ecgi

Cleaning Corporate Governance

Finance Working Paper N° 738/2021
March 2021

Jens Frankenreiter
Columbia University

Cathy Hwang
University of Virginia

Yaron Nili
University of Wisconsin

Eric Talley
Columbia University, Millstein Center for Global Markets and Corporate Ownership, and ECGI

ECGI Working Paper Series in Finance

Cleaning Corporate Governance

Working Paper N° 738/2021
March 2021

Jens Frankenreiter
Cathy Hwang
Yaron Nili
Eric Talley

The authors gratefully acknowledge comments from and discussions with Emiliano Catan, Quinn Curtis, Robert Daines, Steven Davidoff Solomon, Allen Ferrell, Jill Fisch, Jeff Gordon, Michael Klausner, Yair Listokin, Joshua Mitts, and Holger Spamann. The production of this article involved a research team too large to mention in this footnote, and we therefore list them (with our great appreciation) in Appendix A. We also thank the librarians at the Columbia Law School, the University of Virginia School of Law, and the University of Wisconsin Law School. All errors are ours.

Abstract

Although empirical scholarship dominates the field of law and finance, much of it shares a common vulnerability: an abiding faith in the accuracy and integrity of a small, specialized collection of corporate governance data. In this paper, we unveil a novel collection of three decades' worth of corporate charters for thousands of public companies, which shows that this faith is misplaced.

We make three principal contributions to the literature. First, we label our corpus for a variety of firm- and state-level governance features. Doing so reveals significant infirmities within the most well-known corporate governance datasets, including an error rate exceeding eighty percent in the G-Index, the most widely used proxy for "good governance" in law and finance. Correcting these errors substantially weakens one of the most well-known results in law and finance, which associates good governance with higher investment returns. Second, we make our corpus freely available to others, in hope of providing a long-overdue resource for traditional scholars as well as those exploring new frontiers in corporate governance, ranging from machine learning to stakeholder governance to the effects of common ownership. Third, and more broadly, our analysis exposes twin cautionary tales about the critical role of lawyers in empirical research, and the dubious practice of throttling public access to public records.

Keywords: Corporate Governance; G-Index; Corporate Law; Corporate Finance

JEL Classifications: G3, K22

Jens Frankenreiter*
Researcher
Columbia University Law School
435 W 116th St,
New York, NY 10027, USA
phone: +1 (929) 855-4577
e-mail: jgf2132@columbia.edu

Cathy Hwang
Professor of Law
University of Virgina School of Law
580 Massie Road
Charlottesville, VA 22903, USA
phone:+1 (434) 243-8543
chwang@law.virginia.edu

Yaron Nili
Assistant Professor of Law and Smith-Rowe Faculty Fellow in Business Law
University of Wisconsin Law School
975 Bascom Mall,
Madison, WI 53706, USA
phone: +1 (857) 222-9342
e-mail: nili@wisc.edu

Eric Talley*
Isidor and Seville Sulzbacher Professor of Law
Columbia University, Columbia Law School
435 West 116th Street
New York, N.Y. 10027-7297, United States
phone: +1 212 854 0437
e-mail: etalley@law.columbia.edu

*Corresponding Author

CLEANING CORPORATE GOVERNANCE

Jens Frankenreiter,[*] *Cathy Hwang,*[**] *Yaron Nili*[***] & *Eric Talley*[****]

170 U. PENN. L. REV. 1 (*Forthcoming* 2021)

Although empirical scholarship dominates the field of law and finance, much of it shares a common vulnerability: an abiding faith in the accuracy and integrity of a small, specialized collection of corporate governance data. In this paper, we unveil a novel collection of three decades' worth of corporate charters for thousands of public companies, which shows that this faith is misplaced.

We make three principal contributions to the literature. First, we label our corpus for a variety of firm- and state-level governance features. Doing so reveals significant infirmities within the most well-known corporate governance datasets, including an error rate exceeding eighty percent in the G-Index, the most widely used proxy for "good governance" in law and finance. Correcting these errors substantially weakens one of the most well-known results in law and finance, which associates good governance with higher investment returns. Second, we make our corpus freely available to others, in hope of providing a long-overdue resource for traditional scholars as well as those exploring new frontiers in corporate governance, ranging from machine learning to stakeholder governance to the effects of common ownership. Third, and more broadly, our analysis exposes twin cautionary tales about the critical role of lawyers in empirical research, and the dubious practice of throttling public access to public records.

[*] Postdoctoral Fellow in Empirical Law and Economics at the Ira M. Millstein Center for Global Markets and Corporate Ownership, Columbia Law School.

[**] Professor of Law, University of Virginia Law School.

[***] Assistant Professor of Law, University of Wisconsin Law School.

[****] Isidor & Seville Sulzbacher Professor and Faculty Co-Director of the Ira M. Millstein Center for Global Markets and Corporate Ownership, Columbia Law School. The authors gratefully acknowledge comments from and discussions with Emiliano Catan, Quinn Curtis, Robert Daines, Steven Davidoff Solomon, Allen Ferrell, Jill Fisch, Jeff Gordon, Michael Klausner, Yair Listokin, Joshua Mitts, and Holger Spamann. The production of this article involved a research team too large to mention in this footnote, and we therefore list them (with our great appreciation) in Appendix A. We also thank the librarians at the Columbia Law School, the University of Virginia School of Law, and the University of Wisconsin Law School. All errors are ours.

Table of Contents

Introduction

Corporate governance lapses are blamed for some of the most ignominious business catastrophes in recent history, from Enron's epic collapse,[1] to Wells Fargo's $3 billion fine,[2] to the implosions of WeWork[3] and Theranos.[4] And in the wake of each debacle, legions of empirically-minded researchers soon followed,[5] marshaling mountains of quantitative data to unpack lessons about where governance failed, and how we can improve it.[6] Their collective efforts have met with a ravenous reception: Empirical corporate governance research now dominates the law and finance landscape,[7] routinely informing government policy,[8] real-world practice,[9] and vigorous academic debate.[10] By any reasonable accounting, the topic is a major success story in the interdisciplinary study of law.

---

[1] Woodrow W. Clark & Istemi Demirag, *Enron: The Failure of Corporate Governance*, 8 J. CORP. CITIZENSHIP 105, 105 (2002).

[2] Press Release, Department of Justice, Wells Fargo Agrees to Pay $3 Billion to Resolve Crim. & Civ. Investigations into Sales Pracs. Involving the Opening of Millions of Accts. Without Customer Authorization (Feb. 21, 2020), https://www.justice.gov/opa/pr/wells-fargo-agrees-pay-3-billion-resolve-criminal-and-civil-investigations-sales-practices.

[3] Michael Peregrine, *WeWork and the Value of Effective Governance*, FORBES (Sept. 17, 2019), https://www.forbes.com/sites/michaelperegrine/2019/09/17/wework-and-the-value-of-effective-governance.

[4] Pamela Wasley, *The Theranos Crisis: Where Was the Board?*, FORBES (Apr. 27, 2016), https://www.forbes.com/sites/groupthink/2016/04/27/the-theranos-crisis-where-was-the-board.

[5] *See generally* Robert Bartlett & Eric Talley, *Law and Corporate Governance*, *in* 1 THE HANDBOOK OF THE ECONOMICS OF CORPORATE GOVERNANCE 177 (Benjamin E. Hermalin & Michael S. Weisbach eds., 2017) (illustrating the importance of empirical data to corporate governance research).

[6] *See, e.g.*, Paul Gompers, Joy Ishii & Andrew Metrick, *Corporate Governance and Equity Prices*, 118 Q.J. ECON. 107 (2003); Rafael La Porta, Florencio Lopez-de-Silanes, Andrei Shleifer & Robert W. Vishny, *Law and Finance*, 106 J. POL. ECON. 1113 (1998).

[7] La Porta et al.'s interrelated article has been referenced 23,266 times in academic journals, including over 650 citations in law review articles. SSRN indicates that Gompers et al.'s article was cited 9,485 times, including over 175 law review articles. *See infra* Part I.

[8] *See, e.g.*, Good Faith Determinations of Fair Value, 86 Fed. Reg. 748, 795 (Jan. 6, 2021) (to be codified at 17 C.F.R. pts. 210, 270); Amendments to Financial Disclosures About Acquired and Disposed Businesses, 85 Fed. Reg. 54,002, 54,043 n.442 (Aug. 31, 2020) (to be codified at 17 C.F.R. pts. 210, 230, 239, 240, 249, 270, 274); Form CRS Relationship Summary; Amendments to Form ADV; Required Disclosures in Retail Communications and Restrictions on the Use of Certain Names or Titles, 83 Fed. Reg. 21,416, 21,442 nn.233–38, 21,485 nn.575–76 (proposed May 9, 2018) (to be codified at 17 C.F.R. pts. 240, 249, 275, 279).

[9] *See, e.g.*, CREDIT SUISSE RSCH. INST., HOW CORPORATE GOVERNANCE MATTERS 18 (2016), https://www.credit-suisse.com/media/assets/corporate/docs/about-us/research/publications/how-corporate-governance-matters.pdf (advertising an investment strategy using the well-known "G-Index" as a factor for picking high performing stocks).

[10] *See, e.g.*, Martijn Cremers & Allen Ferrell, *Thirty Years of Shareholder Rights and Firm Value*, 69 J. FIN. 1167 (2014) (discussed at greater length below); Amir N. Licht, Chanan Goldschmidt & Shalom H. Schwartz, *Culture, Law, and Corporate Governance*, 25 INT'L REV. L. & ECON. 229 (2005); William A.

And yet a potentially fatal flaw has long lurked just beneath this seemingly resplendent facade: shallow data. Many of the preeminent contributions in empirical corporate governance depend commonly (and critically) on a surprisingly slender stockpile of datasets whose provenance is frustratingly obscure. But virtually no one has seriously attempted to gauge the integrity of these pivotal inputs.[11]

Until now. In this article, we unveil a new resource that allows researchers—for the first time—to investigate the fidelity of foundational corporate governance metrics. And the results aren't pretty. We demonstrate that several of the most heavily relied-upon governance datasets suffer from inaccuracies so extensive as to call into question some of the landmark insights of the field.

The resource we unveil is anchored by a first-of-its-kind textual corpus representing over a quarter-century's worth of corporate charters for S&P 1500 listed issuers.[12] We hand-label[13] a significant subset of these full-text documents for characteristics that feature prominently in the governance literature. And, rectifying a longstanding deficit in the field, we make the corpus publicly available as open-source, in the hope that it will catalyze and improve future research. Collectively, we refer to

---

Reese Jr. & Michael S. Weisbach, *Protection of Minority Shareholders Interests, Cross-Listings in the United States, and Subsequent Equity Offerings*, 66 J. FIN. ECON. 65, 78–79 (2002); Holger Spamann, *The "Antidirector Rights Index" Revisited*, 23 REV. FIN. STUD. 467 (2010); Miroslava Straska & H. Gregory Waller, *Antitakeover Provisions and Shareholder Wealth: A Survey of the Literature*, 49 J. FIN. & QUANT. ANALYSIS 933 (2014).

[11] Most researchers have by and large presumed the integrity of the data, focusing instead on new ways to analyze, interpret, or critique its use. *See, e.g.*, Sanjai Baghat & Roberta Romano, *Empirical Studies of Corporate Law*, in HANDBOOK OF LAW & ECONOMICS, 945-1012 (2007); Lucian Bebchuk, Alma Cohen & Allen Ferrell, *What Matters in Corporate Governance?*, 22 REV. FIN. STUD. 783 (2009) [hereinafter Bebchuk et al., *What Matters*]; Lucian A. Bebchuk, Alma Cohen & Charles C.Y. Wang, *Learning and the Disappearing Association Between Governance and Returns*, 108 J. FIN. ECON. 323 (2013) [hereinafter Bebchuk et al., *Disappearing Association*]; Lawrence D. Brown & Marcus L. Caylor, *Corporate Governance and Firm Valuation*, 25 J. ACCT. & PUB. POL'Y 409 (2006); John E. Core, Wayne R. Guay & Tjomme O. Rusticus, *Does Weak Governance Cause Weak Stock Returns? An Examination of Firm Operating Performance and Investors' Expectations*, 61 J. FIN. 655 (2006); Cremers & Ferrell, *supra* note 10; Michael Klausner, *Fact and Fiction in Corporate Law and Governance*, 65 STAN. L. REV. 1325 (2013); Jonathan M. Karpoff, Robert J. Schonlau & Eric W. Wehrly, *Do Takeover Defense Indices Measure Takeover Deterrence?*, 30 REV. FIN. STUD. 2359 (2017).

[12] The S&P Composite 1500 Index is a broad-based stock index of U.S.-traded equities designed to represent a broad-based market portfolio. It is the aggregation of the S&P 500, the S&P MidCap 400, and the S&P SmallCap 600, covering approximately 90% of the market capitalization of U.S. stocks. *See* PHILLIP BRZENK, HAMISH PRESTON & AYE SOE, S&P DOW JONES INDICES, THE S&P COMPOSITE 1500: AN EFFICIENT MEASURE OF THE U.S. EQUITY MARKET 3 (Dec. 11, 2020), https://www.spglobal.com/spdji/en/documents/research/research-the-sp-composite-1500-an-efficient-measure-of-the-us-equity-market.pdf.

[13] Labeling is a procedure whereby a third party (typically a natural person with relevant expertise) evaluates, ranks, and/or categorizes the substantive content of documents in a corpus. *See* Todd Kulesza, Saleema Amershi, Rich Caruana, Danyel Fisher & Denis Charles, Structured Labeling to Facilitate Concept Evolution in Machine Learning 3075 (2014), https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1050.1212&rep=rep1&type=pdf.

our raw corpus and labels as the "Cleaning Corporate Governance" (or CCG) database. The database provides researchers with an unprecedented capability to analyze the composition and structure of the very textual heart of corporate governance—certificates of incorporation—across firms, industries, jurisdictions, and over time.

But it is substantially more than that. The CCG also allows researchers—for the first time—to reassess foundational insights from law and finance. We use it, for example, to show that the ingredients of the most renowned corporate governance index in the field, the "G-Index,"[14] are riddled with inaccuracies, resulting in an estimated error rate exceeding *80%*—a rate that gets *worse* over time. And these inaccuracies are not simply garden-variety statistical anomalies. Rather, we demonstrate that they unsettle even one of the most beatified results in the field: that systematically investing in firms with "good governance" delivers returns that significantly eclipse the market. When reanalyzed with corrected data, this result changes appreciably. To the extent any part of it survives, it does so in a materially attenuated form.

The value of the CCG is not limited to reassessing prior results in the corporate governance literature, however. It also helps lay the foundation for the next chapter of corporate governance research at a critical moment, when we stand at the crossroads of several new and exciting directions the field might pursue. Machine learning and computational text analysis, for example, are becoming increasingly prominent in many areas of legal scholarship[15] but have yet to gain a significant foothold in corporate governance.[16] The CCG is ideal for these methodologies, and

---

[14] *See* Gompers et al., *supra* note 6.

[15] *See, e.g.*, Jens Frankenreiter & Michael A. Livermore, *Computational Methods in Legal Analysis*, 16 ANN. REV. L. & SOC. SCI. 39 (2020); Kellen Funk & Lincoln A. Mullen, *The Spine of American Law: Digital Text Analysis and U.S. Legal Practice*, 123 AM. HIST. REV. 132 (2018); Michael A. Livermore, Allen B. Riddell & Daniel Rockmore, *The Supreme Court and the Judicial Genre*, 59 ARIZ. L. REV. 837 (2017); Jonathan Macey & Joshua Mitts, *Finding Order in the Morass: The Three Real Justifications for Piercing the Corporate Veil*, 100 CORNELL L. REV. 99 (2014); Marian Moszoro, Pablo T. Spiller & Sebastian Stolorz, *Rigidity of Public Contracts*, 13 J. Empirical Legal Stud. 396 (2016); Julian Nyarko, *Stickiness and Incomplete Contracts*, 88 U. CHI. L. REV. 1 (2021); David E. Pozen, Eric L. Talley & Julian Nyarko, *A Computational Analysis of Constitutional Polarization*, 105 CORNELL L. REV. 1 (2019); Eric L. Talley, *Is the Future of Law a Driverless Car?: Assessing How the Data-Analytics Revolution Will Transform Legal Practice*, 174 J. INSTITUTIONAL & THEORETICAL ECON. 183 (2018); Eric L. Talley & Drew O'Kane, *The Measure of a MAC: A Machine-Learning Protocol for Analyzing Force Majeure Clauses in M&A Agreements*, 168 J. INSTITUTIONAL & THEORETICAL ECON. 181 (2012).

[16] The literature applying machine learning to governance is still thin, and very little of it focuses on foundational corporate governance documents themselves (due in part to the absence of a corpus like the CCG). *Cf.* Gabriel Rauterberg & Eric L. Talley, *Contracting Out of the Fiduciary Duty of Loyalty: An Empirical Analysis of Corporate Opportunity Waivers*, 117 COLUM. L. REV. 1075 (2017) (building a targeted corpus of corporate opportunity waivers from public filings); Elvis Hernandez-Perdomo, Yilmaz Guney & Claudio M. Rocco, *A Reliability Model for Assessing Corporate Governance Using Machine Learning Techniques*, 185 RELIABILITY ENG. & SYSTEM SAFETY 220 (2019) (marshaling select financial disclosure items

we deploy several of them here. In particular, we use them to corroborate our error-correction efforts and to shed light on a host of deeper governance questions—including whether legal origins matter and how governance evolves during periods of disruption like the Financial Crisis. The emergent scholarly literature on "common ownership" can also benefit from the CCG.[17] While this literature raises troubling questions about whether large passive investors are conduits for anti-competitive behavior, its proponents still struggle to pin down the precise mechanism through which passive ownership translates into conscious parallelism.[18] The CCG provides an intriguing tool for smoking out such a mechanism (if one exists): by dusting for fingerprints left at the scene of the crime, as manifested in stockholder rights and governance structures in our corpus. Similarly, the CCG can help reveal how governance shapes (and is shaped by) the very purpose of the corporation itself, particularly as scholars and policy makers take the concept of *stakeholder* governance more seriously.[19] Pre-existing governance metrics—which tend to focus exclusively on *shareholder* interests—have little to say about this topic, but the CCG is a ready resource for generating new measures that bear directly on non-shareholder constituencies.

More broadly, this article exposes two systemic issues that should concern empirical researchers of all stripes. The first is that corporate governance research has a critical need for lawyers and lawyerly judgment. We conjecture that a principal reason that data errors have propagated for so long in this field is that lawyers were exiled (or relegated themselves) to the back seat of the data aggregation project. In their absence, non-lawyers were left to do much of the work, proceeding—best they could—to dispense judgments about the effects of formal legal documents, statutes, case law, and the like. While perhaps a commendable first effort, such casual empiricism no longer suffices. Lawyers can and must play a more central role in empirical corporate governance research, reclaiming the function for which they are professionally trained.

---

related to corporate governance to assess "systems failure" in firms); Ryan Bubb & Emiliano Catan, *The Party Structure of Mutual Funds* (Eur. Corp. Governance Inst., Working Paper No. 560, 2020), https://ssrn.com/abstract=3124039 (using machine learning techniques to study mutual fund voting patterns).

[17] *See* José Azar, Martin C. Schmalz & Isabel Técu, *Anticompetitive Effects of Common Ownership*, 73 J. FIN. 1513 (2018); Einer Elhauge, *Horizontal Shareholding*, 129 HARV. L. REV. 1267 (2016); Eric A. Posner, Fiona Scott Morton & E. Glen Weyl, *A Proposal to Limit the Anti-Competitive Power of Institutional Investors*, 81 ANTITRUST L.J. 669 (2017); Fiona Scott Morton & Herbert Hovenkamp, *Horizontal Shareholding and Antitrust Policy*, 127 YALE L.J. 2016, 2034-35 (2018).

[18] *See* Scott Hemphill & Marcel Kahan, *The Strategies of Common Ownership*, 129 YALE L.J. 1392 (2020).

[19] *See, e.g.*, Elizabeth Warren, Opinion, *Companies Shouldn't Be Accountable Only to Shareholders*, WALL ST. J. (Aug. 14, 2018), https://www.wsj.com/articles/companies-shouldnt-be-accountable-only-to-shareholders-1534287687; Robert J. Rhee, *A Legal Theory of Shareholder Primacy*, 102 MINN. L. REV. 122 (2018); Cathy Hwang & Yaron Nili, *Shareholder-Driven Stakeholderism*, 2020 U. CHI. L. REV. ONLINE 1 (2020).

Second, our enterprise underscores the seemingly banal observation that *data availability matters*. A lot. Another likely reason for poor data quality in this area is that corporate governance documents are surprisingly difficult to collect, organize, and analyze. Many notable jurisdictions (such as Delaware) actively throttle public access to their rich documentary trove, tossing in exorbitant access fees for good measure. Federal regulators (such as the SEC) provide several governance documents for free, but only in highly disorganized form. And, the few private enterprises that have attempted to organize them also protect their creations aggressively with paywalls, user restrictions, and ominous litigation threats.[20] Although the CCG partially unshackles the next generation of corporate governance scholars from these restraints, we nonetheless join with others (in law and elsewhere) in calling for better and less restrictive public access to public documents.[21]

The remainder of this article proceeds as follows. Part I assesses the most important empirical corporate governance studies to date, and the role of the most critical datasets within them. We also observe that because of the prohibitive challenges in obtaining underlying textual data, most researchers have relied on commercial third-party sources. Part II describes our research design and data collection protocols, providing a descriptive snapshot of the size, reach, and scope of the CCG. We then demonstrate that corporate charters are highly dynamic documents, amended with increasing frequency.[22] Yet they have also progressively become more "lawyered," growing longer, more technical, and less readable than their forebears of a quarter century ago. More provocatively, this Part uses the CCG to document the alarming inaccuracy of prominent corporate governance indices, showing that even one of best-known results in the field attenuates considerably in the presence of cleaned data. Part III explores important future uses of the CCG, including its ability to generate novel insights about the state and evolution of corporate charters. Among

---

[20] *See* Alaina Lancaster, *ROSS Intelligence Accuses Thomson Reuters of Crushing Competitors With 'Sham Copyrights and Intimidation Tactics'*, THE RECORDER (Jan. 25, 2021), https://www.law.com/therecorder/2021/01/25/ross-intelligence-accuses-thomson-reuters-of-crushing-competitors-with-sham-copyrights-and-intimidation-tactics/.

[21] *See* Julie Sobowale, *The Battle to Free Legal Information*, NATIONAL MAGAZINE (Feb. 4, 2021), https://www.nationalmagazine.ca/en-ca/articles/legal-market/legal-tech/2021/the-battle-for-legal-information; Adam R. Pah, David L. Schwartz, Sarath Sanga, Zachary D. Clopton, Peter Dicola, Rachel Davis Mersey, Charlotte S. Alexander, Kristian J. Hammond & Luís A. Nunes Amaral, *How to Build a More Open Justice Science System*, 369 SCIENCE 134 (2020) (chronicling restrictions of the PACER system over federal judicial records). Use of our corpus is free to all, governed by a Creative Commons license. *See infra* note 189.

[22] This contrasts with the usual perception that certificates of incorporation are relatively slow to change. *See, e.g.*, Lucian Arye Bebchuk, *The Case for Increasing Shareholder Power*, 118 HARV. L. REV. 833 (2005); Lucian Arye Bebchuk, *Limiting Contractual Freedom in Corporate Law: The Desirable Constraints on Charter Amendments*, 102 HARV. L. REV. 1820 (1989); Frank H. Easterbrook & Daniel R. Fischel, *The Corporate Contract*, 89 COLUM. L. REV. 1416 (1989). *But see* Geeyoung Min, *Shareholder Voice in Corporate Charter Amendments*, 43 J. CORP. L. 289 (2018) (documenting an uptick in amendment frequency for the top 200 companies in U.S. markets after 2005).

other things, we illustrate how the database lends itself to a wide variety of emergent computational and machine learning techniques, spotlighting several applications. Part IV discusses the broader implications of our study, situating it within the larger enterprise of empirical legal studies. A final section concludes.[23]

I. The State of Play in Empirical Corporate Governance Research

This article puts forward, for the first time, a clean, open-source, researchable corpus of corporate charters—the documentary DNA of corporate governance. But before proceeding to describe the CCG database itself, it is important to underscore *why* this data resource is so important. While there are many moving parts, two forces predominate: supply and demand. We discuss each below, followed by a discussion of the practical constraints that face researchers who endeavor to collect raw corporate governance documents.

A. Demand

The field of law and finance is, in relative terms, extremely young. Until about 25 years ago, finance and business law researchers typically sailed on scholarly ships that passed in the night: Financial economists gravitated toward theoretical models and data-driven explorations, while legal scholars immersed themselves in institutional detail, exploring rich contextual structures that seemingly defied quantification.

Change began to take hold, however, with a series of seminal articles in the mid-1990s. A collection of prominent finance scholars set about exploring how legal institutions affect profit generation, market value, and other relevant corporate outcomes. At the vanguard of the effort were several provocative papers by La Porta, Lopez-de-Silanes, Shleifer & Vishny (LLSV).[24] LLSV explored how formal country-level shareholder protections are correlated to and/or predict several important measures of company and shareholder value. To quantify their analysis, LLSV canvassed inter-jurisdictional protections to formulate an "antidirector rights index"—a country-level proxy for shareholder rights. Their index led to several provocative findings, including that countries with stronger investor protections are more likely to

[23] Several appendices (both attached and online at www.publiccompanycharters.com) provide additional details about our study design, results, robustness checks, and access to the corpus itself.

[24] *See* La Porta et al., *supra* note 6; Rafael La Porta, Florencio Lopez-de-Silanes, Andrei Shleifer & Robert Vishny, *Legal Determinants of External Finance*, 52 J. FIN. 1131 (1997); Rafael La Porta, Florencio Lopez-de-Silanes & Andrei Shleifer, *Corporate Ownership Around the World*, 54 J. FIN. 471 (1999); Florencio Lopez-de-Silanes, Andrei Shleifer & Robert Vishny, *Investor Protection and Corporate Valuation* (Nat'l Bureau of Econ. Rsch., Working Paper No. W7403, 1999), http:// papers.ssrn.com/paper.taf?abstract_id=227583; Florencio Lopez-de-Silanes, Andrei Shleifer & Robert Vishny, *Agency Problems and Dividend Policies Around the World*, 55 J. FIN. 1 (2000).

have common-law legal origins,[25] more advanced capital markets,[26] more ownership dispersion,[27] higher firm valuations,[28] and less earnings manipulation.[29]

LLSV's collective contributions were an instant classic, and for good reason: They demonstrated concretely how law "mattered" for nearly all aspects of finance, ranging from firm value, to ownership composition, to market risk. Tens of thousands of articles have cited LLSV, in both widely-respected law and finance/economics journals.[30] Using the index, LLSV and hundreds of others generated a laundry list of provocative and influential findings.[31] And legions of other articles to date have incorporated LLSV's index or its underlying data as inputs[32] to establish connections between shareholder protection and the size of capital markets,[33] ownership dispersion,[34] firm valuation,[35] and earnings management.[36]

As influential as LLSV's contributions were,[37] it soon became evident that their approach was just the tip of a much larger corporate governance iceberg. As lawyers know all too well, much of contemporary business law consists of a set of *background* rules that can give way if firms take steps to modify their application or opt out completely.[38] Setting the jurisdiction as the unit of analysis inevitably misses (or mushes) this firm-level heterogeneity.

That lacuna was soon to be filled by another watershed contribution, this time courtesy of Gompers, Ishii, and Metrick (GIM).[39] GIM introduced a then-novel third-party dataset created by the Investor Responsibility Resource Center (IRRC), which

---

[25] La Porta et al., *supra* note 6, at 1.

[26] Spamann, *supra* note 10, at 468.

[27] *Id.*

[28] *Id.*

[29] *Id.*

[30] According to JSTOR, LLSV's article was referenced 23,266 times, including in some of the highest-profile journals, such as the *Quarterly Journal of Economics*, *Journal of Financial Economics*, and the *Review of Financial Studies*.

[31] *See supra* note 24.

[32] *See* Spamann, *supra* note 10. For instance, Licht et al. used it to analyze the relationship between culture and the level of minority shareholders' and creditors' protection, finding that national cultural priorities consistent with public acceptance of litigation correlated with indices of creditor and shareholder voting rights. *See* Amir N. Licht et al., *supra* note 10. Reese and Weisbach also used the index, finding that companies in legal systems with less shareholder protection were more likely to cross-list in the United States. *See* Reese & Weisbach, *supra* note 10.

[33] Spamann, *supra* note 10, at 480.

[34] *Id.* at 468.

[35] *Id.*

[36] *Id.*

[37] We note that LLSV's contributions, like much of empirical corporate finance, could only suggest causal connections, but did not have an "identification strategy" to test such claims.

[38] Gillian Hadfield & Eric Talley, *On Public Versus Private Provision of Corporate Law*, 22 J.L. ECON. & ORG. 414, 418 (2006).

[39] Gompers et al., *supra* note 6.

purported to quantify shareholder protections at the individual firm level, accounting for both jurisdiction-level differences and firms' private ordering decisions. Their data tracked a cross section of large U.S.-traded issuers over several years. Consequently, the governance data that GIM marshaled included much of the granularity and panel structure that LLSV lacked, facilitating a far richer analysis of the interaction between governance and outcomes.

Not to be out-Mamboed in the governance index dance-off, GIM proposed an index of their own—the "G-Index"—which represented the sum of 24 binary variables from their dataset related to shareholder protections, antitakeover measures, and governance rights. They offered the G-Index as a rough proxy for *good governance*: Lower G-Index scores corresponded to more "democratic" or shareholder-friendly firms, while higher scores corresponded to "dictatorial" or management-friendly firms. And when the authors projected these scores onto several financial performance metrics, their findings were noteworthy: They showed that companies with relatively democratic governance profiles outperformed their more dictatorial counterparts along multiple dimensions, including firm value, profitability, and growth.[40] But one result in particular stood out: that good governance was also a financial arbitrage opportunity.[41] GIM estimated that a "long-short" investment strategy of (a) buying companies with the most democratic profiles, and (b) selling short the most dictatorial ones delivered a risk-adjusted return that outperformed the market *by an eye-popping 9% per year*, a wedge that confounded explanation by accepted theories in finance.[42]

If LLSV was an instant classic, then GIM was a mic drop. Notwithstanding its more recent vintage, GIM has been cited by more than 9,000 academic articles,[43] and it is the seventy-third most downloaded paper of all time on the Social Science Research Network.[44] Scores of follow-on papers have either employed the G-index directly, have attempted to build upon it, or have attempted to test it in other settings.

The decades since GIM's contribution, in fact, have spawned an alphabet soup of governance indices, all derived directly from the same foundational data used to construct the G-Index. These include the "E-index" (a subset of the G-index

---

[40] Gompers et al., *supra* note 6, at 121-29; Bernard Black, Antonio Gldeson de Carvalho, Vikramaditya Khanna, Woochan Kim & Burcin Yurtoglu, *Corporate Governance Indices and Construct Validity,* 25 CORP. GOVERNANCE 1 (2017).

[41] Gompers et al., *supra* note 6, at 139-42.

[42] *Id.*

[43] The Social Science Research Network indicated that Gompers's article was cited 9,485 times. Articles citing Gompers have appeared in multiple volumes of some of the most cited journals, such as the *Quarterly Journal of Economics*, the *Journal of Finance*, the *Journal of Financial Economics*, and the *Review of Financial Studies*.

[44] Paul A. Gompers, Joy L. Ishii & Andrew Metrick, *Corporate Governance and Equity Prices*, SOC. SCI. RSCH. NETWORK, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=278920 (last visited Feb. 1, 2021).

measuring management entrenchment),[45] the "O-index" (a subset of the G-index that does not include the E-index),[46] and the "D-index" (the Deterrence index, measuring takeover defense),[47] among others.[48] Others have used the G-Index (and/or its variations) as a jumping-off point for new empirical corporate governance research. One study, for example, relied on the same data to argue that firms with weaker governance structures have smaller cash reserves.[49] Another contribution used the G-index to test whether weak governance causes diminished stock returns.[50] And yet another used the G-Index as a proxy for the strength of a firm's other governance mechanisms to show the significant impact that female directors had on a board's inputs and firm's outcomes.[51] Several critics have also emerged, too, questioning the generality and longevity of the G-Index's relationship to outcomes, and observing that such effects appear to change materially in the periods after the publication of GIM's study.[52]

While these various follow-on contributions differ in many respects, they have one thing in common: They all place abiding faith in the integrity of the data that impelled GIM. And they have done so—across disparate areas of law, finance, accounting and economics—with considerable zeal.[53] Even those who have come out

---

[45] Melih Madanoglu & Ersem Karadag, *Corporate Governance Provisions and Firm Financial Performance*, 28 INT'L J. CONTEMP. HOSP. MGMT. 1805, 1806 (2015).

[46] Bebchuk et al., *What Matters*, *supra* note 11.

[47] Karpoff et al., *supra* note 11.

[48] *See generally* Straska & Waller, *supra* note 10, at 933.

[49] Jarrad Harford, Sattar Mansi & William F. Maxwell, *Corporate Governance and Firm Cash Holdings in the US*, 87 J. FIN. ECON. 535, 541 (2008), http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.708.533&rep=rep 1&type=pdf.

[50] Core et al., *supra* note 11.

[51] *See* Renée B. Adams & Daniel Ferreira, *Women in the Boardroom and Their Impact on Governance and Performance*, 94 J. FIN. ECON. 291 (2009).

[52] *See, e.g.*, Klausner, *supra* note 11; Robert M. Daines, Ian D. Gow & David F. Larcker, *Rating the Ratings: How Good Are Commercial Governance Ratings?*, 98 J. FIN. ECON. 439 (2010); Bebchuk et al., *Disappearing Association*, *supra* note 11.

[53] *See* Bebchuk et al., *What Matters*, *supra* note 11, at 784 (calling Gompers et al. paper "influential"); Genc Alimehmeti & Angelo Paletta, *Corporate Governance Indexes: The Confounding Effects of Using Different Measures*, 4 J. APPLIED ECON & BUS. RSCH. 64, 65 (describing Gompers et al.'s paper as a "landmark" paper); Beth Ahlering & Simon Deakin, *Labour Regulation, Corporate Governance and Legal Origin: A Case of Institutional Complementarity?* (U. of Cambridge Centre for Bus. Rsch., Working Paper No. 312, 2005) (noting that La Porta et al.'s research has "inform[ed] the policy and working methods of the World Bank and other international financial institutions"); John C. Coffee, Jr., *The Rise of Dispersed Ownership: The Roles of Law and the State in the Separation of Ownership and Control*, 111 YALE L.J. 1 (2001) (referring to La Porta et al.'s law-and-finance scholarship as "seminal"); Sofie Cools, *The Real Difference in Corporate Law Between the United States and Continental Europe*: Distribution of Powers, 30 DEL. J. CORP. L. 697, 699 (2005) (same); Nicholas Thompson, *Common Denominator*, LEGAL AFF. (2005) (describing La Porta et al.'s impact as influential).

as critical of the governance-index enterprise have based their arguments largely on the indices' predictive qualities, presuming the accuracy of the underlying indices.[54]

The demand for data-driven corporate governance insights, moreover, transcends academia. It also extends to professional governance advocates, Wall Street investors, and even government regulators. The U.S Securities and Exchange Commission, for example, routinely uses the empirical corporate governance literature—including LLSV and GIM—in rulemaking. A recent proposal to amend federal proxy rules, for example, cites both studies, interpreting them to demonstrate that "[s]trong shareholder rights have been associated with higher firm valuations and better developed equity markets."[55] The SEC has cited empirical literature in a proposed rule on investment advisors and broker-dealers,[56] in a report on M&A disclosure requirements,[57] and in a proposed rule on accelerated filers.[58] Earlier this year, the SEC cited LLSV again in a final rule on good faith determinations of fair value.[59]

Clearly, empirical corporate governance research, and the abecedarian conga line of indices it spawned, have become a centerpiece of both academic discourse and regulatory decision making. That attention, in turn, has increased the demand for more resources (quantitative data in particular) that could power additional insights to help adjudicate policy debates. But a demand for empirical corporate governance resources would remain unrequited without a corresponding supply. As we detail in the next subsection, that supply chain has proven to be limited, expensive, and undependable.

B. Supply

The seemingly insatiable demand for quantitative corporate governance resources has always faced serious supply shortages. Indeed, as provocative as the findings of LLSV, GIM and their progeny were, perhaps their most enduring contributions were the new data they brought to the table, quantifying governance for the first time.

But the sheer novelty of these efforts was also their Achilles' heel. Turning nebulous bodies of prolix corporate governance texts into concrete, measurable, usable data requires an unusual mélange of quantitative skill, economic intuition,

---

[54] *See, e.g.*, Daines et al., *supra* note 52; Bebchuk et al., *Disappearing Association*, *supra* note 11.

[55] Universal Proxy, 81 Fed. Reg. 79,122 (Nov. 10, 2016) (to be codified at 17 C.F.R. pt. 240).

[56] Form CRS Relationship Summary; Amendments to Form ADV; Required Disclosures in Retail Communications and Restrictions on the Use of Certain Names or Titles, 83 Fed. Reg. 21,416, 21,442 nn.233–38, 21,485 nn.575–76 (proposed May 9, 2018) (to be codified at 17 C.F.R. pts. 240, 249, 275, 279).

[57] Amendments to Financial Disclosures About Acquired and Disposed Businesses, Release No. 33-10786, 2020 WL 5096804, at 54043 n.442 (Aug. 31, 2020).

[58] Form CRS Relationship Summary; Amendments to Form ADV; Required Disclosures in Retail Communications and Restrictions on the Use of Certain Names or Titles, Release No. 34-83063, 2018 WL 2114080, at 21442 nn.233-238 (May 19, 2018).

[59] Good Faith Determinations of Fair Value, 86 Fed. Reg. 748, 795 (Jan. 6, 2021) (to be codified at 17 C.F.R. pts. 210, 270).

and—most importantly—lawyerly chops. Corporate governance regimes are typically conjured from a dense thicket of documents, statutes, legislative histories, case law, and a superstructure of interpretive canons. Parsing these inputs into usable data is all but impossible without legal training. Even today, very few possess the requisite skills to peel back the layers of this institutional onion. This skill set was rarer still a quarter-century ago.

The difficulties of "coding law" were immediately apparent even in the early studies that analyzed jurisdictional corporate governance regimes. These studies, starting with LLSV's pioneering work, had to quantify country-level legal protections. To go about doing so, the authors needed to assess—across over four dozen national jurisdictions—six mechanisms for investor protection.[60] Having little or no legal training themselves, the authors eventually farmed out the work, surveying local lawyers in each jurisdiction to identify whether it had the various enumerated investor protections of interest.

The resulting methodology was innovative for its time, but it was also frustratingly opaque. There is scant information about how the authors identified their respondents or the respondents' expertise. There is also no information on how the authors dealt with inter-respondent inconsistencies. As others soon noticed, these types of inconsistencies proved commonplace.

No one seriously endeavored to interrogate the underlying anti-directors rights index itself until 2010, when Holger Spamann began kicking its tires in a replication study. Spamann used a more systematic approach to recruit and orient foreign-trained lawyers to recode the majority of LLSV's primary jurisdictional data, taking significant care to ensure inter-coder reliability. When the dust settled, he found that many of the features contained within the original index were incorrect, and—more importantly—that certain of the most provocative results could not be replicated once the data were corrected.[61]

Spamann's findings were careful, systematic, and ultimately devastating. More generally, his analysis exposed a larger problem that continues to vex empirical corporate governance: The challenge of quantifying *jurisdictional* legal factors that are themselves somewhat nuanced, the ambiguity that occurs when a statutory mandate is overridden by firms that opt out of it, and the deployment of personnel who had not been trained to assess these factors consistently. Challenges like these suffused LLSV's data, and Spamann surmises that the ensuing reliability issues may have been the

[60] They were shareholder voting by mail, voting without blocking of shares, the limits on shareholders' ability to call a special meeting, whether minority shareholders had proportional board representation or cumulative voting, whether existing shareholders had a preemptive right to buy new issuances of stock, and the kinds of judicial remedies available to shareholders. Gompers et al., *supra* note 6, at 114-15.

[61] Spamann, *supra* note 10, at 469.

byproduct of differences in corporate practice[62] and the fact that LLSV's original survey respondents were unclear about whether their remit was to answer questions about formal legal mandates, prevailing private-ordering norms, or their own personal experiences.[63]

By the time Spamann's deconstruction of LLSV hit the presses, many of the cool kids of corporate governance were already chasing the next rainbow: *firm-level* governance. Zeroing in on a more granular unit of analysis constitutes a considerable improvement (for the reasons detailed above); yet at the same time, the firm-level approach seems susceptible to many of the same vulnerabilities that plagued LLSV. Maybe even worse: Coding law consistently at the jurisdiction level is hard enough; layering on firm-level governance can complicate matters considerably. Not only must one correctly interpret when and how companies have attempted to tailor their internal governance affairs, but one must do so against the backdrop of statutory and jurisdictional rules.

Figure 1 conceptualizes some of these difficulties using a planetary metaphor for corporate governance choices. A statute/regulation (represented by the black hole at the system's center) represents a fixed background rule on corporate governance that applies to companies incorporated in the jurisdiction. Should an entity (represented by the various planets) wish to replace that rule and with its own self-styled regime, it is as though it needs to break the gravitational pull of the statutory mandate.

For certain governance mandates, opting out is impossible, as depicted by the inner red planet. Here, the statutory rule is *immutable*, with a gravitational pull so strong as to trap all objects within its event horizon. If a company within this zone wished to embrace a different regime, its efforts would be null and void.[64] In other situations (represented by the successively more distant planets), the gravitational pull of the mandate is weaker. Here, the statute specifies a *default rule,*[65] theoretically permitting firms to embrace self-styled regimes through private ordering. But here too, differing requirements may apply if a firm is to achieve the needed escape velocity to break

---

[62] *Id.* at 473.

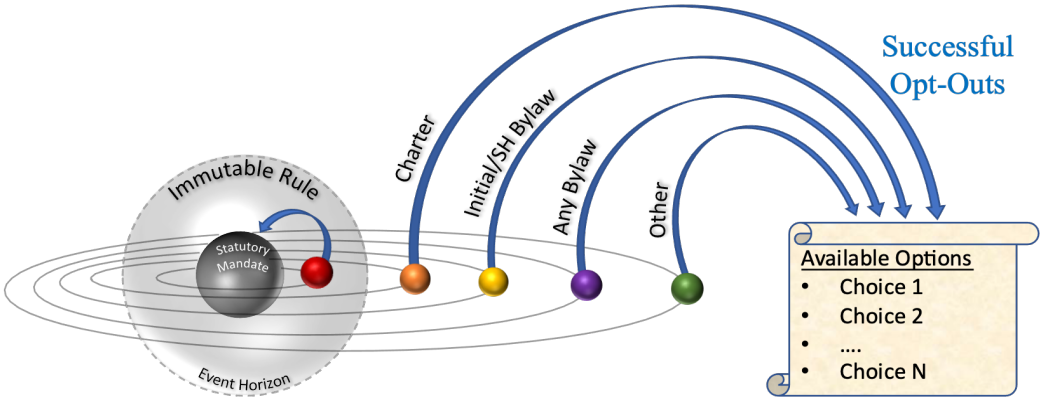
[63] As Spamann notes in his 2010 review of La Porta et al., *supra* note 6, practitioners inconsistently interpreted La Porta et al.'s questions. For example, neither Finland nor the United States defaults to cumulative voting, but La Porta et al. coded Finland as zero and the United States as one. *See id.* at 472.

[64] Pennsylvania's famous constituency statute, for example, does not allow companies to waive or avoid the statutory mandate that the board must account for multiple stakeholders' interests. *See* 15 PA. CONS. STAT. ANN. § 1715(b); *see also* Robert Goodyear Murray, *Money Talks, Constituents Walk: Pennsylvania's Corporate Constituency Statute Can Maximize Shareholders' Wealth*, 48 BUFF. L. REV. 629, 643 (2000) (opining that the constituency statute "is a specific grant of discretion to directors to determine which constituency group's interests to elevate above others, ranking shareholders as only one of the interests and not giving them a priority interest"). Similarly, Delaware corporations can extend bylaw amendment power to directors only if done through a provision in the articles of incorporation. *See* DEL. CODE ANN. tit. 8, § 109.

[65] *See generally* Ian Ayres & Robert Gertner, *Filling Gaps in Incomplete Contracts: An Economic Theory of Default Rules*, 99 YALE L.J. 87 (1989).

away: The most tenacious types of default rules (conceptualized through the orange planet) require nothing less than a charter provision to opt out.[66] Other default mandates are less sticky, giving way as well to contraventions in "lower level" corporate documents—such as a shareholder-enacted bylaw (yellow planet), an ordinary bylaw (purple),[67] or a simple board resolution (green).[68]

Figure 1: Statutory Mandates and Achieving Opt-Out Escape Velocity



Even when a state law provision permits opt-outs, and even if the corporation takes the requisite steps to achieve escape velocity, another layer of complexity awaits: the corporation may still not be free to explore all parsecs of the corporate governance universe: state law often deems certain types of self-styled regimes to be off limits.[69] The upshot of this discussion is that for many dimensions of corporate governance, the regime that ultimately applies to the firm requires understanding (a) what the state's substantive background mandate is; (b) whether that mandate permits opting out; (c) what the lowest level document is for executing an opt out; (d) what the constrained

---

[66] Delaware's staggered board statute, for example, requires that any board stagger be effectuated through the charter, initial bylaw, or a shareholder-promulgated bylaw. *See, e.g.*, DEL. CODE. ANN. tit. 8, § 141(d) (2010).

[67] In this category, and in the remaining ones, the stated means for opting out is sufficient, but any "higher level" document can generally accomplish the task too. Consequently, because charters are at the top of the corporate governance pecking order, a charter provision would also be sufficient to opt out of a state mandate that allows opt-outs through a "lower level" document, such as a shareholder approved bylaw.

[68] For example, Delaware permits corporations the option to provide for proxy expense reimbursement to activists through ordinary bylaw provisions. *See* DEL. CODE. ANN. tit. 8, § 113.

[69] For instance, while Delaware permits firms to include a forum selection provision in its bylaws, it prohibits a corporation to *exclude* Delaware courts from hearing "internal corporate claims." DEL. CODE. ANN. tit. 8, § 115. And Maryland gives shareholders a default right to convene a special meeting with the support of 25% votable shares. Corporations are allowed through their bylaws to increase that threshold, but they are not allowed to increase it beyond 50%. MD. CODE ANN., CORPS. & ASS'NS § 2-502.

choices are for the opting-out entity; and (e) whether the corporation has succeeded in opting out in a manner that complies with (a) through (d). In short, it *is* logical, but it's also complicated.

So how does the firm-level data GIM relied on fare in this more complicated environment? Unfortunately, we simply do not know at first blush. Even as compared to LLSV's index, the IRRC data used by GIM are surprisingly opaque and poorly documented. Little remains (if it ever existed) about what went into it. In fact, the dataset appears to have had no detailed manual, but it instead refers interlocutors to the appendix of the GIM paper,[70] a curious move since that appendix only cursorily describes the variable definitions, with little mention of data gathering and quality-control measures.[71] And, as detailed above, there does not appear to have been any researcher with the time, resources, and risk tolerances to interrogate the firm-level IRRC data used by GIM.[72]

Consequently, today's researchers have scant information about how the IRRC constructed their labels.[73] We know little about what documents they consulted—state law, charters, bylaws, or something else entirely. There is no information about how coders resolved inconsistencies between the documents, if they considered multiple documents at all. There is no indication about the credentials of the coders themselves, or measures to ensure inter-coder consistency. And there is only a small amount of information about the nuances of state laws. For example, when the coders noted that a state law existed on a particular topic, did they assess whether the state law immutably *required* something of corporations? Or was the state law a default that allowed corporations to opt out? Or was the state law silent, and allowed corporations to opt-in?

The IRRC data's issues are compounded further by a rotisserie of corporate ownership changes: In 2005, IRRC was acquired by ISS.[74] And two years later, RiskMetrics acquired ISS, changing data gathering protocols and retiring the IRRC

---

[70] *See* Overview of IRRC Governance Database in WRDS, *available at* https://wrds-www.wharton.upenn.edu/documents/718/Overview_of_IRRC_Legacy_Governance_Definitions.pdf. Our review of the IRRC Corporate Takeover Defenses data did not reveal a methodology section.

[71] Gompers et al., *supra* note 6, at 145-50.

[72] The original companion manuals to the data set are surprisingly hard to find. We searched in every law library in the United States and mobilized the combined forces our respective institutional librarians—including contacts at law firms. After weeks of searching, we managed to a few examples, whose methodology descriptions were frustratingly opaque. *See* note 73, *infra.*

[73] Although an annual IRRC publication described the various label categories, it does not touch on the methodology of data collection or the training/expertise of labelers themselves. *See, e.g.,* Virginia K, Rosenbaum, CORPORATE TAKEOVER DEFENSES ix (1998) (devoting all of three short sentences to data collection protocols).

[74] Robert Kropp, *SRI Field Continues to Shift with RiskMetrics' Acquisition of KLD*, GREENBIZ (Nov. 6, 2009), https://www.greenbiz.com/article/sri-field-continues-shift-riskmetrics-acquisition-kld

data into "Legacy" status.[75] The contemporary ISS dataset has improved documentation, but it is still slim, and in any event it covers only a loosely overlapping set of variables with IRRC's, excluding critical ingredients in the G-Index. As a result, it is now impossible to extend G-index data computations beyond 2006.[76] For those interested in slicing and dicing the G-Index, then, they must largely do so inside a time capsule from the 1990s and early 2000s.

Nevertheless, the robust demand for governance studies has induced legions of contemporary scholars to return to the original wellspring of the IRRC (and the associated G-index) to study governance, assuming those data to be accurate and hoping to say something generalizable to contemporary settings. Top finance journals continue to publish research that is based on those early data.[77] One notable contribution[78] even extended the IRRC data and G-Index going backwards in time (from 1978-1989) using a subsample of companies, but in doing so it also largely presumed the integrity of the IRRC database itself. And while the moving tectonic plates of empirical corporate governance literature are increasingly surfacing concerns regarding methodological designs,[79] data integrity[80] and empirical design,[81] the IRRC and the G-Index have largely remained uninterrogated.

Although there are a few alternatives to the IRRC index, those that exist are problematic too. The contemporary ISS dataset[82] offers a variety of governance metrics, but it does not offer enough of them to replicate well-known indices, and its protocols appear somewhat different from IRRCs even for the same variables it traces. It too, has relatively poor documentation, and seems potentially vulnerable to similar

---

[75] *See* ISS, *Changes in ISS (formerly RiskMetrics) Governance Database for 2007*, https://wrds-www.wharton.upenn.edu/pages/support/manuals-and-overviews/iss-formerly-riskmetrics/changes-iss-formerly-riskmetrics-governance-database-2007/.

[76] *Id.*

[77] *See, e.g.*, Karpoff et al., *supra* note 11 (employing an instrumental variables strategy on the G-Index in an attempt to tease out causal inference). When a variable being instrumented for is subject to measurement error, however, it can generate spurious results). *See* Dan A. Black, Mark C. Berger & Frank A. Scott, *Bounding Parameter Estimates with Nonclassical Measurement Error*, 95 J. AM. STAT. ASSOC. 739 (2000).

[78] *See* Cremers & Ferrell, *supra* note 10. We discuss this important paper in the context of our project at greater length *infra* note 107.

[79] *See* Yair Listokin, *Interpreting Empirical Estimates of the Effect of Corporate Governance*, AMERICAN LAW AND ECONOMICS REVIEW 90 (2008)

[80] *See* Emiliano M. Catan, *The Insignificance of Clear-Day Poison Pills*, 48 JOURNAL OF LEGAL STUDIES (2019); Emiliano M. Catan & Marcel Kahan, *The Law and Finance of Antitakeover Statutes,* 68 Stanford. L. Rev. 629 (2016).

[81] *See* Robert Bartlett & Frank Partnoy, *The Misuse of Tobin's q*, 73 Vand. L. Rev. 353 (2020).

[82] *See* ISS, *supra* note 75.

coding errors.[83] Another popular data provider, Factset,[84] tracks even fewer governance metrics than ISS, and it is difficult to obtain data for historical years. And another, Compact Disclosure, is poorly organized for this particular task, and it ceased all updates in 2006.

Without cataloguing the remaining (modest) list of other governance datasets, none of them can be easily quality checked without access to the underlying documents from which they are purportedly built. And yet, none of these sources (that we are aware of) allows users to access the texts on which their data labels are based. To audit accuracy, then, one must recover these documents independently, read them for legal import, and confirm whether the assigned label was correct.

And this is where the challenge *really* begins. For even if one possessed the skills (and resources, and patience) to weed through mountains of raw governance documents for substantive content, simply gaining access to an organized corpus of them is surprisingly hard. In *theory,* of course, lots of corporate governance documents are in the public domain, and state and federal governments have the means to provide organized access to them. Moreover, statutes and case law are but a quick internet search away. Harvesting this information should not be all that difficult, should it?

Yet, irritatingly, it is. This article focuses on what would seem to be the easiest of targets—articles of incorporation (also known as charters), a corporation's first and most important corporate governance document.[85] The charter is critical, for it is both a corporation's birth certificate and its constitution: To form a corporation in any U.S. jurisdiction, an incorporator must first file a charter (including within it a host of necessary ingredients) with a state secretary of state, who in turn maintains repositories of such documents.[86] And for publicly traded companies, charters also must be filed with the SEC.[87] In theory, then, public company charters should be readily available from both state *and* federal sources.[88] In practice, however, extracting the text of contemporary charters on a wide-scale basis is tricky, expensive, and time-consuming.

---

[83] In Online Appendix C, we show that the contemporary ISS governance data also appear to be hampered by significant errors as judged by our newly compiled data (though not as severe as IRRC's).

[84] *See Data Solutions*, FACTSET, https://www.factset.com/solutions/business-needs/data-solutions (last visited Feb. 10, 2021).

[85] Although corporations often have numerous governance documents, the most important one is the charter. In addition to charters and bylaws, corporations generally have additional governance documents, such as committee charters, corporate governance guidelines, and a variety of other documents that corporations adopt to meet stock exchange, regulatory, and other requirements. *See* Sarah C. Haan, *Shareholder Proposal Settlements and the Private Ordering of Public Elections*, 126 YALE L.J. 262 (2016) (showing that many corporate disclosures about campaign finance are made as the result of negotiated private settlements with shareholders); Yaron Nili & Cathy Hwang, *Shadow Governance*, 108 CALIF. L. REV. 1097, 1107-09 (2020).

[86] *What are Articles of Incorporation?*, HARBOR COMPLIANCE, https://www.harborcompliance.com/information/what-are-articles-of-incorporation (last visited Feb. 5, 2021).

[87] 17 C.F.R. § 229.601(b)(3) (2021).

[88] 15 U.S.C. § 78(m).

Consider what might be the most obvious strategy: approach relevant state governments to gain access to their primary documents. Good luck: Delaware, where the majority of public companies are incorporated,[89] makes it risibly difficult to obtain corporate charters in native form. By way of example, consider the task of assembling the chartering history of Google's parent company, Alphabet Inc. Although it is a Delaware-incorporated entity, searching on the Delaware Secretary of State's website yielded no results: the entity search function returns only the first 50 hits matching "Alphabet," and Alphabet Inc. was not among them. But even if Alphabet Inc. had been among the first 50, obtaining information about whether the entity is active requires one to pay a $10 fee. To obtain an inventory of all documents filed in the state by that entity would cost an additional $175 for each registrant.[90]

All this, of course, still falls short of producing the raw texts themselves. For that, one would additionally have to make a formal document request for each individual entity with the Delaware Department of State, incurring a per-document fee of $10 for the first page plus $2 for each additional page.[91] After some period of days or weeks, a hard copy packet would arrive in the mail, whereupon the researcher would need to use character recognition technology to scan and digitize its contents.[92] The costs quickly add up: There are 1,713 Delaware-incorporated issuers in our corpus, comprising about 60% of the total number we track across all U.S. states. Delaware entities have a mean chartering history of 3.3 documents per issuer, and an average length of 12.4 pages per document.[93] All told, if one wholly disregards labor costs, and

---

[89] *About the Division of Corporations*, DEL. DIV. OF CORPS., https://corp.delaware.gov/aboutagency/ (last visited Jan. 27, 2021) (noting that more than 66% of the Fortune 500 are incorporated in Delaware).

[90] *See Accessing Corporate Information*, DEL. DIV. OF CORPS., https://corp.delaware.gov/directweb/ (last visited Jan. 27, 2021) (describing the process of requesting documents). While on first blush there appears to be a cheaper $20 option to access a list of filed documents, that list only shows the last five documents filed. (To discover this informational nugget, a research assistant spent 20 minutes on the phone with the Delaware Secretary of State's office.

[91] For the current fee schedule, *see* Delaware Department of State, Division of Corporations Fee Schedule (2020), https://corpfiles.delaware.gov/Augustfee2020.pdf. Although Delaware evidently maintains the entire collection in digital form, the cumbersome process described in the text appears non-negotiable, even for pure researchers. Early in this project, and armed with the written endorsement of Hon. Leo Strine (the then-sitting Chief Justice of the Delaware Supreme Court), we approached Delaware's Deputy Secretary of State and Director of Division of Corporations requesting access to the state's corpus (on a confidential basis). Our attempt was quickly rebuffed. *See* E-mail from Deputy Secretary Kristopher Knight to Eric Talley (Aug. 31, 2017) (on file with authors) ("I appreciate the offer, but the Division has a long-standing practice of not participating in such arrangements").

[92] Alternatively, one could take a quick sojourn to Dover, Delaware—about an hour's drive south of Interstate 95, and coincidentally abutting the scenic Bombay Hook National Wildlife Refuge—where one could then enjoy the privilege of queuing up for one of the state-issue public terminals to access and print documents (for one hour at a time as we understand, if other users are waiting). *Id.*

[93] This figure is based on a mean word count of 4,571 at an average of 368 words per page. *See infra* Part II for a more fulsome analysis of these measures.

further assumes no expediting costs ($1,000 per document for same-day service[94]), we estimate that the Delaware Department of State would charge *no less than $485,190*,[95] simply to replicate three-fifths of the textual corpus we unveil (free of charge) in this article.[96]

Similar attempts to obtain primary documents from New York, California, Nevada, and Massachusetts fared little better: In each state, we grappled with decades-old computer systems, spent hours on the phone, and were offered seemingly random collections of charter documents to be dribbled out over the course of days or weeks, usually for hundreds of dollars. In any realistic sense, then, seeking out governance documents from state repositories is a non-starter.

Those interested in publicly traded companies have two other possible avenues. First, many companies' investor relations websites contain charters. Their digital format is far from uniform, however, and they typically contain only *current* (but not historical) versions, frustrating researchers who wish to study governance both cross-sectionally and over time. Consequently, such sources have limited value.

The other option (and the one we ultimately pursued here) is to go to the SEC, which maintains a repository of current and historical filings that users can access for free—depending on how one defines *free*. To be sure, public companies are required to file up-to-date charters with the SEC, and the Commission duly records all public filings from the last twenty-five years or so on its online EDGAR database.[97] That said, EDGAR proves to be a cumbersome hunting ground for governance documents. The interface is notoriously hard to filter and search,[98] and locating charters is particularly challenging. Although components of EDGAR filings have become predictable and regularized over the years, corporate charters and bylaws have not, and their disclosed content is often squirreled away in odd and irregular places.[99] And,

---

[94] *see* Delaware Department of State, Division of Corporations Fee Schedule (2020), https://corpfiles.delaware.gov/Augustfee2020.pdf.

[95] This figure ignores the $10 fee to gauge whether the firm is active, but includes the Long Form certificate of filings, resulting in total a cost computed by: [$175 + $10(3.3)+$2(3.3)(12.7-1)]×1713.

[96] We were unable to determine how much revenue the Delaware Secretary of State's office generates in charging for access to these ostensibly public documents. Consequently, we cannot estimate how our efforts to make a sizable corpus of them freely available to the public may cut into these revenue margins.

[97] 15 U.S.C. § 78(m).

[98] Several practitioners have even authored how-to articles on EDGAR use. *See, e.g.*, Duff McDonald, *Unscrambling Edgar: The SEC's Database Is Torture to Use but Help Is Out There*, 28 MONEY 175 (1999).

[99] In theory, issuers are supposed to tag their charters as "Exhibit 3" in the context of periodic filings (10-K, 10-Q) and current reports (8-K). But in practice, these tags are applied with varying degrees of consistency. For example, Biglari Holdings Inc.'s 2018 S-4 registration statement contains the text of its charter in a section labeled "Annex II," while Parkway Property Inc.'s 1996 charter is in "Exhibit B" to the companies' preliminary proxy statement. These tagging inconsistencies appear also to frustrate the search algorithms of commercial services. *See also* John Gerdes Jr., *EDGAR-Analyzer:*

because EDGAR only has filings from the mid-1990s, locating pre-1990s materials requires submitting a records request to the SEC—an exercise that, reminiscent of state regulators, requires Byzantine paperwork,[100] a 20-day processing period, and an hourly processing fee for the lucky employee charged with hunting down the documents.[101]

A related strategy for the exhausted researcher might be to leverage commercial search platforms, such as Westlaw, LexisNexis, and Bloomberg, which also track EDGAR filings but purport to offer user-friendly search conduits. Our research suggests that Westlaw is perhaps the most comprehensive for a project like ours, allowing researchers to search for "articles of incorporation/bylaws" for individual companies. However, Westlaw seems to compile its underlying data in a way that succumbs to the EDGAR filing irregularities described above.[102] Consequently, extracting a firm's chartering history on Westlaw frequently results in troubling gaps in coverage. Holding that issue aside, Westlaw's search results often require extensive post-processing by researchers seeking to build a usable panel of chartering histories. Westlaw does provide researchers with unfiltered lists of up to 1,000 documents that may contain charters (interlaced with bylaws and many other texts). While potentially useful, these lists usually contain duplicates[103] and are ordered by filing date (which can be years or even decades after their effective dates).

Even if one is lucky enough to locate comprehensive chartering histories from a commercial provider like Westlaw, she will likely be prohibited from using them for a project of any scale. Users can download no more than 100 text documents at once. With a combination of patience, ingenuity, and web-scraping technology, one might be able to work around some of these *technical* speed bumps; but doing so would almost certainly violate Westlaw's user agreement, subjecting the user to a lock on her account. Posting the results publicly might also trigger the litigious wrath with which the company enforces its user agreements.[104] If one hopes to access *and* share the full

---

Automating the Analysis of Corporate Data Contained in the SEC's EDGAR Database, 35 DECISION SUPPORT SYS. 7, 8 (2009) (describing SGML tags and identifying challenges posed by improper tagging).

[100] *See Public Documents*, U.S. SEC. & EXCH. COMM'N, https://www.sec.gov/fast-answers/answerspublicdocshtm.html (last visited Feb. 10, 2021).

[101] *See Schedule of Fees for Record Services*, U.S. SEC. & EXCH. COMM'N, https://www.sec.gov/foia/feesche.htm (last visited Feb. 10, 2021).

[102] Evidently, Westlaw (like other commercial providers) focuses exclusively on EDGAR filings located under the "Exhibit 3" heading—the exhibit category supposedly designated for charters and by-laws. In practice, as noted above, many companies disregard this heading mandate, squirrelling away their governance documents under different exhibit numbers (or none at all).

[103] For example, Phase Forward Inc adopted a new charter in the context of its IPO in July 2004. Different versions of this document can be found in four different filings from around the time of the IPO, and all filings are represented as separate entries in Westlaw's search results.

[104] *See Terms of Use*, THOMSON REUTERS, https://legal.thomsonreuters.com/en/legal-notices/terms-of-use (last visited Feb. 10, 2021) (stating "you will not reproduce, duplicate, copy, download, store, further transmit, disseminate, transfer, or otherwise exploit this website, or any portion hereof without Thomson Reuters' prior written consent"). The company has been aggressive of late in

textual chartering histories for thousands of public companies, then, commercial data providers offer little refuge.[105]

II. Reclaiming Corporate Governance

Given the paucity of existing data and the shortcomings in strategies for gathering the data from scratch, researchers enjoy few attractive options, and they are usually left to muddle through in an *ad hoc* way while adjusting priorities and limiting their research designs.

This article endeavors to sweep away these obstacles, reclaiming the textual DNA of corporate governance for all researchers in the process. In this Part, we begin by unveiling the results of this multi-year enterprise: the CCG, which involved harvesting by hand thousands of corporate charters spanning the better part of three decades, cleaning them, labeling them, and making the resulting corpus an open-source public good (as it always should have been). We then spotlight several immediate payoffs of this effort, ranging from intriguing descriptive accounts of the corpus, to reassessing heretofore accepted wisdoms in law and finance, to marshaling the emergent tools of machine learning and computational text analysis to unpack the myriad stories that these critical documents tell.

A. Charter Texts

Over the course of several years, we have been assembling a comprehensive textual dataset containing present and historical charters of almost 3,000 of the largest publicly traded companies in the United States. Our dataset is based exclusively on digitized filings with the SEC made available on the EDGAR database, therefore ensuring our ability to share it as an open-source resource.

To ensure accuracy and comprehensive coverage, we harvest charters manually, with the help of a small army of research assistants. Doing so allows us to avoid the many pitfalls of the automated approaches evidently used by commercial services. We use a formal organizational hierarchy and numerous quality-control measures to exercise quality control over our collection efforts. Senior members of the team (law school graduates and advanced JD candidates) cross-checked most information assembled by junior research assistants.[106]

---

enforcing these provisions. *See, e.g.*, Lancaster, *supra* note 20 (documenting the case of a legal-tech startup that was driven from business from a lawsuit filed by Westlaw's parent company).

[105] Our investigation of Westlaw's main competitors revealed nearly identical terms-of-use prohibitions, but with inferior search functionality. Lexis, for example, seems to require users to search for specific terms or companies, which curtails the ability to even pull up all NASDAQ companies at once.

[106] More details on our data gathering protocols can be found in Appendix B. Although we consider our collection protocol to be a significant improvement over standard commercial providers, the biggest advantage of our dataset is its open-source nature. We make the dataset available to the

The resulting dataset contains the chartering histories for 2,899 companies, starting with the first fully restated charter the company uploaded onto EDGAR. For many companies, this is the first full charter that was filed after EDGAR went live in 1995. For a majority of companies (around 58%), however, we are able to trace their charter history well back into the 1990s.[107] In some cases (approximately 12%), we successfully extract a chartering history to a year prior to 1990. Based on this harvested information, we construct a textual corpus that treats *an issuer's charter document in a given measuring year* as the observational unit. In other words, our dataset has a "panel" structure, observing the charter text(s) that governed the internal affairs of each active company as of January 1 of each year between 1990 and 2019.[108]

Even without elaborate data crunching, our corpus renders some interesting insights about public company charters over time. Consider *charter length*. In principle, certificates of incorporation could be quite short, with most of the nitty-gritty baked into other governance documents (such as bylaws).[109] Indeed, most state statutes require charters to have only a few informational ingredients,[110] and they can be shorter than 75 words.[111] Public company charters are typically longer, but are still relatively

---

public, and we invite others to contribute to it *and* correct any mistakes that escaped our quality control measures. We return to this point *infra* Part IV.

[107] A notable paper started down a route similar to ours, but ultimately chose a different path for a different purpose. *See* Cremers & Ferrell, *supra* note 10. There the authors tracked approximately 1,000 companies from the IRRC data set *backwards* in time to the 1978-89 era. The authors did not attempt to audit the G-Index itself over the IRRC years, but instead used a random sample of IRRC observations from 1990 in order to emulate the labeling conventions of the IRRC. *Id.* at 1172. They do not report on error rates they discovered in the IRRC, nor do they attempt to assess whether the IRRC's conventions were consistent with objective legal judgments. Unlike this paper, the authors did not collect the raw corpus of governance documents; but instead focused on generating labels only. (Neither their labels nor their constructed index is available to the public, though they evidently have made their constructed index available to select researchers.)

[108] Although we relegate most of our data collection protocols to Appendix B, one detail warrants attention here. When a typical issuer amends its charter, its disclosure frequently takes the form of a focused statement of the amendment, unaccompanied by a *de novo* restatement of the entire amended charter. In fact, several such piecemeal disclosures will often stack up before an issuer wrangles them into a full restatement. Manually interlacing such amendments into the pre-existing charters proved infeasible. Instead, we aggregated full restatements and partial amendments as follows: For each issuer in any year, we consider its charter to consist of (a) the most recently disclosed full charter, appended by (b) all disclosed partial amendments executed after the restatement in (a) but before the observation year. This protocol preserves information, but also may lead to some distortions (such as the measured length of a charter). Therefore, where appropriate (such as in Figure 2 below), we limit attention to only the most recent full restatement (suppressing any intervening partial amendments).

[109] *See generally* Nili & Hwang, *supra* note 85.

[110] Typically, they require (a) the name of corporation; (b) its purpose or nature of its business; (c) its official address; (d) a description of corporation's capital structure; and (e) duration of corporation's existence. Some statutes require descriptions of the incorporators, paid-in capital, and initial board structure. *See* 1 Corp. Forms § 2:2, Marvin Hyman and Publisher's Editorial Staff (2020).

[111] *See, e.g.*, Form 1.01: Articles of Incorporation (Legal Minimum), *in* Robert Brown, Jr., Herbert B. Max & Stacey L. Bowers, Raising Capital: Private Placement Forms & Techniques (3d ed. 1995).

brief. In practice, however, the (full versions of) charters of publicly traded companies are far longer, typically ranging between 1,000 and 12,000 words.

Figure 2: Mean Word Length (Full Charter Restatements)

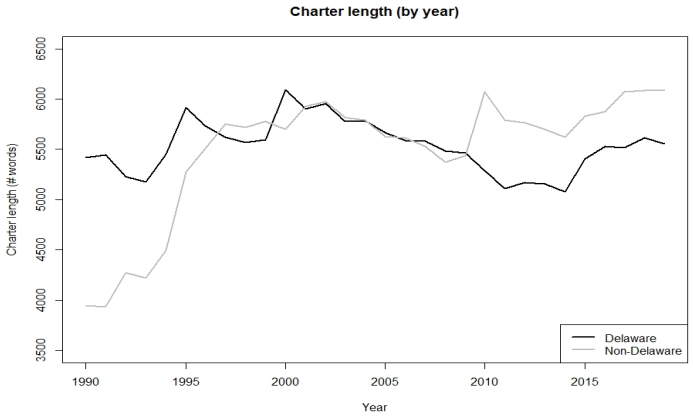
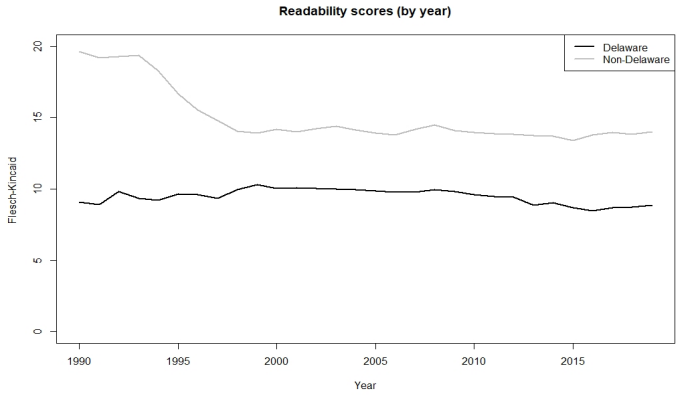


Figure 2 reports how the average length of charters has changed over time, with Delaware-incorporated issuers in black and non-Delaware companies in gray. Interestingly, the mean charter length for firms incorporated in Delaware has held steady (or even modestly shrunk) over the past 30 years, while that of non-Delaware firms has grown precipitously. In the early 1990s, a typical non-Delaware charter was 30% shorter than that of a typical Delaware-based corporation. In the last decade, in fact, this gap has not only closed but even been reversed slightly.[112]

[112] In Figure 2, we report the mean based on the length of all full charters in the dataset for active companies at a given point in time. This implies that the sample of companies (both in Delaware and elsewhere) changes over time. However, the results reported above remain substantially similar when we restrict the sample to companies that appear in the full panel of the dataset.

Figure 3: Mean Flesch-Kincaid Readability Scores (Full Restatements)



Another constructive measure of our charters corpus explores their overall readability. To what extent can a layperson read and understand the content of this foundational governance document? To get a handle on this question, we assessed our corpus of charters against the well-known Flesch-Kincaid (F-K) scale. Originally developed by the U.S. military to assess the content of mechanical instructional manuals, F-K scores are calculated on the basis of the average length of words and sentences in a document. The score proxies proportionally to readability, so that the higher the score of a document, the easier it is to read. F-K scores below a score of 10.0 are considered to be the most challenging, appropriate to a professional trained in the field. (Obvious candidate groups here might be lawyers, board members, and executives.)[113]

For the most part, as Figure 3 suggests, the charters in our corpus are not breezy page-turners. In fact, corporate charters in Delaware in particular have *always* been within the hardest tranche of the F-K scale. Perhaps more interesting is the fact that much like with length, non-Delaware charters in our dataset have been closing the readability gap, too. One potential explanation is more heavy "lawyering" of public company governance documents during the 1990s and early 2000s—a time period coinciding precisely with the rise of quantitative governance research and enhanced shareholder activism.

Although we will circle back to explore several of these (and other) textual attributes of our corpus later in this article, it warrants noting that this is the first time (to our knowledge) metrics like these have even been possible on a widespread panel of corporate charters.[114] That observation alone underscores the great potential of the CCG as a tool to unlock empirical governance along untold dimensions.

---

[113] For more detail on the F-K scores, *see infra* Online Appendix C.

[114] Nili and Hwang, for instance, used this technique to analyze audit committee charters. *See* Nili & Hwang, *supra* note 85.

B. Data Labels

Notwithstanding the several interesting acontextual measures of our corpus, significant additional work is required to distill the substantive legal content from the textual contents of documents (a process that is often referred to as "labeling" the corpus). Here, there is no substitute for reading the documents and deploying lawyerly judgment (an exercise that lawyers do quite well). Thus, in a parallel effort to the harvesting and cleaning of the raw charter corpus, we also develop two related labeled datasets.[115] The first (and most painstaking to produce) involves manually labeling the content of the harvested charters along several dimensions.[116] We developed a detailed common rubric that requires a variety of quantitative and textual inputs. The first few of these inputs relate to document "meta-data" (such as execution date, company identifiers, state of incorporation and whether the document was a full restatement or a partial amendment). The remaining pertain to substantive governance choices as reflected in the text of the charter. Our rubric requires the coder to read, identify, label, and extract relevant language from the charter pertaining to 28 governance provisions in each chartering document.[117] We took great care in both designing uniform labeling protocols and training our research team. We double- and triple-assigned identical labeling tasks to our less experienced coders to detect and redress labeling inconsistencies. Senior members of the team also acted as supervisors to adjudicate differences in coding styles and to flag challenging issues with the entire team in weekly progress meetings.[118]

Second, we supplement the *firm*-level observations with a *state*-level labeled dataset tracking sixteen statutory governance rules[119] across all 50 states (and the District of Columbia) from 1990 to 2019. Here the then-prevailing statutory mandates are labeled not only for their substantive content, but also whether companies are permitted to opt out of the statute, what sort of measures are required to effectuate an opt out, and whether companies opting out are required to choose from a constrained set of

---

[115] The term "labeled" refers to using human judgment to rank, classify, or assess the content of a text (or portion thereof). *See* JIAFENG GUO & YANYAN LAN, INST. OF COMPUT. TECH., CHINESE ACAD. OF SCIS., TOP-K LEARNING TO RANK: LABELING, RANKING AND EVALUATION 751-52 (2015), http://www.bigdatalab.ac.cn/~lanyanyan/papers/2012/SIGIR2012-niu.pdf. Nearly all existing corporate governance databases are themselves labeled databases. The CCG, in contrast, includes both unlabeled content (the raw corpus) and labeled content (described herein).

[116] We take some care to elucidate these steps here for the sake of future researchers who will use this database, and in light of the relatively opaque documentation that attends other corporate governance datasets. *See supra* Part I.

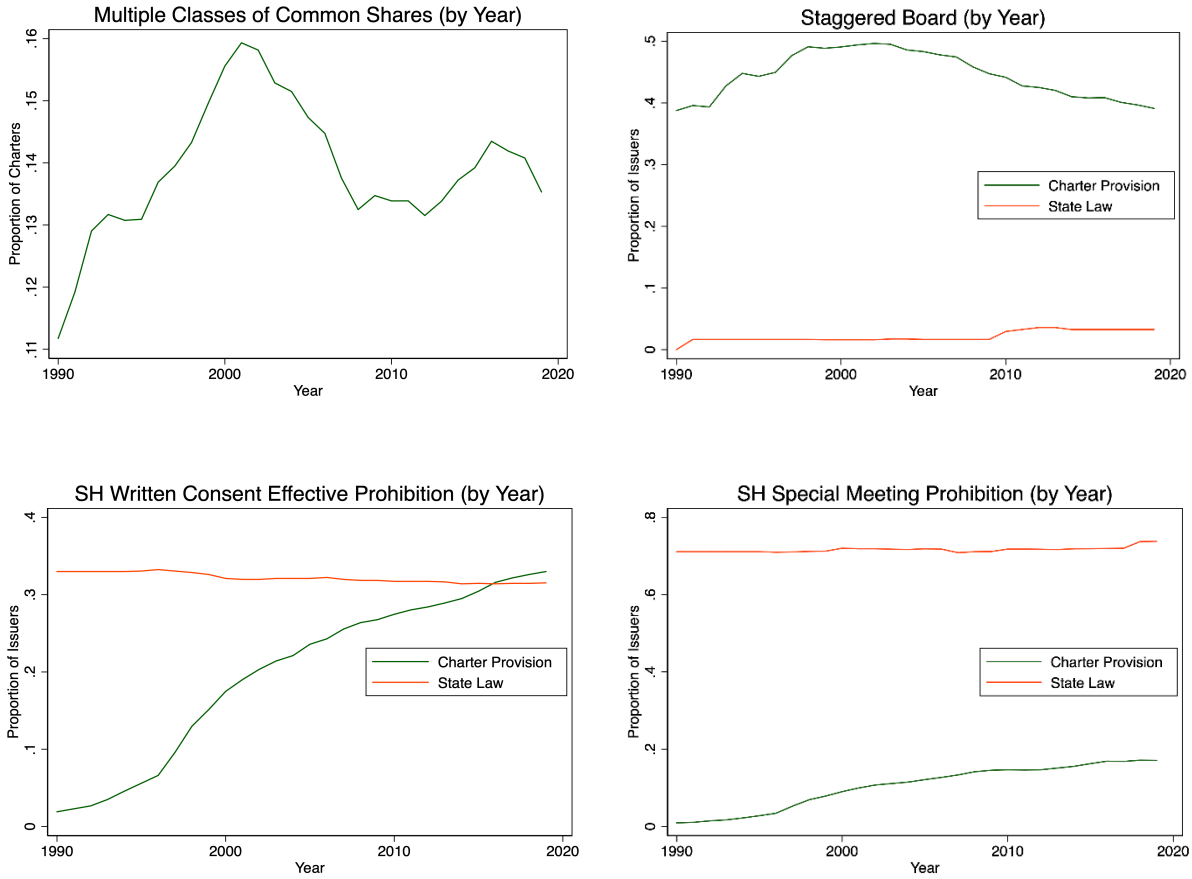
[117] Many of these pieces of information later made it possible to match our documents to issuer information from external datasets, as demonstrated below. A description of our labeling protocol is included in Appendix B.

[118] We used a formal organizational hierarchy similar to the one employed for charter harvesting. *See infra* Appendix B.

[119] *Id.*

options.[120] In all cases, state law provisions were labeled by either the authors or advanced law students under the direct supervision of the authors.

Figure 4: Mean Governance Characteristics of Charters and State Statutes[121]



When combined with the textual corpus, the associated labeled datasets allow us to track dozens of governance characteristics, across companies and over time. Although space constraints prevent us from highlighting every single facet here, we highlight a few interesting trends in the four panels of Figure 4. Each panel of the

[120] *See supra* Figure 1 and related text.

[121] Although three of the panels in the Figure reflect the substantive state background rule, statutory heterogeneity precludes illustrating other dimensions of the background rule, such as whether it is a default or immutable rule, what type of document (if any) is required to opt out, and what a company's opt-out choices are. These considerations, however, will come into play in the next subsection, when we use our database to assess (and correct) the contents of existing governance datasets, such as the IRRC.

figure tracks the extent to which charters in our corpus reflect one of four different types of governance provisions: (i) Multiple classes of common stock; (ii) staggered boards; (iii) prohibitions on shareholder action by written consent; and (iv) prohibitions on special meetings. In the latter three categories, we also track the substantive background content of the relevant state law provision(s) for each issuer (based on the substantive content of the then-prevailing statute for the state of incorporation). The interaction between state law and our charter coding is important, since state law may provide these governance provisions even if they are not specifically elected in the charter.

Consider first multi-class (or dual-class) common stock provisions. These provisions allow founders to maintain control of their companies even after their equity stake is dilute, and are a hot topic of debate among investors and governance experts.[122] Charters are required to spell out the capital structure of the corporation, and the provision for multiple classes of common stock is easily tracked. Issuers with multi-class common oscillate somewhat over our sampling period, rising continuously to a peak at just under 16% in 2001, generally declining thereafter. Although not pictured here, it also bears noting that when multiple classes of common stock are authorized in the charter within our sample, it is overwhelmingly likely that they will not carry equal voting rights and privileges on a per-share across the board. Around 78% of the issuers with multi-class common stock in our sample each year articulate unequal per-share voting rights, a ratio that remains roughly constant over time.

Moving to the second panel, consider staggered board composition—a structure that typically designates three overlapping classes of directors whose terms interlace (much like the US Senate), so that only one third of the directors are up for re-election in any given year. Staggered boards are typically considered to be a key way to delay and/or deter a hostile takeover or proxy contest.[123] Express provisions in the charter that stagger the board are common in our data, and we find them in around 45% of our charters overall. That said, the frequency of charter-authorized board staggering provisions has declined discernibly over the last several years, and by 2019 they were a clear minority (around 39%). This trend is no doubt due to the significant pressure that proxy advisers and other shareholder watchdogs have placed on board staggering

---

[122] *See e.g.*, Lucian A. Bebchuk & Kobi Kastiel, *The Untenable Case for Perpetual Dual-Class Stock*, 103 VA. L. REV. 585 (2017) (reviewing prior literature highlighting the costs of dual class and making the case against perpetual dual class shares); Lucian A. Bebchuk & Kobi Kastiel, *The Perils of Small-Minority Controllers*, 107 GEO L.J. 1453 (2019); Daniel R. Fischel, *Organized Exchanges and the Regulation of Dual Class Common Stock*, 54 U. CHI. L. REV. 119, 136–39 (1987) (arguing that dual-class stock facilitates long-term planning); Zohar Goshen & Assaf Hamdani, *Corporate Control and Idiosyncratic Vision*, 125 YALE L.J. 560, 566–67 (2016) (arguing that dual class could be value enhancing); John C. Coffee, Jr., Dual Class Stock: The Shades of Sunset, CLS BLUE SKY BLOG (Nov. 19, 2018), http://clsbluesky.law.columbia.edu/2018/11/19/dual-class-stock-the-shades-of-sunset/(describing academics and practitioners as "polarized" over dual class structures).

[123] *See* Gompers et al., *supra* note 6, at 145-52.

provisions in recent years.[124] Additionally, the background law of four states require (or at some point required) staggering of public companies,[125] and about 3% of our firms (in an average year) are incorporated in such states.

A topic of growing importance in contemporary governance debates is the extent to which shareholders enjoy significant latitude to engage (or be engaged) in activism, via written consent rights or the ability to call special meetings.[126] Each of these rights substantially tips the balance of power in control away from the board and towards the hands of shareholders. Such governance devices are reflected in the third and fourth panels of the Figure. With respect to written consent rights, our charters manifest a growing proclivity to either prohibit such actions outright or to effectively do by imposing a requirement that *all* shareholders entitled to vote must act unanimously via written consent (a functional impossibility for public companies). The bottom left panel lumps these two effective prohibitions together, and it shows that effective prohibitions on written consents are expressly provided for in about 16% of our charters overall; that fraction, however, has been growing dramatically, from next to nothing at the beginning of our sample to fully a third of issuers by the end. In addition, about a third of the issuers in our sample are subject to state laws that effectively prohibit written consent actions (all by imposing a unanimity requirement).[127]

Now consider provisions that prohibit shareholders from forcing the convening of special meetings. Here we once again see a clear increasing trend toward express prohibition through charter provisions over time, growing from near zero to around one fifth of the charters by the end of our sample period. However, in those cases explicitly granting shareholders a right to convene a special meeting, the mean triggering percentage is around 33% on average (and has been falling since the 1990s). In addition, the state law of several states does not grant shareholders the right by

---

[124] *Id.*

[125] *See* IOWA CODE § 490.806A; IND. CODE ANN. § 23-1-33-6(c); MASS. GEN. LAWS ch. 156B, § 50A, ch. 156D, § 8.06(b)–(g); OKLA. STAT. tit. 18, § 1027(D)(2) (repealed effective Mar. 5, 2013).

[126] *See* Gompers et al., *supra* note 6, at 145-52.

[127] *See* ALASKA STAT. § 10.06.423; ALA. CODE § 10A-2A-7.04; ARK. CODE ANN. § 4-26-710; ARIZ. REV. STAT. ANN. § 10-704; CAL. CORP. CODE § 603; COLO. REV. STAT. § 7-107-104; CONN. GEN. STAT. § 33-698; D.C. CODE § 29-305.04; FLA. STAT. § 607.0704; GA. CODE ANN. § 14-2-704; HAW. REV. STAT. § 414-124; IDAHO CODE § 30-29-704; IND. CODE ANN. § 23-1-29-4; KY. REV. STAT. ANN. § 271B.7-040; LA. STAT. ANN. § 12:1-704; MASS. GEN. LAWS ch. 156D, § 7.04; MD. CODE ANN., CORPS. & ASS'NS § 2-505; ME. STAT. tit. 13-C, § 704; MICH. COMP. LAWS § 450.1407; MINN. STAT. § 302A.441; MO.REV. STAT. § 351.273; MISS. CODE. ANN. § 79-4-7.04; MONT. CODE ANN. § 35-14-704; N.C. GEN. STAT. § 55-7-04; N.D. CENT. CODE § 10-19.1-75; NEB. REV. STAT. § 21-256; N.H. REV. STAT. ANN. § 293-A:7.04; N.J. Stat. Ann. § 14A:5-6; N.M. STAT. ANN. § 53-18-8; N.Y. BUS. CORP. LAW § 615; OHIO REV. CODE ANN. § 1701.54; OR. REV. STAT. § 60.211; 15 PA. CONS. STAT. § 2524; 7 R.I. GEN. LAWS § 7-1.2-707; S.C. CODE ANN. § 33-7-104; S.D. CODIFIED LAWS § 47-1A-704; TENN. CODE ANN. § 48-17-104; TX BUS. ORGS. § 6.201-202; UTAH CODE ANN. § 16-10a-704; VA. CODE ANN. § 13.1-657; VT. STAT. ANN. tit. 11A, § 7.04; WASH. REV. CODE § 23B.07.040; WIS. STAT. § 180.0704; W. VA. CODE § 31D-7-704; WYO. STAT. ANN. § 17-16-704.

default to convene special meetings, and such statutes affected about 70% of our firms (a number that includes Delaware corporations—at nearly three fifths of our sample).[128]

It is important to keep in mind that the interplay between state law and private ordering can sometimes be subtle in ways not fully captured in Figure 4. As we noted earlier, corporate governance documents can (and frequently do) attempt to opt out of the background state rule, which can (and frequently does) give way if the opt-out is executed appropriately. It is critical to keep track of all of these factors in assessing whether a particular governance device is present (or absent) in a company in any given observation year. For example, state law affects whether a board-staggering is already the presumptive rule in the jurisdiction. And state law similarly dictates whether opting out is possible, and if so, whether opting out must be done via the charter, or instead could be done in a lower-level document (shareholder approved bylaws, ordinary bylaws, board resolutions, and the like). And for those companies that opt out, state law may further constrain the number of overlapping classes of directors that are permitted when a firm opts out (often a maximum of three). The panels of Figure 4 account for only charter contents and background state rule as self-contained matters, with none of the additional interplay. In later sections, however, we take pains to carry through this interplay when we compare the CCG database to other existing corporate governance data.

C. Reassessing What We Know (or Thought We Knew) about Corporate Governance

The CCG database—including both the raw corpus and the labeled datasets—gives us a powerful set of new tools to analyze governance characteristics at the firm level. This ability, in turn, also makes it possible (for the first time) to tabulate side-by-side comparisons of the CCG database with other oft-used governance metrics. One that merits particular attention—and the most renowned source of firm-level corporate governance metrics in the law and finance literature—is the "ISS Legacy" (aka IRRC) database, discussed at length in the prior section. In the pages that follow, we set about comparing the CCG database to the individual items in the IRRC, in order to assess their accuracy. At the risk of issuing an academic spoiler alert, the results of this exercise reveal that our worst fears about data integrity have come to pass: as explained below, we uncover an alarming pattern of miscodes in numerous of the governance dimensions that comprise the G-Index (and its variations). The errors are so widespread, in fact, that *even under a conservative error-detection protocol, we estimate that the G-index is coded incorrectly more than 80% of the time.*

---

[128] *See* ALA. CODE § 10A-2A-7.02; DEL. CODE ANN. tit. 8, § 211; IND. CODE § 23-1-29-2; KAN. STAT. ANN. § 17-6501; MO. REV. STAT. § 351.225; N.C. GEN. STAT. § 55-7-02; N.Y. BUS. CORP. LAW § 602; OKLA. STAT. tit. 18, § 1056; OR. REV. STAT. § 60.204; 15 PA. CONS. STAT. § 2521; 7 R.I. GEN. LAWS § 7-1.2-701; S.C. CODE ANN. § 33-7-102; VA. CODE ANN. § 13.1-655.

But before going there, it is instructive first to illustrate with a specific example how we assess and classify potential errors: *director exculpation provisions,* one of the 24 variables included in the G-Index. Lawyers, professors, and corporate law students know these provisions well. In all states that permit and/or imply them (and by 2003 all did), an exculpation provision shields directors (and in some cases officers) from monetary liability for breaching their fiduciary duty of care. Such statutes do not typically deny injunctive relief, nor do they permit exculpation for conduct that (among other things) would be disloyal, would lack good faith, or would constitute corporate waste.[129] There is nonetheless some substantive variation among states' mandates. For example, in five states, the statutory rule exculpates directors automatically without needing an implementing charter provision.[130] And two of those states go even further, making director exculpation immutable.[131] Among the states where exculpation remains a default rule (in either direction), some permit firms to opt out of the rule via one or more governance documents that sit at a "lower" echelon than the charter (such as a bylaw provision).[132]

As far as we are able to discern, the IRRC database never considered much (if any) of the statutory heterogeneity described above, ignoring (for example) whether the state's background rule already exculpates directors, or how/whether a firm might opt out of that mandate. Instead, the IRRC seems to have limited its labeling attention to a counting exercise based on an issuer's corporate governance documents (which we conjecture focused on corporate charters, though we do not know for sure). Our approach, in contrast, pays close attention to the interplay between statutory mandates and governance documents.[133]

---

[129] Delaware's famous "102(b)(7)" statutory provision provides the most common template. DEL. CODE ANN. tit. 8, § 102(b)(7). There, as in the vast majority of other states, exculpation for a breach of duty of care is not preordained by statute, but instead may be adopted by the corporation through express exculpation provision if done via the charter. West Virginia was the last to add an exculpation statute, following this same enabling template in 2002, *see* W. VA. CODE § 31D-2-202. *But see* IND. CODE § 23-1-35-1 (where exculpation also covers loyalty and good faith).

[130] *See* FLA. STAT. § 607.0831; IND. CODE § 23-1-35-1; NEV. REV. STAT. § 78.138; OHIO REV. CODE ANN. § 1701.59; WIS. STAT. § 180.0828.

[131] *See* FLA. STAT. § 607.0831; IND. CODE § 23-1-35-1.

[132] These include Pennsylvania, 15 PA. CONS. STAT. § 513 (shareholder-promulgated bylaw); Minnesota, MINN. STAT. § 302A.111 (bylaw); Ohio, OHIO REV. CODE ANN. § 1701.59 (bylaw/regulation) and Utah, UTAH CODE ANN. § 16-10a-841 (bylaw and/or shareholder-promulgated resolution).

[133] To put a finer point on it, a corporation's directors may enjoy exculpation protection in one of multiple scenarios: (1) the state's background rule already prescribes exculpation subject to an immutable rule; (2) the state's background rule prescribes exculpation as a default rule, and the corporation has not attempted to opt out with using statutorily prescribed means; or (3) the state's statute prescribes a default rule that does not prescribe exculpation, but allows the corporation to contract out pursuant to a statutorily prescribed means and the corporation has done so. The IRRC database does not appear to have used any of these criteria, but instead uses a less-specific version of

To assess the accuracy of the IRRC against our CCG database, we entertain alternative strategies for identifying mis-codes, which we define as "permissive" and "conservative" strategies. The permissive approach would simply ask whether there is an inconsistency between what we observe in a company's charter and what is contained in the IRRC. Under this approach, any inconsistencies are deemed to be errors by the IRRC. The conservative approach, in contrast, identifies an inconsistency as a miscode only if it is *impossible* (in light of charter text and the underlying statutory framework) for the IRRC's designation to be correct as a legal matter.

To see how these criteria work in practice, consider again the exculpation example. Suppose ABC Inc. is a Delaware corporation whose 2006 charter comprises part of our dataset. Suppose further the IRRC represents that the company exculpated directors in 2006, but our corpus does not reflect any such provision in the charter for that year. The permissive approach would immediately deem this inconsistency to be a miscode. The conservative approach requires more steps, taking account of the fact that, as of 2006: (a) Delaware law *did not* grant default exculpation to directors; (b) Delaware *did* allow opt-outs through an exculpation provision; and (c) any attempt to opt out *must have been* reflected in the *charter* to have legal effect. Applying these criteria, the conservative approach would also register an error: for the lack of an express exculpation term in ABC's 2006 charter, combined with the contours of Delaware law as of 2006, necessarily imply that the IRRC label could not possibly be correct.[134]

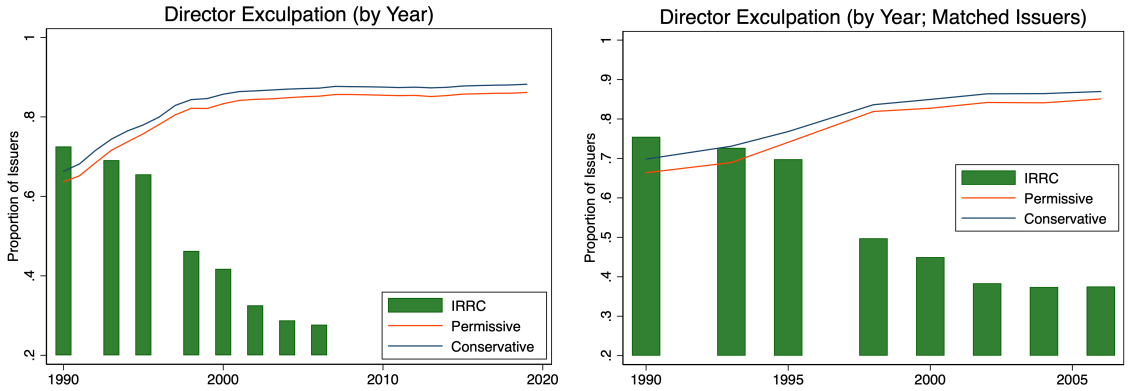
The two panels of Figure 5 compare the IRRC's exculpation designation to the CCG dataset over time, under each of these two aforementioned approaches. The left panel compares the IRRC designations for all issuers by year (bars) against all issuers of the CCG dataset (lines). The red line corresponds to the permissive approach, tracking the mean number of issuers per year that have an express exculpation provision in the company charter. The blue line tracks the conservative approach, factoring in the statutory regime of each firm's state of incorporation as well. The right panel renders a similar comparison, but it confines attention to only the set of companies where we have a positive match between the CCG and IRRC datasets (a limitation that drops all but the observation years covered by the IRRC, scattered sporadically between 1990 and 2006).

---

criterion (3) with no attention to the statutory background requirements. Our approach, in contrast, marshals all three steps described above.

[134] Notice that our conservative error-detection protocol might reach a different outcome from the permissive one if ABC were incorporated in a different state—such as Ohio—which allows corporations to opt out through a lower "echelon" document (i.e., a bylaw provision). In this case, if the IRRC reflects exculpation but we do not observe such a provision in the charter, we cannot deem a miscode to have occurred under the conservative rubric, since it is at least *possible* that the firm executed its exculpation regime through a bylaw provision (which our chartering corpus and labels do not track).

Figure 5: CCG-IRRC Comparison of Director Exculpation Provisions



As one can see from the CCG trendlines—and consistent with the longstanding view of judges and practitioners[135]—exculpation provisions have grown close to ubiquitous, and they were already on a strong growth trajectory by 1990, shortly after Delaware began by statutory change to allow them in 1986.[136] By 2006, around 85% of all issuers (and 96% of Delaware corporations) had such provisions in their charter. In contrast, and for reasons that would likely befuddle most corporate lawyers, the IRRC data suggest a strong opposite trend, implying that only 27% of all issuers (and 32% of Delaware corporations) exculpated directors by 2006. This striking divergence is present in both the unmatched and matched subsamples, and it persists even if we ignore state law super-structures, limiting our attention to express provisions in the charter.[137] Note further that the IRRC's miscoding problem appears to grow *worse* (not better) in time. Lacking a helpful description of the IRRC's labeling protocols, we can only speculate why the dataset appears so alarmingly inaccurate on exculpation. One possibility—consistent with machine learning text analysis we describe in the next section—is that drafting protocols for charter provisions likely became more "lawyerly" and technical during much of the early 2000s—a transition that may have caused readers with limited legal training to overlook exculpation terms that featured technical language.

---

[135] *See* Randy J. Holland, *Delaware Directors' Fiduciary Duties: The Focus on Loyalty*, 11 PA. J. BUS. L. 675, 691-93 (2009).

[136] DEL. CODE ANN. tit. 8, § 102.

[137] Tabulating at the company level, the IRRC misclassifies exculpation rights 51% of the time using the permissive approach and 52% of the time using the conservative approach. Note that the conservative approach yields a higher error rate (at least on this dimension), because by construction it takes into account state-level statutory provisions—which IRRC appears to ignore completely.

D. Aggregating G-Index Errors

Director exculpation provisions are but one example of seventeen individual governance characteristics on which the CCG enables us to audit the accuracy of the IRRC and G-Index. We conducted a similar set of comparisons to other analogous variables in our labeled dataset, and the results of this comparison are captured in Table 1. To simplify, the table reports error rates using the permissive rubric, and it defines a provision to be a "Positive" when the IRRC reflects it to be present and a "Negative" otherwise. If the IRRC coding matches our charter coding in the CCG database, we further deem the IRRC coding to be "True" and otherwise "False". The table thus tracks "True Positive" ("True Negative") designations—where our data and the IRRC agree about the presence (absence) of a provision—as well as "False Positive" ("False Negative") designations, in which the IRRC indicates a provision to be present (absent) and our labeled data reveal the opposite. The table then lays out correct classification rates, error rates, and F1 scores (a conventional way to assess classification accuracy balancing false positives and false negatives[138]).

Because of lack of documentation on definitions in the IRRC data, the Table defines certain features according to multiple criteria. We found that blank check preferred stock could fall into one of several categories. We also found that supermajority charter amendment provisions could be construed narrowly (only if they pertain to the entire charter) or broadly (also if they pertain to the whole charter or enumerated portions of the charter). Note from the table that while some of the metrics are relatively sound, others are particularly problematic. Averaging across all listed dimensions, our data suggest that the IRRC data errs at least 20% of the time.

[138] The F1 designates the harmonic mean between "precision" (the fraction of true positives to all *classified* positives) and "recall" (the fraction of true positives to all *actual* positives). The score is bounded between 0 and 1, with higher scores suggesting a more accurate classification. F1 is a commonly used metric in text analysis and binary classification. For more, *see* Pozen, Talley & Nyarko, *supra* note 15, at 33.

Table 1: Inconsistencies Between the IRRC and CCG Data Labels

	True +	True -	False +	False -	Correct %	Error %	F1
Unequal Vote	0.45%	98.60%	0.82%	0.14%	99.05%	0.95%	1.00
Merger Supermajority	14.16%	54.73%	2.64%	28.47%	68.89%	31.11%	0.78
Written Consent	19.29%	33.81%	18.40%	28.50%	53.10%	46.90%	0.59
Special Meeting	28.78%	6.80%	11.59%	52.83%	35.58%	64.42%	0.17
Lim Amend Chtr (whole + part)	2.56%	35.96%	0.49%	61.00%	38.51%	61.49%	0.54
Lim Amend Chtr (whole)	0.87%	93.60%	2.18%	3.36%	94.46%	5.54%	0.97
Dual Class	10.36%	85.04%	0.80%	3.80%	95.40%	4.60%	0.97
Director Liability Waiver (DoC)	38.08%	11.30%	6.30%	44.32%	49.39%	50.61%	0.31
Director Indemnification	6.63%	54.68%	20.84%	17.86%	61.30%	38.70%	0.74
Cumulative Voting	4.52%	86.94%	7.22%	1.32%	91.46%	8.54%	0.95
Classified Board	45.70%	37.54%	14.05%	2.71%	83.24%	16.76%	0.82
Blank Check (Simple)	86.39%	8.23%	3.24%	2.13%	94.62%	5.38%	0.75
Blank Check (Any)	50.08%	11.07%	37.61%	1.24%	61.15%	38.85%	0.36
State Law: Return of Profits	14.32%	82.91%	0.90%	1.87%	97.23%	2.77%	0.98
State Law: Fair Price	30.72%	65.82%	2.85%	0.61%	96.54%	3.46%	0.97
State Law: Director Duties	4.03%	71.23%	0.34%	24.40%	75.26%	24.74%	0.85
State Law: Cont. Share Acq.	26.94%	71.28%	1.12%	0.66%	98.22%	1.78%	0.99
State Law: Cash-Out	3.21%	96.58%	0.04%	0.18%	99.79%	0.21%	1.00
State Law: Business Comb.	89.68%	8.61%	0.99%	0.72%	98.29%	1.71%	0.91

These all seem like relatively large overall error rates. What does it imply for the G-Index, which amalgamates all of them? As discussed earlier, the G-Index has become easily the most prominent corporate governance index in the literature, spawning litany of variations. And since the index is computed by summing the several indicator variables in the IRRC database, we are also in a unique position to use our labeled dataset to reevaluate the G-Index, correcting it on an item-by-item basis in situations where we found evidence of a clear miscode.

To implement our corrections, we utilized the conservative approach described above to detect and then correct miscodes. For each firm observed in each year of the IRRC, we began with the value of the G-Index as reported in the IRRC dataset. For the issuers we could match, we moved incrementally through the binary governance variables one at a time, determining (per the conservative approach) whether there was an *unmistakable coding error* in the IRRC. If none was found, we moved onto the next variable. If there was an unmistakable coding error, we manually corrected the value of the index by one point upwards or downwards (depending on the variable). We repeated this process *ad seriatim* for all matched issuers and all years in the IRRC

database until we had exhausted the list of labeled variables tracked in the CCG data that corresponded to elements of the G-Index.[139]

By way of illustration, recall that in our director exculpation example above, it became clear that the IRRC erroneously indicated that ABC Inc. had exculpated directors when in fact it had not. Because the existence of an exculpation provision would ordinarily imply an increase in the firm's G-Index during that year, it follows that the G-Index for ABC must have been erroneously increased by one, reflecting this faulty exculpation designation. Our protocol corrected that mistake, reducing the G-Index of the firm in that year by one point. A similar process—crossing charter contents and state law—applies to all other analogous variables comprising the index.[140]

We reiterate that we take care to make these adjustments only when there is a *clear error* in the IRRC, as per the conservative approach. Thus, in some cases where our labeled datasets and the IRRC conflict, we might still refrain from making an adjustment, because we cannot satisfy the clear error standard. We thus give the IRRC the benefit of the doubt even in cases where we have grounds to suspect a coding error.

Consider first our assessment of errors along a very simple metric: For those firms and years where a comparison was possible, how frequently was the reported G-Index score incorrect? As illustrated in Table 2, the answer is alarming. Averaging over all years and all matched companies, *we find the G-index to be inaccurate over four-fifths of the time* (82.95%). More disconcertingly, as with director exculpation, we find the incidence of error *grows* in magnitude over time (from 73.68% in 1990 to 88.58% in 2006), even as the database was generating increased attention among academics, regulators and practitioners.

We emphasize that this error rate is almost certainly a lower bound, since we deployed the conservative error correction rubric, intervening only for unambiguous errors; we made no corrections to *probable* errors when it was still possible that the IRRC reflected a provision not in our corpus (such as bylaws, contracts, and so forth). The estimated error rate also errs on the conservative side because our labels do not track every single one of the elements that comprise the index. Notwithstanding these constraints that bias our miscoding estimates downward, this is a distressing error rate for a core dataset that has long been the very foundation of empirical corporate governance research.

---

[139] When we were unable to match an IRRC firm to the CCG, we left the G-Index intact. We exclude these unmatched firms in the discussion that immediately follows, but we include them in several of our replications of the GIM results in the subsequent subsection.

[140] A description of our protocol is included in Appendix B.

Table 2: G-Index Coding Error Incidence, by Year (Matched Issuers)

Year	No Error (%)	Error (%)
1990	26.32	73.68
1993	23.22	76.78
1995	22.47	77.53
1998	20.8	79.2
2000	15.74	84.26
2002	13.75	86.25
2004	12.99	87.01
2006	11.42	88.58
Total	17.05	82.95

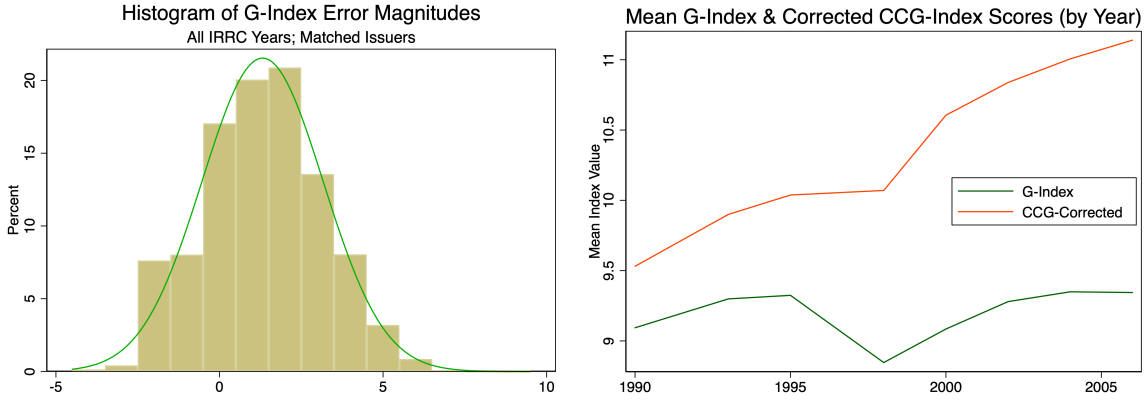
Of course, the mere existence of an error in the index may not be fatal if its magnitude constitutes something akin to rounding error. In other words, if the overall size of the observation error remains modest, perhaps the noise detected above might not be too troubling.[141] Our data permit analysis here as well; and we also find substantial cause for concern. The histogram in Figure 6 demonstrates the overall magnitudes of the G-Index errata (even with our conservative rubric for correction). It exhibits considerable variation in our matched firms, with an overall standard deviation of 1.83. Relative to the variability of the G-index as a whole (whose standard deviation is 2.71[142]), that is a distressing noise to signal ratio.

The aggregated errors we detect using the CCG database not only introduce considerable *noise*, but also a discernible *bias*. Specifically, our corrections reveal an additional downward bias in the G-Index of around -0.75 points overall, one that *grows worse* (not better) over time, as the right-hand panel of Figure 6 illustrates. By 2006, even under our conservative re-coding protocol, G-Index retains roughly the same degree of measurement error variability, but compounds it with a downward bias of nearly 2 points.

[141] We allow for this possibility with a scholarly grain of salt, given the 80%+ error rate reported in the text. Indeed, given this error rate, the best-case scenario for salvaging the G-Index against the woes of measurement error would be if it were exactly right 20% of the time, too low by one point 40% of the time, and too high by one point the remaining 40%. In that case, the overall index would be unbiased on average, but the errors would still have a standard deviation of around 0.9, fully one-third the size of the standard deviation in the G-Index itself (of 2.71 in our matched data).

[142] As noted above, the standard deviation of the G-Index for our matched company-years is also 2.71, suggesting that our matching protocol rendered a representative set of matches.

Figure 6: Miscode Error Magnitudes and CCG Corrections to G-Index



At the risk of some overkill, we emphasize once again that our error-correction approach errs strongly on the side of conservatism in several ways. First, we re-assessed only 17 of the 24 criteria in the GIM database that are confirmable from our charter texts and labeled datasets. Moreover, we avoided hazarding a guess (even an educated one) on whether various dimensions were likely miscoded, and we only corrected those inputs that we could be certain were mis specified, leaving in place designations where we could not identify an error with certainty.

It warrants noting that our choice to highlight flaws in the G-Index and IRRC specifically is strictly a matter of authorial choice, warranted (in our view) by the index's prominence in the literature and its centrality to appreciable follow-on research (such as the E-Index, O-Index, and D-index to name a few). That said, we are by no means limited to this particular single comparison: The CCG can easily be recruited into a quantitative battle of the bands with other well-known governance databases. In Online Appendix C, for example, we show how the CCG stacks up against a different source: the *contemporary* ISS governance database (2007-present)—one that itself has attracted considerable academic attention (but contains insufficient data to reassemble the G-Index). There, we find a disconcerting pattern of misclassification rates too, comparable to those illustrated above, reinforcing our concerns about data integrity and accuracy.

E. The Arbitrage Value of Good Governance Revisited

Having shown that the CCG dataset exposes disconcerting errors in popular corporate governance indices, an immediate next question concerns assessing what that implies for the field more broadly. As noted above, scores of "folk wisdoms" from empirical corporate governance were generated from these data, many of which are now among the most well-known in the field. Do they hold up?

Testing all the implications would take more time, space, and reader patience than is achievable here; but for the sake of illustration, we return once again to the G-Index, using our CCG corpus to reassess one of the most famous empirical findings in all of law and finance, from GIM.[143] As noted above, their seminal examination introduced the G-index, for the first time peering into granular, firm-specific governance practices and linking them to various financial performance metrics. Perhaps the most famous finding was that "good governance" is also an arbitrage opportunity. More precisely, GIM showed that a "long-short" investment strategy of buying "democratic" firms (in the lowest decile of the G-index) and short-selling "dictatorial" firms (in the highest decile) would systematically outperform the market by 71 basis points *per month* (the equivalent of an astonishing 8.9% per year[144]). This finding was (and remains) a striking result, a serviceable cocktail-party table topic, and a veritable modern-day slogan that informs policy and academic debates in corporate governance the world over.[145]

A now-obvious red flag, however, is that the result depends critically on the contents of the IRRC database—the very same resource that we showed to be alarmingly inaccurate. Might this inaccuracy have infected their ultimate results (not to mention the hundreds of scholarly contributions that followed after)?

The answer is not clear on *a priori* grounds. On the one hand, the GIM results could grow even stronger once the miscodes are rectified: This is what one might ordinarily expect when an independent variable (like G-Index) is muddled by random observation error.[146] In that case, cleaner data would be expected to sharpen and amplify the original results. On the other hand, data infirmities can sometimes be the root cause of an evident statistical result, typically when such infirmities are not the artifact of random noise. In this case, the effect of cleaner data can weaken the results. And for the G-Index errors, it is not clear *a priori* which of these stories is most likely to hold.[147]

---

[143] Gompers et al., *supra* note 6.

[144] For those keeping score, a basis point is simply 1/100 of a percentage point, and thus conversion into an excess return is given by $(1+0.0071)^{12} - 1 = 0.089$, or 8.9%.

[145] *See supra* notes 45-59 and accompanying text.

[146] This effect is sometimes referred to as "classical" measurement error, and it represents the case where the measurement error of a variable is independent of the variable's "true" value. *See* John Bound, Charles Brown & Nancy Mathiowetz, *Measurement Error in Survey Data*, *in* 5 HANDBOOK OF ECONOMETRICS 3705–843 (James J. Heckman & Edward Leamer eds., 2001); Darren H. Lubotsky & Martin Wittenberg, *Interpretation of Regressions with Multiple Proxies*, 88 REV. ECON. & STAT. 549 (2006); Aaron Chalfin & Justin McCrary, *Are U.S. Cities Underpoliced? Theory and Evidence*, 100 REV. ECON. & STAT. 167 (2018).

[147] Our best a priori guess would be on the latter, however. Because the G-Index is the aggregation of two dozen dummy variables, it is a likely candidate for *non-classical* measurement error. The most extreme version of this effect is for binary variables (where measurement error can never be classical by definition), but this problem afflicts bounded discrete variables too. Black et al., *supra* note 77, at 743. In fact, it merits observing that the most well-known result in GIM—which focuses on the two most

To investigate these questions, we set about reanalyzing the GIM governance-arbitrage result using the CCG data. The fruits of this effort are contained in Table 3 (which is fashioned after the portion of GIM that reports on the arbitrage result[148]). The first column of the table simply reproduces their reported results, with each row associated with an approximate decile of governance-indexed portfolios, ranging from no greater than 5 (democratic firms) to no less than 14 (dictatorial firms). The estimates reported in the table represent the unexplained return (or "Alpha") generated by an investment in a value-weighted portfolio drawn from that decile,[149] and consequently taking a "short" position in that portfolio would generate the same return with the opposite sign. Consequently, the difference between the most democratic portfolio's Alpha and most dictatorial portfolio's Alpha represents the unexplained return of the long-short investment described above. That difference generates their celebrated 71-basis-point monthly return. The second column represents our best effort at an exact replication of GIM's results with historical data from the IRRC. The remaining columns represent a re-estimation of their results using our CCG-corrected data under a variety of approaches.

As one can see from the table, our exact replication (column 2) is nearly spot on with the original, bolstering confidence that we can, in fact, substantially replicate GIM's findings with their own data. The third column represents our replication using the original GIM dataset, but one where we substituted the CCG-Index (i.e., the 'corrected' G-index value as described above) for the reported one whenever we were able to generate a correction. It bears noting that during this period of years (1990-1998), we faced limitations in matching up our dataset, and thus only about 42% of the GIM firm-years could be checked and corrected. In unmatched issuers, we simply continued to use their originally reported scores. Even with this modest correction, however, we estimate a discernible lower unexplained return to 59bps (representing a one-sixth attenuation from the original, the equivalent of a reduction in the unexplained annual spread from 8.9% to 7.3%). In the fourth column, we perform a similar analysis, but with a recently downloaded version of the IRRC database (which appears to have been modified slightly in 2010, subsequent to GIM's original analysis), and the unexplained return dropped even further to 49.9bps (representing a reduction in the unexplained annual spread from 8.9% to 6.2%). Finally, in the fifth column we returned to the original IRRC dataset and ran our results with only our matched

extreme deciles in the G-Index—is *especially* likely to exhibit non-classical measurement error. *See* Bound et al., *supra* note 146.

[148] *See* Gompers et al., *supra* note 6, at 123 tbl.VI.

[149] Specifically, the reported alpha values corresponded to the constant term of a four-factor Fama-French-Carhart asset pricing estimation on monthly returns for each decile's value-weighted portfolio. By construction, this constant represents the average return that cannot be explained by conventional asset pricing factors. *Id.*

firms.[150] Here, we estimate an unexplained alpha of 55.9bps (representing a reduction in the unexplained annual spread from 8.9% to 6.9%).

Table 3: Replication and Robustness of Good Governance as Arbitrage

	Original GIM	Exact Replication (Historical IRRC)	Replication with CCG-Corrections (Historical IRRC)	Replication with CCG-Corrections (Updated IRRC)	Replication With Matched Firms (Historical IRRC)
G ≤ 5 (Democracy)	0.29*	0.26	0.175	0.118	0.334
	0.13	*0.14*	*0.17*	*0.15*	*0.20*
G = 6	0.22	0.189	-0.005	-0.082	-0.021
	0.18	*0.19*	*0.17*	*0.18*	*-0.21*
G = 7	0.24	0.234	0.161	0.112	0.285
	0.19	*0.19*	*0.21*	*0.19*	*0.24*
G = 8	0.08	0.017	0.264	0.264	0.382
	0.14	*0.14*	*0.16*	*0.15*	*0.19*
G = 9	-0.02	-0.066	-0.173	-0.185	-0.203
	0.12	*0.12*	*0.13*	*0.13*	*0.16*
G = 10	0.03	0.012	0.134	0.154	0.246
	0.11	*0.11*	*0.14*	*0.13*	*0.18*
G = 11	0.18	0.137	0.043	0.051	0.142
	0.16	*0.16*	*0.14*	*0.14*	*0.20*
G = 12	-0.25	-0.283	-0.167	-0.172	-0.253
	0.14	*0.15*	*0.15*	*0.16*	*0.19*
G = 13	-0.01	-0.066	-0.09	-0.106	-0.195
	0.14	*0.14*	*0.15*	*0.14*	*0.21*
G ≥ 14 (Dictatorship)	-0.42*	-0.438*	-0.415*	-0.381*	-0.225
	0.19	*0.18*	*0.17*	*0.17*	*0.20*
Democracy-Dictatorship (bps)	71.0	69.8	59.0	49.9	55.9
Attenuation from Original (in %)	-	-1.69%	-16.90%	-29.72%	-21.27%
Implied Annual Excess Return	8.9%	8.7%	7.3%	6.2%	6.9%

Performance attribution regression of Democracy - Dictatorship Portfolios; 1990-1998. The first column restates the estimates from Table VI of Gompers, Ishii & Metrick ("GIM" 2003). The second column reports our attempt at an exact replication. The remaining three columns are replication robustness checks using CCG-corrected data for a variety of comparison samples. Coefficient estimates reflect unexplained return (α) values from Fama-French four-factor portfolio regressions. Standard Errors in italics. (* = 0.05 significance; ** = 0.01 significance)

All told, our results strongly suggest that not only do the errors in the G-Index affect results, but that they do so in a disconcerting way. In each of the reported replications that use our corrected CCG-Index, the estimated abnormal return *grows weaker*—exactly the opposite movement from what one would expect had the G-Index merely been hamstrung by garden-variety measurement errors. Rather, each replication drives appreciably downward the estimated extent to which "good governance" predicts abnormally good returns to investors. Averaging across the columns, approximately a quarter of the 8.9% premium reported by GIM dissolves in the presence of corrected data.

[150] We confirmed that the GIM analysis with the uncorrected G-index data and matched firms delivers estimates almost exactly on par with the results reported in column 1.

It is worth noting that even with our corrections, the GIM arbitrage result does not "go away" completely. And accordingly, one could entertain the possibility that the original result—albeit reliant on imperfectly coded data—merely generated a result that was still 'real' but just a little too rosy. While we cannot rule out this possibility, we are skeptical. The fact that the attenuation effect is discernible after correcting only the matched firm-years (42%) with a deliberately conservative rubric raises serious concerns that Table 3 vastly understates the effect of the problem. And, reiterating our point above, the directionality of the change is particularly concerning, since it moves counter to what one would expect to see with classical measurement error adding just mere noise to the effect.

Unpacking and testing these possibilities are beyond scope of this paper, so we leave that project to future research (at least for now). And there may be considerable unpacking left to do: as noted above, a substantial number of empirical corporate governance contributions of the last two decades rely on the same data sources as did GIM. At the same time, the open-source nature of the CCG database means that many scholars can participate in this enterprise. The above exercise (or something close to it) can be used to revisit the results of dozens of well-known empirical corporate governance results in the literature.[151]

III. Corporate Governance as "Big Data"

Although the CCG provides a powerful and novel way to reevaluate several old chestnuts of corporate governance, its primary—and most exciting—use is prospective. This section spotlights some of the ways that future scholars and policymakers can make use of the CCG. In particular, the CCG's underlying textual corpus allows for emergent techniques that uses machine learning and computational text analysis.[152] We explore some preliminary findings of our own using those techniques here. Such techniques, when applied to the CCG and its underlying corpus, have the potential both to improve on the accuracy of conventional empirical practices and to broaden the horizons of corporate governance research.

A. Document-Level Trends

First, the CCG corpus allows us to identify some interesting document-level trends. In Part II, we noted that some of the most rudimentary measures of our charter documents helped show that charters of non-Delaware companies have become longer and less readable over time, effectively converging with those of Delaware-

[151] *See* Bebchuk et al., *What Matters*, *supra* note 11; Karpoff et al., *supra* note 11; Madanoglu & Karadag, *supra* note 44; Straska & Waller, *supra* note 10; Harford et al., *supra* note 49; Core et al., *supra* note 11; Adams & Ferreira, *supra* note 51; Bebchuk et al., *Disappearing Association*, *supra* note 11; Daines et al., *supra* note 52.

[152] *See* sources cited *supra* note 15.

incorporated companies. Here, we use computer processing of these documents to explore several other metrics.[153]

Type-Token Ratios. The "Type-Token Ratio" (TTR) is a common metric that represents the ratio of unique terms divided by the total number of words in the document. This metric helps researchers understand a document's repetitiveness and redundancy.

We illustrate our TTR findings in Figure 7. The top two panels show the mean TTR for all charters that were in place in a given year, distinguishing between Delaware and non-Delaware issuers, and also subdividing between the full restatements only, and the complete set of effective charter documents in place in a given year (tacking on any partial amendments).[154]

Our analysis shows that the overall TTR ratios of charters have generally declined over time under either measure. As was the case for changes in charter length and readability, most of this change has been outside of Delaware, with non-Delaware issuers converging with their Delaware counterparts by the end of our sample. This trend suggests that much of the growth in length of charters outside Delaware is accompanied by a greater tendency towards repetitiveness and/or redundancy.

Syntactic Similarity. Another interesting machine-learning measure concerns inter-document comparisons, which focus on assessing the similarity between two or more documents. It is unclear what trends to expect here: the increasing attention on "tailored" corporate governance regimes might lead one to predict that governance documents have grown less similar to one another over time. However, because shareholder activists, arbitrageurs, and proxy advisory firms—among others—have become more sophisticated in recent years, we might also expect that results lead in the other direction.

One common technique in machine learning of assessing document similarity is to measure the extent to which a numerical representation of one document (represented by a mathematical vector) aligns with that of another document.[155] Many anti-plagiarism and e-discovery algorithms use this approach, and it is often captured through a "cosine similarity" statistic that ranges from 0 (reflecting utter dissimilarity)

---

[153] The use of these techniques requires some preprocessing that is by now pretty standard in the field of computational text analysis. All measures presented below are based on the text of charters from which we stripped any content other than words as well as so-called stop words. After that, we used a technique called stemming to avoid treating simple inflected variations of words as different words.

[154] For more details on the treatment of piecemeal amendments within the corpus, *see supra* note 108; *infra* Appendix B.

[155] Jeremy McClane, for example, has used this technique to consider the role of boilerplate in securities disclosure. *See* Jeremy McClane, *Boilerplate and the Impact of Disclosure on Securities Dealmaking*, 72 VAND. L. REV. 191 (2019).

to 1 (reflecting complete similarity).[156] We compute this measure and use a set of plots similar to Figure 7 to track it over the span of the CCG dataset.[157] The Figure illustrates the mean cosine similarity of each charter document with the charter that is most similar to it during that same measuring year (its "nearest neighbor"). As shown in the figure, inter-document similarity has also grown discernibly during the span of the CCG dataset (both inside and outside of Delaware).[158]

Figure 7: Mean Type-Token Ratio and Nearest-Neighbor Cosine Similarity

[156] Cosine similarity is a popular similarity measure because it is independent of document length.

[157] We employed a standard set of computational text analysis transformations before making these computations. *See infra* notes 162-163 and accompanying text for greater elaboration.

[158] One potential explanation for the growth in similarity is the fact that the numbers of charters available in the CCG grows over time. However, this effect cannot fully explain the growth in similarity. Most importantly, while the numbers of charters of active companies peaked in 2006, similarity scores continue increasing after that.

Document measures such as these are helpful to explore industry- and firm-level characteristics. Do these types of measures reflect (or predict) attributes about the nature of the company's business, size, and profitability? Table 4 highlights a variety of document measures across these three slices of the CCG.

Table 4: Charter Content Means, by Industry, Asset Value & Tobin-Q Categories (Observations at the Firm-Year Level)

	Charter Length	F-K Readability	Type-Token Ratio	Cos-Sim Nearest Neighbor
Construction	3,545.78	9.0497	0.2581	0.5156
Finance Insurance & RE	8,272.08	9.3909	0.2244	0.5189
Manufacturing	5,033.04	11.1322	0.2542	0.4745
Mining	4,840.77	9.8136	0.2595	0.4586
Retail Trade	4,905.25	13.2327	0.2600	0.4678
Services	4,611.40	14.1941	0.2603	0.4748
Transp Comm Elec	6,031.25	9.7958	0.2460	0.4391
Wholesale Trade	6,059.22	10.6836	0.2453	0.4844
Other	2,849.93	11.8280	0.3135	0.4709
Total	5,631.00	11.1363	0.2496	0.4785

Panel A: SIC Industry Group

Quartile 1	3,270.17	19.1155	0.3141	0.4416
Quartile 2	3,930.50	17.4098	0.2845	0.4508
Quartile 3	4,561.56	13.3036	0.2678	0.4657
Quartile 4	6,345.16	9.3656	0.2365	0.4879
Total	5,634.15	11.1338	0.2495	0.4786

Panel B: Asset Value Quartile

Quartile 1	8,240.46	8.5174	0.2277	0.4966
Quartile 2	5,666.74	10.1874	0.2393	0.4752
Quartile 3	5,057.70	11.2232	0.2526	0.4770
Quartile 4	4,161.73	13.8544	0.2718	0.4854
Total	5,632.49	10.9320	0.2478	0.4815

Panel C: Tobin's Q Quartile

Consider first how our corpus breaks down by industry. Here, the content of corporate charters appears to fluctuate considerably, both across industries and (in some cases) over time. Panel A reports on the mean of several of the metrics discussed above, but this time disaggregated across different 1-digit SIC sectors. Several characteristics stand out. Most notably, issuers in the Finance sector stand out across

all of the aforementioned measures: they are long (over 8,000 words on average), repetitive (scoring lowest on unique word ratios and type-token ratios) complex (scoring second lowest in F-K readability), and overall, quite similar (scoring highest in cosine similarity). On the other end of the spectrum, the service sector typically has charters that are relatively short, tailored, easy to read, and less emulatory.

Larger companies are generally thought to be more complex, and Panel B of the Table bears this intuition out, disaggregating the population into size quartiles according to total asset value (as reported in the issuer's 10K for the year observed). Remarkably, *all four document metrics* change monotonically as one moves through the size quartiles: larger firms have longer, less readable, more repetitive, and more emulatory charters on average than do their smaller counterparts.

Panel C does something similar, but here we subdivide companies up into quartiles based on their Tobin's-Q value each year, which is thought to measure the "value added" of the firm's operation.[159] Here, we once again see a trend that exhibits substantially *the reverse monotonic relationships* that we saw with asset value. Here, as we move into higher Q ratio quartiles, mean charter lengths decline, readability increases, and redundancy falls. There does not appear to be a strong trend in inter-document similarity, however.

There is much more one can do with these sorts of measures. But even with this cursory appraisal, much real economic activity within firms leaves footprints in corporate governance documents (or perhaps vice versa). This observation suggests intriguing possibilities for researchers who wish to track whether and how a variety of political and economic phenomena (such as common ownership patterns) interact with the distribution of authority and control rights in firms.

B. Latent Semantic Content

A second exciting aspect of the CCG is its potential for unleashing a rich array of tools from computational textual analysis that allow deeper inquiry into document substance and structure.[160] Below we report on a few such applications, relating to legal origins and sectoral effects, and we end by demonstrating how text analysis helps tell part of the story of the evolution of an industry during moments of upheaval.

To the legal traditionalist, many of the tools we discuss below may seem (for want of a better term) un-lawyerly. After all, most of them begin by taking the texts of admittedly complex and nuanced legal documents, distilling them into numeric vector representations, and manipulating those representations to isolate "clusters" of

---

[159] The table uses the ratio of the issuer's market value to its book value to measure Tobin's Q. While not precisely equal to Q, the market/book ratio is a widely accepted approximation. *See* Tim Adam & Vidham K. Goyal, *The Investment Opportunity Set and Its Proxy Variables*, 31 J. FIN. RSCH. 41 (2008).

[160] *See, e.g.*, sources cited *supra* note 15.

affiliated or similar documents. While such mathematical renderings would seemingly be at odds with traditional legal analysis, these tools are surprisingly powerful and many parts of legal practice have long embraced computational techniques to augment traditional approaches.[161] It is in that spirit that we employ them below.

We transform each of the charters in our corpus into a vector representation based on the vocabulary used.[162] Because our corpus allows us to observe firms' charters multiple times, we develop two alternative representations, which we refer to as the "snapshot" and the "mashup" versions. The snapshot treats each year the company is publicly traded as a single observation, delivering a vector representation of the company's charter as it existed on January 1 of that year. Consequently, any company observed over several years in our dataset will (by definition) be associated with several snapshots of its charter. The mashup combines the various snapshots together into a single composite for the company, taking the mean values of vector elements for years in which we observe snapshots.[163]

In many applications (including ours), a vector representation of a document may have dozens, hundreds, or even thousands of dimensions—far too many to illustrate graphically. Nevertheless, the dimensions are derived using a technique designed to ensure that each successive dimension has diminishing explanatory power.[164] Consequently, by limiting attention to just the first two dimensions of our vectorized texts, we can retain the most important sources of variation while still enabling us to represent the "location" of each document in two-dimensional space. The panels of Figure 8 do just that, for the mashup versions of company charters. Charters that bear strong similarities to one another are located in close proximity, and accordingly if there are *several mutually similar* documents, they will tend to cluster in tight local neighborhoods; documents that are highly dissimilar, on the other hand, will be far apart, and several mutually dissimilar documents will scatter untidily about the plot, exhibiting no obvious clustering pattern.

Even in this low dimensional setting, the panels from Figure 8 show discernible evidence of clustering—patterns that directly bear on whether jurisdiction and/or legal origins leave their marks on firm governance. The left panel of the figure utilizes color

---

[161] These include discovery and motion practice, transactional due diligence, and predicting outcomes of legal disputes. *See* sources cited *supra* note 15.

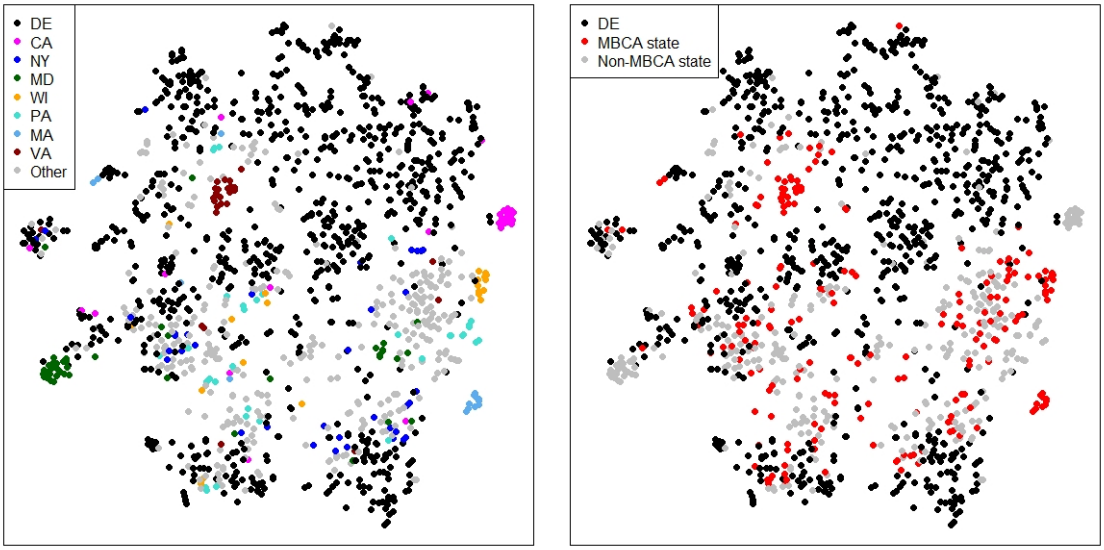
[162] More specifically, each document was distilled into a vector of ones and zeros depending on whether a word was featured in the document or not, and then rescaled using a familiar term-frequency/inverse-document frequency (tf-idf) transformation. For details, *see* Frankenreiter & Livermore, *supra* note 15; Talley, *supra* note 15.

[163] More precisely, the mashup vector for each company consists of the averaged tf-idf scores across all observed years. There are, of course, other possibilities for combining documents, but we opted for this because of its ease of use and the fact that the main alternative (appending all years' charters into a "mega-charter") might introduce unwanted distortions.

[164] For a more detailed explanation, *see* Talley, *supra* note 15.

codings to differentiate between each company's state of incorporation,[165] and highlighting eight states that are well represented incorporation homes within our dataset.[166]

Figure 8: Governance Clustering and State / Legal Origin (in 2-dim space)



Several interesting features of the left-hand panel stand out. First, there appear to be relatively tight clustering patterns for many states, though this does not appear to be true categorically. An example of a state whose companies exhibit tight clustering is California, depicted with the dense colony of magenta dots on the right side of the panel. The fact that the charters of almost all public California companies fall into this tight neighborhood suggests that their charters are very similar to one another, and very *dissimilar* to the charters of companies incorporated elsewhere. Maryland, represented by forest green dots on the left side of the panel, shows an analogous pattern. Maryland has a considerable share of the incorporation market for real estate investment trusts, and virtually all of the Maryland issuers in this cluster are, in fact, REITs.

California and Maryland's patterns stand in stark contrast with New York, represented in dark blue. New York companies are scattered haphazardly with no strong pattern, suggesting comparably low levels of intra-state similarity. And, in some ways similar to New York, the dominant majority of Delaware incorporated firms in

[165] We tracked the incorporation date as of the date of filing, thereby picking issuers who reincorporated out of state. For such issuers, the Figure classifies them for which state they were incorporated in the longest amount of time.

[166] *See* Online Appendix D for a state-by-state breakdown.

our dataset (58% of the observations, depicted as black dots) also appear to sprawl entropically across all quadrants of the diagram, indicating substantial governance heterogeneity (at least as measured by the latent semantic content of chartering documents).

The right-hand panel reproduces the identical geographical layout as the left-hand one, but it color-codes charters differently, based on the legal origin of the state's corporate code. Here we subdivide issuers into three groups, depending on whether they are incorporated in Delaware (black dots), incorporated in states that adopted MBCA in substantial part (red dots), or (iii) incorporated neither in Delaware nor in any jurisdictions that embraced the MBCA (gray dots).[167] In contrast to the right panel, there appears to be very little evidence that a common "legal origin" of the state's corporate statute matters much for determining or predicting the contents of charters, at least as measured by whether the state built its law on the basis of the MBCA. Even the few apparent red clusters in MBCA states appear to be artifacts of *intra-state* clustering, since those same local neighborhoods are clearly associated with distinct states, such as Massachusetts and Virginia, in the left-hand panel. These figures are consistent with the view that state of incorporation can and often does shape the content of charters, but common statutory origin appears to have little if any explanatory power.[168]

Another facet of the textual corpus is related to industry group effects. Figure 9 once again reproduces the same two-dimensional representations from Figure 8 above, this time color coding by industry sector (at the 2-digit SIC group). Here, we see notable evidence of clustering for certain sectors, particularly in finance (pictured in red, once again channeling those Maryland REITs identified above) and manufacturing (which manifests several sub-sector clusters).[169] This pattern is also consistent with Table 3, where recall that the finance sector was an outlier in all pertinent document metrics (length, readability, TTR, and cosine similarity).

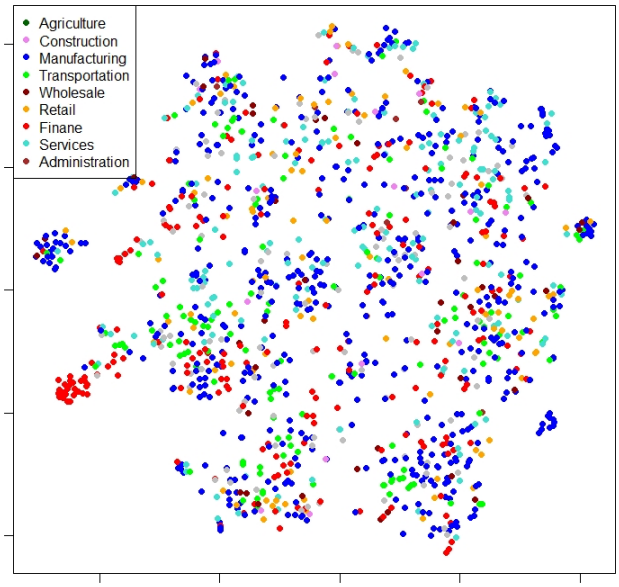
---

[167] We impute MBCA legal origins from the 50-state survey produced by USLegal.com. *See State Corporation Laws*, USLEGAL, https://corporations.uslegal.com/basics-of-corporations/state-corporation-laws/ (last visited Feb. 4, 2021).

[168] *Cf.* Jeffrey M. Gorris, Lawrence A. Hamermesh & Leo E. Strine, Jr., *Delaware Corporate Law and The Model Business Corporation Act: A Study in Symbiosis*, 74 L. & CONTEMP. PROBS. 107, 108 (2011) ("[T]here have been occasions on which [the MBCA's] hallmark precision has impaired its utility as a model, and its assertions of superiority have been overblown"). In Online Appendix E, we report several additional analyses that investigate this question from different angles. None of these approaches provides any appreciable evidence that charters from two different jurisdictions sharing an MBCA origin would be more similar to each other than charters from that did not share this origin.

[169] In a separate set of robustness checks, we used a more general statistical test for whether the clustering of charters of companies from the same industry that can be observed in Figure 9 could be explained as a result of chance. The results from this analysis indicates that this is not the case. This result suggests that a firm's corporate governance is at least partly a function of its area of business.

Figure 9: Governance Clustering by Industry (in 2-dim space)

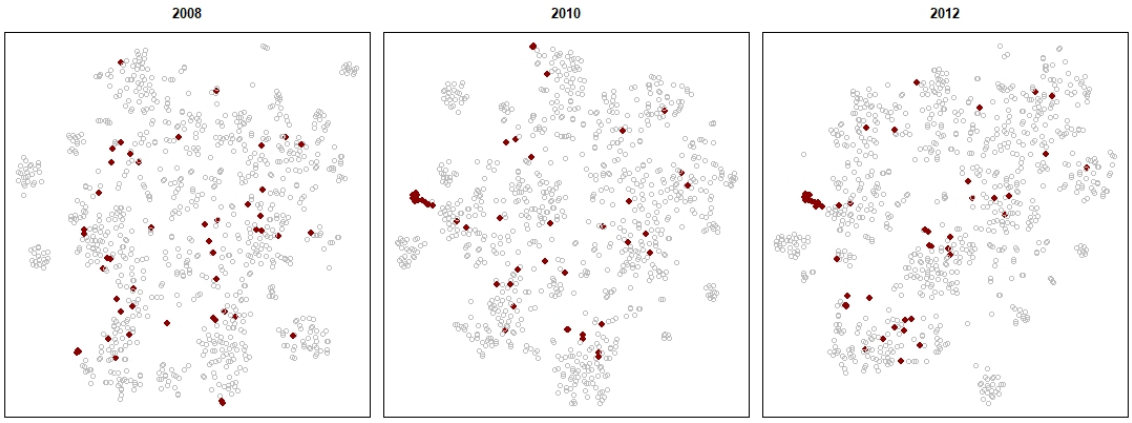


Although our focus thus far has been trained on the "mashup" versions of company charters (blending all years of a company's charter), the snapshot versions of our corpus are also well situated to unlock interesting dynamic clues about how governance evolves over time. The transformation of the banking sector (a subset of the finance industry) during the Financial Crisis provides an interesting example of this story.[170] Figure 10 illustrates diagrams similar to those above, but each is now separately generated for differing snapshot years (representing 2008, 2010, and 2012, respectively). Banks are represented with red dots,[171] while all other companies are in gray. In 2008, the chartering contents of bank charters appeared far from homogenous, with few (if any) clustering neighborhoods. By 2010, however, this pattern changed dramatically, with a tight cluster of banks emerging—and this cluster clearly persisted into 2012.

[170] While governance studies no doubt feature prominently in the study of banking history, we are not aware of work that uses computational analysis to accomplish the task. *Cf.* Charles W. Calomiris & Mark Carlson, *Corporate Governance and Risk Management at Unprotected Banks: National Banks in the 1890s*, 119 J. FIN. ECON. 512 (2016) (analyzing the impact of corporate governance structure on banking policy and practice in the 1890s).

[171] *I.e.,* SIC code 60.

Figure 10: Transformation of Bank Governance During the Financial Crisis



It does not take a rocket scientist (or rocket-science attorney) to make an educated guess as to why bank charters began clustering at this moment in time. In October 2008, the United States government interceded at the height of the financial crisis, infusing Troubled Asset Relief Program (TARP) funds into the coffers of dozens of large and medium-sized banks. These infusions typically took the form preferred equity purchases that gave the government considerable shareholder rights—rights that were formally recorded at some point in early 2009; accordingly, insofar as they affect the text of corporate charters (as new share issues must), they largely show up in the CCG on January 1, 2010 (depicted in the middle panel). Since the US government was effectively a large and powerful horizontal shareholder across multiple banks, it is not surprising that TARP administrators used that power in bargaining, so that the basic terms and conditions of the preferred share issuances stayed relatively uniform. Indeed, many bank recipients of TARP funds simply inserted a verbatim version of a standard provision into their corporate charters, resulting in a discernible clustering of charters within this industry.[172] As the recipient banks progressively paid off their TARP investments and left the program, this clustering pattern steadily began to dissipate, and it had largely disappeared by 2018.

Although computational text analysis is surely overkill for documenting the banking sector's well-known transformation during the Financial Crisis, it is comforting to see that a tool seemingly as blunt as document vectorization (into two

[172] The dark red cluster in the middle and right panel of Figure 10 includes the following firms: Bank of America; First Financial Bancorp; Firstmerit Corp; BB&T Corp; Fifth Third Bancorp; Suntrust Banks Inc; Old National Bancorp; First Midwest Bancorp; Keycorp; Umpqua Holdings Corp; Valley National Bancorp; and M&T Bank Corp. According to ProPublica, all these banks at one point received TARP money, and 10 of the 12 were among the first beneficiaries of the government's equity purchase program in late 2008. *See Bailout Tracker: Bailout Recipients*, PROPUBLICA, https://projects.propublica.org/bailout/list (last updated Nov. 9, 2020).

dimensions, no less) can still capture a notorious wave of governance convergence. Perhaps more exciting is the use of these and similar techniques to tease out other, less clear-cut evolutionary trajectories. For example, there is now a growing and controversial literature positing a provocative circumstantial argument suggesting that passive investing is anticompetitive, since it results in large common/horizontal ownership in the ownership blocks held by index funds across sectors. Several of the combatants in this area have raised important questions about identifying the causal mechanism (if any) that converts common ownership into oligopolistic power (e.g., compensation, lobbying, intra-firm governance, etc.).[173] A natural way to test whether common ownership affects governance might attempt to measure if and to what extent the emergence of common / horizontal ownership blocks also predicts patterns of convergence in the content of governance documents (such as charters and bylaws).

C. Supervised Learning Tools

Our textual corpus and labels are not only interesting for their descriptive applications: They also allow us to make some predictions. For example, as companies continue to file charters over time, they add to the corpus, and we can then use our labels to train a machine-learning classifier to quickly label the new filings. Doing so not only allows us to absorb the new additions into our database more quickly, but it also facilitates error detection and correction in all our existing labels (a task we have already implemented in part for this paper). Moreover, future researchers will also be able to use our corpus to generate new labels, indices, and evaluative metrics that are not currently part of any major governance data collection enterprise, including features such as forum selection provisions, board diversity provisions, stakeholder provisions, and the like.

Although space limitations preclude us from demonstrating the full range of the conceivable supervised-learning applications, one that is directly relevant to our analysis above concerns our evaluation of the G-Index and corrections thereto embodied in the CCG-Index. As discussed in Part II, the components of the G-Index continue jointly and severally to be used by corporate governance researchers to inform critical academic and policy debates. Because those components are purportedly derived in large part from governance documents themselves, our corpus (and labels) should bear a natural relationship to them, effectively allowing us to use the CCG database to "predict" the G-Index score. Similarly, our corpus can also allow us to predict the CCG-Index, giving us an indirect measure of the fidelity of each index to underlying governance documents and statutory structure.

To explore these possibilities, we used each of the G-Index and CCG-Index to calibrate a machine learning classifier, generating predicted values of each index based *solely* on the semantic content of our corpus of charters and their associated labels. We took care to use identical, well-established estimation techniques to calibrate each

[173] *See supra* notes 17-18.

classifier.[174] The performance of our classifiers against their "target" index is depicted in the scatter-plot panels of Figure 11, with the G-Index on the left and the CCG-Index on the right. Our calibrated classifier for the G-Index works reasonably well, successfully predicting 45% of the variation in the index.[175] In some respects, this result bolsters one's confidence that the G-index reflects *something* about company-level governance choices as manifested in the charter (as it purports to do). By the same token, the classifier *fails* to predict the remaining 55% of the variation in the G-Index. Based solely on this analysis, one cannot tell how much of the prediction noise is due to (a) a limitation on what such classifiers can offer, or (b) inaccurate labeling of the underlying data used to formulate the G-Index.

The right panel uses the identical approach, but this time as applied to the predicted and actual values of the CCG-Index. It is immediately clear that this classifier also has good predictive power. *More* than good, in fact: the CCG-Index predictions are a much tighter fit than those of the G-index, and the scatter-plot "cloud" is substantially more centered around our prediction line. Indeed, calibrating to the CCG-Index substantially increases our predictive power, *from 45% to 69% of the variance in the respective indices*. Put another way, when compared to the G-Index predictions, our CCG-Index predictions *are more accurate by half*, a result that is consistent with the conclusion that our corrections indeed remediated substantial coding errors in the G-Index.[176]

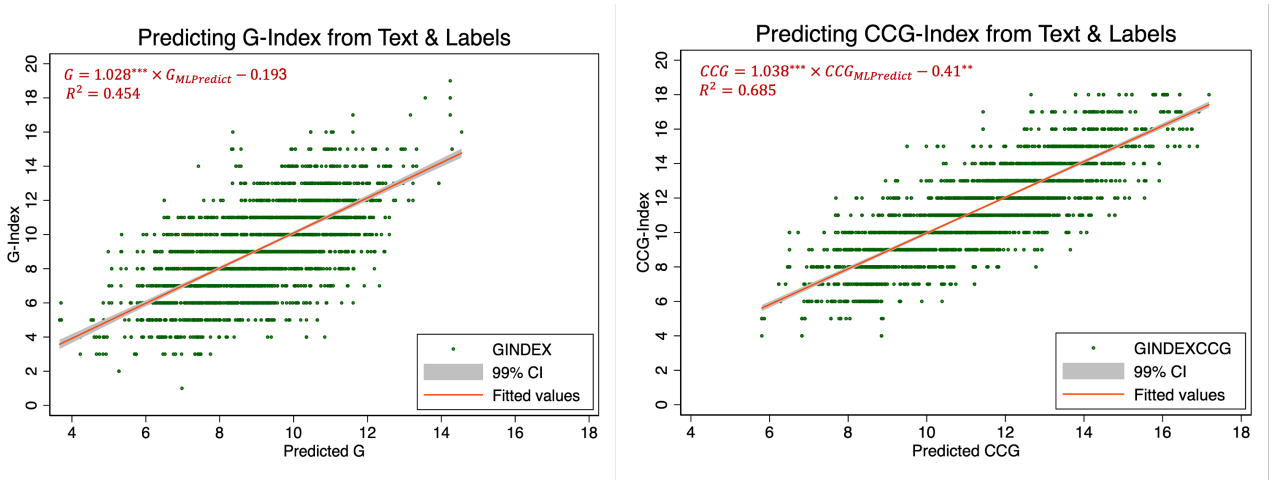
---

[174] For the technically-minded reader, we deployed lasso regressions on the principal components of the vectorized texts to generate all predicted values. This approach counteracts over-fitting risks by imposing a multiplicative penalty ($\lambda>0$) on the sum of the absolute value of estimated coefficients. It thus automatically shrinks the set of non-zero coefficients, retaining only the most explanatory ones. In our case, the penalty parameter (λ) was calibrated to minimize the sum of squared residuals in a 10-fold cross-validation, and the resulting coefficients were then used to generate predicted values as described in the text. For more on this utterly scintillating technique, *see* Robert Tibshirani, *Regression Shrinkage and Selection via the Lasso*, 58 J. ROYAL STAT. SOC'Y: SERIES B(METHODOLOGICAL) 267 (1996).

[175] We reiterate that all predicted values were generated on an "out-of-sample" basis using a 10-fold cross-validation: the documents were split into ten arbitrary groups, and for each group the model was calibrated using only the remaining nine partitions, rendering predictions for the held-out group.

[176] Because the CCG-Index corrections were themselves based on a direct labeling of the raw corpus, the alert reader might wonder whether a trained classifier would fare better for purely mechanical reasons in predicting the CCG-Index. We discount this concern on both conceptual and statistical grounds. First, the above approach constitutes a reasonable way to ascertain the reliability of our labeling protocols (described at length in Appendix B) against those used to assemble the IRRC (described nowhere). The G-index is purportedly based on the same documents and statutory structures as is our CCG-Index; and thus if the IRRC were labeled correctly to begin with, our scatter plots manifest trivial differences in predictive power (rather than a nearly 25-percentage-point difference in R^2). Second, in a separate robustness check (see Online Appendix F), we stripped out our labels completely, training the text classifiers *solely* with the raw textual content of corporate charters. While both predictors degrade, the same qualitative result still holds: our classifier explains substantially more variation in the CCG-Index (44.4%) than it does in the G-Index (34.5%).

Figure 11: ML Predictions of G-Index (Left) & CCG-Index (Right)[177]



IV. Implications and the Road Ahead

For many readers, this article may prove to be something of a Pandora's box: Most immediately, it problematizes several corporate governance "folk wisdoms" that have long been considered settled—including a result so beatified in the literature that it has achieved *slogan* status: that good governance translates into good returns. Using a deliberately conservative error-correction protocol, we demonstrated not only that the field's standard metrics for good governance appear grossly inaccurate, but also that the connection between governance and investment returns is materially sketchier than previously thought. Governance may yet matter, but the case for it far less clear than we thought it was.[178]

We still do not know the full implications of the errors our analysis has begun to uncover. Hundreds of studies have incorporated GIM's results or made use of the same questionable data. The various competitor indices are notable examples,[179] but there are scores of others that use the data or indices as inputs and/or controls in their own empirical designs.[180] Regulators, too, have turned to GIM and its progeny to support a litany of governance reforms.[181] Even critics of governance indices have largely presumed that the underlying data are accurate.[182] As a result, errors in popular

[177] Univariate OLS regression estimates of Predicted on Observed G-Index and CCG-Index overlaid on each panel. Significance: * = 0.05; **=0.01; ***=0.001.

[178] Moreover, any more aggressive correction to data errors—including the addition of bylaw data and expanded firm-level matching—might well cause further attenuation.

[179] *See* Bebchuk et al., *What Matters*, *supra* note 11; Madanoglu & Karadag, *supra* note 44; Karpoff et al., *supra* note 11.

[180] *See supra* notes 45-54 and accompanying text.

[181] *See supra* notes 55-59 and accompanying text.

[182] *See supra* note 54.

governance datasets have plausibly propagated throughout much of the corporate governance universe, affecting law, policy, practice, and theory. Reexamining the robustness of these results with more accurate data is a monumental undertaking that can only feasibly be spread across many years and several researchers.

Our Pandora's box also renders two ominous caveats about the critical importance of data availability and the vital role of lawyers in legal empiricism. As to the former, there is significant corporate governance data out there and free for the taking, but accessing it in usable form is fraught with difficulty. Simply put, the task of finding, harvesting, and cleaning fundamental governance documents for thousands of companies over three decades is hard work. The most amenable source—the SEC's quirky EDGAR online filing system—is cumbersome and poorly organized (especially for this task). State regulators are even less helpful, often requiring travel to a physical repository, only to face antiquated extraction technologies and exorbitant access fees. The small number of third-party commercial purveyors also charge fees, throttle downloading activity, and zealously protect their investments with litigation threats. Even for us—motivated though we were—pulling off this enterprise literally required a worldwide pandemic in the summer of 2020, which destabilized the economy and unexpectedly made available dozens of highly qualified research assistants.[183]

It is difficult to understand why public access to public records should be so tough.[184] The end result has been to make comprehensive governance research accessible only to the most well-heeled, well-connected, and patient—an observation that itself sounds a dissonant note about the uneven intellectual and economic playing field. Until now, the rest of us *hoi polloi* have been largely left to make do with the same small number of commercial resources, whose reliability has always been a little suspect, and which we have now shown to be hamstrung with inaccuracies.[185]

The second systemic caveat from our analysis concerns the surprisingly critical role of lawyers in empirical research. Although we cannot know for sure, our results strongly suggest that whoever originally labeled the dominant extant corporate governance databases had limited (if any) legal training. Our educated guess is that much of it was coded by non-lawyers. This approach—while no doubt economical—has a significant and unfortunate shortcoming: As we have shown above, turning corporate governance texts into quantitative data is a big ask. It requires nuanced

---

[183] See *infra* Appendix A.

[184] Corporate governance documents are only one example of such hurdles. *See, e.g.*, Pah et al., *supra* note 21 (chronicling restrictions of the PACER system over federal judicial records).

[185] As noted above, commercial data providers provide two types of data: data scraped from the SEC, or independently gathered and coded data with little to no information about the gathering and coding process (the IRRC and ISS datasets). The former imports whatever inaccuracies were in the SEC data onto a new platform, while layering on a theoretically easier-to-use search function. But while these search functions improve upon the SEC's, they invariably miss important bits of information, which makes data gathered through these types of databases almost certainly incomplete.

domain knowledge, legal judgment, and familiarity with broader principles of law and regulation. Nonlawyers, almost by definition, possess few of these skills.

And therein lies the rub: Lawyers, a professional class defined largely by a common aversion to numbers, appear long ago to have surrendered the project of corporate governance data collection to others. This was a mistake. In our opinion, the time is long past for lawyers to shed our quantitative heebeegeebees, roll up our sleeves, and reclaim the field of corporate governance research (*including* the data bit).

While our Pandora's box unleashes some admittedly negative mojo, it also contains a substantial beacon of hope. The versatility and open-source accessibility of the CCG database holds considerable promise for unlocking future chapters of corporate governance discourse along multiple dimensions. For example, recent years have seen a burgeoning attention to the societal role of corporations, the role for non-shareholder constituencies in corporate governance, and the alternatives to a shareholder primacy view of corporate law. A common rejoinder to the stakeholderism movement is *cost*: that stakeholderism chases marginal or unproven benefits while sacrificing the returns that shareholder primacy is widely known to create.[186] If such widely known folk wisdoms are, in fact, the vestiges of inaccurate data (as we have argued), then this cost-based rejoinder packs a considerably punier punch. More generally, when armed with our open-source resource, empirically-minded corporate governance researchers will be far better equipped to explore both conventional and emergent corporate governance questions.

Even more promising is the potential for the CCG to transform fundamentally the way we "do" corporate governance research writ large. Our corpus is a critical ingredient for harnessing novel techniques from computational text analysis and machine learning, and then applying them to our understanding of how firms are organized. Already, machine learning is fueling exciting results in many legal domains[187]—and we believe that corporate governance is an especially availing (yet still relatively untapped) target. Marshaling these new techniques to complement more traditional methodologies can lead to better empirical understanding, better theory, and better policy. Our discussion above merely scratches the surface of what can be accomplished with these techniques.

Almost as important is the fact that the CCG is effectively future-proof. While the most popular corporate governance datasets today consist solely of (questionable) data *labels*, we provide the underlying textual inputs themselves—the very DNA of corporate governance. This raw textual corpus will empower future researchers to expand the breadth of existing labeled datasets, to correct mistakes in existing ones

---

[186] *See, e.g.*, Lucian Bebchuk, Kobi Kastiel & Roberto Tallarita, *For Whom Corporate Leaders Bargain?* (working paper), 93 S. Cal. L. Rev. (forthcoming 2021), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3677155; Lucian Bebchuk & Roberto Tallarita, *The Illusory Promise of Stakeholder Governance*, CORNELL L. REV. (forthcoming 2021).

[187] *See* sources cited *supra* note 15.

(including ours), and to devise novel analytic measures to take on as-yet-unknown questions. The possibilities are endless, and the potential for intellectual payoff appreciable. With the CCG database, moreover, *we all* will have a running start.

Conclusion

In this paper, we have unveiled the fruits of a multi-year effort to harvest, clean, organize, and make publicly available (for the first time) a corpus of foundational corporate governance documents spanning three decades and thousands of public companies. We have demonstrated the immediate value of this resource by combining it with carefully hand-labeled data, which we then used to uncover a disconcerting pattern of errors within the most widely used corporate governance datasets among researchers. And those inaccuracies are consequential: Even after imposing a deliberately conservative error correction protocol, we have shown that some of the most well-known metrics and empirical insights in the field stand on shaky ground.

In the light of these findings, one might reasonably ask whether we should reassess our claim—made at the onset of this article—that empirical corporate governance is a major success story in the interdisciplinary study of law.[188] In our view, no such reappraisal is warranted. To the contrary: notwithstanding the appreciable infirmity of existing data that we have unearthed, as well as the corresponding state of flux it visits upon well-known folk wisdoms of the field, we view our project as ultimately standing on the shoulders of those early efforts. Those contributions permanently changed the conversation about how to understand corporate governance on a broad scale. In bringing this new resource into the public domain, our ultimate goal (and our accomplishment) is to lay the foundation for advancing the empirical corporate governance project even further. By providing a clean, accessible, primary data source going to the very structure of corporate governance, our project helps to provide a critical resource for unleashing new tools from machine learning and text analysis, taking corporate governance research into its next chapter.

And yet, much of the *current* chapter is still being written. Although our contribution makes a marked improvement over the *status quo*, we are neither so prideful or nor so delusional to believe our offering cannot be improved upon. Though relatively comprehensive, our corpus does not include all companies or all conceivable measurement years. It is likely that we have missed at least some relevant texts even for those issuers we track. And, notwithstanding our collaborative judgments as legal scholars, others will quibble with *our* calls about how to label certain elements of the corpus. All that said, the open-source nature of the CCG transforms each of these bugs into features: We invite all corporate governance researchers, professional or

[188] *See supra* notes 7-10 and accompanying text.

academic, expert or dilettante, U.S.-based or foreign, quant or poet, to contribute to this resource, helping each other collectively to cultivate it further.[189]

We also anticipate that the scope of our own task will expand. As monumental as the present undertaking has proven, our efforts have tackled only select (albeit important) pieces of the corporate governance ecosystem. In ongoing work, we have already begun to take on other foundational governance domains, including bylaws, shareholder agreements, board resolutions, and the like. Each presents an opportunity to correct past mistakes, to deploy new computational tools, to push the boundaries of knowledge, and (most fundamentally) to clean corporate governance for good.

[189] Upon publication, we will enable interested researchers to make use of the corpus and to develop improvements to the CCG database, transforming it into a living resource. The use and adaptation of the corpus will be free of charge, and subject to the Creative Commons Share-Alike License (v. 4.0), *see Attribution-ShareAlike 4.0 International*, CREATIVE COMMONS, https://creativecommons.org/licenses/by-sa/4.0/legalcode (last visited Feb. 10, 2021).

Appendices[190]

Appendix A: Research Assistant Acknowledgements

An undertaking of this magnitude necessarily involved a phalanx of research assistants, from law school graduates, to current law students, to undergraduates. Rather than hiding their names in a starred footnote, we use this Appendix to acknowledge their contributions. We are deeply indebted to them for their dedication and service.

Senior Research Assistants[191]
Nicole Banton
Matthew Cunningham
Deandra Fike
Channing Gatewood
Katie Gresham
Qifan Huang
Elisha Jones
Sami Kattan
Gabrielle Kiefer
Andrew Kim
Adam Mazin
Cameron Molis
Courtney Murray
Doriane Nguenang
Sneha Pandya
Emily Park
Olivia Roat
Bhargav Setlur
Tom St. Henry
Avi Weiss
Gretchen Winkel
Geoffrey Xiao
Ben Zonenshayn

Research Assistants[192]
Nathaniel Barrett
Amanda Cooper
Elif Nazli Hamutcu
Alex Inskeep
Justen Joffe
Alexa Levy
Annabelle Liu
Noam Miller
Emily Moini
Stephen Rothman
Adrien Stein
Max Swan
Agnes Tran

[190] Appendices C, D, E and F are available at www.publiccompanycharters.com.
[191] Senior Research assistants were JD students or recent JD graduates at top law schools.
[192] Research Assistants were undergraduate students at eight top universities.

Appendix B: Data Collection and Cleaning Protocols

This Appendix provides further details about our process for building the panel-structured chartering corpus for the CCG database, as well as our labeling protocols. We subdivide our description into (1) locating charters; (2) text extraction; (3) charter content labeling; and (4) state law content labeling.

1. Locating Charters (Task 1)

Our dataset consists of all charters available on EDGAR for all companies that fulfill at least one of the two following alternative criteria.

- The company was part of the S&P 1500 in any year between 2010 and 2019.
- The company is in the IRRC database for at least three out of the following five years of coverage in the IRRC: 1998, 2000, 2002, 2004, and 2006.[193]

For any issuer satisfying at least one of these inclusion criteria we attempt to extract the most complete chartering history available (even for years not satisfying the above criteria). Applying these criteria, we compile a list of 2,899 companies. For all companies in our dataset, we obtain all of their current and historical charters available on Edgar, and extract, clean and organize their text.

To ensure accuracy in our collection and avoid the pitfalls of some of the current commercial data, we tackle this challenge manually, with the help of a large number of research assistants. Our protocol for locating relevant texts leverages the requirement for companies to include information about their corporate charters with their annual 10-K filing.[194] These filings usually do not contain the text of the charter themselves, but instead incorporate the charter by reference to one or more prior filings, typically filed shortly after the charter was adopted/amended. Our harvesting protocol therefore follows a sequential process: (1) locate the company's most recent annual 10-K filing; (2) determine whether the filing reproduces a full charter restatement or merely incorporates one by reference; and (3) if no full restatement is found, use the exhibit references to identify the dates and locations of prior filings that contain the

[193] The ISS Legacy/IRRC database does not cover all years, but instead it observes S&P 1500 issuers periodically, in 1990, 1993, 1995, 1998, 2000, 2002, 2004, and 2006. EDGAR filings were voluntary in 1995, and they did not become compulsory until May 1996. For a history of EDGAR's roll-out, *see History of EDGAR*, EDGAR PRO, https://help.edgar-online.com/edgar/history.asp?site=pro#:~:text=In%201984%2C%20the%20SEC%20allocated,get%20the%20information%20it%20needed (last visited Feb. 10, 2021). Our data collection effort required at least three years of IRRC coverage to focus on issuers that could generate a reasonably reliable panel structure, and we omitted from our search the first three years of the IRRC's coverage (1990, 1993, 1995) since those years pre-dated the full roll-out of the SEC's EDGAR service (our primary data collection source).

[194] 15 U.S.C. § 78(m).

full text of the current charter as well as any intervening amendments.[195] Once those new texts are located and added to our database, the search protocol repeats, with the next iteration starting with the 10-K filing that immediately precedes the filing date of the full restated charter located from the prior iteration. For each issuer, we loop through these sequential steps repeatedly, working backwards in time until the trail runs cold and no more responsive documents can be located on EDGAR.[196] We spot-checked several companies' charters against inventory lists from commercial providers to confirm that our manual collection efforts avoided the aforementioned coverage gaps that befall commercial providers (they substantially do).

Work on this part of the project started in October 2019; overlapping cohorts of research assistants and law students assisted us in different periods of time, but we were fortunate enough to bridge the transition periods with high-quality legacy personnel to help train the next group. We harvested most of the chartering histories for companies in the current S&P 1500 in October/November 2019. Information obtained by research assistants during this phase of the project was later verified by a different set of research assistants—all assignments that were not completed by either senior RAs or the coauthors underwent this procedure. Information for companies that were not part of the S&P 1500 at the time of the start of the project was assembled starting in the summer of 2020. In this part of the project, we assigned the majority of companies to two research assistants at the same time. In case the information provided by the research assistants was not identical, we sent the information to a third research assistant for verification.

2. Text Extraction (Task 2)

In a second step, we use the information obtained from Task 1 to extract the texts of charters from EDGAR. For this, we employ a custom-made python script. This process allows us to gather charter texts for around 80% of the charters we identify in Task 1. For the remaining 20%, research assistants retrieve the text manually.

3. Charter Content Labeling (Task 3)

The third step involves labeling the contents of charters according to a prespecified coding rubric. Our rubric contains a set of 28 questions about the presence of specific provisions in a charter. Ten questions deal with issues regarding the rights associated with different classes of stock and the power of the board to shift the balance of power among shareholders, in particular in the context of takeover defenses. Another 10 questions concern issues of corporate governance (such as

[195] The recorded information also included helpful document text tags, which allowed us to develop a customized computer program to extract the charter texts. In cases where this automated text extraction failed, we extracted the text manually.

[196] As a result, our strategy materially differs from (and is more robust than) the approach apparently used by most commercial services. *See supra* Part I. More details of our training protocol for tracking filed charters and amendments is given in Appendix B.

special meetings and written consents). A final set of questions concerns the liability of managers and corporate officers. For each of these questions, we ask coders to provide us not just with a binary response if the provision was present, but also with relevant text if they are able to locate a provision in the charter.

We implemented this rubric in an Excel spreadsheet that allows us to code the contents of charters for the same company on one sheet. We also make available to coders "redline" documents that show the changes between different versions of (full restatements of) corporate charters. Our research assistant team convened once per week via Zoom to discuss the labeling process and to tackle any issues that occurred during the previous week. Besides, we set up an online forum where coders had the opportunity to ask questions on an ongoing basis, and that was consistently monitored by one of the senior RAs.

Initially, we assigned the same company to multiple coders in order to track agreement rates and identify the need for additional training. After that, we assigned companies to two research assistants at the same time. Senior research assistants reviewed all questions for which the coders' answers diverged. During this phase of the project, we also tracked rates of agreements between coders. After some weeks, we ceased double-assigning companies to JD research assistants. For other research assistants whose coding appeared particularly reliable, we also incrementally reduced the amount of overlap with other coders. However, we made sure that at least 33% of the companies coded by undergrad coders were double-assigned. Overall, we labeled the contents of all the charters for 1,573 issuers in our dataset. The companies included in Task 3 were chosen as follows. Because one of the goals of our manual coding was to replicate studies relying on the IRRC database, we deviated from random assignment in one important way: Whenever possible, we gave priority to companies that were included in the IRRC database.

4. State Law Content Labeling (Task 4)

In a separate effort, we trained business law students to label a panel data set of laws from all 50 states and the District of Columbia regarding sixteen governance-related issues.[197] (Several of these dimensions appear to have been wrongly neglected in notable databases like the IRRC.)

For state law, labelers tabulated the existence and substantive directionality of the provision (e.g., "required" or "not required/silent"), whether it was a default or immutable rule, the lowest echelon of corporate governance document capable of contracting out of the rule (if it was default), and limitations / constraints placed on available choices for issuers opting out (again if it was a default). These criteria were then employed to implement the "conservative" approach to identifying and correcting errors. Our state-level panel data also includes labels for four additional

[197] For additional details, *see* Online Appendices.

state law provisions that we extracted from pre-existing assorted sources in the literature.[198] Although we did not label these *de novo*, when the designations in the literature conflicted with one another we did primary research to reconcile the differences.

[198] Karpoff et al., *supra* note 11; Matthew D. Cain, Stephen B. McKeon & Steven Davidoff Solomon, *Do Takeover Laws Matter? Evidence from Five Decades of Hostile Takeovers*, 124 J. FIN. ECON. 464 (2017); Michal Barzuza & David Smith, *What Happens in Nevada? Self-Selecting into Lax Law*, 27 REV. FIN. STUD. 3593 (2014).

about ECGI

The European Corporate Governance Institute has been established to improve *corporate governance through fostering independent scientific research and related activities.*

The ECGI will produce and disseminate high quality research while remaining close to the concerns and interests of corporate, financial and public policy makers. It will draw on the expertise of scholars from numerous countries and bring together a critical mass of expertise and interest to bear on this important subject.

The views expressed in this working paper are those of the authors, not those of the ECGI or its members.

ECGI Working Paper Series in Finance

www.ecgi.global/content/working-papers

Electronic Access to the Working Paper Series

The full set of ECGI working papers can be accessed through the Institute's Web-site (www.ecgi.global/content/working-papers) or SSRN:

Finance Paper Series	http://www.ssrn.com/link/ECGI-Fin.html
Law Paper Series	http://www.ssrn.com/link/ECGI-Law.html