1980, 2001) or lemmatization (Bird, Klein, and Loper 2009; Princeton University 2010). Even with these techniques, the dimensionality shortcoming of bag-of-words techniques remain.

Bag-of-word vector spaces have high dimensionality. As each unique token necessitates its own orthogonal base vector, tens or hundreds of thousands of dimensions are common. These large vector spaces may not reflect the true underlying semantics in the corpus. For example, the words "executive" and "manager" may be used interchangeably in a proxy statement to refer to senior employees. "One-hot" encoding will treat each of these as separate words with orthogonal meaning. Thus, large vector spaces obscure semantic similarities among sentences and documents.

Le and Mikolov (2014) propose *doc2vec*, a dimensionality-reducing neural network model that can mitigate issues with high dimensionality. *doc2vec* neural networks use a single hidden layer of n nodes to reduce a sample within an arbitrarily large number of unique bag-of-words tokens to a n dimensional vector space. Intuitively, semantic dimensionality corresponds to number of modeled information types. *doc2vec* translates each sentence into a real-valued vector. Sentence vectors close to one another have similar tokens and, consequently, related semantic contexts. Conversely, sentences that have similar semantic contexts will share many tokens and be close to each other in the vector space.

There are two *doc2vec* paragraph vector (PV) versions: Distributed Bag of Words (PV-DBOW) and Distributed Memory (PV-DM). While the algorithm is designated in terms of paragraphs, it can be applied to sentences, paragraphs, or entire documents. Given our application, we use the term "sentence" throughout this discussion. PV-DBOW provides a simple vector summary statistic for the sentence. PV-DM, on the other hand, captures "what is missing from the current context - or the topic of the paragraph" (Le and Mikolov 2014). All our sentences are in the context of proxy statements. Word meaning does not vary materially across documents and, as a result, there is limited opportunity for a "missing" topic vector suitable for PV-DM. As such, we use the PV-DBOW model in our analysis.[5]

In PV-DBOW, each sentence's unique n-dimension vector representation is the only neural

[5] As noted in Section 2.1.3, we confirm this qualitative reasoning via cross-validation. PV-DBOW provides better accuracy that PV-DM for equivalent throughput for our sentence partitioner. Le and Mikolov (2014) also suggest combining PV-DM with PV-DBOW. However, we find no marginal benefit from this and, as a result, only use PV-DBOW.

network input. For each input sentence vector, the output layer contains several randomly selected tokens from the sentence. Training the neural network yields optimized sentence vectors (in addition to parameters for the hidden layer).

The network learns to predict multiple contextual tokens for each sentence from a single, fixed sentence vector input. To see the benefit of this approach over a bag-of-words model, consider two sentences: (i) The compensation committee believes management exceeded its targets over the last fiscal year and (ii) The compensation committee believes the executive team exceeded expectations over the past fiscal year. The sentences are clearly similar, with minor differences due to word choice (e.g. fiscal vs. financial). As the neural network's hidden layer is shared by all input sentence vectors, the sentence vectors for (i) and (ii) need to be close to one another in order to active similar responses in the hidden layer and thereby predict the common sentence tokens in the neural network's output layer. Thus, the algorithm learns, for example, that the words management and executives fall into similar semantic contexts. With a sufficiently large sample of sentence offering sufficient variants, the algorithm learns what other words have similar semantic meaning (e.g. management vs. executives) and, consequently, what sentences are similar.[6]

The semantic dimension *n* and the number of training passes through the sample (epochs) are hyperparameters selected to suit the environment. Very low dimensionality may obscure differences in meaning between sentences; very high dimensionality can emphasize differences in words, not necessarily meaning, and risks having the *doc2vec* behave like a standard bag-of-words model.

2.1.3. Hyperparameters analysis

We perform an extensive grid search to select hyper parameters for our sentence partitioning algorithm. Partitioner performance can be described by two measures common in the machine

[6] For comparison to PV-DM, consider an illustrative example. The word 'flow' may appear in numerous domains, including finance (e.g. cash flows), ecology (e.g. water flows), or neurology (cerebrospinal fluid flow). In PV-DBOW, the summary vector captures the entirety of the sentence and is fed into the neural network. On the other hand, PV-DM uses a vector for the word 'flow' in addition to paragraph vector as neural network input. Thus, the PV-DM finance paragraph vector represents what is missing from the word 'flow': the relationship between flow and cash. Ecology and neurology PV-DM's would have different paragraph vectors to predict river and cerebrospinal, respectively. This concept of "missing" context is not material in our set of proxy statement sentences.

learning categorization literature: precision and recall. We are only interested in identifying irrelevant sentences. Thus, in our context, precision is the fraction of those sentences categorized as irrelevant that truly are. Precision relates inversely to false positives; it is high when the partitioner removes few relevant sentences. Recall is the fraction of irrelevant sentences that are categorized correctly. It is high when the partitioner removes a high fraction of all possible irrelevant sentences. That is, recall increases as the number of false negatives decreases.

$$Precision = \frac{\textit{Number of irrelevant sentences correctly categorized}}{\textit{Number of sentences categorized as irrelevant}}$$

$$Recall = \frac{\textit{Number of irrelevant sentences correctly categorized}}{\textit{Number of irrelevant sentences}}$$

Categorization problems generally trade off recall and precision. Higher precision typically requires lower recall and vice versa. It is easy to see why. For example, a sentence partitioner that simply removed all sentences would have perfect recall. But, in removing all sentences regardless of relevance, it would have poor precision. A partitioner that simply removed a single irrelevant sentence would have perfect precision, but recall near zero.

Our primary purpose for partitioning sentences is to aid the topic modeling algorithm. Maximizing recall would ensure that the topic modeler encounter a high percent of relevant sentences. But, it may have removed many relevant sentences as well due to low precision. This would be particularly problematic with our use of *doc2vec* semantic vectorization, which may increase the likelihood that erroneously removed information gathering and decision-making sentences have similar content. As such, we emphasize precision over recall in our parameter search. Fortunately, semantic vectorization helps us achieve reasonably high recall relative to more naive methods, such as bag-of-words models, without sacrificing precision.

We use cross-validation to examine the sentence partitioning performance. For each set of parameters, we first estimate *doc2vec* using the full sample of 5.9 million sentences. We then perform one-hundred cross-validation iterations. In each, we create a hold-out evaluation sample of 100 random observations by removing them from our training sample. We then perform *k*-NN with the remaining 9,900 training observation to categorize observations in the hold-out evaluation sample. We compute precision and recall for the test iteration. These statistics are averaged over the 100 cross-validation iterations to determine precision and recall for a particular set of parameters.

The grid search evaluates 42,120 hyperparameter sets, covering all possible combinations of the following four-dimensional hyperparameter space:

1. *doc2vec* semantic dimensions (18 values): We allow for 10 to 100 semantic dimensions, in increments of 10, and from 150 to 500 semantic dimensions, in increments of 50.

2. *doc2vec* training epochs (20 values): We vary the number of training epochs through the sample from 5 to 100 in increments of 5.

3. *k-NN* neighbors in classification set (13 values): We independently evaluate 5, 11, 15, 25, 51, 75, 101, 251, 501, 751, 1001, 2501, and 5001 nearest neighbors partitioning schemes. Odd values are chosen so that a simple majority decision rule does not yield a tie (i.e. an equal number of relevant and irrelevant nearest neighbors).

4. *k-NN* decision rule (9 values): The decision rules considered include a simple majority rule and 8 supermajority rules, which require plurality fractions of 0.55 to 0.90 in increments of 0.05. Supermajority rules that require fractional observations in certain *k*-NN set sizes are rounded up (e.g. 25 nearest neighbors with a 0.7 supermajority fraction requires a plurality of 18).

In general, all four hyperparameters jointly determine precision and recall. However, two univariate relations are clear in the cross-validation results. These are presented in Figure 1. We average over all relevant parameters sets to compute cross-validation averages. This means that the precision and recall values for 5 training epochs is an average over all combinations of *doc2vec* semantic dimensions, *k*-NN set size, and *k*-NN decision rules that have 5 training epochs. The levels presented in the figure may not be representative of a particular parameter implementation, but the trends are.

Panel A of Figure 1 displays the univariate results for *doc2vec* training epochs. Precision and recall are stable from 5 to 100 training epochs, with only minor variation due to randomness in cross-validation. This is a reasonable result given the nature of our data. We have a very large number of sentences in which word meaning is highly consistent. These sentences are short relative to applications with paragraphs and documents. Hence, the first few passes through the data provides a lot of information to the neural network and additional training is relatively uninformative.

The role of the *k*-NN decision rule on precision and recall is shown in panel B. *k*-NN precision should increase as the decision rule becomes more stringent. As expected, precision increases from just below 80% for a simple majority (> 0.5 fraction of neighbors) to approximately 95% for the 0.9 supermajority fraction. However, the chance that *k*-NN is indeterminant (i.e. unable to classify an observation) increases with the supermajority requirement. This is seen in the dramatic decrease in recall. Recall is about 90% for a simple majority and decreases to approximately 20% at the 0.9 supermajority fraction.

Given these trends, we fix the *doc2vec* training epochs and *k*-NN decision-rule hyperparameters in the remainder of our analysis. Partitioner performance is relatively insensitive to training epochs. There is a very modest peak in both precision and recall at 40 training epochs, with both metrics about 0.1% above their respective averages. Therefore, we perform all our remaining analysis using 40 training epochs. We also fix the decision-rule to a supermajority fraction 0.75 (i.e. at least three out of four nearest neighbors need to be irrelevant for a sentence to be classified as such). As discussed previously, we are most interested in maximizing precision without sacrificing recall. The

0.75 supermajority represents a reasonable trade-off. Admittedly, the choice of training epochs and supermajority fraction requires judgment, as there is no unambiguous decision metric. Nonetheless, the discussion in Section 2.4 suggests that the final topic output is reasonable and that our sentence partitioning hyperparameters produce a large, representative sample.

Table 1 examines how precision and recall vary with *doc2vec* semantic dimensionality and the number of nearest neighbors used for *k*-NN classification. All cross-validation results use 40 training epochs and a 0.75 supermajority decision rule. Panel A displays results for 75 and fewer nearest neighbors; Panel B display results for 101 and more nearest neighbors.

Recall decreases materially as the number of nearest neighbors used in *k*-NN classification increases. With 50 semantic dimensions, recall decreases from 0.665 when *k*-NN uses 5 neighbors to 0.276 with 2501 neighbors. This result is expected given our 0.75 supermajority rule. As the number of neighbors increases, *k*-NN moves from capturing local to global phenomena in the data. Less than 75% of our training sample sentences are classified as irrelevant. Hence, a large comparison set in *k*-NN tends to this unconditional sample composition, with an insufficient number of nearest neighbors necessary to categorize an observation, resulting in low recall. Precision, on the other hand does not vary as much as recall. It is relatively stable as the number of nearest neighbors increases.

Semantic dimensionality also affects precision and recall. In panel A, recall tends to peak between 40 and 60 total semantic dimensions for *k*-NN with 75 or fewer nearest neighbors. We expect to find an interior optimum for this algorithm. Precision should be poor when the semantic dimensionality is too low because *doc2vec* has difficulty differentiating sentences. When the semantic dimensionality is too high, *doc2vec* will not be able to identify common meaning and precision should also suffer. Recall decreases monotonically as semantic dimensionality increases beyond 101 nearest neighbors.

Consequently, we search for hyperparameters with *k*-NN with 75 or fewer neighbors (panel A of

Table 1). Overall, *doc2vec* with 50 semantic dimensions seems to yield an optimum precision and recall for most *k*-NN neighbor set sizes. Correspondingly, *k*-NN with 25 nearest neighbors generally maximizes recall. Precision is relatively unaffected by neighbor sizes between values of 15 and 75.

In untabulated results, we verify that *doc2vec* PV-DBOW outperforms PV-DM in this application. We repeat the complete grid search using the PV-DM variant. Cross-validation suggests that PV-DBOW generally yields 5% higher precision than PV-DM for equivalent recall (e.g. PV-DBOW precision is 90% while PV-DM precision is 85%).

2.1.4. Implementation

Given the preceding analysis, we parameterize the sentence partitioning algorithm as follows. We train *doc2vec* PV-DBOW with 40 training epochs over 50 semantic dimensions. Classification uses the 25 nearest neighbors with a 0.75 supermajority decision rule. Cross-validation suggests we can achieve precision of 90% with recall of 61% using these hyperparameters.[7]

The partitioning algorithm removes a total of 2.5 million sentences from our corpus, reducing the total number of sentences from 5.9 million to 3.4 million. The cross-validation precision (0.90) suggests that partitioning removed 2.25 million sentences (2.5×0.9) that were unrelated to information gathering and decision-making. Cross-validation recall (0.60) suggests that 1.5 million sentences ($2.25/0.6 \times 0.4$) unrelated to information gathering or decision-making were not removed and remain in the corpus. Hence, we expect our partitioned corpus of 3.2 million sentences to contain 1.7 million related to information gathering and decision-making. Had we not used the partitioning algorithm, the topic modeling algorithm would have processed nearly 2.5 sentences unrelated to information gathering or decision-making for every related sentence. The partitioner greatly improves the corpus quality, requiring the topic modeling algorithm to process only 0.9 unrelated sentences for each related one.

[7] Kusner, Sun, Kolkin, and Weinberger (2015) note the value of combining *doc2vec* and *k*-NN in classification problems. We tested a bag-of-words naive Bayes classifier as a benchmark. This achieved a maximum precision of 0.75 with recall less than 0.50.

2.2. Topic modeling

Having removed undesired sentences, our corpus is now ready for topic analysis. We approach this problem without an ex ante prior about the number or type of topics that describe information gathering and decision-making. Thus, we avoid keyword approaches. Instead, we use Latent Dirichlet Allocation, which can infer topics from the text. Our approach is careful to consider implementation issues specific to our study of corporate boards.

2.2.1. Latent Dirichlet allocation

Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003) is a popular topic modeling algorithm. LDA posits that each document in a corpus discusses one or more topics, with topic mixtures varying across documents. Topics are defined by the word usage. For example, "athlete" is likely to be relevant to a sporting topic, appearing very often. Yet, it may still appear occasionally in a financial topic discussing marketing sponsorship for an athletics goods company. Thus, each LDA topic needs to define a unique topic-specific conditional probability distribution over words. That is, topic distributions define which words are likely representative of a topic and which words are not. These distributions are fixed for the corpus, not document-specific, ensuring that they are meaningful.

LDA is an unsupervised machine learning algorithm. The researcher does not need to specify the topic mixture for each document nor the word distributions for the topics. Let K be the number of topics and V be the number of unique word in the corpus. D individual documents contain N words.[8] The structural document model used in LDA follows:

1. For each document, the document-specific topic mixture θ^d is drawn from a Dirichlet prior parameterized by α, a K-dimensional vector: $p(word_n^d \in Topic_k) = \theta_k^d$ with $\sum\limits_k \theta_k^d = 1$. The Dirichlet prior helps ensure that documents contain a small selection of available topics.

[8] Blei, Ng, and Jordan (2003) parameterize the document length as a random variable in their derivation of LDA. This random variable does not affect the derivation, and we omit it for brevity.

Further, it is the conjugate to the multinomial distribution used for topic and word selection, simplifying the mathematics of Bayesian inference.[9]

2. For each word location in the document, draw a topic k from the document specific topic mixture θ^d. The word is drawn from the appropriate topic probability distribution which defines the conditional probability of observing word v given topic k: $p(v|k)$.

LDA proceeds by Bayesian estimation of the above structural document model, comparing observed words to those randomly drawn. LDA estimation yields word distributions for topics. These are stored in a $K \times V$ matrix β: $p(word_v|topic_k) = \beta_{kv}$. The estimated model can be applied to the observed documents to yield a posterior topic mixture.

2.2.2. Sentence preparation

We are interested in identifying decision-making and information gathering themes for our research. LDA is a word-based algorithm; it does not differentiate words based on usage. This presents a small problem for our work as the committee names generally describe the committee's responsibilities. Consider two sentences that contain the word root "audit:"

(i) The committee has the responsibility of recommending the firm to be chosen as independent *auditors*, overseeing and reviewing *audit* results, and monitoring the effectiveness of internal *audit* functions.

(ii) The members of the *Audit* Committee are Richard W. Edelman (Chairman), James J. Ellis and Rex C. Bean.

The first sentence describes decision-making and information gathering responsibilities for the firm's annual audit report. The second sentence simply names the members of an Audit committee. Clearly, an "Audit Committee" is responsible for audit. But, not all sentences that discuss the "Audit

[9] These distribution may be seen as a multinomial distribution with a single trial (Blei, Ng, and Jordan 2003) or equivalently as a categorical distribution.

Committee" describe information gathering or decision-making responsibilities relevant to our research.

Consequently, we remove all committee names before processing sentences with LDA. As in the main text, committee names are identified using the Stanford CoreNLP (Manning, Surdeanu, Bauer, Finkel, Bethard, and McClosky 2014) dependencies. This helps ensure that our LDA topics are representative of actual responsibilities and not simply committee names. In other words, not all sentences that describe an "Compensation Committee" will be assigned to an executive compensation topic; the committee name is never seen by the algorithm.

We also remove "stop words," commonly occurring words (e.g. 'an,' 'the') that provide little information, from sentences before LDA analysis. Words are also passed through the WordNet lemmatization algorithm (Princeton University 2010) to remove conjugation differences.

2.2.3. Hyperparameter analysis

LDA requires estimating the distribution of the topic-mix Dirichlet prior and the matrix of topic-conditioned word probabilities. Once these parameters are estimated, it is possible to estimate the most likely topic mix posterior for any given sentence. Thus, the focus of LDA Bayesian inference is topic-, not sentence-, specific.

LDA requires two principle hyperparameters: the number of topics and the number of passes through the sample (epochs) used for inference. LDA is often applied to documents covering a wide variety of topics. These characteristics can necessitate a large number of training epochs, prolonging estimation. Our corpus, on the other hand, has two characteristics that expedite LDA estimation. First, as mentioned in Section 2.1.4, our corpus contains 3.2 million sentences after partitioning. Thus, we have a large number of relatively short observations (i.e. there are not many words per sentence). Second, we estimate that 1.7 million sentences describe information gathering and decision-making and 1.5 million sentences are irrelevant. And, these sentences fall in a single application domain. Thus, LDA is able to infer word probabilities reasonably efficiently. Given this quick convergence, we report results using a single training epoch. In untabulated results, we find virtually no difference between LDA results with one epoch and with more than one epoch.

Choosing the number of topics in LDA typically combines quantitative and qualitative analysis. Researchers use one of several measures designed to correlate with topic quality, with a final subjective evaluation typically required. Blei, Ng, and Jordan (2003) propose a measure, *perplexity*, that has traditionally been used to quantify topic quality. Lower *perplexity* suggests that an estimated LDA model is not surprised by topic data in a hold-out sample. Yet, research has questioned the measure's usefulness. LDA produces a list of words most associated with a topic (i.e. the words with the highest conditional probabilities). If topics are of high quality, one would expect that people would be able to detect when a random word is introduced into a list of a topic's five most frequently observed words. Research shows that people are unable to do this for topics selected on *perplexity* (for example, Chang, Boyd-Graber, Gerrish, Wang, and Blei 2009).

Coherence is an alternative to *perplexity*. *Coherence* rates topic words on their similarity, which may be defined in terms of ontological similarity, word co-occurrences, or other methods. Mimno, Wallach, Talley, Leenders, and McCallum (2011) and Newman, Lau, Grieser, and Baldwin (2010) show that topics with high *coherence* fit human intuition. We use Mimno, Wallach, Talley, Leenders, and McCallum (2011) definition of similarity based on word co-occurrences within the corpus's sentences in our analysis. High *coherence* suggests that topic's words often appear together in sentences.

2.2.4. Implementation

Coherence is computed for each topic individually, with researchers typically examining per-topic values and summary statistics. In many practical implementations, it may be sufficient to select the number of topics that maximizes average *coherence* over all topics. However, this is inappropriate for our setting. By construction, LDA assumes all sentences have a fully allocated topic mix ($\sum \theta_k^d = 1$). We know that the corpus analyzed with LDA still contains sentences unrelated to our research focus from the cross-validation of the sentence partitioning algorithm. Topics frequently found in sentences unrelated to information gathering and decision-making should be irrelevant to our work. Further, when building the partitioner training sample, we discovered that most sentences covering information gathering or decision-making were reasonably similar. Sentences irrelevant to our research were idiosyncratic. Hence, we expect LDA to produce only a few topics of interest and those topics to be the most coherent. Maximizing average *coherence* could possibly lower the *coherence* of topics of interest by raising that of irrelevant topics.

These observations guide our method to pick the number of topics. For each LDA model, we compute the average *coherence* of the *n* highest scoring topics. We vary *n* along with the total number of topics in a two-dimensional grid search. The results of the grid search are presented in Table 2. In column (1), we look at the average *coherence* of the 10 highest scoring topics as the total number of LDA topics changes. *Coherence* peaks when the LDA model has 30 total topics. Columns (2) and (3) examine results for the 20 most *coherent* and 30 most *coherent* topics, respectively. The average over the 20 most *coherent* topics peaks at 40 total topics; the average for 30 most *coherent* topics peaks at 40. The average *coherence* of the 40- and 50-most *coherence* topics is a decreasing function of the total number of topics in columns (4) and (5).

Given the grid-search results, we use the LDA model with 30 total topics and focus on the 10 most *coherent* topics. Our sentence partitioning algorithm suggests this is a reasonable result. As discussed in Section 2.1.4, we expect to have 1.7 million sentences describing information gathering and decision-making and 1.5 million other sentences. Assuming equal *coherence* across sentence types, we would expect to find an optimum when the n most *coherent* topics represent 53% of the the total topics. Given our training sample construction insight that there are more topics that do not describe information gathering and decision-making, we expect the optimum to occur when the top n topics are less than 53% of the total topics.

Topic modeling requires qualitative analysis; review of the topics for relevance to the research objective and overall reasonability is required. Highly *coherent* topics may need to be ignored if, for example, they capture a common theme unrelated to information gathering and decision-making. Alternatively, similar topics may need to be combined. For example, a single topic when LDA has 20 topics may become two separate topics when LDA has 30 topics. These two separate topics may be combined through qualitative analysis. For this reason, we do not undertake a further grid search to further refine the hyperparameters. The qualitative discussion is presented in Section 2.4.

2.3. Sentence assignment

After determining which sentences in proxy statements describe information gathering or decision-making, we need to attribute sentences to either a board or a named committee. Sentences that only mention the board are assigned to the corporate board observation for that proxy statement's sample firm-year.

We use the Stanford CoreNLP dependency parser (Chen and Manning 2014) to attribute sentences to committees. As described in the main paper, CoreNLP marks words that form the committee name in only a few ways. Words may appear as noun compound (*nn*) or adjective (*amod*) modifiers to the governing word "committee." Words may also appear with a prepositional-on relationship (*prep on*), which occurs for a "Committee on Governance." When a committee has multiple roles, all descriptive words link directly from the word committee.[10]

We collect these committee descriptors and find the matching committee from the sample of firm-committee-year observations corresponding to the proxy statement. In a small number of cases, BoardEx contains an out-of-date committee name. This typically occurs when the firm changes the conjugation of the committee name. For example, firms occasionally rename a "Nominating Committee" to a "Nominations Committee." To pick up these changes, we match using word stems

[10] The Stanford CoreNLP approach to committee name identification is not the only option. In earlier versions of this research, we used Named Entity Recognizer (NER) (Finkel, Grenager, and Manning 2005). Like the dependency parser, NER uses a neural-network. It is designed to identify named entities such as people, organizations, or location. However, we do not need to identify many of the types of entities for which it is designed (e.g. people). Our application is very well-defined and only needs to look for committees. In practice, we find that the dependency analysis performs slightly better for this application when committee names are not capitalized. The main disadvantage of dependency analysis is the significant computational time required.

that eliminate conjugational differences (Porter 1980, 2001).[11]

In some cases, a proxy statement may refer to a previously mentioned committee (e.g. this committee or the committee). When this occurs, we use the dependency parser to lookback through preceding sentences and find the name of the most recently mentioned committee.

2.4. Algorithm discussion

The sentence partitioning and topic modeling algorithms need to work harmoniously. Our focus on *precision* over *recall* when configuring the sentence partitioner is driven by an understanding of how topic modeling output varies with corpus integrity. In turn, the topic modeling hyperparameters are informed by cross-validation results from the sentence partitioner.

Ultimately, we must qualitatively check whether the topics sensibly reflect information gathering and decision-making responsibilities undertaken by the board and its committees. LDA provides two useful diagnostic tools for doing so. First, the conditional word probability matrix provides information on what words are most commonly associated with a topic. *Coherent* topics should provide words related to information gathering and decision-making responsibility-types that we know boards and committees perform, such as audit, corporate governance, and executive compensation. Second, a trained LDA model can produce the most likely topic mixture posterior for a sentence. By definition, the topic mixture weights are non-negative and sum to one. Hence, a sentence with close to a singular unit weight is very representative of a topic. Such sentences may be examined to ensure they exhibit consistent themes.

Table 3 presents panels containing this diagnostic data for the ten topics with the highest *coherence*. Keywords are provided in decreasing likelihood. The table also lists five highly representative sentences for each topic. The first sentence listed is the most representative sentence for a topic; it is the sentence with the highest loading θ_i in its topic mix vector. The second sentence

[11] Stemming is similar to the lemmatization algorithm discussed in (Section 2.2.1). Lemmatization is guaranteed to deconjugate to a word, which is useful when interpreting LDA topics. Stemming can be more aggressive than lemmatization in deconjugation, often normalizing words into fragments. For comparison, the stem of 'conjugation' is 'conjug', while its lemma is 'conjugate.' Stemming's extra deconjugation can help in matching problem such as this.

is the next most representative sentence, and so on. In some cases, proxy statement descriptions of information gathering and decision making responsibilities do not change materially from one fiscal year to the next. This can result in repeated sentences in our corpus. Hence, we require that all representative sentences are sufficiently different from those that precede it. We define two sentences as different if they have less than 85% of their lemmatized tokens in common. Thus, for example, the third sentence listed in a panel is the sentence with the highest topic weight of all sentences that share less than 85% of lemmatized tokens with each of the first two sentences.

The results suggest our procedure produces reasonable topics. Panel A looks at the most *coherent* topic. The ten most frequent word lemmas are stock, grant, option, share, award, number, exercise, date, price, and restrict. These lemmas clearly focus on the compensation decision-making authority of the board or a committee, and this interpretation is reinforced by the representative sentences listed. Panel B's topic is about audit information gathering, specifically when board members meets with independent auditors. Keywords include financial, statement, management, report, review, internal auditor, control, and independent. The next seven most coherent topics (Panels C through I) also discuss information gathering and decision-making responsibilities. These include the determination of stock option awards terms, the identification of director candidates, the pre-approval of independent auditor services, strategies to maximize firm value, the development and review of corporate governance practices, the appointment of independent auditors, and the recommendation of independent auditor reports. The tenth most *coherent* topic in Panel J does not appear to provide a consistent theme and is ignored in our analysis.

Overall, these results suggest that our natural language processing approach works well. We are able to identify themes related to information gathering and decision-making by boards and committees. And, this is done without biasing the results through ex ante keyword selection or naive inclusion of committee names.

3. Computational notes

It is important in computational contexts such as this to identify the most appropriate way to solve a problem. The grammatical dependency graph search works best in a language suited for high level abstractions. k-NN is simple to implement, but is numerically intensive. *doc2vec* and LDA fit within an entire machine learning ecosystem of neural networks and Bayesian inference.

This section presents brief application-specific notes on our programming decision-making process. We did our best to choose languages and paradigms that allowed us to implement custom libraries correctly, robustly, and with high performance. One common concern throughout the analysis was maximizing computational resources. All code was multi-core to improve efficiency. We also used a grid engine to run multiples jobs in parallel. The analysis would have required years of computer time without these decisions. Please contact the authors with any questions, and code may be provided on request.

3.1. Grammatical dependency graph search

We opted to write our own embedded domain specific language in Haskell for dependency graph search. Haskell is a functional programming language. Functional programming in this context is not a language with functions. Instead, the term refers to languages in which functions can easily manipulate other functions. This feature allows for high-level abstractions through which complex algorithms may be elegantly expressed.

Our library is based on combinators, simple programming units that may be combined to form complex ones. The base unit takes the current progress of a search and looks one step forward. Combining these basic units allows us to implement any search algorithm. We leverage a Haskell abstraction (alternative functors) that allows the base unit to encompass algorithms with choice among grammars (e.g. evaluating multiple options) and words that may or may not be present. A second abstraction (the list monad transformer) helps the algorithm handle cases with zero, one, or more successful search results. Thus, we can pick up data when a proxy statement combines all information about board and committee meeting frequency in a single sentence. The resulting code allows dependency search algorithms to look nearly identical to their underlying structure.

3.2. doc2vec and latent Dirichlet allocation

These machine learning algorithms are very popular, with numerous implementations in the public domain. We use *gensim* (Řehůřek and Sojka 2010), a Python library, that provides a clean application programmer interface. *Gensim* contains multi-core implementations with the bulk of numerical code implemented in C. As a result, the code runs efficiently.

3.3. k*-nearest neighbors*

While many *k*-NN implementations exist in the public domain, we chose to write our own to ensure computational efficiency. The PV-DBOW partitioner grid-search required computing approximately 4.2 trillion vector distances (an additional 4.2 trillion vector distances were computed for the PV-DM benchmark). Hence, code efficiency was critical.

We use Rust, a modern systems-level programming language with performance equivalent to C for our custom *k*-NN algorithm.[12] Rust permits more straightforward multi-core programming and memory management than C. The language's type system, focus on iterations, and embrace of some functional programming paradigms allows for data-centric code design, which we found particularly conducive for writing concise and efficient numerical code.

4. Variable definitions

4.1. Board characteristics

Board Meetings is the number of regular (non-telephonic) meetings held by the Board of Directors in a fiscal year as reported by the firm in the Definitive Proxy Statement (SEC Form DEF 14A) filed with the U.S. Securities and Exchange Commission. *(Source: SEC Edgar)*

[12] Rust has several *k*-NN implementations in its crates.io package repository. However, using our own implementation allowed us to tailor the code to the application and maximize throughput.

Board Size is the number of directors on the Board. *(Source: BoardEx and ISS)*

Committee Meetings (Average) is the average number of regular (non-telephonic) meetings the Board's directors held each year as reported by the firm in its Definite Proxy Statement. This is computed for each director first and then averaged over all directors. *(Source: SEC Edgar)*

Committee Meetings (Total) is the total number of regular (non-telephonic) meetings held by all the Board's committees. *(Source: SEC Edgar)*

LDA-based Outsider-Only Fraction measures the fraction of stated information gathering and decision-making responsibilities that are allocated to outsider-only committees. In our corpus of board or committee sentences, we identify nine topics using LDA that relate to information gathering and decision-making responsibilities. For each director, we form a set of corpus sentences containing all sentences related to the board and all sentences pertaining to committees on which the director is a member. Each sentence receives a weight equal to the sum of that sentence's information gathering and decision-making topic probabilities. The director-level LDA-based OOF is the fraction of total information gathering and decision-making sentence topic weights that apply to outsider-only committees. We average the director-level LDA-based OOF over all board members to compute the board-level characteristic. *(Source: SEC Edgar)*

Meeting-based Outsider-Only Fraction is a board-level average of the fraction of total annual meetings (board and committees) directors have in committees composed entirely of outside directors. It is first computed for each director and then averaged over all directors to derive the firm-year average. *(Source: SEC Edgar)*

Non-SOX Targeted is a time-constant indicator variable that takes the value of 0 if the firm had a majority of outside directors and had fully-outsider committees for audit, corporate governance/director nominating, and executive compensation as of the fiscal year end immediately preceding SOX (i.e. 2001 fiscal year end), and 1 otherwise. *(Source: BoardEx, ISS, and SEC Edgar)*

Number of Committees is the number of standing committees of the Board. *(Source: BoardEx and SEC Edgar)*

4.2. Director characteristics

Age is the director's age in years. *(Source: BoardEx/ISS)*

Education is the maximum educational qualification a director has earned. It is defined as 3 for directors whose maximum achievement is a Ph.D., 2 for directors with a Masters, 1 for directors that earned a bachelors degree, and 0 otherwise. *(Source: BoardEx)*

Female is an indicator variable that takes the value of 1 if the director is female, and 0 otherwise. *(Source: BoardEx/ISS)*

Number of Private Boards counts the private company boards on which the firm's director serves. *(Source: BoardEx)*

Number of Public Boards counts the publicly listed company boards (excluding the current firm) on which the firm's director serves. *(Source: BoardEx/ISS)*

Tenure is the number of years between the fiscal year-end and the date the director was appointed to the Board. *(Source: BoardEx/ISS)*

Missing characteristic dummy variables are a set of dummy variables (one for each characteristic above) that take the value of 1 if the characteristic is missing in ISS, and 0 otherwise.

4.3. Firm characteristics

Assets is the book value of total assets in billions of dollars. *(Source: Compustat)*

Book Leverage is the total book value of long-term and current debt normalized by the book value of total assets. *(Source: Compustat)*

Firm Age is the number of years between the firm's current fiscal year-end date and its first fiscal year-end date available in Compustat. *(Source: Compustat)*

Number of Analysts counts analysts that provided active Earnings per Share (EPS) forecasts for a fiscal year-end. Active forecasts are those released no more than 300 days before the company's actual EPS announcement. *(Source: I/B/E/S)*

Number of Employees provides employees (in thousands) as reported to shareholders. *(Source: Compustat)*

Number of Segments counts business segments reported in the Compustat segments database. *(Source: Compustat)*

Research and Development is the total costs incurred over the fiscal year in the research and development of new products, normalized by the book value of total assets at the start of the fiscal year. *(Source: Compustat)*

Stock Return is the total compound stock return (including dividends) over the fiscal year. *(Source: CRSP)*

Stock Volatility is the annualized standard deviation of daily stock return over the fiscal year. *(Source: CRSP)*

Tobin's **q** is the market-to-book ratio of asset, where the market value of assets is the market value of common equity plus the book value of total assets less the book value of common equity. *(Source: Compustat)*

4.4. Trade characteristics

Book to Market is the ratio of the company's book value of common equity to its market value of common equity, measured as of the fiscal year-end before the trade. *(Source: Compustat)*

Buy and Hold Return is measured over the six-month period beginning the month after relevant insider or outside directors cumulatively executed a net purchase of shares. Abnormal returns are measured relative to six matched Fama-French portfolios formed as the intersection of two size (small and big) and three book equity to market equity (value, neutral, and growth) portfolios. *(Source: CRSP/Thomson Reuters)*

Cumulative Abnormal Return is measured over the two-day period covering the day the trade was received by the SEC and the following trading day. Abnormal returns are measured relative to six matched Fama-French portfolios formed as the intersection of two size (small and big) and three book equity to market equity (value, neutral, and growth) portfolios. *(Source: CRSP/Thomson Reuters)*

Filing Frequency is the number of days over the preceding July through June year in which the director had a net purchase of shares. *(Source: Thomson Reuters)*

Market Capitalization is the market value of common equity in billions of dollars, measured as of the fiscal year-end before the trade. *(Source: Compustat)*

Strong Buy is the number of directors that purchased more shares than they sold on the trading day. *(Source: Thomson Reuters)*

Trade Size is the cumulative net purchase of shares by a company's director on the trading day as a fraction of the firm's number of shares outstanding. *(Source: CRSP/Thomson Reuters)*

4.5. Acquisition characteristics

All Cash Deal is an indicator variable that takes the value of 1 if the deal is entirely in cash, and 0 otherwise. *(Source: SDC Platinum)*

Announcement CAR (Market Adjusted) is the cumulative five-day abnormal stock return over the CRSP equally-weighted return. The five-day window begins two trading days before and ends two trading days after the deal's announcement date. *(Source: CRSP)*

Announcement CAR (Market Model Adjusted) is the cumulative five-day adjusted stock return over the market model, where the market model is estimated from 230 days to 11 days before the announcement using the CRSP equally-weighted return as the market index. The five-day window begins two trading days before and ends two trading days after the deal's announcement date. *(Source: CRSP)*

Cash Flow is total income before extraordinary items, depreciation, and amortization, normalized by the book value of total assets at the start of the fiscal year. *(Source: Compustat)*

Diversifying Acquisition is an indicator variable that takes the value of 0 if the bidder and target are in the same 48 industry groups as defined by Fama and French (1997), and 1 otherwise. *(Source: SDC Platinum)*

High Tech Deal is an indicator variable that takes the value of 1 if the acquirer and target are both in high tech industries defined by Loughran and Ritter (2004), and 0 otherwise. *(Source: SDC Platinum)*

Hostile Deal is an indicator variable that takes the value of 1 if the bid is classified as hostile, and 0 otherwise. *(Source: SDC Platinum)*

Private Target is an indicator variable that takes the value of 1 if the target company is privately held, and 0 otherwise. *(Source: SDC Platinum)*

Public Target is an indicator variable that takes the value of 1 if the target company is publicly listed, and 0 otherwise. *(Source: SDC Platinum)*

Relative Deal Size is the ratio of the deal value to the acquirer's market capitalization of equity (measured as of the close of the 11th trading day before the deal announcement). *(Source: SDC Platinum)*

Stock Deal is an indicator variable that takes the value of 1 if the deal includes any stock, and 0 otherwise. *(Source: SDC Platinum)*

Stock Runup is the acquirer's buy-and-hold abnormal return relative to the CRSP equally-weighted index beginning 210 trading days before and ending 11 trading days before the deal's announcement. *(Source: CRSP)*

Tender Offer is an indicator variable that takes the value of 1 if the acquisition is a tender offer, and 0 otherwise. *(Source: SDC Platinum)*

Transaction Value is the value of the deal in billions of dollars. *(Source: SDC Platinum)*

References

Bird, S., Klein, E., Loper, E., 2009. Natural language processing with Python. O'Reilly Media, Inc., Sebastopol, CA.

Blei, D. M., Ng, A. Y., Jordan, M. I., 2003. Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., Blei, D. M., 2009. Reading tea leaves: How humans interpret topic models. Advances in Neural Information Processing Systems, 288–296.

Chen, D., Manning, C., 2014. A fast and accurate dependency parser using neural networks. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 740–750.

Fama, E. F., French, K. R., 1997. Industry costs of equity. Journal of Financial Economics 43, 153–193.

Finkel, J. R., Grenager, T., Manning, C., 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL), 363–370.

Kusner, M., Sun, Y., Kolkin, N., Weinberger, K. Q., 2015. From word embeddings to document distances. Internation Conference on Machine Learning, 957–966.

Le, Q., Mikolov, T., 2014. Distributed representations of sentences and documents. International Conference on Machine Learning, 1188–1196.

Loughran, T., Ritter, J., 2004. Why has IPO underpricing changed over time? Financial Management 33, 5–37.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S. J., McClosky, D., 2014. The Stanford CoreNLP natural language processing toolkit. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 55–60.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., McCallum, A., 2011. Optimizing semantic coherence in topic models. International Conference on Machine Learning, 262–272.

Newman, D., Lau, J. H., Grieser, K., Baldwin, T., 2010. Automatic evaluation of topic coherence. Human language technologies: The 2010 annual conference of the north American chapter of the association for computational linguistics, 100–108.

Pelillo, M., 2014. Alhazen and the nearest neighbor rule. Pattern Recognition Letters 38, 34–37.

Porter, M. F., 1980. An algorithm for suffix stripping. Program 14, 130–137.

Porter, M. F. "Snowball: a language for stemming algorithms." http://snowballstem.org (2001).

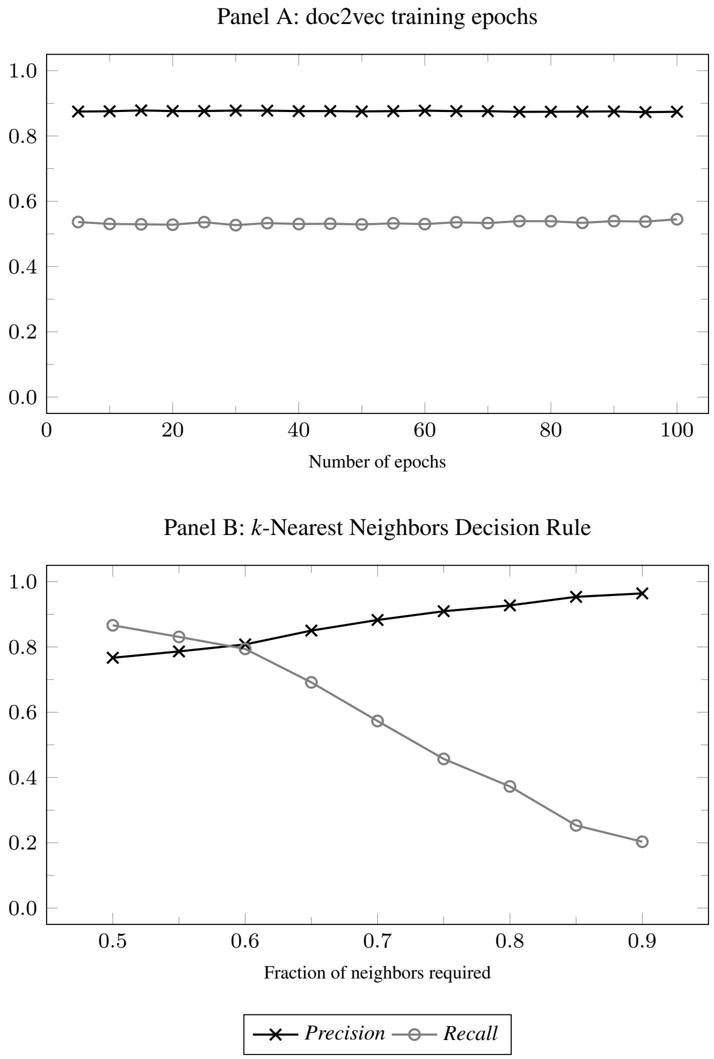
Princeton University. “About WordNet.” *WordNet*. Princeton University. https://wordnet.princeton.edu (2010).

Řehůřek, R., Sojka, P., 2010. Software framework for topic modelling with large corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 45–50.

Figure 1

Sentence partitioning hyperparameter cross-validation: Univariate results

The figure presents results of a grid search that tunes hyperparameters of a partitioning algorithm to remove sentences that do not discuss board or committee information gathering or decision-making from a corpus. We begin by segmenting each sample firm-year's definitive proxy statement (DEF 14A) into sentences and retain only those that mention either the board or a committee. We then translate all sentences into a multi-dimensional semantic vector space using the distributed bag of words of paragraph vector (PV-DBOW) *doc2vec* algorithm (Le and Mikolov 2014). Then, a *k*-Nearest neighbor (*k*-NN) classifier uses a training sample of 10,000 sentences that were manually categorized by the authors as to whether they contained information on information gathering and decision-making. For each point *k*-NN finds the *k* closest points in the *doc2vec* space. Those sentences for which a sufficiently large number of neighbors do not describe board/committee information gathering or decision-making are classified as such. We perform a grid search over 42,120 hyperparameter sets covering all combinations of (i) *doc2vec* semantic dimensions (from 10 to 100 (in increments of 10) and from 150 to 500 (in increments of 50)), (ii) *doc2vec* training epochs (5 to 100 (in increments of 5)), *k*-NN number of nearest neighbors considered (5, 11, 15, 25, 51, 75, 101, 251, 501, 751, 1001, 2501, and 5001), and *k*-NN decision rules (simple majority (>0.50 fraction) and supermajority fractions from 0.55 to 0.90 in increments of 0.90). The figure present two statistics that evaluate the performance of the algorithm using cross-validation of the manually categorized training sample. *Precision* is the fraction of the sentences classified by the algorithm as not related to information gathering or decision-making that are correct. *Recall* is the fractions of the sentences in the evaluation sample that do not relate to information gathering or decision-making that are classified as such by the algorithm. Panel A shows univariate trends for the number of *doc2vec* training epochs. Panel B shows univariate trends for the decision rule. These plots average *precision* and *recall* over all relevant parameter sets in the grid search.



Panel A: doc2vec training epochs

Panel B: *k*-Nearest Neighbors Decision Rule

Table 1

Sentence partitioning hyperparameter cross-validation: Grid search results

The table presents results of a grid search that tunes hyperparameters of a partitioning algorithm to remove sentences that do not discuss board or committee information gathering or decision-making from a corpus. We begin by segmenting each sample firm-year's definitive proxy statement (DEF 14A) into sentences and retain only those that mention either the board or a committee. We then translate all sentences into a multi-dimensional semantic vector space using the distributed bag of words of paragraph vector (PV-DBOW) *doc2vec* algorithm (Le and Mikolov 2014). Then, a *k*-Nearest neighbor (*k*-NN) classifier uses a training sample of 10,000 sentences that were manually categorized by the authors as to whether they contained information on information gathering and decision-making. For each point *k*-NN finds the k closest points in the *doc2vec* space. Those sentences for which a sufficiently large number of neighbors do not describe board/committee information gathering or decision-making are classified as such. We perform a grid search over 42,120 hyperparameter sets covering all combinations of (i) *doc2vec* semantic dimensions (from 10 to 100 (in increments of 10) and from 150 to 500 (in increments of 50)), (ii) *doc2vec* training epochs (5 to 100 (in increments of 5)), *k*-NN number of nearest neighbors considered (5, 11, 15, 25, 51, 75, 101, 251, 501, 751, 1001, 2501, and 5001), and *k*-NN decision rules (simple majority (>0.50 fraction) and supermajority fractions from 0.55 to 0.90 in increments of 0.90). The table present two statistics that evaluate the performance of the algorithm using cross-validation of the manually categorized training sample. *Precision* is the fraction of the sentences classified by the algorithm as not related to information gathering or decision-making that are correct. *Recall* is the fractions of the sentences in the evaluation sample that do not relate to information gathering or decision-making that are classified as such by the algorithm. The table presents statistics using 40 *doc2vec* training epochs and a 0.75 supermajority decision rule. Panel A shows results when 75 or fewer nearest neighbors are used for classification. Panel B shows results when 101 or more nearest neighbors are used.

Panel A

	$k = 5$		$k = 11$		$k = 15$		$k = 25$		$k = 51$		$k = 75$	
doc2vec Dimensions	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
10	0.8521	0.6583	0.8931	0.5694	0.8956	0.5667	0.8920	0.6050	0.9033	0.5756	0.9009	0.5727
20	0.8566	0.6755	0.8955	0.5803	0.8984	0.5815	0.8953	0.6095	0.9079	0.5712	0.9092	0.5615
30	0.8689	0.6642	0.9029	0.5782	0.9107	0.5797	0.9064	0.6093	0.9138	0.5747	0.9152	0.5680
40	0.8736	0.6543	0.9112	0.5503	0.9096	0.5510	0.9115	0.5849	0.9239	0.5516	0.9220	0.5383
50	0.8641	0.6649	0.9067	0.5788	0.9074	0.5806	0.9006	0.6108	0.9122	0.5735	0.9119	0.5689
60	0.8607	0.6650	0.9083	0.5627	0.9045	0.5630	0.9012	0.6057	0.9090	0.5662	0.9102	0.5644
70	0.8666	0.6566	0.9050	0.5557	0.9076	0.5544	0.9003	0.5841	0.9070	0.5541	0.9116	0.5515
80	0.8728	0.6557	0.9064	0.5657	0.9090	0.5645	0.9002	0.5925	0.9142	0.5548	0.9139	0.5472
90	0.8728	0.6400	0.9079	0.5469	0.9067	0.5489	0.8973	0.5806	0.9122	0.5377	0.9127	0.5319
100	0.8531	0.6423	0.9021	0.5534	0.9039	0.5513	0.8990	0.5843	0.9087	0.5457	0.9090	0.5403
150	0.8682	0.6445	0.9114	0.5326	0.9139	0.5344	0.9088	0.5652	0.9150	0.5201	0.9241	0.5104
200	0.8708	0.6553	0.9125	0.5480	0.9121	0.5448	0.9083	0.5717	0.9153	0.5359	0.9194	0.5287
250	0.8655	0.6236	0.9014	0.5199	0.9042	0.5182	0.9007	0.5521	0.9168	0.5128	0.9134	0.4986
300	0.8631	0.6336	0.9020	0.5412	0.9093	0.5442	0.9010	0.5691	0.9198	0.5239	0.9214	0.5144
350	0.8599	0.6306	0.8972	0.5317	0.9068	0.5268	0.9003	0.5708	0.9232	0.5339	0.9193	0.5147
400	0.8688	0.6290	0.9069	0.5333	0.9031	0.5269	0.9059	0.5578	0.9225	0.5113	0.9277	0.4927
450	0.8665	0.6303	0.8989	0.5180	0.9087	0.5174	0.9006	0.5522	0.9228	0.5109	0.9187	0.4935
500	0.8629	0.6272	0.8996	0.5176	0.9056	0.5143	0.9050	0.5334	0.9223	0.4962	0.9274	0.4966

Table 1

Sentence partitioning hyperparameter cross-validation: Grid search results *(continued)*

	Panel B											
	$k = 101$		$k = 251$		$k = 501$		$k = 751$		$k = 1001$		$k = 2501$	
doc2vec Dimensions	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
10	0.8966	0.5786	0.8975	0.5642	0.8908	0.5560	0.8887	0.5341	0.8947	0.5268	0.9058	0.3910
20	0.9082	0.5612	0.9123	0.5310	0.9098	0.5181	0.8930	0.4824	0.9086	0.4896	0.9201	0.3405
30	0.9162	0.5740	0.9126	0.5464	0.9051	0.5344	0.8982	0.4981	0.9102	0.4998	0.9204	0.3275
40	0.9196	0.5448	0.9216	0.5124	0.9115	0.4997	0.8927	0.4915	0.9103	0.4745	0.9189	0.3057
50	0.9098	0.5694	0.9074	0.5328	0.9050	0.5196	0.8998	0.4969	0.9075	0.4861	0.9168	0.2762
60	0.9067	0.5699	0.9127	0.5378	0.9064	0.5252	0.9017	0.5039	0.9039	0.4943	0.9105	0.3052
70	0.9085	0.5555	0.9118	0.5259	0.9050	0.5135	0.8923	0.5046	0.9085	0.4822	0.9150	0.3094
80	0.9057	0.5437	0.9091	0.4994	0.9092	0.4709	0.9155	0.4601	0.9098	0.4288	0.9328	0.2407
90	0.9113	0.5330	0.9115	0.4895	0.9105	0.4749	0.9139	0.4630	0.9092	0.4352	0.9258	0.2499
100	0.9072	0.5434	0.9146	0.5020	0.9077	0.4817	0.9170	0.4661	0.9050	0.4415	0.9191	0.2562
150	0.9206	0.5128	0.9321	0.4655	0.9309	0.4484	0.9201	0.4297	0.9250	0.3996	0.9468	0.1851
200	0.9060	0.5212	0.9216	0.4698	0.9190	0.4416	0.9142	0.4014	0.9254	0.3760	0.9371	0.1455
250	0.9254	0.4955	0.9346	0.4451	0.9300	0.4160	0.9289	0.4021	0.9246	0.3626	0.9428	0.1489
300	0.9178	0.5154	0.9273	0.4721	0.9306	0.4407	0.9283	0.4173	0.9302	0.3812	0.9415	0.1181
350	0.9174	0.5158	0.9239	0.4652	0.9299	0.4366	0.9315	0.4058	0.9399	0.3766	0.9430	0.1423
400	0.9247	0.4879	0.9294	0.4329	0.9359	0.3971	0.9256	0.3522	0.9376	0.3165	0.9322	0.0659
450	0.9179	0.4997	0.9261	0.4468	0.9252	0.4141	0.9348	0.3729	0.9354	0.3325	0.9269	0.0861
500	0.9243	0.4895	0.9347	0.4417	0.9361	0.4062	0.9421	0.3697	0.9427	0.3256	0.9431	0.0918

Table 2
Latent dirichlet allocation: Topic coherence

The table presents *coherence* statistics on topics identified through Latent Dirichlet Allocation (LDA) analysis of company definitive proxy statements (DEF 14A) filed with the U.S. Securities and Exchange Commission (SEC). We segment each proxy statement into sentences and retain only those that mention either the board or a committee. These sentences are then translated into a vectorized semantic representation using the distributed bag of words of paragraph vector (PV-DBOW) *doc2vec* algorithm (Le and Mikolov 2014). A k-nearest neighbor classifier uses an author-supplied training sample to remove sentences unrelated to information gathering or decision-making by boards or committees. The excluded sentences typically relate to descriptive characteristics, such as board or committee membership. The remaining sentences are then processed as follows. We remove stop words and committee names identified via Stanford CoreNLP dependency analysis (Chen and Manning 2014). A committee's name may describe the organization's purpose and, thereby, result in topic identification based on name, not responsibilities. All remaining words are standardized through lemmatization, a process that deconjugates words to a root form. Finally, LDA topic analysis is performed on the processed sentences. For each topic, we compute *coherence* a numeric metric shown to correlate with human understanding (Mimno, Wallach, Talley, Leenders, and McCallum 2011). We examine LDA models using 10 to 100 total topics, in increments of 10. Columns (1) through (5) presents the average *coherence* of the 10, 20, 30, 40, and 50 topics with the highest *coherence* values, respectively. The average *coherence* of all the topics modeled is shown in column (6).

	n-Most Coherent Topics					All
	n=10	n=20	n=30	n=40	n=50	Topics
Total Topics Modeled	(1)	(2)	(3)	(4)	(5)	(6)
10	0.6947	-	-	-	-	0.6947
20	0.7487	0.6721	-	-	-	0.6721
30	0.7531	0.6993	0.6300	-	-	0.6300
40	0.7511	0.7091	0.6621	0.6068	-	0.6068
50	0.7477	0.6802	0.6351	0.5907	0.5446	0.5446
60	0.7369	0.6730	0.6303	0.5913	0.5549	0.5133
70	0.7391	0.6582	0.6089	0.5695	0.5331	0.4669
80	0.6943	0.6488	0.6100	0.5795	0.5491	0.4686
90	0.6888	0.6354	0.5872	0.5489	0.5183	0.4273
100	0.7034	0.6305	0.5808	0.5427	0.5160	0.4090

Table 3
Latent dirichlet allocation: Descriptive topic data

The table presents descriptive information on topics identified through Latent Dirichlet Allocation (LDA) analysis of company definitive proxy statements (DEF 14A) filed with the U.S. Securities and Exchange Commission (SEC). We segment each proxy statement into sentences and retain only those that mention either the board or a committee. These sentences are then translated into a vectorized semantic representation using the distributed bag of words of paragraph vector (PV-DBOW) *doc2vec* algorithm (Le and Mikolov 2014). A k-nearest neighbor classifier uses an author-supplied training sample to remove sentences unrelated to information gathering or decision-making by boards or committees. The excluded sentences typically relate to descriptive characteristics, such as board or committee membership. The remaining sentences are then processed as follows. We remove stop words and committee names identified via Stanford CoreNLP dependency analysis (Chen and Manning 2014). A committee's name may describe the organization's purpose and, thereby, result in topic identification based on name, not responsibilities. All remaining words are standardized through lemmatization, a process that deconjugates words to a root form. Finally, LDA topic analysis is performed on the processed sentences to identify 30 topics. In Panels A through I, we present, in descending order, the 10 topics with the highest *coherence*, a numerical metric shown to correlate with human understandability (Mimno, Wallach, Talley, Leenders, and McCallum 2011). For each topic, we provide a description of the topic (as determined by the authors), the *coherence* score, the ten most relevant topic keywords determined by LDA (in decreasing order), and five representative sample sentences. Sentences are the most representative sentences for the topic as determined by LDA posteriors provided each has fewer than 85% of lemmatized tokens in common with preceding sentences.

Table 3
Latent dirichlet allocation: Descriptive topic data *(continued)*

Panel A

Description:	Awarding of stock and option grants
Coherence:	0.8408
Keywords:	stock, grant, option, share, award, number, exercise, date, price, and restrict

SAMPLE SENTENCES

Rank	Score	Text
1	0.968817	The price of shares of the Company's Common Stock subject to each option (the ”option price”) is set by the Committee but may not be less than 50% of the fair market value on the date of grant in the case of an option that is not an incentive stock option (a nonqualified stock option”), and not less than 100% of the fair market value in the case of an incentive stock option. *Source: https://www.sec.gov/Archives/edgar/data/890662/0000950149-98-001498.txt*
2	0.965476	The Stock Option Committee will determine the number of shares of Common Stock issuable pursuant to each stock option and the exercise or purchase price per share of each stock option, but the exercise price may not be less than 100% of the fair market value of the Common Stock on the date of the grant. *Source: https://www.sec.gov/Archives/edgar/data/103730/0000922423-98-000410.txt*
3	0.964197	The Compensation Committee may grant NQSOs with an exercise price less than the fair market value of a share of common stock on the date the option is granted, provided that the exercise price per share is not less than 85% of the fair market value of our common stock on the date of grant. *Source: https://www.sec.gov/Archives/edgar/data/1096376/0001193125-06-073793.txt*
4	0.964197	The Amended Plan provides for the grant of ISOs and nonqualified options, and that the Committee will determine the number of Common Shares subject to an option, the exercise period of an option, the purchase price per Common Share subject to an option and the other terms of the option, provided that the purchase price per Common Share is not less than 100% of the fair market value of such Common Share on the date of grant of the option. *Source: https://www.sec.gov/Archives/edgar/data/821130/0001047469-09-004158.txt*
5	0.961333	The Board or the Stock Option Committee, as 13 the case may be, has the discretion to determine the eligible employees to whom, and the prices (not less than the fair market value on the date of grant) at which options will be granted; the periods during which each option is exercisable; and the number of shares subject to each option. *Source: https://www.sec.gov/Archives/edgar/data/792641/0000930413-07-005163.txt*

Table 3
Latent dirichlet allocation: Descriptive topic data *(continued)*

Panel B

Description:	Review of independent auditor reports
Coherence:	0.8245
Keywords:	financial, statement, management, report, review, internal, auditor, audit, control, and independent

SAMPLE SENTENCES

Rank	Score	Text
1	0.959722	The Committee meets with the Companys management and Ernst & Young LLP, with and without management present, to discuss the results of their examinations, Ernst & Young LLPs evaluations of the Companys internal control, including internal control over financial reporting, and the overall quality of the Companys financial reporting, and the overall quality of the Companys financial reporting. *Source: https://www.sec.gov/Archives/edgar/data/921506/0001193125-06-095218.txt*
2	0.957971	The Committee met with the internal auditor and the independent auditor, with and without management present, to discuss the results of their audits; their evaluations of the Companys internal control, including internal control over financial reporting; and the overall quality of the Companys financial reporting. *Source: https://www.sec.gov/Archives/edgar/data/85961/0000950144-08-002187.txt*
3	0.956060	With and without management present, the Committee discussed and reviewed the results of the independent auditors examination of the Companys financial statements and internal control over financial reporting, as well as managements report on internal control over financial reporting. *Source: https://www.sec.gov/Archives/edgar/data/26172/0001104659-07-026086.txt*
4	0.956060	The audit committee meets with the independent auditor, with and without management present, to discuss the results of the independent auditors examinations, its evaluation of Mariners internal control over financial reporting and the overall quality of Mariners financial reporting. *Source: https://www.sec.gov/Archives/edgar/data/1022345/0000950129-09-001159.txt*
5	0.953968	The Committee meets with the independent auditor, with and without management present, to discuss the results of their examination, their evaluation of Granites internal controls, including internal control over financial reporting, and the overall quality of Granites financial reporting. *Source: https://www.sec.gov/Archives/edgar/data/861459/0000950134-06-006653.txt*

Table 3
Latent dirichlet allocation: Descriptive topic data *(continued)*

Panel C

Description: Terms of stock option awards
Coherence: 0.7894
Keywords: shall, time, may, determine, term, deem, condition, appropriate, necessary, and subject

SAMPLE SENTENCES

Rank	Score	Text
1	0.919444	Except as otherwise provided herein, Stock Options shall be exercisable at such time or times and subject to such terms and conditions as shall be determined by the Committee. *Source: https://www.sec.gov/Archives/edgar/data/1044590/0001047469-99-011933.txt*
2	0.919444	All SARs shall be in such form as the Committee may from time to time determine and shall be subject to the following terms and conditions *Source: https://www.sec.gov/Archives/edgar/data/790818/0000926044-03-000102.txt*
3	0.919444	In such an event, no payment shall be made unless the Committee shall have been furnished with such evidence as the Committee may deem necessary to establish the validity of the payment. *Source: https://www.sec.gov/Archives/edgar/data/103730/0001206774-04-000294.txt*
4	0.912121	Stock awards may be subject to other terms and conditions, which may very from time to time and among Participants, as the Compensation Committee determines to be appropriate. *Source: https://www.sec.gov/Archives/edgar/data/1003344/0001193125-10-004434.txt*
5	0.912121	The Committee shall determine the terms and conditions of such Awards and such terms and conditions shall be contained in an Award Agreement which evidences such Award. *Source: https://www.sec.gov/Archives/edgar/data/1388195/0001193125-10-091707.txt*

Table 3
Latent dirichlet allocation: Descriptive topic data *(continued)*

Panel D		
Description:	Identification of director candidates	
Coherence:	0.7597	
Keywords:	director, board, candidate, nominate, nominee, recommend, committee, election, consider, and member	
SAMPLE SENTENCES		
Rank	Score	Text
1	0.943137	The Nominating Committee is responsible for identifying individuals qualified to become directors of Rentrak and recommending to the board of directors candidates for election and for recommending individuals to serve on each board committee. *Source: https://www.sec.gov/Archives/edgar/data/800458/0001193125-09-150188.txt*
2	0.943137	The committee also serves as the Boards nominating committee, responsible for identifying and recommending individuals qualified to become Board members and for evaluating directors being considered for re-election. *Source: https://www.sec.gov/Archives/edgar/data/51434/0001193125-10-076750.txt*
3	0.935555	Our nominating committee identifies individuals qualified to become members of the board and recommends to our board of directors nominees for election as directors. *Source: https://www.sec.gov/Archives/edgar/data/1386198/0001437749-11-003609.txt*
4	0.935555	The Committee is also responsible for identifying and evaluating individuals qualified to become Board members and recommending to the Board candidates to stand for election or re-election as directors. *Source: https://www.sec.gov/Archives/edgar/data/70318/0001047469-06-004476.txt*
5	0.930952	The Nominating and Corporate Governance Committee is responsible for screening potential director candidates and recommending qualified candidates to the Board for nomination. *Source: https://www.sec.gov/Archives/edgar/data/1044435/0001193125-08-080882.txt*

Table 3
Latent dirichlet allocation: Descriptive topic data *(continued)*

Panel E

Description:	Pre-approval of independent auditor services
Coherence:	0.7544
Keywords:	service, audit, approve, pre, provide, approval, fee, auditor, independent, and non

SAMPLE SENTENCES

Rank	Score	Text
1	0.965476	The audit committees policy is to pre-approve all audit and permissible non-audit services provided by the independent auditors. *Source: https://www.sec.gov/Archives/edgar/data/23194/0000023194-04-000021.txt*
2	0.964197	The audit committee pre-approves all audit and permissible non-audit services provided by the independent auditors; these services may include audit services, audit related services, tax services and other services. *Source: https://www.sec.gov/Archives/edgar/data/805326/0001206774-04-000345.txt*
3	0.964197	The Audit Committee pre-approves all audit and non-audit services provided by the independent accountants prior to the engagement of the independent accountants with respect to such services. *Source: https://www.sec.gov/Archives/edgar/data/771266/0001193125-04-044712.txt*
4	0.961333	The Corporation's Audit Committee adopted a policy for engaging its independent auditor, BDO, for audit and non-audit services that includes requirements for the Audit Committee to pre-approve audit and non-audit services provided by the independent auditor. *Source: https://www.sec.gov/Archives/edgar/data/881468/0001144204-04-007372.txt*
5	0.961333	The audit committee has adopted a policy that requires the audit committee to pre-approve all audit, audit-related, and permissible non-audit services performed by the external auditor. *Source: https://www.sec.gov/Archives/edgar/data/768251/0000950123-10-027255.txt*

Table 3
Latent dirichlet allocation: Descriptive topic data *(continued)*

Panel F

Description: Strategies to maximize firm value
Coherence: 0.7288
Keywords: term, long, stockholder, interest, believe, incentive, retain, good, value, and shareholder

SAMPLE SENTENCES

Rank	Score	Text
1	0.903333	The Board believes that it is in the best interests of the Company and its shareholders to provide long-term incentives to its employees. *Source: https://www.sec.gov/Archives/edgar/data/932111/0001188112-05-000768.txt*
2	0.903331	The Compensation Committee also believes that Mr. Swett's 9.2% ownership of the Common Stock provides Mr. Swett with a significant incentive to increase values for all of the Company's stockholders. *Source: https://www.sec.gov/Archives/edgar/data/1009532/0000892569-98-001027.txt*
3	0.892592	The Board of Directors believes that Kimberly-Clark's takeover defenses are in the best short-term and long-term interests of the Corporation and its stockholders. *Source: https://www.sec.gov/Archives/edgar/data/55785/0000950134-03-003644.txt*
4	0.892592	The Committee believes that both of its Option Plans align the interests of the employees with the long-term interests of the stockholders. *Source: https://www.sec.gov/Archives/edgar/data/1002225/0001047469-06-005358.txt*
5	0.892592	The Compensation Committee believes that employees who are owners of Allegiance will focus on its long-term success and on building stockholder value. *Source: https://www.sec.gov/Archives/edgar/data/1058703/0000950134-01-502661.txt*

Table 3
Latent dirichlet allocation: Descriptive topic data *(continued)*

Panel G

Description: Development and review of corporate governance practices
Coherence: 0.7280
Keywords: corporate, governance, committee, charter, process, oversee, review, conduct, adopt, and practice

SAMPLE SENTENCES

Rank	Score	Text
1	0.930952	The Corporate Governance and Nominating Committee reviews, as appropriate, the corporate governance practices and related governance structures of the Company, including the annual review of the Corporate Governance Guidelines, the charters of each Board committee and the Companys Code of Conduct. *Source: https://www.sec.gov/Archives/edgar/data/1005414/0001193125-04-061998.txt*
2	0.919444	The Corporate Governance and Nominating Committee is responsible for developing our corporate governance policies and procedures, and for recommending those policies and procedures to the Board for adoption. *Source: https://www.sec.gov/Archives/edgar/data/1158463/0000950123-11-035723.txt*
3	0.919443	In addition, this committee is responsible for reviewing the Companys corporate governance processes and policies and recommending changes as appropriate. *Source: https://www.sec.gov/Archives/edgar/data/837465/0000936392-03-000534.txt*
4	0.912121	Oversees other corporate governance matters including the evaluation of the functioning of the Board and recommends corporate governance principles. *Source: https://www.sec.gov/Archives/edgar/data/5187/0001193125-05-052173.txt*
5	0.912120	The Board of Directors has adopted a set of corporate governance principles as a framework for the governance of the Company. *Source: https://www.sec.gov/Archives/edgar/data/1265888/0000950133-08-001683.txt*

Table 3
Latent dirichlet allocation: Descriptive topic data *(continued)*

Panel H

Description: Appointment of independent auditors
Coherence: 0.7234
Keywords: independent, firm, accounting, public, register, auditor, accountant, appointment, selection, and ratification

SAMPLE SENTENCES

Rank	Score	Text
1	0.951666	BDO Seidman, LLP served as our independent registered public accounting firm for 2006 and 2007 and has been appointed by our audit committee to serve as our independent registered public accounting firm for 2008. *Source: https://www.sec.gov/Archives/edgar/data/1041954/0001193125-08-081724.txt*
2	0.949122	The Audit Committee is recommending ratification of its appointment of KPMG LLP, which served as our independent registered public accounting firm in 2004, to serve as our independent registered public accounting firm for 2005. *Source: https://www.sec.gov/Archives/edgar/data/878079/0000893220-05-000990.txt*
3	0.946296	Deloitte & Touche LLP served as our independent registered public accounting firm for 2007, and our Audit Committee has selected Deloitte & Touch LLP to serve as our independent registered public accounting firm for 2008. *Source: https://www.sec.gov/Archives/edgar/data/854709/0000950152-08-002231.txt*
4	0.943137	Subject to stockholder ratification, our Audit Committee has appointed Grant Thornton LLP to serve as independent registered public accounting firm for 2005. *Source: https://www.sec.gov/Archives/edgar/data/1089542/0000950144-05-005795.txt*
5	0.943137	Ernst & Young LLP served as the Companys independent registered public accounting firm for 2006 and has been selected by the Audit Committee to serve as the Companys independent registered public accounting firm for 2007. *Source: https://www.sec.gov/Archives/edgar/data/1367396/0001193125-07-050837.txt*

Table 3
Latent dirichlet allocation: Descriptive topic data *(continued)*

Panel I

Description: Recommendation of independent auditor report
Coherence: 0.6983
Keywords: year, fiscal, end, december, discussion, form, statement, recommend, review, and include

Sample Sentences

Rank	Score	Text
1	0.957971	Based on the reviews and discussions referred to above and our review of the Company's audited financial statements for the year ended December 31, 2003, we recommended to the Board that the audited financial statements be included in the Annual Report on Form 10-K for the fiscal year ended December 31, 2003, for filing with the SEC. *Source: https://www.sec.gov/Archives/edgar/data/1005181/0001047469-04-014505.txt*
2	0.949122	Based on its review, our audit committee recommended to our Board of Directors that the audited financial statements for the Company's year ended December 31, 2010 be included in our Annual Report on Form 10-K for its year ended December 31, 2010, which was filed on February 16, 2011. *Source: https://www.sec.gov/Archives/edgar/data/1125376/0001445305-11-000684.txt*
3	0.946296	Based on its review, the Audit Committee recommended to the Board of Directors that the audited financial statements for the Company's fiscal year ended December 31, 2003 be included in the Company's Annual Report on Form 10-K for the Company's fiscal year ended December 31, 2003, for filing with the Securities and Exchange Commission. *Source: https://www.sec.gov/Archives/edgar/data/1085653/0001047469-04-026976.txt*
4	0.946296	Based on these discussions and reviews, the Audit Committee recommended to the Board of Directors (and the Board has approved) that the audited financial statements for the year ended December 31, 2003 be included in the Companys Annual Report on Form 10-K for the year ended December 31, 2003 for filing with the SEC. *Source: https://www.sec.gov/Archives/edgar/data/1111665/0000950133-04-002435.txt*
5	0.946296	Based on the review and discussion referred to above, the audit committee recommended to the Board, and the Board has approved, that the audited financial statements be included in SenoRx's Annual Report on Form 10-K for the fiscal year ended December 31, 2009. *Source: https://www.sec.gov/Archives/edgar/data/1097136/0001019687-10-001616.txt*

Table 3
Latent dirichlet allocation: Descriptive topic data *(continued)*

Panel J

Description:	N/A
Coherence:	0.6840
Keywords:	section, amend, right, respect, outstanding, intend, may, tax, extent, and maximum

SAMPLE SENTENCES

Rank	Score	Text
1	0.806666	On May 5, 2004, the Compensation Committee amended Sections 1.3 and 4.1 and deleted Sections 5.9 and 6.5(e) of the Plan. *Source: https://www.sec.gov/Archives/edgar/data/54480/0000950137-05-004159.txt*
2	0.806666	The Board of Directors may amend or modify the Incentive Plan in any respect. *Source: https://www.sec.gov/Archives/edgar/data/1084755/0001193125-09-010376.txt*
3	0.758333	Subject to the provisions of the Directors Plan, the Board of Directors may amend the Directors Plan. *Source: https://www.sec.gov/Archives/edgar/data/1017259/0001193125-06-091117.txt*
4	0.758327	The Committee has the right to set its own agenda. *Source: https://www.sec.gov/Archives/edgar/data/912513/0001125282-03-002086.txt*
5	0.758327	The Pay Band Target Percents may be modified by the Compensation Committee ”Prior to the Fiscal Year”. *Source: https://www.sec.gov/Archives/edgar/data/78716/0000078716-95-000016.txt*