A Textual-Factor Approach to Measuring Corporate Governance

Lin William Cong, Tengyuan Liang, and Nadya Malenko[1]

Abstract. We apply textual-factor analysis, a latest machine learning tool in natural language processing and AI for social sciences, to extract and summarize information on corporate governance characteristics using unstructured data from regulatory filings such as proxy statements and texts from media/websites such as analyst reports. We then construct the first comprehensive text-based measure of corporate governance and use it to study the most important dimensions of governance and the interaction between governance characteristics. We also analyze the role of management's engagement with shareholders and study how the way the management explains its governance choices affects proxy advisors' recommendations and shareholder voting outcomes.

Motivation. Corporate governance quality is an important characteristic that affects a company's policies and its interactions with financial markets and its investors across many dimensions. Hence, it is very frequently used in empirical research - as the outcome of interest itself, as the key dependent variable, or as a control variable. However, there is no consensus on how corporate governance should be measured. The reason is that governance has several different dimensions – shareholder rights, the composition of the board, protection from takeovers and proxy fights, and the structure of executive compensation.

Accordingly, different papers use very different proxies for corporate governance. A commonly used measure is the G-index (Gompers et al. 2003) or its variation, the E-index (Bebchuk et al. 2009), which measure the number of antitakeover defenses the firm has. While useful, these indices do not capture other important dimensions of governance mentioned above. Other papers use the level of board independence, whether the CEO and chairman positions are combined, the level of institutional ownership, as well as various combinations of all the above.

Given the variety of dimensions and characteristics on which corporate governance can be measured, it is important to understand how these dimensions and characteristics interact with each other and aggregate together. Are there particular dimensions that matter the most? Do certain characteristics become irrelevant if others are in place? And, importantly, how does the way the management explains its governance choices affect investors' and proxy advisors' perception of the firm's governance quality? Indeed, director and management engagement with shareholders on governance topics is becoming an increasingly important feature of the governance landscape, but remains under-explored.

Goals. Our goal is two-fold: (i) to provide guidance on how governance should be measured, which dimensions and characteristics matter the most, and to develop an index of governance across firms; (ii) to understand how management's communication with shareholders affects their perception of governance choices and their voting behavior, as well as proxy advisors' voting recommendations.

We believe that companies' proxy statements (and their amendments and supplements) contain comprehensive information on governance. The text often provides linkages between different dimensions; other sections have management rationalizing their choices and whether they matter. Accordingly, we plan to use a novel natural language processing methodology to analyze these textual documents, in order to answer the questions above.

[1] Cong (will.cong@chicagobooth.edu) and Liang (tengyuan.liang@chicagobooth.edu) are Assistant Professors of Finance and of Econometrics and Statistics respectively at the University of Chicago Booth School of Business; Malenko (nadya.malenko@bc.edu) is Associate Professor of Finance at the Boston College Carroll School of Management.

Methodology and Analysis. Our analysis utilizes the textual-factor framework by Cong and Liang (2018) and Cong et al. (2018). As a pre-dominant form of unstructured data, texts in social communications are informationally rich and offer potential interpretability. Yet prior literature has focused on pre-defining bag of words or simply applying topic models that precede recent developments in natural language processing. The difficulties in analyzing textual data are three-fold: first, language structures are intricate and complex, and representing or summarizing them using simple frequency-based approaches is highly reductive and may lose important informational content; second, textual data are large-scale in nature, and processing and analyzing a large corpus of documents is computationally burdensome; third, there lacks a framework relating textual data to sparse regression analysis traditionally used in social sciences while maintaining interpretability.

We follow Cong et al. (2018) to draw insights and strengths from both neural network models for natural language processing and topic models in statistical machine learning. Specifically, we start with a continuous vector embedding of each word in a large vocabulary using neural networks (word2vec) to construct the semantic and syntactic links of words in the texts. This essentially captures the rich information and complex language structure in texts. We then cluster words based on their vector embedding using a large scale approximate nearest-neighbor search (Local-Sensitive Hashing) --- a recent innovation in theoretical computing that further reduces the dimensionality of our textual universe and groups close vector representations for easy interpretation. Finally, we overlay a topic model to enhance interpretability by describing frequency distributions over textual factors and over the supporting words within each factor.

The textual factors thus generated allow us to (i) project traditional metrics of governance onto interpretable topics/themes; (ii) use seed words to generate richer clusters of words to better capture certain dimensions of governance (a generalized version of the application of word embedding to measure corporate culture in Li et al. 2018); (iii) let data inform us of new governance-relevant themes/factors. We can then construct multi-dimensional (potentially time-varying) measures of a company's governance by projecting the company's documents on the universal textual factors of governance. Once we construct the new measure of governance, we first compare it to existing metrics (such as the G-index or the E-index, as well as other commonly used metrics) to validate it.

Data and Computation. We utilize four main sources of data: (i) proxy statements and amendments from EDGAR related to shareholder meetings; (ii) the ISS Voting Analytics dataset, which contains both aggregate and individual institutions' votes on various proposals on shareholder meeting agendas; (iii) firm outcomes from WRDS; (iv) annual reports, news, and analyst reports that complement the above. In the Appendix, we provide more details on (i) and (ii).

To extract information from (i) and construct textual factors, we need to process large, unstructured, and high-dimensional data efficiently. Our default choice of computation tool is Python, and would require high-performance computing, especially when performing the Local-Sensitive Hashing step.

References

Bebchuk, L., Cohen, A. and Ferrell, A., 2008. What matters in corporate governance? *The Review of Financial Studies*, 22(2), pp.783-827.

Cong, L.W. and Liang, Tengyuan, 2018. Textual Factors: A Scalable, Interpretable, and Data-driven Approach to Analyzing Unstructured Information. *Working Paper.*

Cong, L.W., Liang, Tengyuan, and Zhang, Xiao 2018. Textual Factors and Applications in Finance. *Working Paper.*

Gompers, P., Ishii, J. and Metrick, A., 2003. Corporate governance and equity prices. *The Quarterly Journal of Economics*, 118(1), pp.107-156.

Li, K., Mai, F., Shen, R., and Yan, X., 2018. Measuring Corporate Culture Using Machine Learning. *Available at SSRN 3256608*.

Appendix A. Proxy Statements and Amendments from EDGAR

To measure various dimensions of corporate governance, we extract the proxy statements (DEF 14A) and their amendments filed by firms from the SEC EDGAR database (U.S. Securities and Exchange Commission's Electronic Data Gathering, Analysis, and Retrieval system). The electronic filings on SEC Edgar start in 1994 and provide proxy statements filed between 1994 and 2018. We have 164,158 DEF 14A filings for 22,204 firms during these time periods.

A proxy statement (DEF 14A) is filed in advance of the annual meeting. It should provide shareholders with sufficient information to allow them to make an informed vote at an upcoming meeting or to authorize a proxy to vote on their behalf. Typically, a proxy statement includes information on shareholder proposals, the voting procedure, background information (including potential conflicts of interests) of nominated directors, compensation structure of the board and executives, and information about the auditors. Proposals that are up for shareholder votes include voting for the nominated directors, approval of the executive compensation plan (say-on-pay), proposals submitted by shareholders, and more routine issues, such as the approval of audit fees or the auditing firm.

A recent example of a proposal up for a vote is the $2.6 billion compensation plan awarded to Elon Musk. On Feb 8, 2018, Tesla filed a DEF 14A that describes the details of the plan. The filing starts as follows:

"As Tesla continues to grow, we have created a new 10-year performance award for Elon that incentivizes him to not only continue to lead Tesla over the long-term, but to help the company achieve these great goals. You will find more details about the proposed award in the following pages, but the basic premise is simple — Elon's compensation will be 100% aligned with the interests of our stockholders. Elon will receive no guaranteed compensation of any kind — no salary, no cash bonuses, and no equity that vests by the passage of time. Instead, Elon's only compensation will be a 100% at-risk performance award, which ensures that he will be compensated only if Tesla and all of our stockholders do extraordinarily well. The award consists of stock options that vest only if Tesla achieves specific milestones, which if fully achieved would make Tesla one of the most valuable companies in the world with a market capitalization of at least $650 billion — more than 10x today's value."

Tesla's board of directors recommended the proposed CEO Performance Award, and it was approved by Tesla's shareholders in a special shareholder meeting on March 21, 2018.

Appendix B. Description of Voting Analytics (VA)

The ISS Voting Analytics database covers Russell 3000 firms. For each company and each shareholder meeting of this company, it provides information on all the proposals on the agenda, the number of votes in favor and against the proposal, the number of abstentions, the total number of outstanding shares, the ISS and management recommendations on the proposal, as well as whether the proposal passed or failed.

In addition to the aggregate vote results, the ISS dataset provides information about the votes of individual mutual funds, which have been required, since 2003, to report their votes on form N-PX to the SEC.