# Measuring Corporate Governance from the Source: Language Models and the Full Proxy Corpus

Andrew C. Baker* Paul Goldsmith-Pinkham†

August 27, 2026

**Abstract**

Commercially available corporate-governance data in the U.S. is expensive and covers only the largest corporations, even though the underlying filings used to generate these measures are available for all publicly-traded firms. In this paper, we develop a framework to convert firm proxy statements from the SEC's EDGAR portal into validated panel data using large language models, at negligible expense. We apply our framework to DEF 14A proxy filings from 1997–2026 and construct a new panel of insider voting and cash-flow ownership, dual-class structure, and board classification covering 140,314 filings and 13,790 firms. We document new facts about the level, concentration, and evolution of insider control, and use the panel to separate managers' cash-flow incentives from voting control when testing the effects of the 2003 and 2013 dividend-tax changes.

*UC Berkeley Law
†Yale SOM

# 1 Introduction

Corporate governance deals with the question of how shareholders supplying capital to firms can ensure a return on their investment from managers who cannot be continuously monitored (Shleifer and Vishny, 1997). Given the centrality of the investor-manager conflict to our capital markets, corporate governance has been a core area of research in corporate finance and law for decades. Much of the seminal work is conceptual or theoretical—from Berle and Means on separated ownership and control in the American economy, to Henry Manne on the market for corporate control, and Michael Jensen and William Meckling on how parties contract around the agency problem. However, the field took an empirical turn towards the end of the 20th century, as social scientists and lawyers sought to quantify how contractual and fiduciary relationships in corporations impact outcomes for shareholders and stakeholders (La Porta, Lopez-de Silanes, Shleifer and Vishny, 1998; Gompers, Ishii and Metrick, 2003).

Although the empirical corporate governance literature is influential and well-cited, it relies upon a series of commercially provided datasets that are expensive, opaque in their creation, and limited in coverage. Subsequent work has shown how different vendors provide classifications and ratings that are largely uncorrelated with each other (Daines, Gow and Larcker, 2010; Larcker, Reiss and Xiao, 2015) and that these differences are usually a result of measurement conventions rather than the scope of coverage (Berg, Kölbel and Rigobon, 2022). Moreover, even when academics have self-coded governance provisions, attention is usually restricted to the largest firms that trade in commonly-tracked indexes (Frankenreiter, Talley, Hwang and Nili, 2021). Given that over 75% of the firm-year observations for publicly-traded firms fall outside of the S&P 1500, the vast majority of published results describe (at best) only a quarter of the visible firm universe.

The limitation on creating quality corporate governance data at scale has never been a lack of access to the underlying information. As we discuss in Section 3.2, much of the information required to derive the key governance measures used in the empirical literature are functionally required, under SEC disclosure rules, to be disclosed in the firm's annual proxy statement filed as a Form DEF 14A with the agency. Moreover, the text of current and historical charters are often retrievable as digitized filings through EDGAR as well, which has been used by prior research teams to diagnose coding errors (Frankenreiter et al., 2021). The scarce resource for generating a quality, comprehensive governance dataset has instead been the labor required to hand-code thousands of files a year, especially for challenging edge-cases that naive decision rules will inevitably get wrong.

In this paper, we demonstrate how researchers can use large language models (LLMs) to classify governance provisions from publicly available records, constructing a firm-year panel that can easily be updated and extended, and that covers more firm-year observations than the commercial vendors. Contemporaneous work also proposes LLM-based classification as a substitute for human coding, while warning that provision-level accuracy is variable, especially for provisions that require multiple steps of legal reasoning to produce (Frankenreiter, 2026). Here we show how a

largely automated routine that uses both open-weight and frontier models, along with validation arms and codified decision rules, can compare favorably to human coded data available in the market. Moreover, our process has the benefit of being comparatively cheap and extendable to new measures, datasets, and years.

The key to our process for parsing firm proxy statements is what we call the *transcribe/compute split*. Our protocol only uses open-weight models to extract values and text that are printed in the filing, and downstream code generates the final measures based on the primitive extractions. For example, upon prompting, the model will return the number of share classes and their votes per share, the beneficial ownership for the directors and officers as reported in the annual beneficial ownership table, and verbatim text from the filing used to support whether the board is classified. The subsequent downstream code then calculates our final value for total economic ownership and voting power based on the number of shares outstanding, inside ownership, option amounts due to vest within 60 days, and the class-based votes per share. When LLMs are prompted to derive the final values themselves, they often default to providing an answer even when they cannot confidently determine one. An advantage to splitting up the protocol into two steps is that every resulting value in the panel can be traced to the underlying primitives from the filings, and judgments—either human or machine—are made one time and applied across the corpus.

We also combine our outputs with a small validation sample, following best practice for analyzing large amounts of text at scale with LLMs (Ludwig, Mullainathan and Rambachan, 2026). We first create an adversarial benchmark of 170 filings from 100 dual-class firm-year observations that are deliberately chosen to over-weight challenging classifications. These filings are both audited by a frontier model (Anthropic's Claude Opus 5.2) and adjudicated by hand against the underlying source documents. This benchmark sample operates as a "regression gate", and any future repair rule cannot be adopted unless it meets or beats the benchmark ratings. We also create a 160-filing sample drawn exclusively from small-cap and OTC issuers in the earliest part of the sample and find that our approach scores 99.4 percent on dual-class classification, 99.3 percent on board classification, and 95.9 and 92.5 percent on insider voting power and economic ownership. After updating the prompt in response to the small set of classification errors in this sample, the remediated panel scores 100 percent on the first three fields and 98.5 percent on the fourth. Although machine auditors still generate measurement errors (Frankenreiter, 2026), adjudicating disputed entries by hand increased the accuracy of our measures by three percentage points. Moreover, our agreement with commercial vendors on the overlapping sample is on the order of 96-99% of the observations for board classification and dual class flags.

Our resulting panel reveals a picture of insider control for the full cross-section of U.S. firms that differs in important ways from what can be seen when looking only at the largest firms. First, insider control of public equity is common and much of this control occurs in ordinary single-class firms. For example, in 1997 the median single-class firm's officers and directors held 16.4 percent of the vote, and 13 percent of single-class firm-years were majority-controlled by

their own insiders. Second, much of that control has fallen over time, and by 2025 the single-class median inside voting share fell to 4.2 percent and the majority-controlled share to 5 percent. Third, charter-based entrenchment through dual-class structures has increased markedly over time, doubling from roughly six percent of firms through the financial crisis to 11.6 percent by 2025, with the median dual-class wedge between insider votes and cash-flow rights quadrupling from 4.1 to 16.6 percentage points. Fourth, most of the rise in dual-class companies occurs outside of the large indexes, and thus outside of commercial vendor coverage. Within the S&P 500 the dual-class share increases from 5.6 to 6.5 percent over the full sample, while outside of the S&P 1500 it has more than doubled to 13.3 percent of firms and 35.3 percent of market capitalization. Even where data coverage already exists, conventional measures of insider share can miss effective control. In roughly a third of dual-class firm-years, the directors-and-officers group row from the beneficial ownership table understates true insider control because the controlling holder is a parent company, founder's trust, or former officer.

We use the broader coverage and refined measures from our panel to revisit the effect of the 2003 dividend-tax cut on firm payout decisions. The 2003 dividend-tax cut reduced the highest dividend income rate from 38.6% to 15%. Dividend initiations at firms with high insider ownership followed shortly after, evidence that managerial tax interests shape payout policy (Chetty and Saez, 2005). Chetty and Saez (2010) formalized an agency model in which the manager's ownership stake generates an incentive to change policy, while their voting control supplies the power to do it. For data availability reasons, the original article and responses from Edgerton (2013), Floyd, Li and Skinner (2015), Hanlon and Hoopes (2014), and Yagan (2015), use insider share as a proxy for control, which conflates ownership and voting power.[1]

Once voting power and economic ownership are measured separately, we find that the increase in dividends in response to the 2003 cut is strongest for the interaction of insider ownership and voting control. The effect of insider economic ownership is 2.2 percentage points greater per standard deviation of the voting–ownership wedge, and the main effect for economic ownership is modest and imprecise. Moreover, firms controlled by an individual or family—whose personal dividend-tax rate fell—initiated dividends at a much higher rate than firms that were controlled by a corporate parent, whose tax position remain unchanged. Re-estimating Yagan (2015)'s study design on this controller margin reproduces a null effect for investment and employment, consistent with the tax cut changing the division of corporate cash rather than impacting real firm activity. Although the sample size for the estimates is necessarily small, the pattern is uniformly the one predicted by the agency model.

Finally, we show how this protocol can be used to expand the governance sample to include more firms, additional measures, and future observations. The protocol developed in Section 2.2, including extracting targeted sections of the proxy filings, auditing samples, and adjudicating output with frontier-models and human coding, can be redone for new measures at a low marginal

[1]The only direct evidence on the distinction comes from Swedish registry data (Jacob and Michaely, 2017).

cost that does not scale with the number of filings. A researcher who wants to measure a new provision can explain the variable, the firms to which it applies, and where it is disclosed in the filing. This prompt will be followed by a pilot sample that establishes whether the measure can be derived from already-retrieved primitives or requires new extraction. The process proceeds through edge cases that are ruled on once, in writing, as decision rules in code, after which the measure is generated for the full panel with an audited accuracy rate and, where an external benchmark exists, a reported consistency rate. The same logic can extend the sample to categories of firms that our data does not currently cover, including REITs, foreign-listed firms, and firms that trade over-the-counter. Our hope is that the research community could use this audited extraction protocol, along with a public corpus of available documents, to make governance data that is comprehensive, transparent, traceable, and freely shared.

# 2 Generating Governance Data

In light of the agency problem associated with the separation of ownership and control between managers (agents) and purveyors of capital (principals), the field of corporate governance has been primarily interested in defending shareholder interests (Tirole, 2001). Managers are tasked with the day-to-day operation of the firm, while shareholders (the purveyors of capital) reserve a limited set of narrow but important rights, including electing a board of directors with oversight capacity over the management team and approving a narrow set of corporate transactions. The rules governing the interplay between shareholders, the board of directors, and management, are codified within a corporation's governing documents, namely the corporate charter and its by-law provisions, which are filed with the secretary of state for the state of incorporation. The vast majority of published empirical work in corporate governance, across economics, finance, accounting, and law, uses variation in governance measures from these documents to tease out the implications of the agency relationship.

## 2.1 Commercial Data Sources

Governance researchers have historically been limited to a narrow set of data providers. The first set of large-scale papers on governance provisions used measures from the Investor Responsibility Research Center (IRRC), which coded 24 distinct corporate-governance provisions beginning in 1990 (Gompers et al., 2003). The IRRC data was the foundation for two particularly influential papers in corporate governance: Gompers et al. (2003) used the IRRC measures to derive a "Governance Index" that was shown to empirically drive firm value and profits, and Bebchuk, Cohen and Ferrell (2009) used a smaller subset of the provisions to form an "entrenchment index" that was also "monotonically associated with economically significant reductions in firm valuation as well as large negative abnormal returns". IRRC was later purchased by one of the two dominant

proxy advisory firms in the U.S.—Institutional Shareholder Services (ISS)—which now produces governance data along with their core proxy advisory and voting research products. The ISS governance data is provided for academic use through the Wharton Research Data Services (WRDS), and as a result has been the standard source for governance measurement in academic work.

In part driven by the rise of socially conscious investing, with its focus on environmental, social, and governance ("ESG") investment factors, a number of commercial governance data vendors arose as well. One of the first alternative providers was SharkRepellent (acquired by FactSet in 2005), a specialized governance and takeover defense database that includes over 200 data points on charter/bylaw provisions, poison pills, and state statutes for thousands of U.S. public companies. Academic research has found "substantial differences" in the coding of certain provisions between IRRC/ISS and SharkRepellent (Larcker et al., 2015). Other commonly used datasets and vendors include BoardEdge and Board Brief from Equilar, BoardEx from Diligent, MSCI's ESG Research (which replaced an earlier product called GMI), Sustainalytics from Morningstar, S&P Global, Refinitiv from LSEG, and Bloomberg.

These commercial governance databases share specific limitations that have inhibited the empirical governance literature. First, the measures are usually coded using opaque proprietary processes, and measures of the same provision routinely disagree. Daines et al. (2010) find that the leading commercial governance ratings are largely uncorrelated with each other and have little power to predict the financial outcomes (e.g., accounting restatements, shareholder litigation, and firm performance) that they are designed to capture. Measurement issues are likely even worse for ESG ratings, and the different values across the major ESG ratings services appears to be largely driven by measurement differences rather than differences in coverage or weighting (Berg et al., 2022) . Because coding determinations are made within a black box, researchers confronted with a disagreement between providers cannot trace a given classification back to the underlying disclosure.

Second, access to these databases is expensive and requires ongoing institutional subscriptions. Although governance ratings are constructed almost entirely from publicly available disclosures, including proxy statements, charters, and bylaws filed with the Securities and Exchange Commission (SEC) or state agencies, they are licensed at subscription prices that place them beyond the reach of many researchers. Finally, only the largest firms are covered. The original IRRC dataset covered the S&P 500, with successor products only extending to the S&P 1500, leaving thousands of small-cap and OTC issuers (where insider ownership is most concentrated and governance frictions are likely the most severe) unmeasured. However, most of the source documents are freely available through the SEC's EDGAR portal and are generally standardized by the disclosure requirements imposed by the SEC.

## 2.2 Using Open-Source LLMs for Data Creation

This section describes our framework for converting publicly available firm reporting information into structured governance data with open-source large language models (LLMs). The resulting panel is essentially free to produce, carries audited field-level accuracy, and covers a much broader universe of firms than any commercial database.

### 2.2.1 SEC Data

As our input source, we use the text and data from the annual proxy statement, the disclosure document provided to shareholders whenever a firm solicits votes. The requirement for annual shareholder meetings to elect directors comes from state corporate law—including in Delaware, the state of incorporation for a majority of U.S. public firms—with the major national securities exchanges also imposing a similar requirement as a listing standard.[2] Federal securities statutes generally govern disclosure rules, and Rule 14a-3, promulgated under Section 14(a) of the Securities Exchange Act of 1934, prohibits solicitation of proxies unless each person solicited is furnished a proxy statement containing specified information.[3] Given that a widely held firm cannot functionally assemble a quorum without the solicitation of proxies, effectively every U.S. reporting company is required to file a definitive proxy statement under form type DEF 14A on EDGAR each year.

The proxy statement is a stable extraction target for machine learning classifiers, in part, because the mandatory disclosure information is enumerated by Schedule 14A. Item 6, the section titled "Voting securities and principal holders thereof," requires the firm to state the number of shares outstanding of each class entitled to vote and the number of votes to which each class is entitled, the record date, and beneficial-ownership information required in Item 403 of Regulation S-K. Item 7 covers directors and executive officers, requiring disclosure of director nominees and their biographies, related-party transactions, board leadership structure, committee composition, independence determinations, and—critical for our staggered board measure—whether directors stand for election annually or in staggered classes. Item 8 requires disclosure of the full executive- and director-compensation amounts. The remaining items depend on the matters being voted upon, including meeting logistics, revocability of proxies, appraisal rights, and the interest of insiders in the matters at hand, among other considerations. In addition, companies can always choose to voluntarily disclose additional information not mandated under Schedule 14A.

The beneficial-ownership table required by Item 403 is the most important and challenging object in the filing for our purposes. Whenever directors face election at a meeting, the issuer must report, in tabular form, each person known to beneficially own more than five percent of

[2] 8 Del. C. *S* 211; NYSE Listed Company Manual *S* 302.00; Nasdaq Rule 5620(a).

[3] 15 U.S.C. *S* 78n(a); Regulation 14A is codified at 17 C.F.R. *SS* 240.14a-1 through 240.14a-21, with Schedule 14A at 17 C.F.R. *S* 240.14a-101 and the filing requirements at 17 C.F.R. *S* 240.14a-6. Where directors are to be elected, Rule 14a-3 additionally requires that the solicitation be accompanied or preceded by an annual report.

any class of voting securities, together with the amount and percent of class held; the ownership of each class of equity securities by every director and nominee, by each named executive officer, and by all directors and executive officers as a group (the group row from which our insider voting and cash-flow measures are taken); and any arrangements, including pledges of securities, that could result in a change of control.[4] Under the relevant SEC rules, a person is determined to be the beneficial owner of those shares that they have the right to acquire within sixty days through options, warrants, or conversion rights, with the additional shares added both to that holder's numerator and denominator.[5]

The items measured through our parsing protocol are mandatory for every firm and every year, reported in a stable regulatory framework, and have been publicly and freely available on EDGAR since electronic filing became effective and complete in the mid-1990s. Unfortunately, while the content of this disclosure has been consistent, its form has not. The same required table appears in thousands of distinct layouts, with essential qualifications often found in footnotes, multi-class positions (where relevant) split across columns whose headers might vary by filer, and early-period filings rendered as plain-text tables with no markup at all. Before the advent of LLMs, extracting quality data from these documents for the entire panel of publicly traded firms would have been cost prohibitive for an academic researcher given the unstructured nature of the reporting content.

### 2.2.2 Measuring Staggered Boards, Dual Class Structures, and Beneficial Ownership

We use open-source models and a dedicated parsing framework to produce three validated governance measures as a proof of concept. First, we test whether we can measure the presence of a staggered, or "classified", board from the firm proxy statement. A staggered board partitions the directors into different classes with overlapping multi-year terms, so that only a portion of the board is up for election each year. This election staggering functions as a takeover defense because a hostile bidder or activist will have to prevail in two consecutive proxy contests before gaining control of the board. This process is particularly protective when combined with a poison pill and is among the provisions shown to substantially reduce the probability of a hostile takeover (Bebchuk, Coates and Subramanian, 2002). Consequently, it is one of the most studied anti-takeover provisions, is a component of both the Gompers et al. (2003) G-index and the Bebchuk et al. (2009) E-index, and has been alleged to lower firm value (Bebchuk and Cohen, 2005).

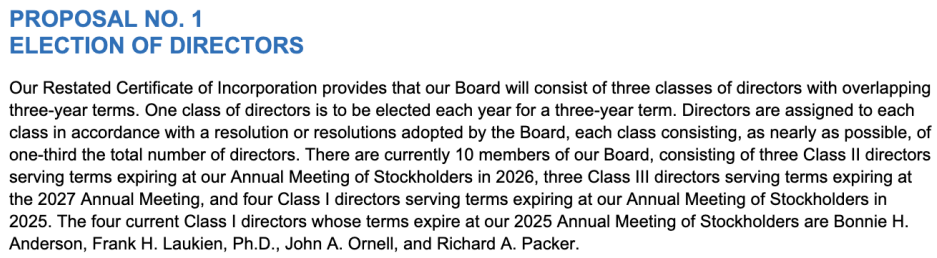
Although no SEC rule explicitly requires that a company state whether its board is classified in their proxy statement, other provisions dictate a level of disclosure sufficient for the classification task. Item 401(a) of Regulation S-K requires the registrant to identify the company's directors and nominees and "state his term of office as director and any period(s) during which he has served as such", along with any arrangement or understanding under which the person was selected. With a

[4] 17 C.F.R. *S* 229.403.
[5] 17 C.F.R. *S* 240.13d-3(d)(1).

classified board, this information will reveal if the company's board of directors is not all standing for election at the same time but instead have staggered terms.

Frequently, a proxy statement will use the words "classified," "staggered," or "Class I/II/III" when describing director elections, which makes the parsing process straightforward for a well-instructed LLM. Figure 1 shows an example from the 2025 proxy statement for Bruker Corporation, an American manufacturer of high-performance scientific instruments and diagnostic solutions. By specifying the classes of directors and that their terms are overlap, the excerpted language clearly indicates that the firm has a staggered board.

Figure 1: **2025 Bruker Corporation: The Easy Case for Staggered Board**



**PROPOSAL NO. 1**
**ELECTION OF DIRECTORS**

Our Restated Certificate of Incorporation provides that our Board will consist of three classes of directors with overlapping three-year terms. One class of directors is to be elected each year for a three-year term. Directors are assigned to each class in accordance with a resolution or resolutions adopted by the Board, each class consisting, as nearly as possible, of one-third the total number of directors. There are currently 10 members of our Board, consisting of three Class II directors serving terms expiring at our Annual Meeting of Stockholders in 2026, three Class III directors serving terms expiring at the 2027 Annual Meeting, and four Class I directors serving terms expiring at our Annual Meeting of Stockholders in 2025. The four current Class I directors whose terms expire at our 2025 Annual Meeting of Stockholders are Bonnie H. Anderson, Frank H. Laukien, Ph.D., John A. Ornell, and Richard A. Packer.

Excerpt from Bruker Corporation Form DEF 14A for the 2025 annual meeting, filed April 11, 2025 (CIK 0001109354). The passage states both that the board is divided into three classes and the term-expiration year of each class. Only the latter is required by Item 401(a) of Regulation S-K; the former is voluntary. Source: SEC EDGAR, accession no. 0000950170-25-053286.

Similarly, there are easy cases for non-staggered, or "unitary", boards of directors. A board where every director stands for election each year has only one term to report, and the disclosure required by Item 401(a) often results in language tracking Delaware General Corporation Law $S$ 141(b): each director serves until the next annual meeting and until a successor is duly elected and qualified. Figure 2 excerpts such a passage from the Tyson Foods DEF 14A proxy statement for the 2026 annual meeting. The asymmetry with the Bruker filing in Figure 1 is worth noting here because it could represent a plausible source of classification error. A filing for a company with a classified board will generate more text keyed to terms such as "class," "term," and "expire" than a unitary board, and a retrieval step that classifies passages on those terms will have more to match on when that language is present than when it is absent. This could make classifying a unitary, non-staggered, board more challenging if the classification protocol is naive.

However, not every partition of a board into different classes necessarily indicates the presence of a staggered board. Proxy statements often use the phrase "classes" to describe directors elected by separate constituencies of shareholders. This can happen, for example, when dual-class firms allow each class to designate a share of the board, or when preferred stockholders acquire the right to name directors if preferred dividends are in arrears. These arrangements are also often displayed in the proxy statement in a manner superficially similar to those of a company with a true staggered board, with the proposal split into panels with section headings that group directors by class, or references to "Class A Directors". A naive classifier trained only on a small set of key

Figure 2: **2026 Tyson Foods Incorporated: The Easy Case for Unitary Board**

**ELECTION OF DIRECTORS**

The number of directors that will serve on the Board following the Annual Meeting is expected to be fifteen (15) but may be changed in the manner provided in the Company's by-laws. Each director is elected until the next annual meeting of shareholders and until such director's successor is duly elected and qualified. Our by-laws provide that no person shall be nominated to serve as a director after they have passed their 72nd birthday (the "Retirement Age By-law"), unless the Board has voted, on an annual basis, to waive or continue to waive the Retirement Age By-law for a nominee. In accordance with the Company's by-laws, this age limitation does not apply to the Chairman of the Board of Directors. The Board has voted to waive the Retirement Age By-law for Ms. Barbara A. Tyson and Governor Mike Beebe.

Excerpt from "Election of Directors," Tyson Foods, Inc. Form DEF 14A for the 2026 annual meeting, made available to shareholders on or about December 17, 2025 (CIK 0000100493). The passage discharges the same Item 401(a) disclosure obligation as Figure 1 but reports a single uniform term: no class designations or differing expiration years appear anywhere in the section. Source: SEC EDGAR, accession no. 0000100493-25-000114.

words would inevitably err in such cases. The E.W. Scripps Company's 2025 proxy statement, reproduced in Figure 3 below, illustrates such a case. Although the company's 12-member board is elected in full every year (i.e., the firm has a unitary board), the holders of "Class A Common Shares" are entitled to elect three of the directors while the holders of their "Common Voting Shares" are entitled to elect nine.

Figure 3: **2025 E.W. Scripps: A Challenging Case**

ANNUAL MEETING OF SHAREHOLDERS
AND PROXY STATEMENT

*The E.W. Scripps Company*
*312 Walnut Street, Cincinnati, Ohio 45202*



**Proposal 1 Election of Directors**

(Item 1 on the Proxy Card)

A board of 12 directors is to be elected, three by the holders of Class A Common Shares voting separately as a class and nine by the holders of Common Voting Shares voting separately as a class. The nominating & governance committee recommended to the Board each of the nominees set forth on the following pages. In the election, the nominees receiving the greatest number of votes will be elected. Directors are elected by the shareholders for terms of one year and hold office until their successors are elected and qualify. Each of the director nominees identified in this Proxy Statement has consented to being named as a nominee in our proxy materials and has accepted the nomination and agreed to serve as a director if elected.

Excerpts from The E.W. Scripps Company's Form DEF 14A, filed March 20, 2025. Class A Common Shares elect "the greater of three or one-third of the directors" and vote on nothing else, Common Voting Shares elect the remaining directors and vote on all other matters. The proxy splits its nominee tables by electing class, with the division by constituency and not by term, but directors are "elected by the shareholders for terms of one year," and the filing never mentions a classified board or Class I/II/III directors. Source: SEC EDGAR, accession no. 0000950170-25-042524.

Second, we measure whether the firm has a dual-class share structure. A dual-class firm is one that has two or more classes of common stock with differential voting rights—typically a low-vote class sold to public investors and a high-vote class retained by founders, their families, or other insiders. This structure enables the firm to raise outside equity capital while allowing insiders to retain voting control in excess of their ownership rights, driving a "wedge" between control over corporate decisions and claims on cash flows (DeAngelo and DeAngelo, 1985; Gompers, Ishii and Metrick, 2010). Given that this wedge effectively insulates management from the market for corporate control and shareholder voting, the dual-class structure is among the more powerful firm entrenchment devices. Although historically rare and concentrated in media and other family-controlled companies, dual-class firms have become markedly more common over

time, with a pronounced rise in dual-class IPOs among founder-led technology firms (Aggarwal, Eldar, Hochberg and Litov, 2022). The trend has generated an active policy debate over mandatory sunset provisions (Bebchuk and Kastiel, 2017) and led S&P Dow Jones to bar newly listed multi-class firms from the S&P Composite 1500 between July 2017 and April 2023.

As with board classification, nothing explicit in Schedule 14A requires a company to disclose whether it is a dual-class company. However, as stated earlier, Item 6 of the schedule requires the registrant to state the number of shares outstanding of each class of voting securities and the number of votes per share that each class is entitled to. The resulting disclosure under Item 6 makes classification trivial for the majority of publicly-listed firms. Figure 4 reports the relevant passage from the proxy statement for Hormel Foods Corporation's 2026 annual meeting. The filing states the number of shares of common stock outstanding on the record date, provides that each share is entitled to one vote, and confirms that the company has "no other class of shares outstanding."

Figure 4: **2026 Hormel Foods: The Easy Case for a Single Share Class**

**PROXY STATEMENT**

HORMEL FOODS CORPORATION
1 HORMEL PLACE
AUSTIN, MINNESOTA 55912

**GENERAL INFORMATION**

The enclosed proxy is solicited by the Board of Directors (the "Board") of Hormel Foods Corporation (the "Company") for use at the 2026 Annual Meeting of Stockholders to be held on January 27, 2026. We are first mailing this proxy statement and form of proxy, or a Notice of Internet Availability of Proxy Materials, to stockholders on or about December 17, 2025. The 2026 Annual Meeting of Stockholders will be held in-person in Austin, Minnesota.

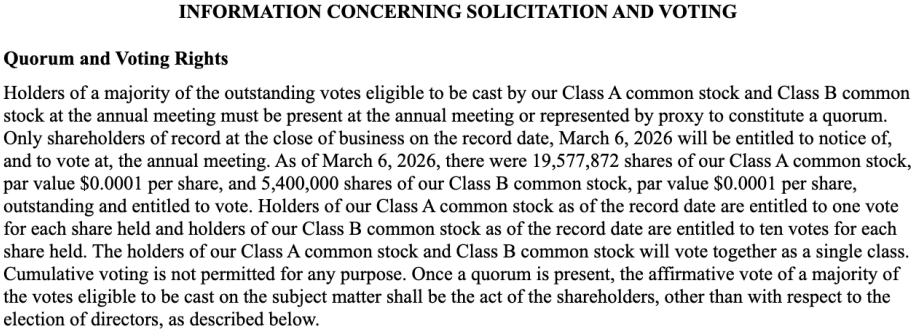
*Voting Securities.* Stockholders of record at the close of business on November 28, 2025 are entitled to vote at the meeting. We had 550,106,937 shares of common stock outstanding as of that date. Each share of common stock is entitled to one vote. There is no cumulative voting. We have no other class of shares outstanding.

Excerpt from the "General Information" section of Hormel Foods Corporation Form DEF 14A for the 2026 annual meeting, filed December 17, 2025 (CIK 0000048465). The voting-securities passage reports 550,106,937 shares of common stock outstanding on the November 28, 2025 record date, states that each share carries one vote, and confirms that no other class of shares is outstanding. Source: SEC EDGAR, accession no. 0000048465-25-000074.

A corresponding easy classification example for a dual-class firm appears in Figure 5, excerpted from the proxy statement for the 2026 annual meeting of Gaia, Inc., a streaming media company. The passage identifies the firm's two classes of common stock, the number of shares of each class outstanding as of the proxy statement's record date, and states how many votes each share type is entitled to—one vote per Class A share and ten votes per Class B share, with the two classes voting together as a single class. A single paragraph in the document, which resides in a similar location across proxy statements, contains the requisite information to easily classify the firm as

dual-class and to compute total votes outstanding.

Figure 5: **2026 Gaia: The Easy Case for Dual Class**



**INFORMATION CONCERNING SOLICITATION AND VOTING**

**Quorum and Voting Rights**

Holders of a majority of the outstanding votes eligible to be cast by our Class A common stock and Class B common stock at the annual meeting must be present at the annual meeting or represented by proxy to constitute a quorum. Only shareholders of record at the close of business on the record date, March 6, 2026 will be entitled to notice of, and to vote at, the annual meeting. As of March 6, 2026, there were 19,577,872 shares of our Class A common stock, par value $0.0001 per share, and 5,400,000 shares of our Class B common stock, par value $0.0001 per share, outstanding and entitled to vote. Holders of our Class A common stock as of the record date are entitled to one vote for each share held and holders of our Class B common stock as of the record date are entitled to ten votes for each share held. The holders of our Class A common stock and Class B common stock will vote together as a single class. Cumulative voting is not permitted for any purpose. Once a quorum is present, the affirmative vote of a majority of the votes eligible to be cast on the subject matter shall be the act of the shareholders, other than with respect to the election of directors, as described below.

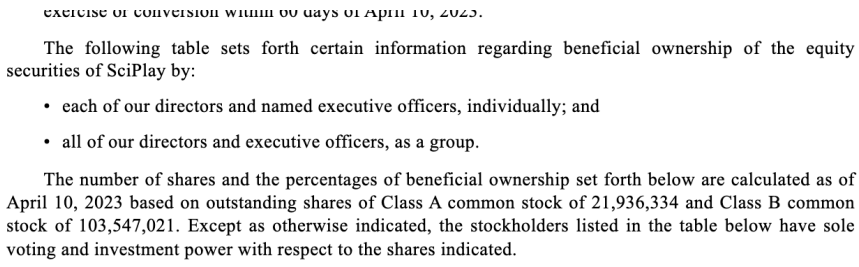
Excerpt from "Information Concerning Solicitation and Voting," Gaia, Inc. Form DEF 14A for the 2026 annual meeting, filed March 27, 2026 (CIK 0001089872). The passage reports the shares outstanding of each class as of the March 6, 2026 record date, the votes to which each class is entitled (one per Class A share, ten per Class B share), and that the two classes vote together as a single class. Source: SEC EDGAR, accession no. 0001140361-26-011724.

Most companies fall into one of these two straightforward categories, although complications can arise. For example, Item 403(b) of Regulation S-K mandates the disclosure of directors' and executive officers' beneficial ownership of equity securities of the filer "or any of its parents or subsidiaries."[6] Figure 6 illustrates how this can generate a classification issue with an excerpt from the 2023 proxy statement of Light & Wonder, Inc., a gaming company with a single class of common stock. The directors and officers of Light & Wonder also held shares of SciPlay Corporation, a majority-owned subsidiary, so the proxy reports a beneficial-ownership table for the equity securities of SciPlay, along with the per-class outstanding-share counts for SciPlay's Class A and Class B common stock. This portion of the proxy includes the indicia of dual class structures described above, including two class labels, per-class share counts, and an accompanying two-column beneficial-ownership table, but the classes belong to the subsidiary rather than to the filer. A classifier that keys solely on the class terminology near an ownership table would inadvertently code Light & Wonder as dual-class, while a correct extraction would pull the class structure for the parent filing entity.

Finally, our most conceptually difficult generated measure is inside ownership, or the amount of the firm's stock (and votes), beneficially owned by the directors and executive officers as a group, or by related entities. Insider ownership conceptually plays two roles within the agency framework. A larger equity stake from management ties their wealth to firm value and ostensibly aligns incentives with those of outside shareholders (Jensen and Meckling, 1976). However, the votes attached to the same stake insulate managers from any disciplinary measures provided by shareholder voting (Morck, Shleifer and Vishny, 1988). Given that matters put to a shareholder vote, first and foremost the annual director elections, are decided by votes cast, a large insider bloc will act as a buffer against outside activist pressure. A dissident slate must overcome the inside

[6] 17 C.F.R. *S* 229.403(b).

Figure 6: **2023 Light & Wonder: A Challenging Case**



exercise or conversion within 60 days of April 10, 2023.

The following table sets forth certain information regarding beneficial ownership of the equity securities of SciPlay by:

- each of our directors and named executive officers, individually; and
- all of our directors and executive officers, as a group.

The number of shares and the percentages of beneficial ownership set forth below are calculated as of April 10, 2023 based on outstanding shares of Class A common stock of 21,936,334 and Class B common stock of 103,547,021. Except as otherwise indicated, the stockholders listed in the table below have sole voting and investment power with respect to the shares indicated.

Excerpt from the beneficial-ownership section of Light & Wonder, Inc. Form DEF 14A, filed May 1, 2023 (CIK 0000750004). Light & Wonder has a single class of common stock; the Class A and Class B common stock described in the passage—including per-class outstanding-share counts as of April 10, 2023—belong to SciPlay Corporation, a majority-owned subsidiary whose securities are reportable under Item 403(b) of Regulation S-K because they were held by Light & Wonder directors and officers. Source: SEC EDGAR, accession no. 0001104659-23-053352.

vote share before reaching a majority, and a sufficiently large bloc potentially allows insiders to foreclose any proposal, including the replacement of the directors themselves. Unlike a staggered board or a dual-class charter provision, this form of protection will not show up in the governing documents filed with the secretary of state for the state of incorporation, but resides entirely in the required proxy disclosures.

The disclosure that identifies insider ownership is the beneficial-ownership table required by Item 403 of Regulation S-K, described in Section 2.2.1. For each class of voting securities, Item 403 requires disclosure of the holdings for each director, nominee, and named executive officer (NEO), the holdings of all directors and executive officers as a group, and the holdings of any outside person known to beneficially own more than five percent of the class. The as-a-group row is our primary extraction target, and in easy cases there is nothing else to reconcile against. Figure 7 reports such a case from the proxy statement of CytoDyn Inc., a biotechnology company whose shares trade over the counter. CytoDyn is the type of issuer that commercial governance datasets can not cover, although because it never listed on an exchange it also lies outside the CRSP-based frame we use in this paper. However, it files the same Item 403 disclosure as every listed firm, and our framework can easily be extended to this population. The company has a single class of common equity, and no other party beneficially owned more than five percent of the outstanding shares. The beneficial ownership table thus only includes the individual directors and NEOs, as well as the group row aggregating both the share count and the percent of class. There is no wedge to compute for single class firms, because the share of votes and ownership percentage are the same.

The corresponding straightforward case for a dual-class firm requires some additional features, most notably the "per-class" columns. Figure 8 reports the beneficial-ownership table from the proxy statement for the 2026 annual meeting of Duluth Holdings Inc., a clothing and apparel retailer that has two classes of common equity. Their Class A common stock receives ten votes per share, while Class B receives only one. For each holder and for the collective group, the table

Figure 7: **2025 CytoDyn: The Easy Case for Beneficial Ownership with a Single Share Class**

**STOCK OWNERSHIP BY PRINCIPAL STOCKHOLDERS, DIRECTORS AND EXECUTIVE OFFICERS**

The following table sets forth the beneficial ownership of common stock as of September 15, 2025, by (i) each of our directors; (ii) each of our named executive officers; and (iii) our current executive officers and directors as a group. To our knowledge, as of September 15, 2025, there were no beneficial owners of more than 5% of our outstanding common stock.

| Name of Beneficial Owner | Amount and Nature of Beneficial Ownership (1) | Percent of Total (2) |
|---|---|---|
| Directors and Named Executive Officers: | | |
| Jacob Lalezari, M.D. (3) | 2,312,500 | * |
| Robert Hoffman | — | * |
| Mitchell Cohen | — | * |
| Tyler Blok(4) | 1,255,186 | * |
| Tanya Durkee Urbach (5) | 2,121,724 | * |
| Lishomwa C. Ndhlovu, M.D., Ph.D. (3) | 2,059,611 | * |
| Karen J. Brunke, Ph.D.(3) | 1,884,611 | * |
| Ryan Dunlap(3) | 1,785,334 | * |
| Stephen M. Simes(3) | 1,778,012 | * |
| Current directors and executive officers as a group (8 persons)(6) | 13,196,978 | 1% |

* Less than 1% of the outstanding shares of common stock.

(1) Beneficial ownership includes shares of common stock as to which a person or group has sole or shared voting power or investment power. Beneficial ownership also includes shares subject to stock options, warrants, or other rights to acquire shares that are exercisable currently or within 60 days following September 15, 2025; such shares are deemed outstanding for purposes of computing the number of shares beneficially owned and percentage ownership of the person or group holding such stock options, warrants or other rights, but are not deemed outstanding for purposes of computing the percentage of any other person. Unless otherwise stated, numbers represent shares of common stock.

(2) Percentages are based on 1,257,158,425 shares of common stock outstanding as of September 15, 2025.

(3) Represents shares subject to stock options.

(4) Includes 1,254,936 shares subject to stock options.

(5) Includes 1,959,611 shares subject to stock options.

(6) Includes 13,034,615shares subject to stock options.

Excerpt from "Stock Ownership by Principal Stockholders, Directors and Executive Officers," CytoDyn Inc. Form DEF 14A for the November 21, 2025 annual meeting, filed September 29, 2025 (CIK 0001175680). The filing states that no person beneficially owned more than five percent of the outstanding common stock as of September 15, 2025, so the table lists only the individual directors and named executive officers and the group row (13,196,978 shares, or approximately 1% of the single one-vote-per-share class). CytoDyn's shares trade over the counter and the company has never listed on an exchange, so it lies outside the CRSP-based frame of the panel described in Section 3.1; the excerpt illustrates the disclosure, which is identical in form for listed and unlisted filers. Source: SEC EDGAR, accession no. 0001104659-25-094518.

reports the number of shares and the percent of each class along with the percent of total voting power. The group row aggregates insider control, with directors and executive officers holding all 3,364,200 Class A shares (through a voting trust held by founder Stephen Schlecht), 12,048,655 Class B shares (34.7% of that class) and a resulting 66.8% of the total voting power. From the aggregate row we can measure the insider economic ownership, with the group's 15,412,855 shares representing 40.4% of the 38,128,698 shares outstanding across both classes. The insider voting share (66.8%) exceeds the cash-flow claim (40.4%) by roughly 26 percentage points. This wedge is calculated entirely from the values printed in the table, and the filing notes that no options or other rights to acquire shares were outstanding, so the 60-day beneficial-ownership adjustments described in Section 2.2.1 are not triggered.

However, beneficial-ownership disclosures are not always as straightforward, even if we constrain our attention to the "as-a-group" row. Figure 9 reproduces the ownership table from the same Gaia, Inc. proxy statement whose Item 6 disclosure simplified the dual-class classification in Figure 5.

Figure 8: **2026 Duluth Holdings: The Easy Case for Beneficial Ownership with Dual Class Shares**

**SECURITY OWNERSHIP OF CERTAIN BENEFICIAL OWNERS**

The following table sets forth certain information with respect to the beneficial ownership of our Class A and Class B common stock as of the record date (except as otherwise noted) for:

- each shareholder known by us to be the beneficial owner of more than 5% of our capital stock;
- each of our directors and director nominees;
- each of our named executive officers; and
- all of our directors and executive officers as a group.

We have determined beneficial ownership in accordance with the rules of the SEC. Under these rules, beneficial ownership includes any shares as to which the individual or entity has sole or shared voting power or investment power, as well as shares that may be acquired within 60 days of the record date. There are no shares of common stock subject to options or other rights outstanding as of the date of this proxy statement. Except as indicated by the footnotes below, we believe, based on the information furnished to us, that persons and entities named in the table below have sole voting and investment power with respect to all the capital stock that they beneficially own, subject to applicable community property laws.

Applicable percentage ownership is based on 3,364,200 shares of Class A common stock outstanding and 34,764,498 shares of Class B common stock outstanding as of the record date, as well as shares which may be acquired within 60 days of the record date. Unless otherwise indicated, the address of each beneficial owner in the table is: Duluth Holdings Inc., 201 East Front Street, Mount Horeb, Wisconsin 53572.

| | Beneficial Ownership of Common Stock | | | | |
|---|---|---|---|---|---|
| | Class A | | Class B | | % of Total |
| | Shares | % | Shares | % | Voting Power |
| *5% Shareholders:* | | | | | |
| Stephen L. Schlecht and Marianne M. Schlecht Descendants Trust[1] | — | — | 8,694,295 | 25.0% | 12.7% |
| Askeladden Capital Management LLC and Samir Patel[2] | — | — | 2,207,287 | 6.3% | 3.2% |
| *Named Executive Officers, Directors and Director Nominees:* | | | | | |
| Stephen L. Schlecht | 3,364,200[3] | 100.0% | 7,119,588 | 20.5% | 59.6% |
| Stephanie L. Pugliese[4] | — | — | 2,460,009 | 7.1% | 3.6% |
| Samuel M. Sato[5] | — | — | 857,966 | 2.5% | 1.3% |
| Heena K. Agrawal | — | — | 396,152 | 1.1% | * |
| David Homolka | — | — | 379,911 | 1.1% | * |
| David C. Finch | — | — | 207,480 | * | * |
| Janet H. Kennedy | — | — | 112,128 | * | * |
| Brett L. Paschke[6] | — | — | 155,314 | * | * |
| Susan J. Riley | — | — | 82,556 | * | * |
| Ronald Robinson | — | — | 74,973 | * | * |
| Scott K. Williams | — | — | 136,629 | * | * |
| All Executive Officers, Directors, and Director Nominees as a Group (12) | 3,364,200 | 100.0% | 12,048,655 | 34.7% | 66.8% |

Excerpt from "Security Ownership of Certain Beneficial Owners," Duluth Holdings Inc. Form DEF 14A for the 2026 annual meeting, filed April 21, 2026 (CIK 0001649744). Class A common stock carries ten votes per share and Class B common stock one vote, voting together as a single class; 3,364,200 Class A and 34,764,498 Class B shares were outstanding on the March 27, 2026 record date. The group row reports 3,364,200 Class A shares (100.0% of the class), 12,048,655 Class B shares (34.7%), and 66.8% of total voting power; the implied economic ownership is 15,412,855 of 38,128,698 shares, or 40.4%, for a voting–economic wedge of roughly 26 percentage points. The proxy reports that no options or other rights to acquire shares were outstanding. Source: SEC EDGAR, accession no. 0001193125-26-165206.

Recall that, under Rule 13d-3(d)(1), a security holder is deemed to beneficially own any shares they have the right to acquire within 60 days, which includes the conversion of another class. Because Gaia's ten-vote Class B stock is convertible into Class A stock at any time, the group's Class A row reports 6,339,721 shares, which implicitly includes all 5,400,000 conversion shares, the same ones also reported again in the Class B group row directly below it.[7]

Dividing the reported share count for the group by the 19,577,872 Class A shares outstanding yields an ownership share of 32.4%, which matches neither the 26.02% in the percent-of-class column nor the 25.38% in the final column, which reconciles only against the enlarged base of 24,977,872 shares assuming full conversion of the Class B. The two measures that our process calculates do not, unfortunately, appear directly anywhere in the document. The insider group holds roughly 715,000 Class A shares along with its 5,400,000 Class B shares, for 74% of the total votes, against an economic ownership claim of 25.4%. This difference nets out to a wedge of nearly

[7]This co-mingling is documented only in the footnotes to the table, which disclose that Mr. Rysavy's Class A figure of 5,768,225 shares consists of 291,682 shares held directly, 76,543 restricted stock units vesting within sixty days, and 5,400,000 shares issuable upon conversion of Class B stock held personally and through a family estate-planning entity, referred to as "Jirka Rysavy, LLC" in the table but "Jirka Rysavy, LLLP" in the footnote.

fifty percentage points that is hidden in the filing. In order to extract these measures from this table, we have to pull the printed share counts and ownership shares as primitives, detect how and why each percentage is computed, and derive voting power and economic ownership downstream in a separate process.

Figure 9: **2026 Gaia: A Challenging Case for Beneficial Ownership**

**BENEFICIAL OWNERSHIP OF SHARES**

The following table sets forth information with respect to the beneficial ownership of our common stock as of March 6, 2026 for (i) each person (or group of affiliated persons) who, insofar as we have been able to ascertain, beneficially owned more than 5% of the outstanding shares of our Class A common stock or Class B common stock, (ii) each director and director nominee, (iii) each executive officer named below in the Summary Compensation Table, and (iv) all current directors and executive officers as a group. We have based our calculation of the percentage of beneficial ownership on 19,577,872 shares of our Class A common stock and 5,400,000 shares of our Class B common stock outstanding on March 6, 2026.

| Title of Class of Common Stock | Name and Address of Beneficial Owner | Amount and Nature of Beneficial Ownership[1] | Percent of Class[2] | Percent of Class A Assuming Full Conversion of Class B Ownership[3] |
|---|---|---|---|---|
| Class A | AWM Investment Company, Inc.[4] | 1,252,090 | 6.40% | 5.01% |
| | BlackRock, Inc.[5] | 1,027,164 | 5.25% | 4.11% |
| | Flint Ridge Capital LLC[6] | 1,180,000 | 6.03% | 4.72% |
| | John P. Szabo, Jr.[6] | 1,950,000 | 9.96% | 7.81% |
| | Koller Capital LLC[7] | 1,267,763 | 6.48% | 5.08% |
| | Jirka Rysavy[8] | 5,768,225 | 23.09% | 23.09% |
| | Kiersten Medvedich[9] | 93,769 | *% | *% |
| | James Colquhoun[10] | 951,203 | 4.86% | 3.81% |
| | Ned Preston[11] | 2,000 | *% | *% |
| | Yonathan Nuta[12] | 5,541 | *% | *% |
| | Paul Sutherland[13] | 347,719 | 1.78% | 1.39% |
| | Kristin E. Frank[14] | 117,467 | *% | *% |
| | Anaal Udaybabu | — | —% | —% |
| | Keyur Patel | — | —% | —% |
| | Kimberly Arem[15] | 5,000 | *% | *% |
| | All directors, director nominees, and officers as a group (9 persons) | 6,339,721 | 26.02% | 25.38% |
| Class B | Jirka Rysavy[8] | 1,400,000 | 25.93% | 5.60% |
| | Jirka Rysavy, LLC[8] | 4,000,000 | 74.07% | 16.01% |
| | All directors, director nominees, and officers as a group (9 persons) | 5,400,000 | 100.00% | N/A |

Excerpt from "Beneficial Ownership of Shares," Gaia, Inc. Form DEF 14A for the 2026 annual meeting, filed March 27, 2026 (CIK 0001089872)—the same filing excerpted in Figure 5. Share counts follow the beneficial-ownership convention of Rule 13d-3(d)(1), so the group's Class A row of 6,339,721 shares includes all 5,400,000 shares issuable upon conversion of the Class B stock reported separately in the Class B panel, and the two group rows overlap. The percent-of-class figures are computed on holder-specific, as-converted denominators rather than the shares outstanding, and no printed number reports the group's voting power (roughly 74% on an unconverted basis) or its roughly 25% economic claim. Source: SEC EDGAR, accession no. 0001140361-26-011724.

This share conversion problem is only one entry in a larger set of complications with beneficial ownership that surfaces across the corpus of filings. While a longer discussion of edge case rule-setting is detailed in Appendix A, we briefly discuss here a set of considerations that our protocol determined based on an empirical review of proxy filings. A first set concerns which shares count as inside ownership. Under SEC reporting rules, shared voting power is attributed to every person who holds it, so that a single ownership block can appear in multiple rows at once. For example, shares held in a voting trust might be reported for the trust as well as for each of its trustees. Shares held by family partnerships, foundations, and estate vehicles will often appear in the individual rows of the family members who control them and in a separate family-group row that partially

overlaps the directors-and-officers group.[8]

There are also considerations regarding how to count votes. Some classes carry votes that are stated only in the footnotes but not in the ownership table, including conduit classes that own a handful of shares but command hundreds of millions of votes, fractional votes-per-share ratios that change for each filing year, and preferred stock that votes alongside the common shares but does not enter economic-ownership measures.[9] We also see more complicated economic ownership forms, including umbrella-partnership ("Up-C") firms that report ownership on both a publicly-available and fully-exchanged basis, frequently without saying which convention the table follows. Finally, some complications are arithmetic rather than conceptual, including percentages reported as fractions, printed percentages that imply a different number of shares outstanding than stated in other parts of the filing, group rows that are missing rather than zero, and early filings that are rendered as plain text with column boundaries only visually demarcated. Although none of these cases arise often, but they are common enough that a naive pipeline would produce errors that are correlated with ownership structure. Appendices A.5 and A.6 catalog the deterministic repairs and conventions that resolve different complications (see, in particular, Table 5).

Given the intricacies of the corporate form, our parsing process cannot resolve every knife-edge issue, and our resulting dataset surely contains residual measurement error. However, the same is true for every attempt to codify governance provisions at scale, including the hand-coded commercial datasets on which the empirical governance literature was built. Gompers et al. (2003, p. 113) acknowledge that the IRRC data underlying their G-Index measure did not update for every company in each edition and was intended to be merely "a starting point for institutional investors", with the result being "a noisy measure of a firm's governance provisions". Over the intervening years, further research has corroborated this concern, with Daines et al. (2010) concluding that the inability of rating providers to agree on measurement "highlights the need for future research on developing reliable and valid measures of the construct 'corporate governance.'" Our framework, though not without cost or error, departs from prior attempts by creating audited accuracy estimates, where every value can ultimately be traced to the underlying disclosure, as we describe in Section 2.2.5.

[8] In Houlihan Lokey's 2022 proxy statement, the identical block of 19,135,484 Class B shares (100% of the class )is reported in four separate rows: once for the HL Voting Trust, once for each of its two trustees, and once for the group (accession no. 0001302215-22-000063). Village Super Market's 2025 proxy reports a "Sumas Family Group" as a five-percent holder alongside individual rows for its four members, each of whom is also a director or officer (accession no. 0001140361-25-039481).

[9] RB Global's subordinate voting shares carried 0.0139696 votes per share at the 2025 record date, a ratio set by an exchangeable-share formula rather than the charter (accession no. 0001140361-25-010020). At the other extreme, some filers print voting percentages that govern only certain matters: John B. Sanfilippo & Son's ownership table reports each holder's "% of Outstanding Votes on Matters Other than Election of Directors," because its two classes elect portions of the board separately (accession no. 0000950170-25-114207).

### 2.2.3 Sample Construction and the Transcribe/Compute Split

We begin our sample with the full merged CRSP/Compustat universe of U.S. common equity (share codes 10 and 11) from 1997 through 2026, without restricting to exchange or index-membership. The resulting sample includes every listed firm, regardless of size or venue, so is inclusive and broader than the S&P 1500 constituents which commercial governance databases cover. Because we then collect every proxy that a firm has ever filed with the SEC, the panel also follows firms into and out of public listing, and includes years where their shares trade over the counter. Issuers that never list on an exchange never enter CRSP, and are therefore outside our frame. We begin with the CRSP/Compustat universe because it is required by our applications, but applying the framework unchanged to the full population of EDGAR proxy filers would be straightforward. Our panel begins in 1997 because that is when EDGAR electronic filing became complete, and extraction quality was low for the earliest 1994–96 filings. Appendix A details the universe construction and every other implementation choice summarized in this section.

We use the historical linking tables available through WRDS to merge stock trading data and firm financial reporting records to the SEC's EDGAR filing portal using the SEC filer identity (CIK).[10] Foreign private issuers are outside the scope of this project because they do not file proxy statements under Schedule 14A. Similarly, government-sponsored enterprises (GSEs) are flagged and excluded from our primary governance measures because their board structure and ownership are fixed by federal charter rather than chosen by the firm. The resulting merge contains 125,817 firm-years for 14,440 CIKs. We then pull the full filing history from EDGAR for every CIK hit, and collect each definitive proxy statement (form types DEF 14A and DEF 14) on record, storing the underlying text only once into a local cache to be used in later stages of our protocol. This process results in a total of 140,314 definitive proxy filings from 13,790 firms.

We refer to the LLM parsing process for the extracted proxy filings as the *transcribe/compute split*. We assign the open-weights model the task of extracting values and text, as printed in the filings, rather than directly coding the measures. Every subsequent measure is then computed after extraction in separate code. The model extracts the firm's share classes (the class label, number of votes per share, and shares outstanding, along with special-case fields for classes whose votes are stated only in the aggregate and for operating-partnership units in umbrella-partnership structures), the values from the rows of the beneficial-ownership table (each holder's name, per-class share count, and the percent-of-class and total-voting-power figures as printed), the group's 60-day option count where the filing discloses it, and the board-classification determination, accompanied by a verbatim supporting quotation from the filing. All final measures, including the total shares and votes outstanding, the group's economic ownership under the SEC's 60-day convention, its

[10]Note that we use historical record linkages here, which is important because CIKs, tickers, and company names can all change over time for a given firm. Using static header information would drop firms that changed identity from our sample, which would disproportionately affect small and delisted firms. Each row of the panel records which link source produced its CIK.

voting power, the wedge between them, the dual-class classification, and the row's quality tier are derived afterwards using the parsed terms and values as primitive measures.

The Gaia ownership table in Figure 9 demonstrates the value of our protocol. The ultimate ownership measures for this observation, including the insider group's voting power of roughly 74 percent and its economic claim of roughly 25 percent, do not appear directly in the document. Calculating them correctly requires recognizing that the insider group's Class A row includes 5,400,000 shares that are only issued upon conversion of the Class B shares. If we simply asked the LLM to derive the quantities directly, it would produce a guess, most often by reporting economic ownership equal to voting power. By splitting transcription and computation, the model is never asked to perform a challenging or opaque arithmetic. Rather, the model simply extracts observations from table rows, including ones with overlapping share counts, and then we resolve the conversion double-count explicitly with one downstream code.

This protocol avoids hallucinations while parsing measures at a scale that would take human coders thousands of hours. Splitting the derivation process in this manner also provides three additional benefits that scale with the project scope. First, individual observations our final panel can be audited because each derived measures traces back to primitives that a human can check against the filing itself. This makes the validation protocol detailed in Section 2.2.5 possible. Second, it makes it extremely cheap, both in time and cost, for us to change how we measure a given governance provision. Determinations that must be made without a mechanical answer and that require judgment, for example whether voting preferred stock should count as insider voting power, or whether an absent group row should be counted as zero or missing, are determined once and in codified in code. Making the decision to revise such a convention will recompute the entire panel in short order without changing the underlying model or re-reading a single filing. Third, the split produces in built-in validation, because we can recompute each beneficial owner's voting power from the reported shares and votes per share and compare it to the proxy's printed value. Material differences between the reported and calculated values signify a process error in our protocol, which is an internal consistency check on the extracted measures.

### 2.2.4 Architecture

We identify and retrieve the underlying text documents only once, pulling each firm's EDGAR history under the SEC's rate limits, and each filing's cleaned text is stored locally. Each subsequent extraction wave and audit reads from the same document cache instead of returning to EDGAR, which streamlines the process. We generate the final firm-year panel from the text corpus after apply coding conventions and quality flags with deterministic code described in Appendix A.5. Because the resulting panel is versioned, each vintage can be reconstructed quickly using vintaged code.

The second organizing principle is a division of labor between two classes of language models.

Our protocol uses cheap open-source models to parse documents at scale, and only uses expensive frontier models to validate this output. The transcription pipeline runs on an open-weights 35-billion-parameter model, which transcribes the panel's 140,314 filings at effectively zero marginal cost. The model runs at temperature zero to be reproducible, and its output is constrained to a typed schema so that unstructured text is never produced. A frontier commercial model (Claude Opus) is used only to construct the audit ground truth described in Section 2.2.5, to adversarially adjudicate disagreements, and to rescue any filing whose layout defeats the standard extraction path. The total frontier-model spend for the entire validated panel is on the order of one hundred dollars.

Although proxy statements can run hundreds of pages long, the disclosures that we need generally arise in similar locations. Our pipeline deterministically extracts relevant context windows, rather than handing the model the full document, or relying on embedding-based retrieval. The model identifies the relevant sections of the proxy by the structure of the mandated disclosures, because the beneficial-ownership table is located by its required "as a group" row, and snippets are identified around the votes-per-share statement required by Item 6 and the director-term disclosure required by Item 401(a). Relying on the structure of the required disclosures rather than on a hand-chosen governance vocabulary addresses some of the ex-ante concerns identified in Section 2.2.2. A unitary board will not necessarily generate text addressing "class" and "term," but the proxy cannot avoid disclosing director terms, which means the relevant information can be recovered either way. The selected portions are also guided by pre-determined layout hints, including share counts parsed from the cover page, the ownership-table column headers in the order they appear, and explicit warnings for known inconsistencies, such as when a subsidiary issuer's securities were reported in the ownership section for Light & Wonder in Figure 6. This guidance is designed to save the model for reading disclosures rather than decoding layout, but inconsistent filing text will still govern in the case of a conflict.

Finally, include a set of pre-worked example to assist the model. The pipeline includes five short demonstrations built off hazards we found to present difficulties for using naive, one-shot extraction processes. These include classes whose vote amounts are stated only in the aggregate, umbrella-partnership (Up-C) units, shares deemed beneficially owned upon conversion, phantom units, and voting preferred stock. Each example is injected only when a pre-determined trigger detects the presence of a hazard in a filing. Most filings therefore carry no examples at all, so to avoid priming the model to detect complex issues that do not exist. When the full standard protocol fails, either because it cannot locate the beneficial ownership table or it returns no owners from a filing, we send the full filing to a frontier model with a larger context window.

This escalation is rare and can be identified in the output. A frontier-model-based extraction will replace the incumbent value only when it strictly improves against ground truth (if available), or it clearly improves upon the incumbent where we do not. Any fix that changes a stored or derived value leaves an audit trace and is flagged in the affected row in the final panel dataset, so

the provenance of any value in the panel can be reconstructed. Appendix A.5 catalogs the repair rules in their application order, and Appendix A.6 lists the ruled conventions that are enforced in the derivation step.

### 2.2.5 Validation Protocol

We use our protocol to generate governance variables for all publicly available proxy statements, and hand-auditing every entry would be infeasible and cost prohibitive. Instead, we build in validation as a component of our pipeline, using two checks for different purposes, and manually check frontier audit determinations.

Our first check conducts an adversarial benchmark for 170 filings from 100 dual-class firms that are deliberately chosen to over-weight the challenging cases found in Section 2.2.2, including low quality tiers, sharp swings in a firm's time series, and complicated ownership structures. Each filing was labeled by a frontier model and adjudicated by hand against the source proxy. Because the sample is designed to be challenging, we use it to stress test the process and, more importantly, as a fixed benchmark that has to be exceeded before any proposed change can enter the pipeline. We refer to this requirement as a "regression gate." A prompt revision or repair rule can only be adopted if it matches or improves the benchmark score, so a change designed to fix one flaw will not silently break other filings that were previously parsed correctly.

For our second check we draw representative random samples and audit them independently to generate a precision estimate for the panel. The frontier model is shown the extracted values for each sampled filing, alongside the classification and source text, and asked to re-derive every field. Disagreements are flagged and passed to a second model that is prompted to refute the auditor, and instructed to side with the pipeline when the answer is not clear. A field counts as an error only when the auditor disagreed and the adversarial verifier upholds the disagreement. On the initial 250-filing sample (150 dual-class, 100 single-class firm-years), dual-class classification was correct in all 248 measurable filings, board classification in 98–100 percent per arm, and the insider voting and economic ownership percentages in 93–96 percent.

Our protocol expands governance measurement beyond the large-cap universe that commercial data providers cover, and includes filings that are older, with more varied layouts, and where the ownership structures are more concentrated. We drew a second sample of 160 filings exclusively from this expanded territory, including small-cap and OTC issuers and the earlier 1997–2002 period. Against an adjudicated truth, the pipeline scored 99.4 percent on dual-class classification, 99.3 percent on board classification, 95.9 percent on insider voting power, and 92.5 percent on insider economic ownership, with the hardest subset of measures—economic ownership among small-cap dual-class firms—at 87.5 percent. We use this check and the resulting review to generate a deterministic set of repairs in Appendix A.5 using the regression gates described above above. The remediated panel scores 100 percent on the three classification and voting fields and 98.5

percent on economic ownership against the same truth.

Although LLM-based extraction has advanced dramatically in recent years, machine auditors still make mistakes (Frankenreiter, 2026). Where a dispute arose regarding an auditor's verdict, it was sent to human adjudication against the source filing rather than being resolved unilaterally. We maintain a standing rule throughout that ascribes $\pm 0.75$ percentage-point differences to rounding and denominator-basis, but a value that lands within this tolerance that is shown to be the result of an improper method will still be flagged as an error. In our data, human adjudication moved the pooled voting-power accuracy up by 3.3 percentage points and the economic-ownership accuracy by 2.9 points relative to the raw frontier model audit, and an uncorrected machine audit would have understated pipeline accuracy.

Finally, to some extent, the panel nature of the output supplies its own validation instrument, because a firm's governance structure is highly persistent. Implausible within-firm changes in the board or dual-class series are flagged for review, although the values are not immediately overwritten based solely on these patterns. We report the agreement with external datasets, including the CII dual-class list and the ISS/RiskMetrics governance files, in Section 3.3.

### 2.2.6 Quality Tiers

Some filings in our sample cannot be reliably extracted, usually because a denominator share count is left unstated, the calculated percentage is impossible, or the printed and recomputed voting figures disagree materially. Rather than drop problematic observations, every row in the resulting dataset carries a quality tier assigned with the output. These include `ok` for rows whose values are internally consistent, `equal-votes` for dual-class firms with identical votes per share, or where control operates through class-specific board-election rights rather than a vote differential so that the per-share wedge is undefined. `partial` where a share-class denominator is missing, `broken` where an extracted value is impossible (a percentage far above 100, which is usually a denominator issue), `suspect` where the recomputed and printed voting powers disagree by more than two percentage points. and `gse` for the government-sponsored enterprises discussed in Section 2.2.3.

The error rate from the audits for flagged rows is roughly seventeen percentage points higher than on clean rows, though just under two percent of the panel is assigned to `partial`, `broken`, or `suspect`. The flagged observations are roughly two and a half times higher among dual-class filings than single-class ones, and three times higher in the earliest filing years than for the most recent years. Harder-to-extract filings are also disproportionately more common in concentrated-ownership, exotic-structure firms that governance research might be particularly interested in. By flagging rather than dropping these observations, we allow subsequent analysis to investigate them further and choose its own exclusion policy. In this paper, our default is to include all tiers in our analysis, excluding flagged tiers only for sensitivity analyses.

### 2.2.7 How Not to Use LLMs

The design rules described above were adopted in response to specific failures. During the parsing process, we learned the following rules for how *not* to use LLMs to generate extracted data.

*Do not ask the model to output something it can't see.* This lesson drove the transcribe/compute split from Section 2.2.3. When asked directly for a measure of economic ownership, the model reliably produced a number regardless of its ability to actually determine it, most often by just equating it with voting power. This error is hard to detect for individual observations but is deeply problematic in aggregate, because it will tend to bias the wedge measure towards zero.

*Examples don't teach concepts.* So-called "few-shot" prompting improves how accurately a mid-sized model can transcribe what it already understands, but it will not add concepts to the model's knowledge base. When we added a schema field for sunset provisions mid-project, the production model populated it in 4 of 10,294 extraction attempts despite worked examples demonstrating exactly what was wanted. To accurately capture the concept would require a deterministic text structure or the power (and cost) of a frontier model.

*No use re-estimating the parsing.* At temperature zero the model is (almost completely) a deterministic function of its context, and re-computing the output will reproduce the same errors. Any computation errors have to be fixed downstream from the transcription, in deterministic code that has generalizable rules, can be unit-tested, and leaves an audit trail.

*Be careful when making changes to the protocol.* A revised prompt that is better for specific extraction terms will often silently break others. Twice during development, a wholesale change lowered accuracy in our adversarial benchmark. A challenger process should replace the incumbent approach only if it can be shown to strictly improve against ground truth where we have it, or to be clearly conceptually superior where we do not.

*You can not always trust the printed numbers.* We treat a filing's printed percentage as authoritative for what it is designed to measure (i.e., we assume that the lawyers did not make mechanical errors), but what a printed figure measures can be ambiguous. When we attempted an aggressive "printed-values-win" rule, it broke more extractions than it fixed. Our final rule adopts a printed figure only when it arithmetically reconciles with the row's transcribed primitives.

*Do not use a keyword rule without checking it.* Proxy statements can reuse identical language for unrelated concepts. "Three-year terms" describes staggered boards and option-vesting schedules with the same vocabulary. A global keyword search for staggered boards frequently leads to incorrect classifications unless it is was limited to the small window of board-related language. This is one reason a simple "bag of words" approach has not succeeded at generating governance data at scale.

*Do not take the machine audit at face value.* A frontier model used to audit open-weights output unfortunately still make mistakes, and human review of disputed verdicts increased our measured accuracy by roughly three percentage points (Section 2.2.5), Reflexively accepting a

machine audit would have understated the pipeline's accuracy and generated phantom errors.

The primary takeaway of these rules is that a language model inside a data pipeline is a not magic but just another tools, albeit one whose errors are systematic and reproducible. The process that makes the resulting output trustworthy is the structure around the model, including the context construction fed to the model, the deterministic computation rules on the transcriptions, regression gates on changes, and validation with committed evidence at different steps. These rules are not specific to proxy statements, and we expect the same lessons to transfer to other settings with partially structured data from standardized public documents.

# 3 Governance Panel: 1997–2026

## 3.1 Universe and Coverage

The resulting panel contains one row per definitive proxy filing, derived from 140,314 filings for 13,790 firms, of which 137,336 are proxy statements (DEF 14A) and 2,978 are information statements (DEF 14C). Each row is dated by the calendar year of the filing. A proxy filed in year $t$ produces ownership and board structure measures as of a record date early in year $t$, and maps to the fiscal year immediately preceding it. In addition, a firm occasionally files multiple proxies in a year, with an amended or a special-meeting proxy reported in addition to the annual meeting proxy filing. This affects 6,049 of the panel's 133,569 distinct firm-years (4.5 percent), contributing 6,745 additional rows. Each separate filing appears as its own row, and analyses done at the firm-year level keep only the latest-filed definitive proxy.

After keeping the validated quality tiers (`ok` and `equal-votes`) and dropping government-sponsored enterprises, our analysis sample consists of 128,552 firm-years for 13,728 firms over filing years 1997–2025. The remaining observations include 1,873 rows tiered `partial`, 690 tiered `broken`, 66 tiered `suspect`, and 32 government-sponsored enterprise filings, which we retain in the released dataset with their tier designations, per Section 2.2.6. The flagged observations represent just under two percent of the sample, and their share falls over time—from 3.2 percent of filings in 1997–2002, to 2.0 percent in 2003–2010, 1.2 percent in 2011–2018, and 0.9 percent thereafter. The flagged share is roughly two and a half times higher among dual-class filings (4.5 percent) than among single-class filings (1.7 percent).

Table 1 reports the coverage metrics over time using our protocol. Across the full 1997–2024 frame we recover at least one definitive proxy for 114,104 of 125,817 firm-years, or 90.7 percent. Coverage is lowest in the earliest part of the sample, at 85.6 percent, when a larger share of documents were plain-text. The rate rises to 91.9 percent in 2003–2010, 94.3 percent in 2011–2018, and 92.5 percent for the most recent period. The uncovered sample consists primarily of firm-years without an annual meeting or where no proxy was solicited. This occurs for newly listed firms whose first proxy season falls in the following year, firms that enter bankruptcy or

liquidation, and firms that are acquired before their meeting date. Coverage varies within firm size and time period. For 2003–2024, the smallest within-year market-capitalization decile of the frame is covered at 91 percent against 96 percent for the largest, and the corresponding figures for 1997–2002 are 79 and 84 percent. Because the pipeline treats every filing the same, the size gradient largely reflects the meeting and solicitation behavior of small firms rather than extraction differences.

Assigning each observation to an index tier as of its filing date shows[11] that 10.5 percent of firm-years are S&P 500 constituents, 13.6 percent are in the S&P 400 or 600, and 75.9 percent lie outside the S&P 1500 altogether (the population that ISS, IRRC, and Execucomp-based governance datasets do not observe). The outside share has fallen over time as the market has shifted towards larger firms, from 85.7 percent of firm-years in 1997–2002 to 66.6 percent in 2019–2024. Weighted by market capitalization, the share of observations that is uncovered by commercial data providers is only 13.0 percent. How much this coverage gap matters depends on the research question and underlying measure of interest. However, any fact about insider ownership drawn from S&P 1500 data pertains to only a quarter of the firm-years in which U.S. reporting companies disclose this information.

The panel only includes domestic reporting companies because foreign private issuers file annual reports on Form 20-F and are exempt from the proxy rules. As noted above, domestic issuers that never list on NYSE, AMEX, or Nasdaq do not enter the CRSP universe and so are not retrieved, even though they file proxy statements. In this paper, we build off the CRSP/Compustat sample because it connects most directly to the applications in the literature, but extending the data set to the full population of EDGAR proxy filers is a simple matter of pulling additional filings, and a researcher interested in the over-the-counter market could do so with the pipeline unchanged. We return to this in Section 6.

## 3.2 Measured Objects

Every row of the final dataset includes three groups of variables: the primitives transcribed directly from the filings, the derived governance measures, and flags that record how each derived value was produced. Table 2 lists the principal fields. Here we define the derived measures and report how choices affect the resulting values.

**Board classification.** A firm-year is coded as having a classified (staggered) board when

[11]The tier shares are computed within the panel, which does not contain every index constituent. The CRSP share-code restriction to codes 10 and 11 excludes real estate investment trusts (code 18) and some firms incorporated outside the United States (code 12), both of which file DEF 14A and are otherwise ordinary domestic filers. At mid-2024, 443 of the 503 S&P 500 constituents had a resulting observations, and of those, 431 (97 percent) have a recovered 2024 proxy, so the shortfall is a function of the sample rather than the extraction process. The gap widens over the sample as both categories increase, from roughly 5 percent of S&P 500 constituents in 2000 to 12 percent in 2024. Nothing in the pipeline prevents parsing these filings, and extending the frame to additional share codes does not require a different method.

Table 1: Panel coverage relative to the CRSP/Compustat frame, by era

| | Frame | Filings | Firms | Coverage | Tier composition (%) | | |
|---|---|---|---|---|---|---|---|
| Filing years | firm-years | recovered | | (%) | ok | equal-votes | flagged |
| 1997–2002 | 37,420 | 37,409 | 8,975 | 85.6 | 95.5 | 1.3 | 3.1 |
| 2003–2010 | 35,766 | 39,932 | 7,055 | 91.9 | 96.8 | 1.1 | 2.0 |
| 2011–2018 | 29,167 | 31,424 | 5,518 | 94.3 | 97.4 | 1.3 | 1.2 |
| 2019–2026 | 23,464 | 31,549 | 5,273 | 92.5 | 95.4 | 3.7 | 0.9 |
| All | 125,817 | 140,314 | 13,790 | 90.7 | 96.3 | 1.8 | 1.9 |

*Notes:* Frame firm-years are the CRSP/Compustat-merged universe of U.S. common stocks (share codes 10 and 11, no exchange filter) linked to SEC filer identifiers as described in Section 2.2.3; the frame ends in fiscal 2024, and filings extend through the 2026 proxy season, so the final row's frame count and coverage rate cover 2019–2024 only. Filings recovered counts definitive proxy statements (DEF 14A and DEF 14C) with an extracted row in the panel, including filings by firms in years outside their CRSP tenure. Firms is the number of distinct SEC filer identifiers with at least one filing in the era. Coverage is the share of frame firm-years with at least one recovered filing. Tier composition is the share of filings in the era by quality tier; flagged combines the `partial`, `broken`, and `suspect` tiers, and the 33 government-sponsored enterprise filings account for the remainder.

the director-election disclosure from the proxy statement reports that directors serve multi-year terms that expire in different years, regardless of whether the filing uses the words "classified" or "staggered," and as unitary when every director stands for election annually. The dataset also stores a verbatim supporting quote from the filing so that researchers can audit this determination. A board classification is available for 128,569 of the 128,574 firm-years, 47.8 percent of which are classified.

**Dual-class status.** We distinguish in the data between *structural* and *canonical* dual-class structures. A firm-year is structurally dual-class when the proxy reports two or more classes of common stock outstanding, regardless of voting rights, which covers 9,765 firm-years from 1,636 firms. The canonical measure requires that the classes carry differential voting rights, and excludes three structures where multiple classes do not represent a control choice by the firm's insiders. These include special-purpose acquisition companies (SPACs), whose founder shares are temporary and automatically convert to ordinary common stock once the SPAC completes a merger; tracking stocks, whose classes track separate businesses and carry no control wedge over a common enterprise; and GSEs. Canonical dual-class structures cover 9,006 firm-years from 1,336 firms. Preferred stock does not count for either measure, and a firm whose only second class is a series of preferred is measured as single-class.

**Insider voting power, economic ownership, and the wedge.** The primary insider ownership measures come directly from the directors-and-officers group row of the beneficial-ownership table. Voting power is the group's share of total votes outstanding, taking the filing's printed total-voting-power figure where one is printed, and otherwise recomputing it from per-class shares and votes per share. Economic ownership is the group's beneficially owned shares as a percentage of shares outstanding, using the SEC's reporting convention in which shares acquirable within sixty days enter both numerator and denominator. The wedge is calculated as voting power minus

economic ownership, in percentage points. Voting power is available for 120,012 analysis firm-years and economic ownership for 120,511. In a single-class, one-vote-per-share firm the two coincide and the wedge is zero by construction.

**Variants, and how much they matter.** Four alternative constructions are provided alongside the headline measures, which a researcher might be interested in.

*Direct versus beneficial economic ownership.* Following Gompers et al. (2010), the literature generally measures insider cash-flow rights from shares actually held, excluding options that are exercisable within sixty days. We provide this direct measure wherever it can be constructed. For example, where the filing discloses the group's option count in a footnote we subtract it (53,897 firm-years), and where it does not, we back the count out from the printed percentage and share count (52,729), reporting which method was used. Across firm-years with both measures the median gap between beneficial and direct economic ownership is about one percentage point, but it is largest where option-heavy compensation makes the beneficial figure a poor proxy for wealth at stake. The right panel of Figure 30 plots the distribution by method.

*Umbrella-partnership basis.* In Up-C structures, the firm's Class A shares are only a fraction of the economic enterprise, with insiders holding the remainder in operating-partnership units exchangeable into Class A. Economic ownership on the Class A basis alone overstates insider ownership, so the canonical ownership measure is computed on a fully exchanged basis, with the raw Class A figure stored separately. Among dual-class firm-years, roughly 4,500 are flagged as basis-divergent in this manner, and the raw figure understates the canonical one by a median of about nine and a half percentage points (Figure 30, left panel). The divergence, in part, is what allows us to identify an Up-C structure.

*The family bloc.* The group row from the beneficial ownership table only counts directors and officers, and omits family members, trusts, foundations, and family holding entities that are aligned with management but are not themselves directors or officers. In family-controlled firms, this reporting practice can substantially understate the insider bloc. We therefore provide alternative measures that add non-insider holders tied to the D&O group by surname or by the company's own name. We also develop an aggregation rule for holders reporting shared voting power over the same shares, and a check that a family entity is not already subsumed in any individual's row. Among dual-class firm-years, the family bloc adds more than five percentage points of voting power in 1,211 of 8,630 cases and more than five points of economic ownership in 1,600 of 8,845 (Figure 29).

*Controllers.* Where the largest single holder in the ownership table is a parent company, founder's trust, or former executive that commands more voting power than the directors-and-officers group, ownership row will understate who controls the firm. We report the identity and voting power of the top controller as separate columns, and in 2,547 of 8,581 dual-class firm-years the top controller's voting power exceeds the group's, and in 856 the controller holds a majority while the group row is below fifty percent (Figure 28).

**Quality tier and provenance.** Each row carries the quality tier from Section 2.2.6, the maximum discrepancy between the printed and recomputed voting-power figures for any holder in the table, and a set of flags recording which derivation repairs, overrides, or conventions factored into the row's values, so that any number can be backed out directly from the transcribed primitives.

## 3.3 Agreement with Outstanding Datasets

External datasets provide a complementary check on the panel, even though, as mentioned in Section 2.1, commercial governance datasets frequently disagree with each other and none can be considered ground truth. However, external comparison allows us to validate our output on a sample of tens of thousands of firm-years. Table 3 summarizes the comparisons.

**ISS and its predecessors.** The longest-running and most commonly used external series is the ISS governance data on WRDS, comprising the legacy IRRC/RiskMetrics volumes through 2006 and the ISS series from 2007 onward. Matching to the firm's historical CUSIP, we compare 3,936 firms over roughly 39,900 firm-years for both dual-class status and board classification. Agreement on dual-class status is 98.4 percent year-by-year and 98.8 percent when the external value in an adjacent year is allowed to match, a tolerance that accommodates the irregular update cycle of the legacy volumes. Agreement on board classification is about 97 percent year-by-year and about 99 percent with the same one-year tolerance.

**The IRRC dual-class list.** The replication file for Bebchuk et al. (2009) separately lists the firms that IRRC classified as dual-class in each volume from 1990 to 2006, which are excluded from the entrenchment-index sample. Our canonical dual-class measure agrees with this list in 97.0 percent of firm-years, and in 97.6 percent with a one-year tolerance.

**The CII dual-class list.** The Council of Institutional Investors (CII) maintains a public list of U.S. dual-class companies. The list deliberately excludes foreign private issuers, SPACs, firms below $200 million in market capitalization, and multi-class structures without a voting wedge. Of 225 CII-listed firms we can match to the panel, 223 are dual-class in our data, a recall of 99.1 percent; the two differences are a firm whose second class was outstanding only in years outside CII's snapshot and a firm whose high-vote class is held by a parent and appears in the filing only as a footnoted parent-held class.

**Disagreements.** The ISS comparison registers a disagreement in 3.0 percent of firm-years in our sample through 2006 for dual-class, and in 0.6 percent thereafter. The pre-2007 disagreements arise in one frequent setting: the legacy IRRC volumes code as dual-class a set of firms—Sturm Ruger, Cardinal Health, Ethan Allen, and Old Republic among others—whose proxies over the same years report a single class of common stock outstanding with one vote per share, and whose commercial coding reverts to single class as soon as the ISS series begins in 2007. This suggests that the pre-2007 results are likely erroneous. With board classification, the disagreements are closer to symmetric and concentrated in a small number of firms: 683 firms disagree in at least one

Table 2: Principal fields of the released panel

| Field | Definition | Origin |
|---|---|---|
| *Identifiers* | | |
| `cik`, `accession`, `form`, `filing_date`, `filing_year` | SEC filer identifier, filing accession number, form type (DEF 14A / DEF 14C), and filing date and year | EDGAR |
| *Share structure* | | |
| `classes`, `n_classes` | Each common class with label, votes per share, and shares outstanding; count of classes | Transcribed |
| `is_dual_class` | Structural: two or more common classes outstanding | Derived |
| `dual_class_ex_spac` | Canonical: structural with differential votes, excluding SPACs, tracking stocks, and GSEs | Derived |
| `is_spac`, `tracking_stock` | Structural exclusion flags | Derived (text scan) |
| *Board* | | |
| `board_classified`, `board_note` | Classified (staggered) board indicator with verbatim supporting quotation | Transcribed |
| *Insider ownership* | | |
| `insider_group_vote_pct` | Directors-and-officers group share of total votes (printed where available, else recomputed) | Derived |
| `insider_group_econ_pct` | Group share of shares outstanding, SEC beneficial-ownership convention; Up-C on fully exchanged basis | Derived |
| `insider_group_wedge` | Voting power minus economic ownership, percentage points | Derived |
| `insider_group_econ_pct_direct`, `..._direct_method` | Economic ownership excluding sixty-day options; method (options disclosed / backed out) | Derived |
| `insider_group_econ_pct_raw`, `econ_basis_divergent` | Class-A-basis economic ownership and Up-C divergence flag | Derived |
| `insider_group_vote_pct_incl_family`, `..._econ_pct_incl_family` | Family-inclusive variants of the insider measures | Derived |
| `top_controller`, `top_controller_vote_pct` | Largest single holder by voting power and its share | Derived |
| *Structural flags* | | |
| `has_class_board_election`, `has_voting_agreement`, `has_reduced_cashflow_class`, `is_up_c_signal`, `controlled_company` | Class-specific director-election rights; voting agreements or irrevocable proxies; a class with no or reduced dividend rights; umbrella-partnership language; "controlled company" listing-standard disclosure | Derived (text scan) |
| *Quality and provenance* | | |
| `quality` | Tier: ok, equal-votes, partial, broken, suspect, gse | Derived |
| `max_validation_delta_pp` | Largest gap between printed and recomputed voting power across holders | Derived |
| `board_text_repair`, `group_row_text_repair`, `denominators_backsolved`, `votes_per_share_backsolved`, `aggregate_vote_repair`, `audit_value_override`, `dual_class_source` | Indicators for each derive-time repair, override, or convention that changed the row's values | Derived |

*Notes:* "Transcribed" fields are returned by the language model as printed in the filing; "Derived" fields are computed by deterministic code from transcribed primitives, cached filing text, or both. The full field list, including per-owner rows and per-class positions in the underlying store, is documented in the data release. Field definitions follow the conventions in Appendix A.6.

year, but most in only one or two, consistent with transition timing. Importantly, unlike black-box commercial ratings, our audit trail allows each disagreement to be identified, manually inspected, and resolved by reading the filing.

Table 3: Agreement with external governance datasets

| | | | | | Agreement (%) | |
|---|---|---|---|---|---|---|
| Source | Measure | Years | Firms | Firm-years | Strict | Lenient |
| ISS/RiskMetrics/IRRC | Dual-class | 1997–2024 | 3,936 | 39,927 | 98.4 | 98.8 |
| ISS/RiskMetrics/IRRC | Classified board | 1997–2024 | 3,936 | 39,924 | 96.7 | 98.9 |
| IRRC list (BCF file) | Dual-class | 1998–2006 | 2,478 | 7,648 | 97.0 | 97.6 |
| CII list (2019) | Dual-class recall | 2017–2020 | 225 | — | 99.1 | — |

*Notes:* Strict agreement is year-by-year, while lenient allows the external value in an adjacent year to match, accommodating the external series' update lag. ISS/RiskMetrics/IRRC is the ISS governance data on WRDS, combining the legacy governance volumes through 2006 with the ISS series from 2007, matched on historical CUSIP. The IRRC list is the set of firms excluded as dual-class in the replication file of Bebchuk et al. (2009) (BCF). The CII comparison treats dual-class status as persistent within 2017–2020 and reports recall because CII's list excludes foreign private issuers, SPACs, firms below $200 million in market capitalization, and no-wedge structures by construction. Board-classification agreement is computed against the panel before the board repairs of Table 5; on the remediated panel, strict agreement is 97.1 percent.

# 4 Inside Ownership and Control in the Aggregate

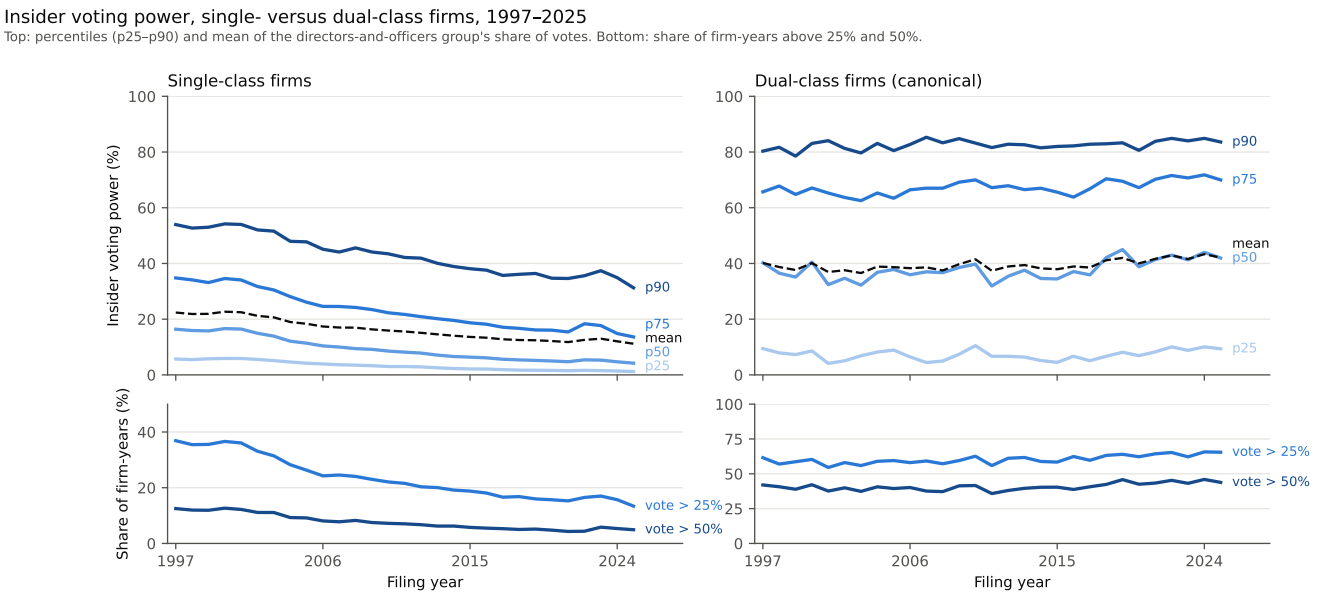
The final panel dataset allows us to describe insider control across the population of U.S. reporting companies. Insider control of public firms is common and much of it resides in single-class firms. Consistent with prior work, dual-class structures have become more prevalent, driven both by new listings and sustained by a low rate of unwinding. The gap between control and cash-flow rights has widened, and it is produced by a varied set of mechanisms, not only through a high votes-per-share ratio. Finally, most of this activity occurs outside of the S&P 1500 which has not been historically covered in the literature.

## 4.1 The Level of Insider Control

Figure 10 plots the insider voting share for the directors-and-officers group, as defined in Section 3.2 by filing year, separately for single-class and canonical dual-class firms. Among single-class firms the distribution has trended steadily downward for three decades, with median insider voting power falling from 16.4 percent in 1997 to 4.2 percent in 2025, the 90th percentile from 54.0 to 31.3 percent, and the share of firm-years in which insiders control more than a quarter of the votes from 37 to 13 percent. The mean insider percentage falls above the median throughout, from 22.4 against 16.4 percent in 1997, and 11.2 against 4.2 percent in 2025, given the distributions long

right tail. In every year of the sample, some single-class firms are majority-controlled by their own officers and directors, although that share has fallen from 13 to 5 percent over the panel.

Figure 10: Insider voting power, single- versus dual-class firms

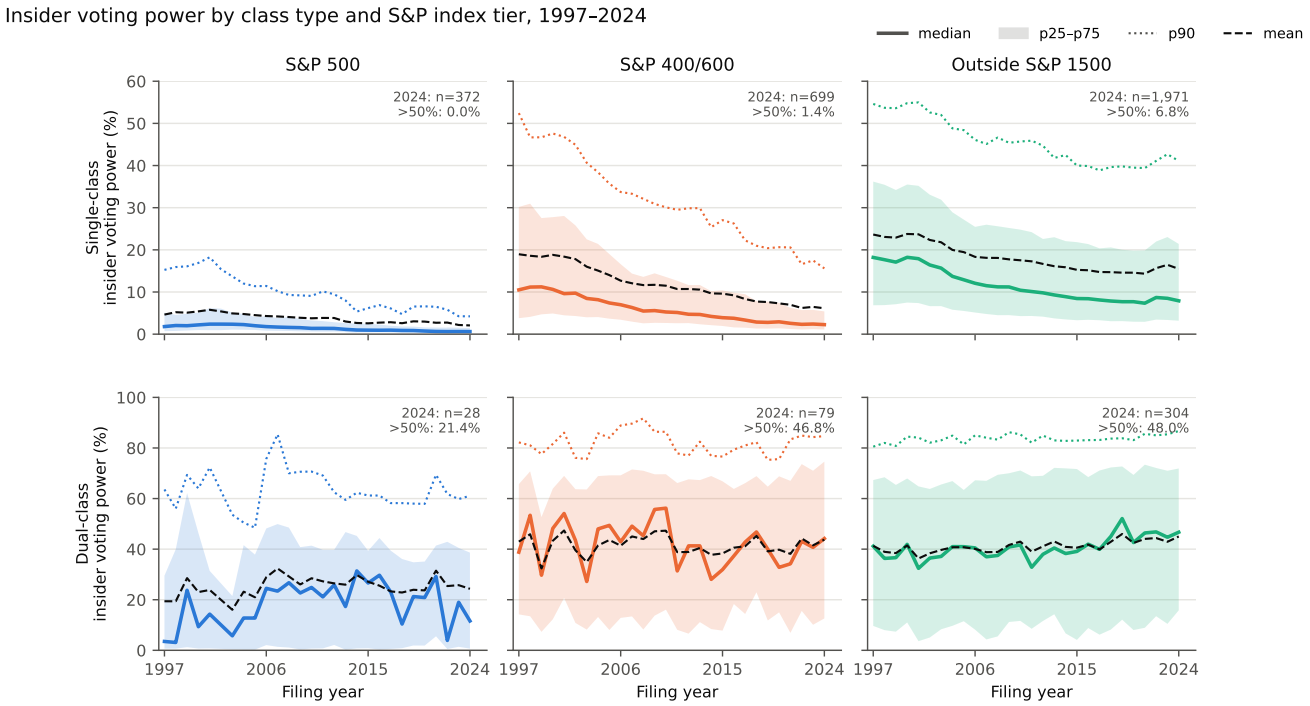



*Notes:* The top panels plot the 25th, 50th, 75th, and 90th percentiles (solid lines, light to dark) and the mean (dashed) of insider group voting power by filing year, separately for single-class firms (left) and canonical dual-class firms (right). The bottom panels plot the share of firm–years in which the insider group holds more than 25 percent and more than 50 percent of votes; note the different vertical scales across the two bottom panels. Over the full sample, 73 percent of firm–years with insider voting power above 50 percent, and 84 percent of those above 25 percent, are single-class firms.

Among dual-class firms the time trend is flat rather than declining. Median insider voting power was 40.2 percent in 1997 and 41.9 percent in 2025, the 90th percentile falls above 80 percent in every year of the sample, and the share of firm-years with majority insider control increased slightly from 42 to 44 percent. Whatever has eroded insider ownership in the single-class population has not similarly affected the dual-class population. Because single-class firms outnumber dual-class firms by nearly thirteen to one in our sample—111,345 against 8,624 firm-years–the population of closely-held public companies is actually overwhelmingly single-class. Of the 12,663 firm-years in which the officer-and-director group holds outright majority voting power, 72 percent are single-class firms, and of the 32,794 firm-years in which the group holds more than a quarter of the votes, 84 percent are. This suggests that for the cross-section of public reporting firms in the United States, insider control is actually an ownership rather than charter phenomenon.

Figure 11 plots these same time series distributions separately by S&P index tier. In 2024 the median S&P 500 single-class firm reported insider voting power of only 0.6 percent, and none of those firm-years were majority-controlled. In the S&P 400 and 600 the median was 2.3 percent, with 1.4 percent of firm-years majority-controlled. Outside the S&P 1500 the median was 7.9 percent, the 90th percentile 41.1 percent, and 6.8 percent of the firm-years were majority-controlled.

Figure 11: Insider voting power by class type and S&P index tier



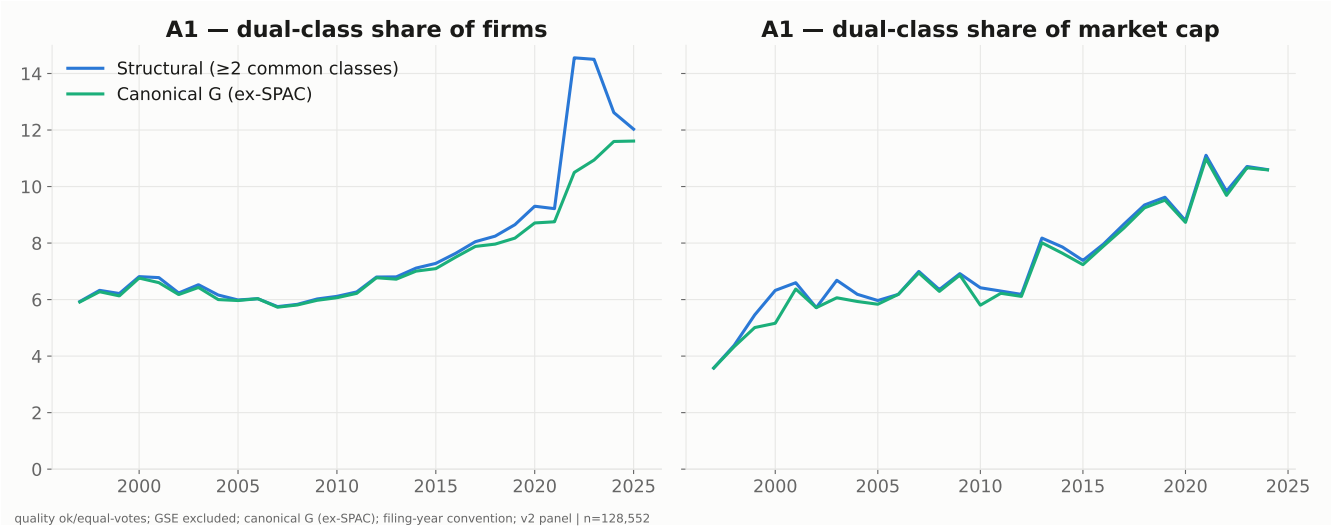

*Notes:* Each panel plots, by filing year, the median (solid), the interquartile range (shaded), the 90th percentile (dotted), and the mean (dashed) of insider group voting power for one class type (the rows are single-class and canonical dual-class) and one S&P index tier (the columns are the S&P 500, S&P 400/600, and outside the S&P 1500). In-panel annotations report the 2024 cell size and the share of 2024 firm–years with insider voting power above 50 percent. Tier assignment combines CRSP daily S&P 500 membership with Execucomp S&P-index codes at fiscal year $t-1$ and the series end in 2024, the last year with full tier coverage.

The differences in inside ownership by firm size is almost entirely a single-class firm feature. Among dual-class firms the ownership shares across firm size are similar and consistent over time, with 46.8 percent of S&P 400/600 firm-years and 48.0 percent of firm-years outside the index being majority-controlled in 2024, against 21.4 percent within the S&P 500. Where the charter separates votes from cash-flow rights (through a dual-class structure), control looks much the same wherever the firm trades. Where it does not, control dynamics look very different by market capitalization, and a literature assembled from index constituents therefore only observes the part of the distribution in which insider control is nearly absent.

## 4.2 Prevalence Over Time

Figure 12 reports the share of firms with a canonical dual-class structure by filing year, as a share of firms (left panel) and market capitalization (right panel). The firm-share series is flat at roughly six percent from the beginning of the panel through the financial crisis (5.9 percent in 1997 and 5.7 percent in 2007) before rising rapidly, reaching 7.9 percent in 2017 and 11.6 percent in 2025. The value-weighted series also increases this period, from 3.6 to 10.6 percent, so the trend does not only reflect small firms adopting this structure, consistent with large technology firms increasingly adopting dual-class structures.

Figure 12: Dual-class prevalence, 1997–2025

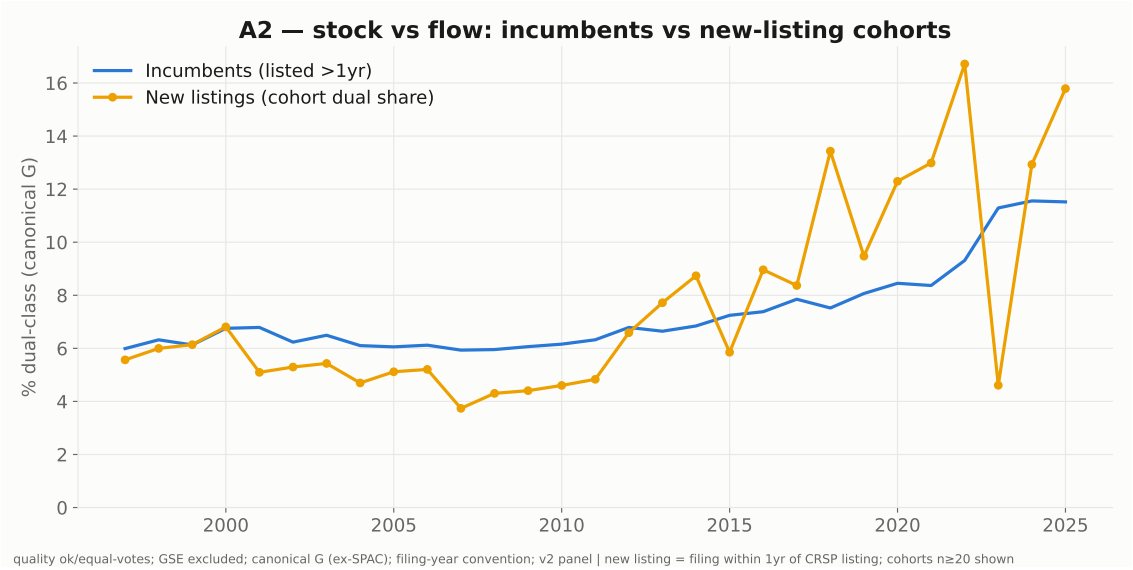



*Notes:* This figure plots the share of firms (left panel) and the share of aggregate CRSP market capitalization (right panel) with a dual-class structure, by filing year. Each panel shows two definitions: structural (two or more common classes outstanding) and canonical (structural, excluding SPACs and tracking stocks). The value-weighted panel omits years in which CRSP market capitalization covers fewer than half of filers.

Recent work on this increase in entrenchment structures measures the flow rather than the stock of such firms, documenting the rising share of dual-class IPOs (Aggarwal et al., 2022). The panel measures that do exist are either older and narrower—Gompers et al. (2010) hand-code filings for the 1995–2002 period, and IRRC-based measures cover roughly the S&P 1500—or they infer share

structure from CRSP, which contains only traded securities and cannot identify dual-class firms where one of the classes does not trade. Given that most founder and family shares don't trade publicly, recovering the remainder requires either hand collection (Kim and Michaely, 2019) or (as here) parsing the disclosure. Figure 13 shows how we can use our data to separate the two margins that affect the stock of dual-class firms. Incumbent dual-class firms, those more than a year past their listing date, increase from 6.0 to 11.5 percent of the sample. Newly listed cohorts increase more, from five percent through 2012 to between 12 and 17 percent in most years after 2017. The rise is substantially a cohort phenomenon, consistent with the founder-led technology listings that motivated the policy debate described in Section 2.2.2.

Figure 13: Stock versus flow: incumbents and new-listing cohorts

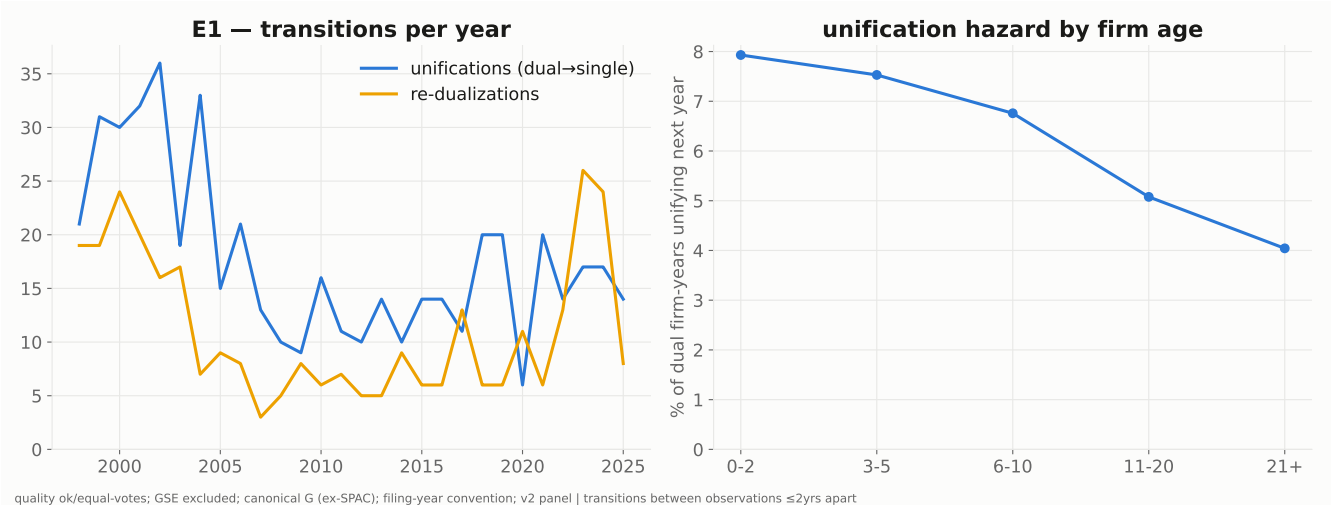



*Notes:* This figure plots the canonical dual-class share separately for incumbents (firm–years more than one year past the firm's CRSP listing date) and for new-listing cohorts (firm–years within one year of listing).

Figure 14 documents transitions in and out of dual class (left) and the annual rate at which dual-class firms unify, by firm age (right). Over the full panel we observe 498 unifications against 312 re-dualizations, and the unification rate falls monotonically with age. Within two years of listing 7.9 percent of firms unify per year, which drops to 7.5 and 6.8 percent over the next eight years, falling to 4.0 percent for firms more than twenty years past listing. Dual class structures that persist past a firm's initial period as a public company will likely remain permanent, with later cohorts of dual-class listings accumulating rather than turning over. The firms that ultimately unify are disproportionately those whose structures had already ceased binding. For firms observed in their final dual-class year, 66.9 percent carried an insider wedge within 2.5 percentage points of zero, compared to 38.4 percent of all dual-class firm-years (Figure 31). Unification functions less as a total surrender of control than than an attempt to simplify the capital table.

The dual-class population has also diverged from single class along another governance dimension. Figure 15 shows the share of firms with a classified board separately for dual-class and

Figure 14: Unifications and re-dualizations




*Notes:* The left panel reports, by year, transitions out of (unifications) and into it (re-dualizations) dual-class status between a firm's consecutive proxy observations at most two years apart. The right panel plots the annual unification hazard rate, or the share of dual-class firm–years that are single-class at the next observation, by time since CRSP listing.

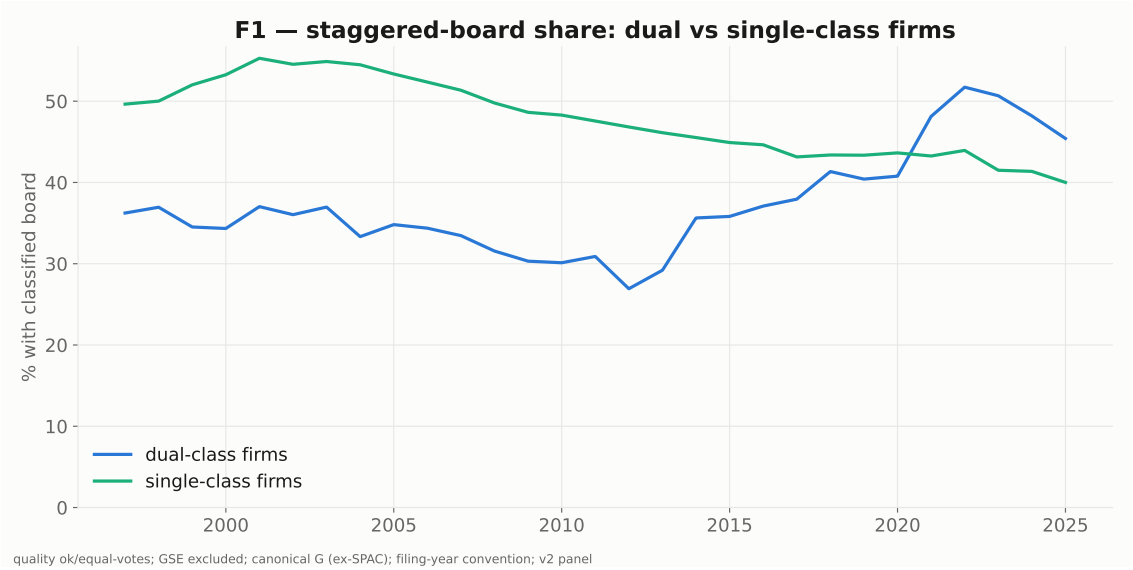
single-class firms. In the late 1990s single-class firms were the more likely to stagger their boards, but by the end of our sample period this pattern reversed. The declassification wave that the governance literature documents, which has been driven by shareholder proposals, proxy-advisor voting policies, and index-fund guidelines, reduced the presence of staggered boards at single-class firm but did nothing for the dual-class population untouched. If anything their rate of staggering has increased over time. This suggests that staggered boards and dual-class charters may be complements, patched together by insiders who want more than one defense.

Figure 16 reveals that dual-class prevalence is also non-monotone in firm size. In the smallest two market-capitalization deciles 4.7 percent of firms have multiple share classes, with roughly 9 percent having two or more classes in deciles six through eight, and 7 percent in the largest decile. Dual class shares are concentrated in telecommunications (35.8 percent) and non-durable consumer goods (19.8 percent), is comparatively rare in health care (2.9 percent) and utilities (2.2 percent), and is U-shaped in age. Both young and old firms, more than twenty years old, are more likely to be dual-class than those in the middle.

## 4.3 The Wedge

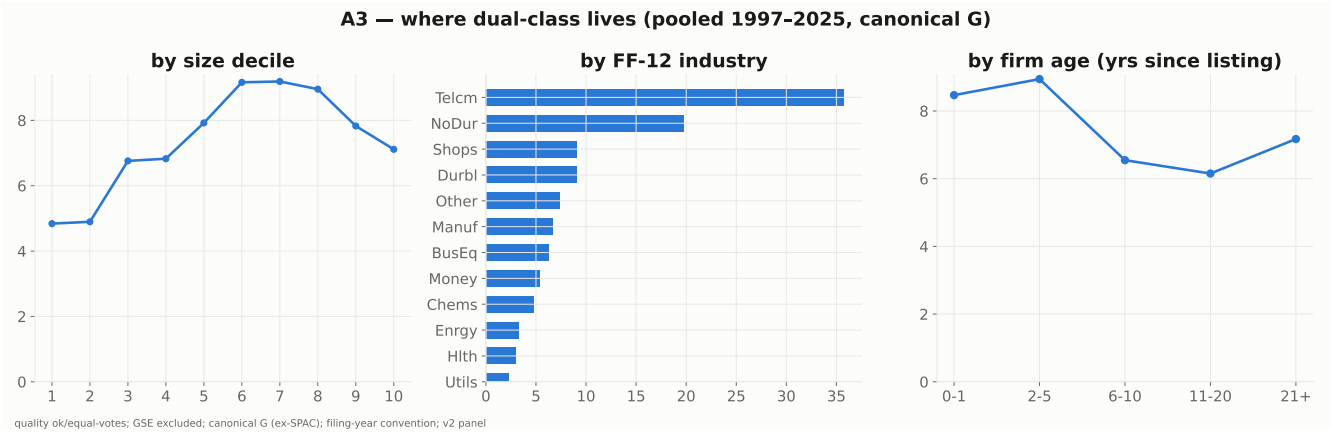
The typical dual-class structure has also become stronger over time. The median insider wedge, calculated as voting power minus economic ownership, among dual-class firm-years increased from 4.1 percentage points in 1997 to 16.6 in 2025, and the 90th percentile rose from 37.3 to 48.4 percentage points, as shown in Figure 17.

Figure 15: Governance bundling: staggered boards and dual-class status



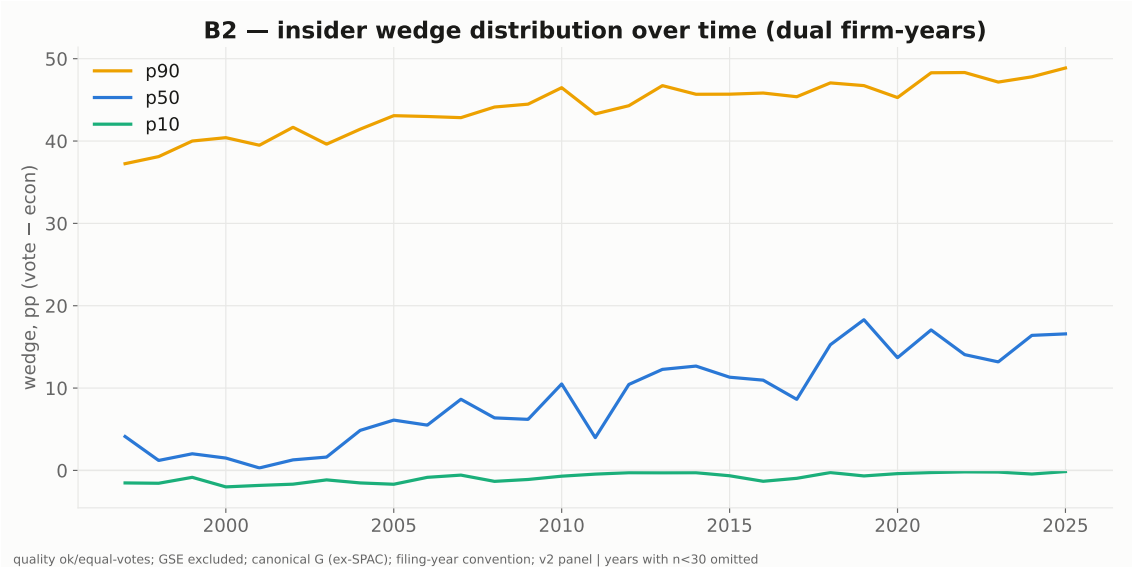

*Notes:* This figure plots the share of firms with a classified (staggered) board, separately for dual-class and single-class firms, by filing year. Board classification is extracted from the proxy's director-election section; firm–years without a board call are excluded, and years with fewer than 30 observations in a group are omitted.

Figure 16: Where dual-class firms are: size, industry, and age




*Notes:* This figure plots the canonical dual-class share, pooled over 1997–2025, by within-year CRSP market-capitalization decile (left panel), Fama–French 12 industry assigned from the filing's SIC code (middle panel), and firm age in years since CRSP listing (right panel).

Figure 17: The wedge distribution over time




*Notes:* This figure plots the 10th, 50th, and 90th percentiles of the insider wedge (voting minus economic ownership, percentage points) among dual-class firm–years, by filing year. Years with fewer than 30 dual-class observations are omitted.

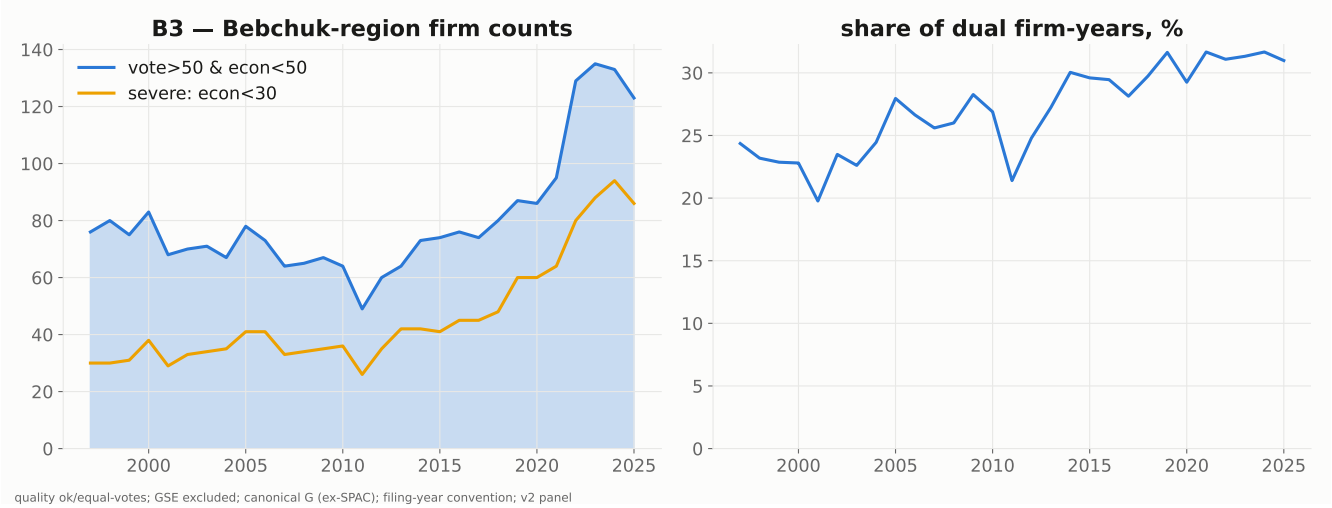
Figure 18 reveals the frequency of firm-year observations where insiders hold majority voting power while bearing a minority of the cash-flow consequences of their decisions.[12] In 1997, 76 dual-class firms had such an ownership structure, 30 of them severely, defined as having an insider economic ownership below 30 percent. By 2024 the corresponding numbers were 133 and 94, and the share of dual-class firms of both types had risen from 24.4 to 31.7 percent. Thus, the most extreme form of managerial entrenchment from multi-class share structures were the fastest growing part of the dual-class market.

Figure 19 breaks down the relative proportion of entrenchment mechanisms that are identified in our data over time. Supervoting structures with a ratio of ten votes to one (or fewer) are the most common type, covering 5,169 firm-years across 716 firms, followed by voting agreements and trusts (1,334 firm-years, 408 firms), class-specific board-election rights (1,069; 152), non-voting classes (1,066; 238), supervoting ratios above ten to one (879; 154), majority corporate holders (718; 213), and umbrella-partnership structures (608 firm-years combined). These categories often overlap and several would not be apparent from only a per-share vote calculation. For example, the votes-per-share ratio can be one to one with control still being complete where one class elects a fixed fraction of the board, or where an irrevocable proxy assigns votes that are not attributable to the beneficial ownership table in the proxy statement.

Moreover, the vote differentials within the dual-class structures have also increased over the sample period. Ratios above ten to one grew from 8.0 percent of dual-class firm-years in 1997 to

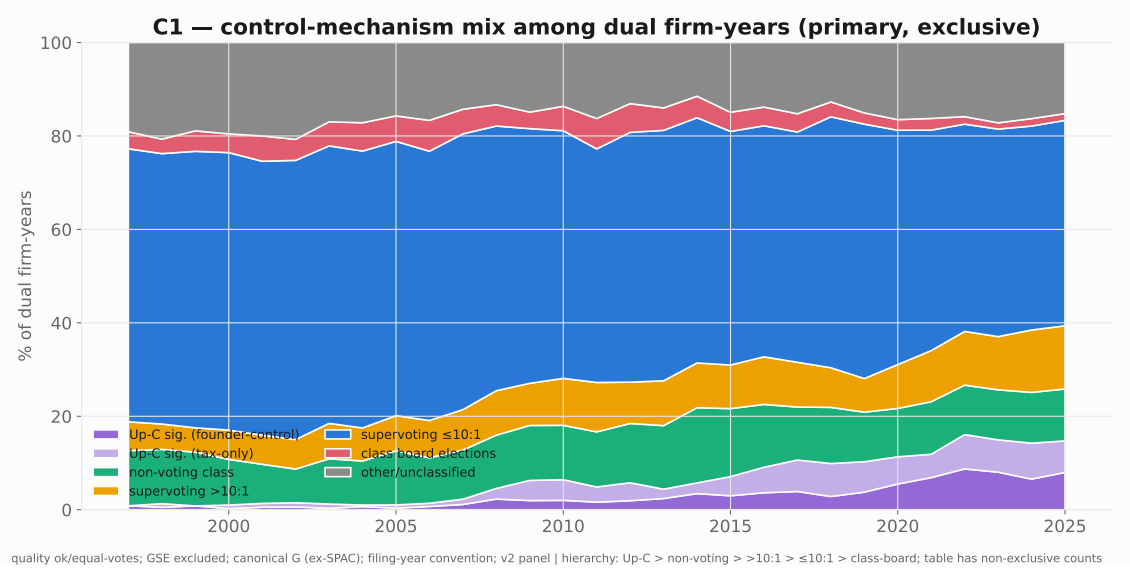
[12]The region is the one emphasized by Bebchuk, Kraakman, and Triantis (2000), whose analysis of controlling-minority structures treats the separation of votes from cash-flow rights, rather than the concentration of votes in aggregate, as the measure of agency cost.

Figure 18: Minority-exposure region counts




*Notes:* The left panel counts dual-class firm–years in which the insider group holds majority voting power with a minority economic stake (vote > 50% and economic ownership < 50%), the region emphasized by Bebchuk, Kraakman, and Triantis (2000), and the subset with economic ownership below 30%. The right panel expresses the region as a share of dual-class firm–years with both measures available.

Figure 19: Control-mechanism mix among dual-class firm-years




*Notes:* This figure assigns each dual-class firm–year a single primary control mechanism using the exclusive hierarchy Up-C structure > non-voting class > voting ratio above 10:1 > ratio at most 10:1 > class-elected board (a class right to elect a specified fraction of directors), and plots the mix by filing year. A firm–year matching several mechanisms is counted once, at the highest-ranked match; the counts reported in the text are non-exclusive and therefore overlap.

18.8 percent in 2025, while structures containing a class of shares without any voting rights at all have held steady at roughly 15 percent of the total. Finally, the level of insider voting power changes at relevant thresholds, suggesting that it is almost surely endogenously chosen over time rather than an inherited stock. Figure 20 shows the distribution of the voting power over the 30-to-90-percent ownership range. Insiders whose holdings put them near the majority line are visibly more likely to end up over these thresholds than below them.

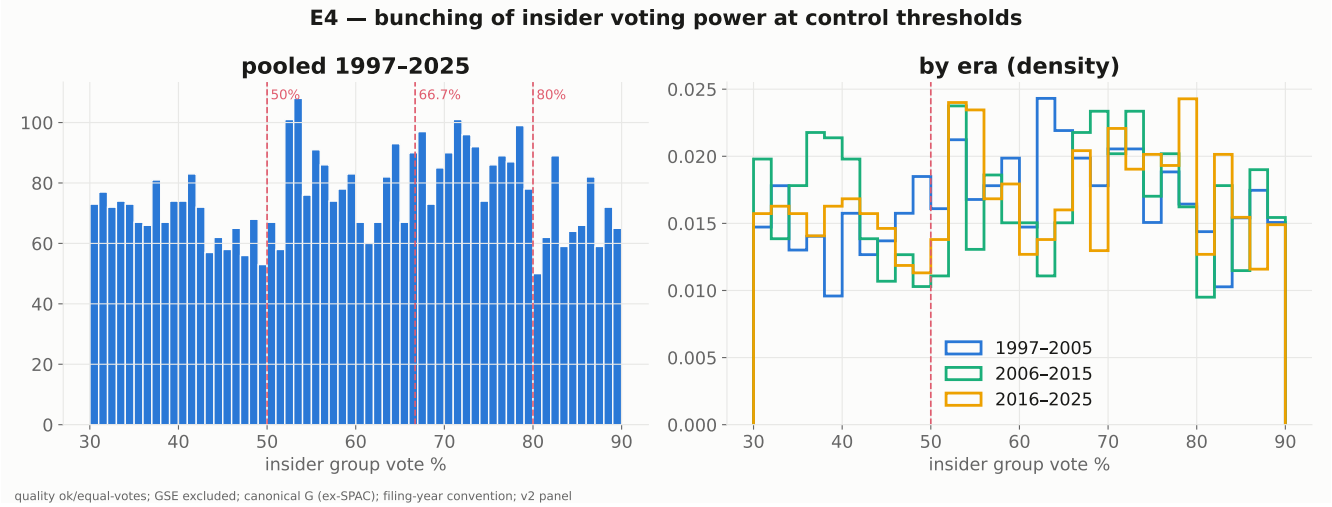
## 4.4 Where Control Hides

Managerial control can be concealed within the reporting process, falling both outside the row of the ownership table that can be easily accessed or parsed in the proxy statement, and outside the sample of firms that commercial datasets cover. The directors-and-officers group row by definition omits holders who are not themselves current officers or directors at the firm. In 2,547 of 8,581 dual-class firm-years with both measures available, the largest single holder in the ownership table commands more voting power than the entire group row, and in 876 of those firm-years the controlling shareholder has an outright majority while the director and officers group falls below a majority threshold. These are usually firms where the controller is a parent company, a founder's trust, or a former executive, which still generate true economic control even if not required to be reported within the table under SEC rules. Family entities aligned with management raise the group's voting power by more than five percentage points in 1,211 of 8,630 firm-years, and its economic ownership by more than five points in 1,600 of 8,845. In roughly a third of dual-class firm-years the conventional measure from the table would understate true insider control.

This rise in dual-class structures has also differed by population size which correlates with coverage by governance databases. Figure 21 plots dual-class prevalence within each index tier over time. Inside the index the series are close to flat, with the S&P 500 seeing a rise in the canonical dual-class share from 5.6 percent in 1997 to 6.5 percent in 2024, and among S&P 400 and 600 firms from 8.3 to 10.2 percent. Outside the S&P 1500 it has more than doubled, from 6.0 to 13.3 percent, and on a value-weighted basis it rose from 6.2 to 35.3 percent. In other words, more than a third of the market capitalization outside the S&P 1500 now consists of dual-class firms. Thus, the rise in the aggregate share of dual-class firms documented in Figure 12 is largely an out-of-index phenomenon that has occurred in the population that the standard governance datasets do not sample.

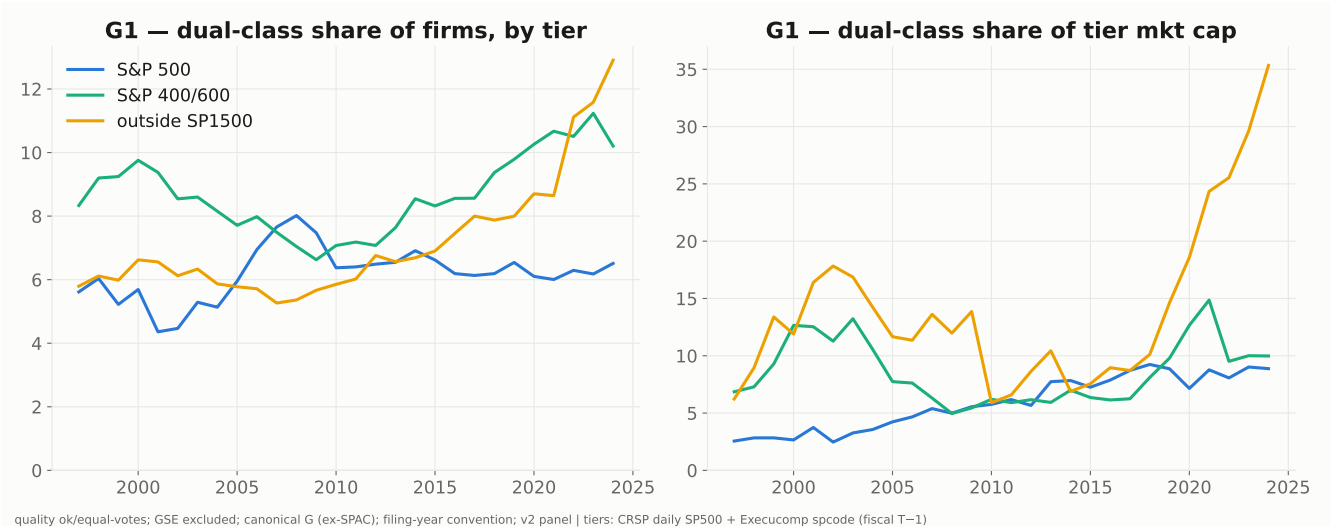
Figure 22 breaks down this share of dual-class activity sitting outside the S&P 1500, and therefore absent from ISS, IRRC, and Execucomp-based datasets. Roughly three quarters of all dual-class firm-years are outside of the index in every year since 2007—86.8 percent in 1997, 73.3 percent in 2007, and 74.4 percent in 2024—with the majority-vote, minority-economics firm-years from Figure 18 covering roughly the same share, 72.9 percent in 2024. Given that these are, by nature, smaller firms, the outside share is only 20.6 percent when weighted by market capitalization,

Figure 20: Bunching of insider voting power at control thresholds




*Notes:* The left panel plots the histogram of insider group voting power among dual-class firm–years over the 30–90% window, with reference lines at the 50% control, 66.7% supermajority, and 80% consolidation thresholds. The right panel plots the same distribution as densities by era.

Figure 21: Dual-class prevalence by index tier

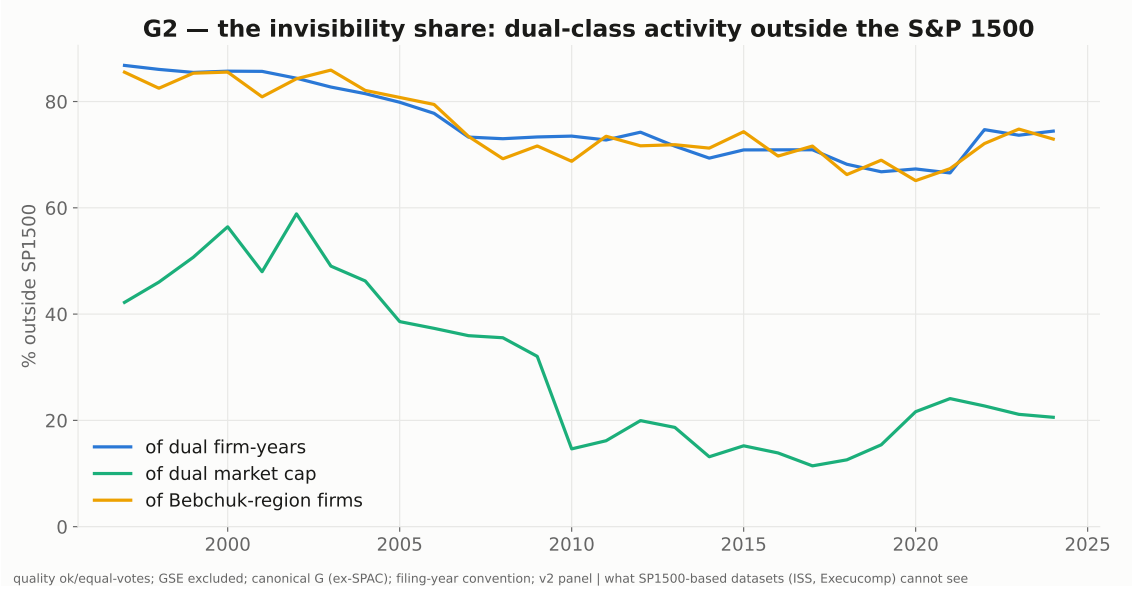



*Notes:* This figure plots the canonical dual-class share within each index tier — S&P 500, S&P 400/600, and outside the S&P 1500 — equal-weighted (left panel) and value-weighted within tier (right panel). Tier assignment combines CRSP daily S&P 500 membership with Execucomp S&P-index codes at fiscal year $t-1$; the series end in the last year with full tier coverage (2024), and cells with fewer than 30 observations are omitted.

down from 42.2 percent in 1997. This fact is consistent with the shrinking number of firms going and staying public over time with the increase of private capital. In recent years, the firms missing from the commercial governance data sample are still a large share of all publicly traded firms, but represent a smaller share of total market capitalization.

Whether these missing firms matter depends on the question the data are used to answer. For an exercise asking about total economy effects, or anything inherently weighted by market size, the omitted population may not necessarily be material. For questions about the overall incidence of governance arrangements, how controlled firms operate, or the relationship between control and corporate decisions, the omitted population is most of the sample of publicly available firms. Moreover, the omitted firms are likely correlated with governance differences, with both the ownership prevalence and wedges being distributed differently across tiers.

Figure 22: Dual-class activity outside the S&P 1500




*Notes:* This figure plots the share of dual-class activity that sits outside the S&P 1500 — and is therefore invisible to S&P-1500-based governance datasets such as ISS and Execucomp — by filing year, on three definitions: the share of dual-class firm–years, the share of dual-class market capitalization, and the share of majority-vote/minority-economics firm–years (vote $> 50\%$, economic ownership $< 50\%$). Tier assignment combines CRSP daily S&P 500 membership with Execucomp S&P-index codes at fiscal year $t-1$, as in Figure 21.

Overall, we find in our data a population of controlled public companies that is bigger and more varied in entrenchment mechanisms than the record built on index constituent alone would suggest. Our increased sample size also allows us to analyze more closely the set of firms where managers, directors, and controllers hold voting power and cash-flow rights in different proportions.

# 5 Application: Dividend Taxes, Incentives, and Control

In this section, we show how our extended governance measures can be used to help adjudicate outstanding research questions. The Jobs and Growth Tax Relief Reconciliation Act of 2003 lowered the top federal tax rate on dividend income from 38.6 to 15 percent. The following year saw a large increase in firms initiating new dividends, and Chetty and Saez (2005) showed that this decision was driven by firms where the executives held a large inside stake in the company. This was argued to be evidence that corporate payout policy, a core area of interest in corporate finance, was driven by the personal tax considerations of managers and directors. Chetty and Saez (2010) then developed an agency-theoretic model the combination of cash-flow stakes and allocation of voting control drive payout policy rather than the cash-flow stake along. In this framework, managers holding large economic interests with unchallenged control respond largely to their own tax rate, whereas a manager that has the same interest but lacks voting control will have take into consideration shareholder interests.

Subsequent research challenged parts of the findings and interpretations in the article. Edgerton (2013) argues that the increase in dividend announcements is attributable to firms becoming more profitable during the sample period rather than due to the tax change, while Floyd et al. (2015) argues that the results are consistent with the general secular payout trend predating the reform. Similarly, Hanlon and Hoopes (2014) shows how dividend changes are concentrated in special dividends timed around rate changes, and Yagan (2015) does not find any effect on real firm outcomes such as investment or employment, calling into question the importance of the agency problem.

All of these papers use "insider ownership", as measured by share count, for firms in the S&P 1500. This share count aggregates two distinct quantities from the Chetty and Saez (2010) model, namely management's cash-flow incentives as well as their voting control of the company. While in single-class firms these two items will be the same, this is not true for more complicated firm structures. As shown in Section 4, dual class firms are not uncommon, especially outside of the larger public firm universe analyzed in prior work. The only direct evidence of the differential impacts of ownership and control comes from Swedish firms (Jacob and Michaely, 2017). Our governance dataset measures voting control and economic ownership as separate measures for close to the full U.S. reporting universe, and therefore allows us to separate the manager's cash-flow incentive from voting control in the U.S. dividend-tax natural experiments.

## 5.1 Empirical Design

In this section we analyze a pair of federal tax changes that both increased and reduced the top dividend tax rates in the United States. The 2003 reform decreased the top dividend rate from 38.6 to 15 percent, while the American Taxpayer Relief Act of 2012, along with other contemporaneous

moves, increased the rate from 15 to 23.8 percent effective January 2013. If the payout responses to the 2003 tax cut reflect a tax incentive rather than being a function of contemporaneous profit changes or pre-existing payout trends, then we should see a similar, but opposite, response to the 2013 tax change. Meanwhile, an explanation based on contemporaneous or pre-existing trends would suggest that we would not expect to see a consistent response to the latter reform.

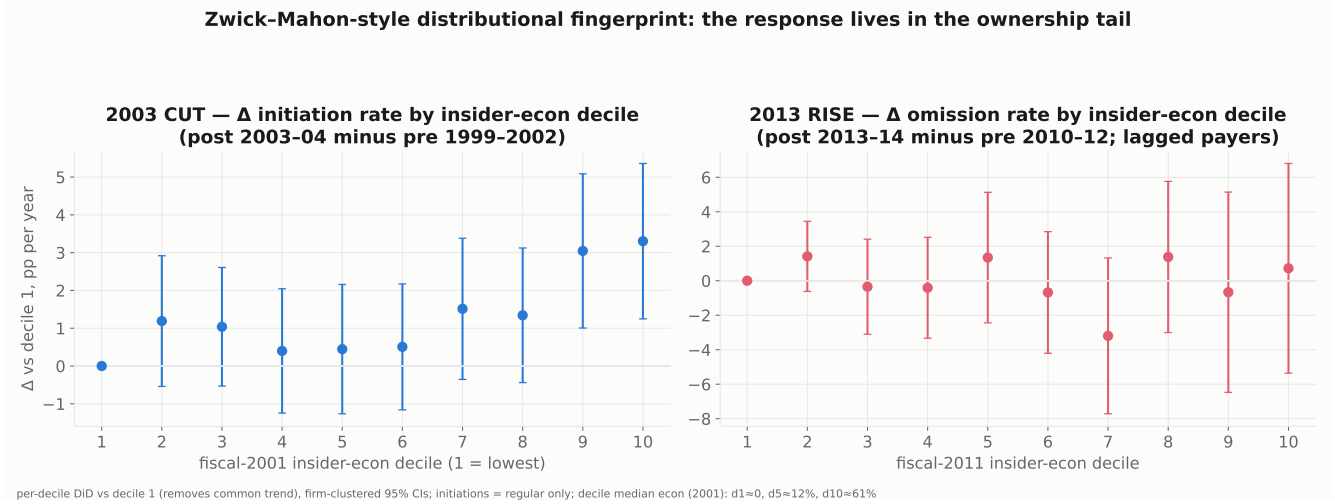
We follow the literature cited above to define our payout measures, except that we focus only on regular dividend initiations, excluding special and one-time distributions following Hanlon and Hoopes (2014). We date dividend terminations to the trading day using the "ex-dividend" date of a firm's final regular payment, and we measure dividends on both the intensive and extensive margins. The measures used to test the owner and control features come from Section 3.2, including the reported directors-and-officers group row for economic ownership and voting power, and our event-time estimates are reported quarterly.

## 5.2 Incentive Gradient

The left panel of Figure 23 shows the change in the initiation rate of regular dividends around the 2003 reform, from the pre-period (1999–2002) to the post-period (2003–04), by decile of insider economic ownership, all relative to the bottom decile. The increase in regular dividends occurred primarily in the far right tail of economic ownership. The increases for the middle deciles of economic ownership are less than a percentage point and are statistically indistinguishable from the bottom decile, while deciles nine and ten, where median insider economic ownership is 43.5 and 61.5 percent respectively, saw initiation rates 3.0 and 3.3 percentage points per year higher. The quarterly event-time estimates show that the increase largely occurred within the enactment year, between the January 2003 announcement for the proposal and the quarters following the May signing. This appears difficult to reconcile with explanations operating through slow-moving profitability or investor-demand channels. Sorting the same exhibit by insider voting power instead of economic ownership produces a nearly identical gradient, which is unsurprising given the correlation between economic ownership and voting power.

The right panel estimates the same analysis for the 2013 rate increase, measuring the change in the number of firms who previously paid a dividend and now omit to pay one, again by ownership decile. The decile-level analysis here is admittedly underpowered because the population of payers is smaller. The 2013 data for the right panel conditions on firms being a dividend payer rather than non-payer, which drops the decile cells from roughly 2,200–2,400 firms per decile at the 2003 cut to 414–915 at the 2013 increase. Standard errors in those deciles roughly triple, and no individual decile change is distinguishable from zero.The count of firms within each decile fell from roughly 2,200–2,400 at the 2003 cut to only 414–915 at the 2013 rate increase, and no decile change is individually distinguishable from zero.

Figure 23: Payout responses by decile of insider economic ownership




*Notes:* The left panel plots the change in the regular-dividend initiation rate around the 2003 dividend-tax cut (post 2003–04 minus pre 1999–2002) by fiscal-2001 decile of insider (directors-and-officers group) economic ownership; the right panel plots the change in the dividend omission rate among lagged payers around the 2013 rate increase (post 2013–14 minus pre 2010–12) by fiscal-2011 decile. Both panels are per-decile difference-in-differences against the bottom decile, which removes the common trend; bars are firm-clustered 95 percent confidence intervals. Initiations count regular dividends only. Median insider economic ownership in 2001 is approximately zero in decile 1, 12 percent in decile 5, and 61 percent in decile 10.

## 5.3 Separating Incentive from Control

While the gradient in Panel A of Figure 23 could be interpreted as the incentive result, the Chetty and Saez (2010) model is strongest on the complementarity between a manager's cash-flow incentive and the power to act on it. Table 4 estimates the interaction in a linear probability model of initiation on year effects and standardized pre-treatment economic ownership, wedge, and their product, with the base year set to the year before each event and standard errors clustered by firm.

The important result is the interaction for the 2003 impact: the initiation response attributable to insider economic ownership is larger by 2.2 percentage points per standard deviation of the voting-ownership wedge ($t = 2.6$). In the joint specification, the wedge main effect implies an increase of 2.0 points ($t = 2.8$), while the economic-ownership main effect is a modest and imprecise 0.5. Thus, the 2003 response was centered where control and its interaction with economic ownership, not economic incentive alone was largest. This result produced the comparative static directly for the Chetty and Saez (2010) model on U.S. data, rather than inferring it from a share-count proxy alone. Among firms with insider control (fiscal-2001 voting power of at least 30 percent), the 1,088 firms whose controllers also held economic ownership above the group median jumped from a 0.4 percent initiation rate in 2002 to 5.1 percent in 2003, while the 21 "wedge controller" firms—votes without proportionate economics, the group the tax cut barely touches—initiated at zero throughout. While the samples are small here the direction is consistent with the model's prediction.

Table 4: Initiations around the 2003 and 2013 tax events: economic ownership, the wedge, and their interaction

| | 2003 (rate cut) (1) | 2013 (rate rise) (2) |
|---|---|---|
| Econ. ownership (per s.d.) | 0.54 | −1.15 |
| | (0.43) | (0.42) |
| Wedge (per s.d.) | 2.02 | 0.29 |
| | (0.71) | (0.61) |
| Econ. × Wedge | 2.22 | −0.01 |
| | (0.87) | (0.31) |

*Notes:* Impact-year coefficients, in percentage points, from linear probability models of regular-dividend initiation on year effects and year-interacted standardized pre-treatment insider economic ownership, wedge (voting power minus economic ownership), and their product, estimated separately around each event on firms at risk of initiation (non-payers in the base year) with pre-treatment ownership data. Ownership measures are the headline D&O group-row constructions, standardized in the pre-treatment year (fiscal 2001 and 2011). Base years are 2002 and 2012; column (1) reports the 2003 coefficients and column (2) the 2013 coefficients; standard errors, clustered by firm, in parentheses.

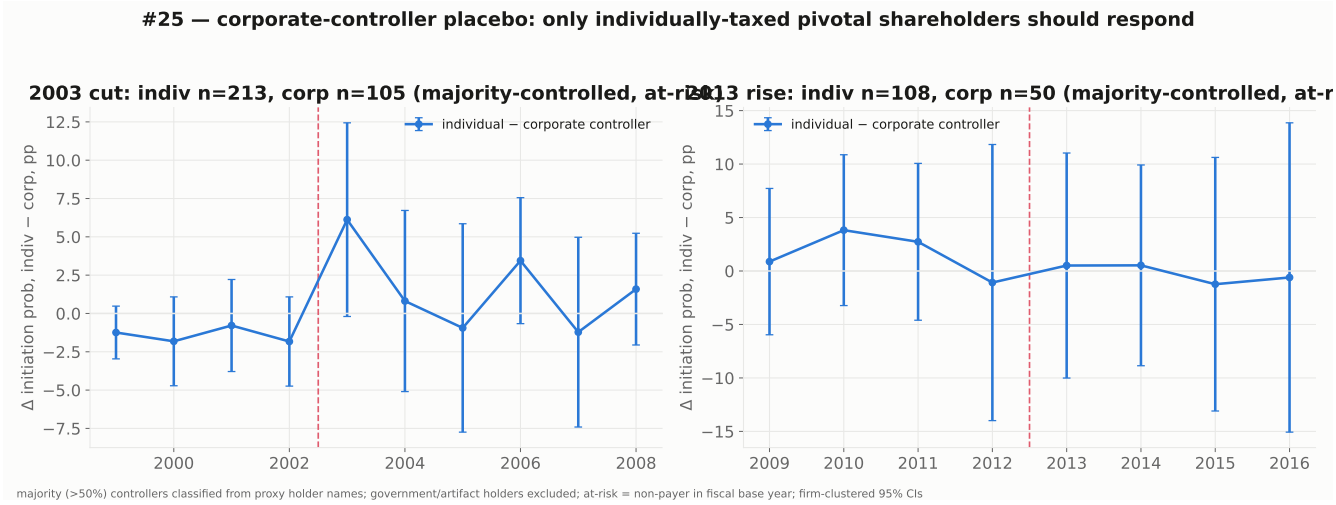
The 2013 rate increase completes the symmetric design on the incentive margin: the economic-ownership gradient reverses sign, with initiations falling by 1.1 percentage points per standard deviation ($t = 2.7$) at the 2013 impact. The interaction, by contrast, is near zero and imprecise at the rise. We read the 2013 event as confirming the sign flip the tax channel requires—no secular-trend story delivers it—while leaving the amplification margin estimated only at the cut, where the at-risk sample is several times larger.

## 5.4 Controller Placebo

One prominent objection to the tax interpretation argument for the 2003 event is from Edgerton (2013), who argued that the initiations rose because business conditions recovered, with the cross-section of insider ownership being correlated with business conditions. Our panel data allows for a placebo design that holds business conditions fixed within the pool of public corporations. Among majority-controlled firms who could initiate a dividend, we compare firms whose controlling shareholder is an individual or family (and thus whose personal dividend tax rate fell from 38.6 to 15 percent) to firms where the controller is instead a corporate parent (for whom the controller's tax position remained essentially unchanged). Both groups are majority-controlled operating companies facing similar economic environment and investor demand, but who differ on the dimension of their controlling shareholder's personal tax exposure.

At the 2003 impact, the 213 individually-controlled firms who were not paying a dividend initiated at 7.8 percent rate against 3.0 percent for the 105 corporate-parent-controlled firms—a 6.1 percentage-point differential ($t = 1.9$) in the event-study specification, without a divergence in any pre-period year (Figure 24). The comparison is different from that of Edgerton (2013), which

Figure 24: Initiation differentials between individually- and corporate-controlled firms



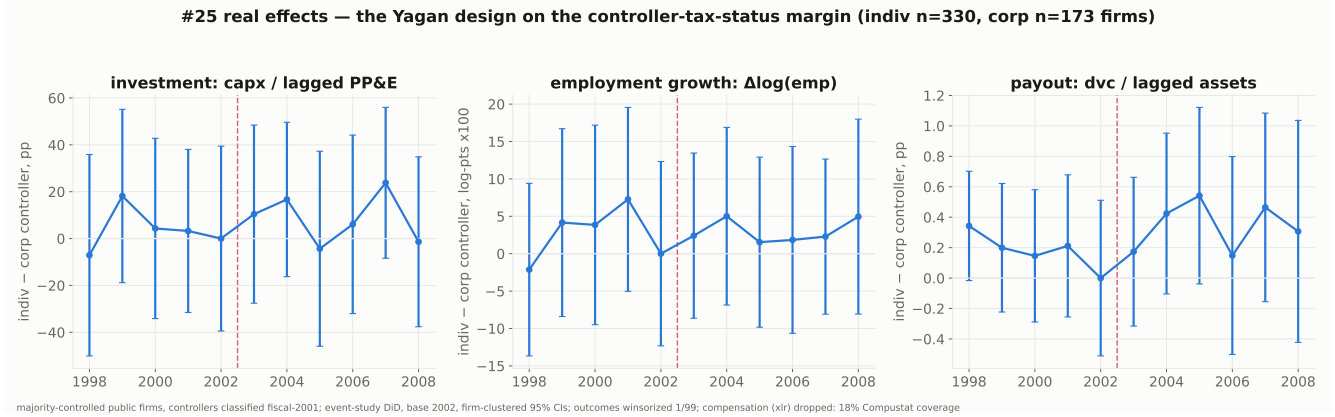

*Notes:* This figure plots year-by-year differences in the regular-dividend initiation probability between majority-controlled firms whose controlling shareholder is an individual or family and those whose controller is a corporate parent, among firms at risk of initiation (non-payers in the base fiscal year). Controllers are classified from proxy holder names, with family and trust markers taking precedence over corporate suffixes; government holders and group-row artifacts are excluded. The left panel is the 2003 rate cut (213 individually-controlled, 105 corporate-controlled firms; base year 2002); the right panel is the 2013 rate increase (108 and 50; base year 2012). Bars are firm-clustered 95 percent confidence intervals.

compared the change for taxable corporations to REITs—entities with different business models, payout mandates, and investor clienteles. Our analysis here holds the entity type, listing status, and control structure fixed and varies only whose ultiamte tax rate changed. The differential is marginally significant with an admittedly small sample, but in a setting where the profitability story predicts no effect. However, at the 2013 increase, a similar comparison on the smaller at-risk sample (108 individually-controlled against 50 corporate-controlled firms) shows no detectable divergence.

## 5.5 Real Effects

Yagan (2015) finds that the 2003 cut left investment and employee compensation unchanged, comparing the effect for C- and S-corporations. We re-estimate a similar design on the controller margin within our panel using an event-study difference-in-differences analysis between the 330 individually-controlled and 173 corporate-parent-controlled majority-controlled firms, with a base year 2002, on investment (capital expenditure over lagged property, plant, and equipment), employment growth, and dividend payout (Figure 25). The investment and employment differentials are noisy zeros through the event window, and no post-event coefficient approaches its standard error in either series. The payout differential turns positive after the cut and persists, although given the small sample size the confidence intervals are wide. The two results suggest that the tax change moved the division of the firm's cash between retention and distribution, in the direction that the agency model would suggest, without detectably moving the real activity of the firm.

Figure 25: Real effects on the controller margin: the Yagan (2015) design




*Notes:* This figure plots year-by-year differences between individually-controlled (330 firms) and corporate-parent-controlled (173 firms) majority-controlled public firms in investment (capital expenditure over lagged property, plant, and equipment), employment growth ($\Delta$ log employees, log points $\times$ 100), and payout (common dividends over lagged assets), in an event-study difference-in-differences with base year 2002. Controllers are classified from fiscal-2001 proxy holder names as in Figure 24. Outcomes are winsorized at the 1st and 99th percentiles; bars are firm-clustered 95 percent confidence intervals.

# 6 A Governance Database Protocol

In this section, we propose how our protocol can be used to create additional measures, which can be done cheaply because the filings are saved once and the transcription process uses open-weight models. The transcription/validation process that we describe in Section 2.2.4, including the use of audit sampling, frontier models for adjudication, and conventions based on regulatory filing guidelines, can all be re-used in the same process for new constructs. During the course of writing this paper we followed this protocol to generate output beyond the initial targets described Section 2.2.2, including a calculation of sixty-day insider option holdings, class-specific rights to elect a subset of the board, and a controlled-company indicator field based on exchange rules.

The protocol for adding a new governance measure could operate in the following manner. First, a researcher would specify the measure of interest and explain the which firms the measure might apply to, where in the filing the identifying information exists, and any anticipated challenging cases. A small sample of proxies would be drawn using this input to establish whether the variable can be generated from primitives already in the corpus or whether it requires a new set to be extracted. In some cases, the new measure may be challenging enough to code that it would require the assistance of a frontier model, requiring some minimal level of API access resources. Edge case rulings and adjudications are made once and saved as a decision rule in the project's parsing code to be used for all subsequent determinations.

After settling on these decision rules, the extraction, transcription, and coding would be run over the full set of saved filings. In addition to this output, the protocol would create a random sample of observations to validate against. This validation sample would then be independently coded by a combination of human coders and frontier models provided with a detailed prompt and

the underlying SEC rules. Where external benchmarks of the measures on other samples exist, as with the CII dual-class website or pre-existing measures of classified boards from academic work, the protocol would test and report the agreement between the new derived measure and the benchmark. Finally, as the parsing and validating process iterates, version control would require each output to be dated and tagged so that researchers can cite to specific releases.

As an example, consider the presence of cumulative voting, which is not yet included in our governance database. Cumulative voting allows a shareholder to cast all of their votes for one nominee, rather than spreading them over the full slate of directors up for election. This increases the voting power of small shareholders and makes managerial entrenchment more difficult. A researcher wanting to add a cumulative voting indicator to the governance database would prompt its inclusion, as well as any related statutes and administrative rules concerning required disclosure. The extraction protocol would then scan the voting-rights sections of a representative sample of proxy filings to determine whether the language is sufficiently clear and uniform to generate a deterministic ruling. After benchmarking against hand-coded samples, and generating any required bespoke coding rules, the measure would be populated for the panel, along with a measure of audited accuracy and agreement rate with outstanding measures (e.g. with IRRC's firm-year coding for the S&P 1500 sample). Moreover, this process can be extended in later years for the same set of measures as each season's proxy filings are made publicly available on EDGAR.

Finally, researchers can request that the sample be broadened to include currently dropped observations (e.g. REITs and foreign-incorporated filers), and the corpus could even be extended to include time-dated charters and company by-laws. This contrasts with a status quo that relies on commercial data vendors who only provide pre-determined variables for a narrower set of firms, and who generate their coding in opaque and inconsistent ways. Our protocol could be customized to the requests of the research community with low marginal cost and full transparency. Moreover, as the data changes, we will maintain archives of releases that can be cited to and compared against as the literature proceeds over the years.

# 7 Conclusion

The empirical corporate governance literature is colored by the underlying data that researchers can access. Commercial data vendors provide a fixed set of governance provisions for the largest publicly-traded firms, using opaque processes that routinely differ across vendor. In this paper, we show how a defined protocol that uses open-weight language models with audit sampling can create a validated panel of insider voting power, economic ownership, dual-class structure, and board classification for hundreds of thousands of filings at low cost. The resulting dataset largely tracks existing measures on large publicly-traded filers, and where it does not, the disagreements can be traced back to quoted figures and language from the underlying filings.

Our data allows us to examine how corporate governance has changed over time for the near-

universe of publicly traded firms. Insider voting power in single-class firms has decreased by three-quarters since our data begins in 1997, while dual-class firms and the wedge between voting rights and economic ownership have increased, largely outside of the S&P 1500. Dual-class firms now account for more than a third of total market capitalization for firms outside the major indexes, and true insider control is understated in table disclosures. Our more detailed measures allow us to tease out the interaction between ownership and voting power when analyzing firm payout decisions in response to the 2003 dividend-tax cut. We confirm the prediction of the model in Chetty and Saez (2010) that both cash-flow rights and voting control matter independently for firm payout decision making.

Data created in this manner is not a finished product but an ongoing creation. The constraint on financial measurement in the age of AI is measurement specification and data validation. Section 6 describes how the research community can use our protocol to extend the corpus to new measures, firm populations, and source documents, at a low cost that doesn't scale with sample size.

# References


**Aggarwal, Dhruv, Ofer Eldar, Yael V. Hochberg, and Lubomir P. Litov**, "The rise of dual-class stock IPOs," *Journal of Financial Economics*, 4 2022, *144* (1), 122–153.

**Bebchuk, Lucian A. and Alma Cohen**, "The costs of entrenched boards," *Journal of Financial Economics*, 11 2005, *78* (2), 409–433.

**Bebchuk, Lucian A and Kobi Kastiel**, "The Untenable Case for Dual-Class Stock," *Virginia Law Review*, 6 2017, *103* (4).

**Bebchuk, Lucian, Alma Cohen, and Allen Ferrell**, "What Matters in Corporate Governance?," *The Review of Financial Studies*, 2 2009, *22* (2), 783–827.

**Bebchuk, Lucian Arye, John C. IV Coates, and Guhan Subramanian**, "The Powerful Antitakeover Force of Staggered Boards: Further Findings and a Reply to Symposium Participants," *Stanford Law Review*, 2002, *55*.

**Berg, Florian, Julian F. Kölbel, and Roberto Rigobon**, "Aggregate Confusion: The Divergence of ESG Ratings," *Review of Finance*, 11 2022, *26* (6), 1315–1344.

**Chetty, Raj and Emmanuel Saez**, "Dividend Taxes and Corporate Behavior: Evidence from the 2003 Dividend Tax Cut," *The Quarterly Journal of Economics*, 8 2005, *120* (3), 791–833.

**_ and _**, "Dividend and Corporate Taxation in an Agency Model of the Firm," *American Economic Journal: Economic Policy*, 2010, *2* (3), 1–31.

**Daines, Robert M., Ian D. Gow, and David F. Larcker**, "Rating the ratings: How good are commercial governance ratings?," *Journal of Financial Economics*, 2010, *98* (3), 439–461.

**DeAngelo, Harry and Linda DeAngelo**, "Managerial ownership of voting rights: A study of public corporations with dual classes of common stock," *Journal of Financial Economics*, 3 1985, *14* (1), 33–69.

**Edgerton, Jesse**, "Four facts about dividend payouts and the 2003 tax cut," *International Tax and Public Finance*, 10 2013, *20* (5), 769–784.

**Floyd, Eric, Nan Li, and Douglas J. Skinner**, "Payout policy through the financial crisis: The growth of repurchases and the resilience of dividends," *Journal of Financial Economics*, 11 2015, *118* (2), 299–316.

**Frankenreiter, Jens**, "Measuring Corporate Governance with Large Language Models," 7 2026.

**_ , Eric Talley, Cathy Hwang, and Yaron Nili**, "Cleaning Corporate Governance," *University of Pennsylvania Law Review*, 1 2021, *170* (1), 1–70.

**Gompers, Paul A., Joy Ishii, and Andrew Metrick**, "Extreme Governance: An Analysis of Dual-Class Firms in the United States," *The Review of Financial Studies*, 3 2010, *23* (3), 1051–1088.

**Gompers, Paul, Joy Ishii, and Andrew Metrick**, "Corporate governance and equity prices," *Quarterly Journal of Economics*, 2003, *118* (1), 107–155.

**Hanlon, Michelle and Jeffrey L. Hoopes**, "What do firms do when dividend tax rates change? An examination of alternative payout responses," *Journal of Financial Economics*, 10 2014, *114* (1), 105–124.

**Jacob, Martin and Roni Michaely**, "Taxation and Dividend Policy: The Muting Effect of Agency Issues and Shareholder Conflicts," *The Review of Financial Studies*, 9 2017, *30* (9), 3176–3222.

**Jensen, Michael C and William H Meckling**, "Theory of the Firm: Managerial Behavior, Agency Costs, and Ownership Structure," *Journal of Financial Economics*, 1976, *3*, 305–360.

**Kim, Hyunseob and Roni Michaely**, "Sticking around Too Long? Dynamics of the Benefits of Dual-Class Voting," *SSRN Electronic Journal*, 1 2019.

**Larcker, David F, Peter C Reiss, and Youfei Xiao**, "Corporate Governance Data and Measures Revisited," 2015.

**Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan**, "Large Language Models: An Applied Econometric Framework," *Annual Review of Economics*, 8 2026, *18* (Volume 18, 2026), 283–316.

**Morck, Randall, Andrei Shleifer, and Robert W. Vishny**, "Management ownership and market valuation: An empirical analysis," *Journal of Financial Economics*, 1 1988, *20* (C), 293–315.

**Porta, Rafael La, Florencio Lopez de Silanes, Andrei Shleifer, and Robert W. Vishny**, "Law and finance," *Journal of Political Economy*, 1998, *106* (6), 1113–1155.

**Shleifer, Andrei and Robert W. Vishny**, "A Survey of Corporate Governance," *Journal of Finance*, 1997, *52* (2), 737–783.

**Tirole, Jean**, "Corporate governance," *Econometrica*, 2001, *69* (1), 1–35.

**Yagan, Danny**, "Capital Tax Reform and the Real Economy: The Effects of the 2003 Dividend Tax Cut," *American Economic Review*, 12 2015, *105* (12), 3531–3563.