

Introducing a New Corpus of Definitive M&A Agreements, 2000-2020

Peter Adelson* Matthew Jennejohn† Julian Nyarko‡

Eric Talley§

February 2024

Abstract

Contract design and architecture is an immensely important topic within economics, finance, and law. However, attempts to study it are significantly hamstrung by the limited availability of publicly available, high quality data. In this paper, we introduce a new corpus of 7,931 Definitive Merger Agreements submitted to the SEC between 2000 and 2020 involving a transaction in excess of \$100 million. Through a combination of machine learning and human evaluation, we are able to associate these agreements with other metadata, such as deal size, industry classification, and advising law firms. In addition, we identify and make available the text of individual clauses contained in these agreements. In a final step, we provide an illustration of how these data can be used to generate novel insights into M&A contract design and drafting practices.

1 Introduction

The now-mature field of contract design dates back nearly a century (Coase, 1937), and it now features myriad rich and varied contributions seeking to characterize and/or test theories of how parties organize private law to shape and enhance their economic environments. In turn, this literature now has spawned legions of intellectual descendants, including efforts to explore the inherent incompleteness that transaction costs impose on contractual structures (Hart and Moore, 1988), the critical importance of governance and control of economic exchange within the contractual boundaries of the firm (Williamson, 1985; Grossman and Hart, 1986), the default rules of contract law (Ayres and Gertner,

*Stanford University.

†U.S. Department of Defense, Office of Strategic Capital; Johns Hopkins University Applied Physics Lab; Brigham Young University.

‡Stanford University.

§Columbia University.

1989), and the role of extra-legal norms and enforcement mechanisms (Landa, 1981; Greif, 1993). A related literature has invited creative academic thought about how to best organize the “production” of contractual provisions themselves, and the choice between designing new bespoke language and repurposing terms from legacy deals (Choi et al., 2021; Choi et al., 2022). And yet a third set of contributions attends the somewhat surprising global properties that ensue when contractual emulation becomes systemic. For example, the rote usage of “boilerplate” provisions may lead to the lock-in of inefficient terms or even the settled meaning of contractual language being lost to memory (Hill, 2001; Gulati and Scott, 2012). In a similar vein, contractual boilerplate may become “sticky” (Nyarko, 2021), resisting change even when the underlying economics strongly favor it. And, the serial layering on of boilerplate terms from different legacy transactions can spawn emergent phenomena that give rise to a complex landscape of hidden traps, loopholes, and counter-loopholes, all conspiring to make the enforcement of boilerplate contracts difficult to predict (Pandya and Talley, 2023; Fontenay, 2020).

Each of the above contributions advances (among other things) important testable predictions about contract practices, and a large cohort of scholars in both law and finance have increasingly turned their attentions to this task. This empirical project is critically important, not only for adjudicating between existing theories, but also for generating new ones about how, when, and whether contract design helps or hinders allocative and productive efficiency. Moreover, the emergence of generative AI and machine learning techniques to study legal text has raised the stakes of empirical contract testing and measurement even further. Now, empirically minded researchers are free to experiment with teasing out new and creative features of contractual texts, rather than depending on the (somewhat arbitrary) labels that prior researchers constructed. In short, data analytics have become highly democratized.

But data availability still lags far behind. It remains difficult for empirically-minded contract scholars to access clean, readable corpora of relevant contractual texts. Those who have built such corpora have at times been reluctant to share them. On other occasions, authors report that they are prohibited from doing so by restrictive licensing practices (Frankenreiter et al., 2021). These constraints have placed a barrier to entry within the field of empirical contracts that makes it difficult to accomplish reproducibility, a core condition for the advancement of the field.

One significant domain of high-stakes private contracting is in the field of mergers and acquisitions. Such contracts are amongst the most lengthy, complicated, and important constructs of modern financial markets, making them eminently worthy of attention. In prior work, three of us introduced a unique corpus of publicly disclosed merger agreements spanning twenty-one years (2000-2020) (Jennejohn et al., 2022). At publication, this corpus was the largest open-access resource of its kind, and we are unaware of any other contributors who have produced a more comprehensive one since.

That said, data collection and cleaning remains an ongoing, laborious process—one that is never truly complete. In the pages below, we illustrate this claim

by correcting and improving on our own prior work (Jennejohn et al., 2022). Specifically, we introduce a far larger, far more comprehensive contract corpus for definitive merger agreements (the "DMA Corpus"). We release this improved data set now, in the hopes that doing so will catalyze and further advance much-needed empirical research on contractual design and evolution.

Our corpus increases the size of Jennejohn et al. (2022) from 2,141 to approximately 7,931 definitive merger agreements signed between 2000 and 2020. The corpus, sourced from publicly available SEC filings, includes the complete text of the agreements and also extracts key provisions, identified through machine learning techniques, within each contract. Parsing agreements in this fashion allows researchers to study dynamic patterns of both the entire agreements and individual terms within a larger agreement, further enhancing the utility of our corpus for a wide variety of uses. We are able to increase the size of the corpus by relaxing a prior requirement from Jennejohn et al. (2022) that each deal tracked *must* appear in both FactSet and in the SDC Platinum Financial Securities Data to be included. This constraint significantly limited the size of our prior endeavor, principally due to the sporadic and inconsistent tracking of SDC Platinum.¹ The effort described below required us to develop a different process that combines both manual labels and machine learning to reliably identify the relevant contracts, a process that can easily be extended to other legal domains of data collection.

The full DMA Corpus (and documentation) is available for download at https://github.com/padelson/dma_corpus. Our sole request to those who wish to make use of the data is to acknowledge this study as the source of the data.

This paper unfolds as follows. Part 2 describes the methods used to collect, clean, and parse the data in the DMA corpus, including a variety of machine learning techniques used to parse the definitive merger agreements. Part 3 orients the researcher through a brief description of the DMA Corpus' salient characteristics. Part 4 demonstrates with a replication and correction of our earlier M&A study reported in Jennejohn et al. (2022), which involved a dataset a fraction of the size of the DMA Corpus introduced here. Finally, in Part 5 we conclude with recommendations for future data collection and research.

2 The DMA Corpus

The DMA Corpus is based on a combination of two data sets: First, deal data from FactSet provides high-level information about deal characteristics, such as announcement date, completion date, and transaction value. Second, we rely on the SEC's EDGAR data base to incorporate the text of the DMA, as well as meta-information about the DMA such as the filing date. The main challenge in creating the DMA Corpus is to accurately merge these two large data sets. As explained in greater detail below, we achieve this goal through a combination

¹As detailed below, the current data set continues to remove contracts if the deal was not tracked by FactSet.

of machine learning algorithms and manual labeling. In particular, we train an intentionally overinclusive algorithm to generate plausible candidate filings for each entry in FactSet. Afterwards, we manually remove false matches. The resulting matches are comparable in quality to those that could be achieved if an attentive human matched each entry manually.

2.1 Source Data

The DMA Corpus builds on two primary sources of data. Contract text is extracted from EDGAR filings as detailed below. Merger information is collected from FactSet, filtered for mergers between January 1, 2000 and December 31, 2020 with at least \$100M in transaction value. This totaled 40,759 merger agreements. According to FactSet, 34,001 of these agreements had completed, 3,412 had been cancelled, and the remaining 3,346 were rumored, pending, or otherwise uncertain.

2.2 Agreement Text

To obtain the text of the agreement, we turn to the SEC's EDGAR data base. The SEC requires publicly registered companies to submit their "material contracts" to EDGAR, a centralized data repository. A "material contract" is an agreement "not made in the ordinary course of business that is material to the registrant."² The contracts are submitted as exhibits to disclosure forms.³ In effect, EDGAR represents the largest known collection of publicly available merger agreements, although it certainly is not complete.⁴

To collect the agreements, we begin by obtaining all 3,207,357 EDGAR filings from January 1st, 2000 to December 31st, 2020. For each filing, we obtain exhibits numbered 10 ("Material contracts") or 2 ("Plan of acquisition, reorganization, arrangement, liquidation or succession").⁵ We also downloaded exhibits numbered 99 ("Additional exhibits") when their EDGAR description contained a word indicating that the exhibit was a contract, agreement, or a plan of merger. Our efforts resulted in 1,150,299 extracted documents. The vast majority of exhibits are filed either in .txt, .html or .htm format. A small number of exhibits 3,835 are filed as .pdf. Due to difficulties posed in processing the .pdf format, we omit these agreements, yielding a final corpus of 1,146,464 candidate documents.

For each agreement, we also retain the meta-information associated with its relevant EDGAR filing. This includes the filing company's name and central index key (CIK), the filing type, the filing date, and exhibit type. The resulting

²17 C.F.R. § 229.601(b)(10)(i).

³Forms 10-K, 10-Q, 8-K, 20-F, 6-k and S-4

⁴For instance, when Google acquired Zagat Survey, the restaurant rating service, in 2011 for \$125M, it did not file the merger agreement with the SEC, presumably because it did not deem the acquisition "material" Efrati, 2011.

⁵17 CFR § 229.601

documents contain (close to) the universe of all contracts filed with the SEC, although merger agreements are representing only a small fraction.

2.3 Merger Data

To obtain information on mergers, we rely on data from FactSet. FactSet maintains proprietary data sets for a variety of business-related fields, such as corporate governance, stock performance, and equity ownership. We use the FactSet merger dataset FactSet, n.d. This dataset includes over 500,000 announced mergers, acquisitions, and spin-offs across the world. The data set reports relevant deal terms, as deal size. It also includes an indicator for the presence of certain legal provisions, such as top-up provisions. We filtered for mergers announced between January 1, 2000 and December 31st, 2020. Consistent with common practice in the literature Quinn, 2010; Jennejohn et al., 2022; Coates, 2016, we subset to mergers with a transaction value of at least \$100M. This totaled 40,759 mergers. According to FactSet, 34,001 of these agreements had completed, 3,412 had been cancelled, and remaining 3,346 were rumored, pending, or otherwise uncertain.

2.4 Matching Contracts with Mergers

In a third step, we attempt to match each entry from our FactSet data base to one or more agreement texts. In developing our matching procedure, our goal was to be intentionally over-inclusive, allowing for false positives at the cost of creating potential matches when in truth, no DMA was filed. To evaluate the matching procedure, we hand-labeled a set of 400 randomly selected rows of FactSet data.

Our process for matching contracts in the corpus with the appropriate FactSet data proceeds as follows:

(1) For a given FactSet deal, we began the process by filtering EDGAR filings around the FactSet deals' announcement date and completion date, looking for matches in filings one day prior to and up to 45 days after the announcement date or up to one day prior and up to 45 days after the completion date.

(2) To search for a given deal, we first tried to obtain the Central Index Key (CIK) for the acquiring company and the target company based on the information available in the FactSet deal data. We first attempted to match the CUSIP in FactSet deal data with an EDGAR CIK. To find an associated CIK, we used a CIK-CUSIP mapping available at: <https://github.com/leoliu0/cik-cusip-mapping/blob/master/cik-cusip-maps.csv>. If the CUSIP did not map onto a CIK, if there were no possible DMAs, or if no filings were perfect fits, we attempted to identify CIKs based on company name information. Within the appropriate filing window, we compared all filing names to the listed names within the FactSet data, using Levenshtein Distance (edit distance) to assess string similarity.⁶ We took at most one matching CIK for every name listed

⁶A company was deemed a sufficient match if its edit distance was less than one third of

with a target or acquiring company. However, if there are multiple names listed, there may be multiple candidate CIKs, up to one for each name.

(3) After obtaining a CIK, we searched within the filing window for all filings made by a company with that matching CIK. This provided a candidate list of documents likely associated with the acquiring or target company, filed around the announcement or completion date of the deal.

(4) Each document was then assessed according to a manually coded set of rules. First, the code extracts the text from the candidate contract. Next, the code attempts to determine the type of the contract using keyword matching, searching for contracts likely related to merger deals. Next, the code analyzes the contract text to assess if it is likely an agreement between the parties listed in the FactSet data. To do this, the code analyzes the text for the company names listed in the FactSet data. The code uses various rules to search for sub-strings of the relevant company name and provides an assessment ranging from zero to one indicating confidence that the contract refers to the entity in question. A name matching score is calculated for both the target and acquirer.

(5) Based on the type of the contract and the name matching score, the code calculates a score for the contract. Certain penalties are applied if the contract indicates it is an amendment to a prior contract (with lesser penalties if the contract is also restated). A contract is only considered a perfect match if it is not an amendment, if it has a name match score of one for the non-filing entity, and if it contains the appropriate keywords indicating that it is likely a DMA.

(6) The algorithm searches first for a match based on the acquirer's CUSIP, then the target's CUSIP, then the acquirer's name, and finally the target's name. Candidate filings are reviewed in date order. The first filing with a matching score of 1 is taken as the matching DMA. If no candidate filing has a matching score of 1, the filing with the greatest matching score is taken.

(7) The performance of the algorithm was assessed against 400 hand-labeled randomly selected rows of FactSet data. The goal of our matching algorithm was to be over-inclusive. That is, if a FactSet entry has a corresponding agreement text, we wanted to be sure that the entries are matched, even if that would come at the cost of sometimes creating matches when a FactSet entry had no accompanying match in our agreement text corpus. In the context of information retrieval, our goal was to optimize *Recall* at the cost of *Precision*. *Recall* is defined as the proportion of true positives that the algorithm correctly identifies. *Precision* is defined as the share of true positives among the positive predictions. We also report the F_1 -Score, which is a weighted mean of *Recall* and *Precision*.

Performance results are shared in Table 1. As can be seen, our matching procedure leads to a *Recall* of 0.96, suggesting that there is a 96% chance of identifying the correct agreement text for a FactSet entry if there is one. Our precision is 0.68, suggesting roughly a third of the matches we generate are incorrect. We then applied the algorithm to the whole FactSet dataset, with

the character length of the company name (or less than 3 for company names shorter than 9 characters). The edit distance must also be less than 11 to account for long names that may have edit distances shorter than one third of the length.

the algorithm ing for each merger in the dataset (1) if there existed a matching DMA in the candidate documents, and (2) if so, which document was the best match for that merger. The algorithm identified a DMA for 11,698 mergers in the FactSet data.

	Recall	Precision	F1 Score
All Identified DMAs	0.96	0.68	0.80

Table 1: Algorithmic Performance

2.4.1 Further Processing

Manual Filtering: After matching, our data set accurately matches almost every FactSet entry to its corresponding DMA if it exists, but it also creates a number of matches when in fact there is no associated DMA. To identify the false matches, a team manually reviewed each of the 11,698 matched entries with the sole task of eliminating false matches. This has the effect of optimizing the precision of the matches within the margin human error without affecting recall. After manually verifying performance, the data set consists of 7,931 merger and DMA pairs, with reviewers rejecting the remaining 3,767 labels. These results are consistent with the reported 0.68 precision score on the sample of 400 mergers.

Kira Processing We submitted each DMA to Kira Systems (Kira), a machine learning contract analysis platform. Kira provides pre-trained classifiers that associates sections of contractual text with a contractual provision. For example, Kira can identify a block of text as identifying the parties of the contract or the governing law of the contract. We used Kira to label text for the following provisions: 'MeToo' representations, pandemic Material Adverse Event (MAE) carveouts, Choice of Law provisions, Choice of Forum/Venue provisions

2.5 Data Format

We share the data as a .csv of metadata about the DMAs and a .json file of DMA text. The text of the DMAs are provided in as a text field in the .json file. The 'url' field present in the metadata .csv and the text .json file permits cross-references between the two. Further information about the data format is available in an accompanying codebook.

2.6 Limitations and Extensions

The data set is limited in two key ways. First, it relies on FactSet to identify merger agreements. If FactSet did not record a merger, the merger will not be in our data set. Second, we rely on EDGAR filings to identify the associated DMA. While the data set has a high likelihood of finding an associated EDGAR exhibit if one exists (recall of 0.96), based on manual investigation, we estimate that

about 70% of FactSet mergers have no associated DMA exhibit on EDGAR. This could be, for instance, because a merger wasn't deemed to be material, because a company simply did not comply, or because the merger did not involve public US companies subject to SEC regulation. The data are skewed to agreements likely to be posted on EDGAR - mergers where the target is a public entity or where the acquirer is a public entity and the deal is sufficiently material to warrant public reporting.

3 Description of the DMA Corpus

	Value
Number of Agreements	7,931
Number of Deals with Public Target	5,192
Number of Deals with Public Acquirer	6,837
Number of Deals between Public Companies	4,177
Number of Deals with DE Target	1,677
Number of Deals with DE Acquirers	1,065

Table 2: Summary Statistics of DMA Corpus

	Average	Median	25th	75th	Min	Max
Deal Size (\$M)	1,900	434	200	1,304	100	127,788
Length in Words	41,384	39,513	31,259	48,612	291	413,192
Estimated Pages	82.8	79.0	62.5	97.2	0.6	826.4
Year	2010.7	2011	2006	2016	2000	2020
Type Token Ratio	0.12	0.11	0.10	0.13	0.04	1.00
Readability	19.8	19.8	18.3	21.3	8.1	47.1

Table 3: Summary Statistics of DMA Corpus

The DMA corpus consists of 7,931 unique merger agreements from 2000 to 2020. Figure 1 provides details on how many agreements exist for each year, how many of these details were marked as completed in the FactSet data, and the median deal value in millions of dollars. Tables 2 and 3 provide summary counts and statistics for the corpus. The appendix provides further summary statistics related to industry-representation in the corpus (Tables 4 and 5). Figure 9 provides a heat map assessing the interaction between acquirer and target SIC codes.

Figure 2 provides information about the length and readability of the DMAs. DMAs in the corpus have increased in length over time, with median word count rising from about 27.5K in 2000 to over 50K in 2020. DMAs have also been getting more difficult to read, according to the Flesch-Kincaid Grade Level

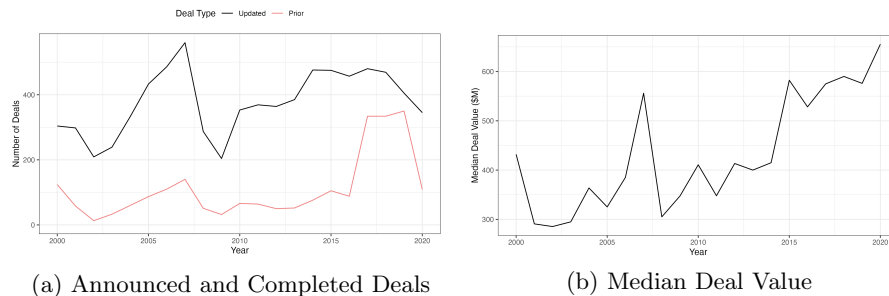


Figure 1: Trends in Agreement Counts and Size

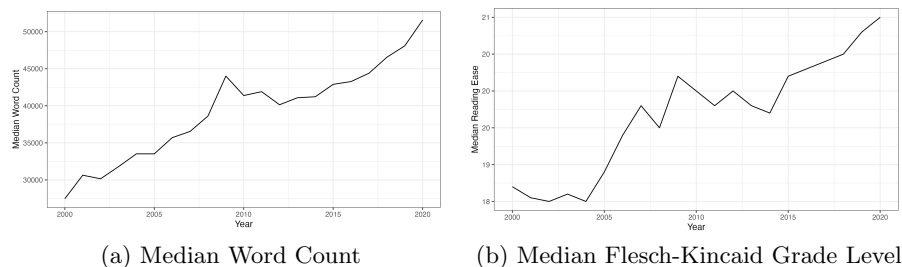


Figure 2: Trends in Agreement Text

metric Kincaid et al., 1975, a metric which assesses the equivalent grade-level of a piece of text (higher numbers indicate lower readability). The median Flesch-Kincaid score rose from 18.7 in 2000 to 21 in 2020. We do highlight, however, that Flesch-Kincaid scores have not been directly validated on legal texts, which is why our evidence is somewhat speculative. Together, these trends indicate that DMAs are getting longer and may be getting harder to read.

The appendix also provides information on the law firms who advised the greatest aggregate deal value on the buy- and sell-sides. Either full deal value is given to any law firm that advised the client (Tables 6, 8) or the deal is apportioned evenly across the law firms who advised the client (Tables 7, 9); for example, if there are two law firms both advising the target, they would each get 1/2 of the deal value. This data is also taken from FactSet. By either, the three firms with the highest volume are, in order, Skadden, Arps, S., M. & Flom; Wachtell, Lipton, R. & Katz; and Sullivan & Cromwell.

4 Analyzing Merger Agreements with the DMA

In this section, we replicate and, in one respect, correct one of our own studies, Jennejohn et al. (2022). Using the labels provided either by Kira or Factset, we track the share of contracts that have the following contractual provisions: 'MeToo' representations, pandemic Material Adverse Event (MAE) carveouts, top-up options, reverse termination fees, New York Choice of Law provisions, and New York Choice of Forum provisions. Our original study used 2,141 M&A deals as the basis for its empirical analysis. We follow the same analysis but instead use the DMA dataset of 7,931. Compared to our earlier study, all results are essentially the same except, as discussed more fully below, we find a materially higher number of reverse termination fees in the new corpus. The reason for the divergence is that the original study filtered for deals that were present both on FactSet and in the SDC Platinum Financial Securities Data. However, we later learned that SDC Platinum is significantly underinclusive, and thus do not apply the filter here. That said, the expanded data set does not change other findings presented in the original study. We note any relevant divergences below.

4.1 #MeToo Representations

#MeToo representations are representations within DMAs that relate to sexual harassment allegations against company executives. These provisions typically involve a representation by the seller that during some look-back period (such as the past 3 years), there have been no sexual harassment allegations against company executives Burnett, 2019. We labeled whether a contract contains a #MeToo provision using Kira. Figure 3 demonstrates that these provisions have suddenly become more prevalent shortly after the Weinstein scandal increased awareness for sexual harassment and abuse in the workplace. Figure 3 is similar to the Figure 10 in the prior analysis Jennejohn et al., 2022, provides an example of an external shock suddenly changing contract drafting practices.

4.2 Pandemic MAE Carveouts

Material Adverse Event (MAE) clauses allocate risk that some event may substantially change the value of the target company between the signing and closing of a DMA. These clauses generally allow a buyer to cancel the acquisition in the event of a MAE, but the clauses are also subject to "carveouts." Monson, 2015 These carveouts are events that are not sufficient grounds to terminate the acquisition. Using Kira, we labeled when a contract contains a pandemic-related carveout to its MAE clause. Figure 4 shows that pandemic-carveouts experienced a modest but steady increase in prevalence starting with the 2009 H1N1 outbreak. As pointed out in Jennejohn et al. (2022), this pattern is suggestive of common diffusion practices that take place under a common learning process. The prevalence of carve-outs subsequently increased dramatically after news of the COVID-19 pandemic broke in 2020, suggesting a deviation from the regular

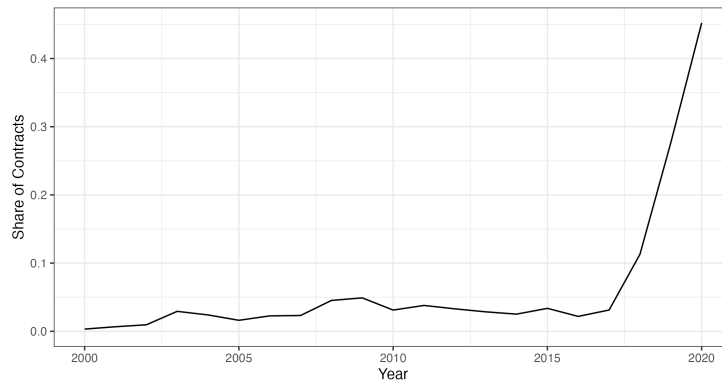


Figure 3: Contracts with #MeToo representations over time

learning process induced by a more significant, external shock in the form of the coronavirus. Overall, results are consistent with Figure 7 in Jennejohn et al. (2022).

4.3 Top-Up Options

Top-up options are options issued by the board of a target company to the buyer. These options allow a buyer who is using a tender-offer to cross the 90% ownership threshold required for a statutory short-form merger Davidoff, 2007. A Delaware General Corporate Law revision in 2013 decreased the significance of the 90% ownership threshold in mergers. Fisher, 2013. Figure 5 shows how Top-Up options became much less popular in the time following this statutory revision. It is an example illustrating the impact of a regime shock on a provision that was previously adopted at an accelerated pace, and the pattern is consistent with Figure 11 in Jennejohn et al. (2022).

4.4 Reverse Termination Fees

Reverse Termination Fees, or RTFs, are fees the buying party may be obligated to pay to the seller if the deal fails to close for certain reasons. Afsharipour, 2010. These fees may come in different flavors, but they generally operate to assign at least some risk that the deal may fail to close to the buyer, often covering financing and regulatory risks. Jennejohn et al., 2022. Figure 6 shows that the prevalence of RTFs increased modestly from 2000-2020, rising from about 14% of contracts to nearly 24%. The patterns differ somewhat from Figure 9 in Jennejohn et al. (2022), where we concluded, albeit with a lot of caution and qualifiers, that RTFs show “very weak discernable patterns” and

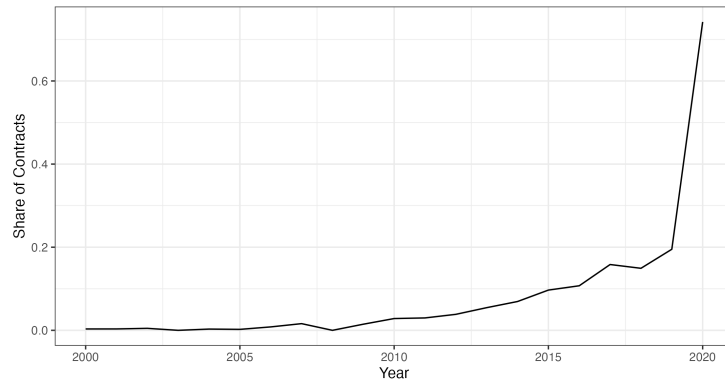


Figure 4: Contracts with Pandemic MAE Carveouts

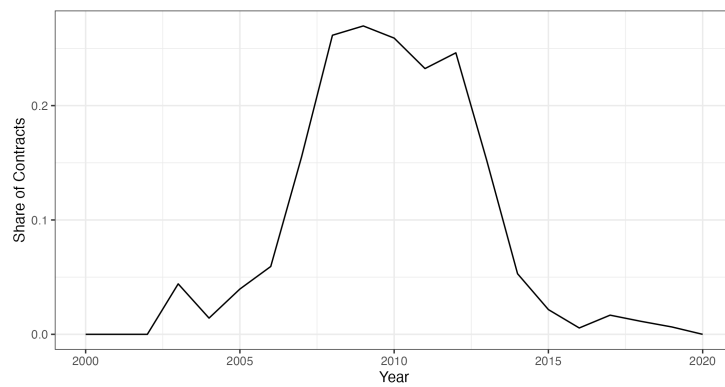


Figure 5: Contracts with Top-Up provisions over time

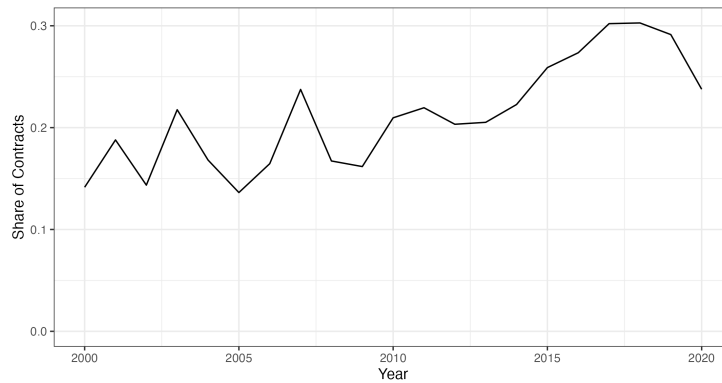


Figure 6: Contracts with Reverse Termination Fee provisions over time

found that “around 2016, RTFs have become mildly less popular in the deals we have tracked.” (Jennejohn et al., 2022, p. 950).

4.5 New York and Delaware Choice of Law and Choice of Forum

Choice of Law (CoL) and Choice of Forum (CoF) provisions are two classic contractual provisions. The Choice of Law provides the parties’ choice on which jurisdiction’s substantive law should apply to the interpretation and enforcement of their contract (Hoffman, 2014). The Choice of Forum provision describes in which district parties will litigate a dispute. Although very common, contract parties need not choose the same jurisdiction for both law and forum; courts of one jurisdiction can apply the substantive laws of another jurisdiction. Two popular jurisdiction for both CoL and CoF provisions in business agreements are New York and Delaware. Figure 7 and Figure 8 track what proportion of contracts have chosen New York Delaware in CoL and CoF respectively.⁷ Jennejohn et al. (2022) only examined choice of forum provisions that select New York, and the patterns we see in 8 are consistent with Figure 12 in Jennejohn et al. (2022), showing an increase in New York court utilization around 2010. This change coincides with a change in Supreme Court precedent that makes it more

⁷Some contracts elect to have substantive law and fora for different disputes regarding the same contract. For example, a contract may subject a financing provision to New York law and courts while selecting Delaware as the governing law and forum for the rest of the contract. The contractual shares reported here are broad, counting such a case as a contract with both a Delaware and New York CoL and CoF provision. Because of this, the ‘share’ of contracts summed across different jurisdictions may be greater than one.

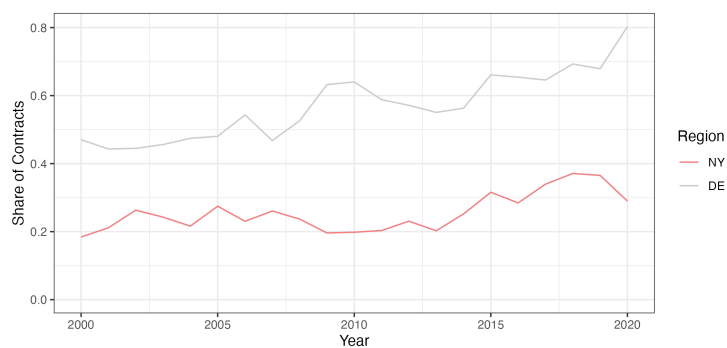


Figure 7: Contracts with New York or Delaware Choice of Law

difficult for parties to access the courts of a state they are not incorporated in.⁸ The case generally increased contracting parties' awareness of issues surrounding forum choice (Nyarko, 2021).

⁸Goodyear Dunlop Tires Operations, S.A. v. Brown, 564 U.S. 915, 919 (2011); Daimler AG v. Bauman, 571 U.S. 117, 139 (2014).

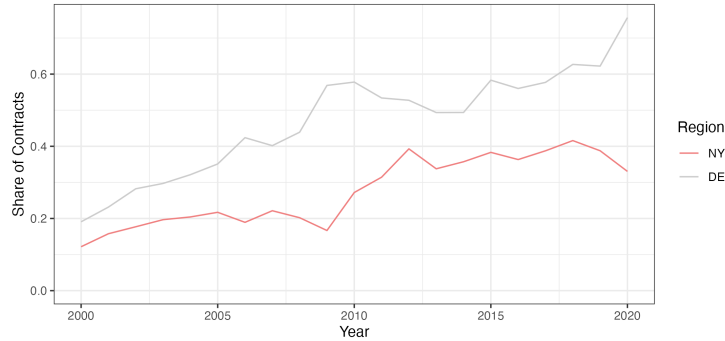


Figure 8: Contracts with New York or Delaware Choice of Forum

5 Conclusion

In this paper, we introduce a large and detailed corpus of nearly 8,000 definitive merger agreements. The corpus, drawn from publicly available SEC filings, captures deals from the years 2000 to 2020 that involve a public company on at least one side of the transaction. This corpus builds upon our earlier work Jennejohn et al., 2022, and replication of our earlier analyses produces substantially similar results. Our hope is that this new corpus, shared as a common resource with the academic community, sparks a new wave of empirical research on the details of contract design and encourages other researchers to make their data publicly available in similar fashion.

In addition, we hope that the methodology we employed to identify DMAs can be adopted to other scenarios in which researchers seek to identify a small number of relevant documents within a larger corpus. In our tiered classification approach, the first step is to remove obvious irrelevant documents from the data set. We do this by comparing structured information of the contracts against our data base. And one can easily imagine other forms of initial filtering. For instance, in other work, a subset of us have studied climate risk disclosures and used simple keywords to remove clearly irrelevant documents from a corpus of 10-K filings Nyarko and Talley, 2022. Other filters might use a simple, computationally inexpensive classifier, like a logistic regression classifier (Li et al., 2023). More generally, the use of this first filter can significantly reduce the number of documents that need to be labeled without significantly increasing error rates. This is especially valuable in contexts where drawing a random sample from the unfiltered corpus would yield few relevant documents.

6 Appendix

Industry	SIC Code	# Deals
Agriculture, Forestry, And Fishing	A	30
Mining	B	587
Construction	C	76
Manufacturing	D	2616
Trans., Comms., Elec., Gas, And Sanitary Svcs.	E	856
Wholesale Trade	F	278
Retail Trade	G	290
Finance, Insurance, And Real Estate	H	1340
Services	I	1851
Unknown	NA	7

Table 4: Number of Deals by Target's SIC Division

Industry	SIC Code	# Deals
Agriculture, Forestry, And Fishing	A	24
Mining	B	492
Construction	C	65
Manufacturing	D	2448
Trans., Comms., Elec., Gas, And Sanitary Svcs.	E	787
Wholesale Trade	F	192
Retail Trade	G	196
Finance, Insurance, And Real Estate	H	2412
Services	I	1292
Public Administration	J	16
Unknown	NA	7

Table 5: Number of Deals by Acquirer's SIC Division

		A	B	C	D	E	F	G	H	I	J
Target SIC Code	-	7	0	0	0	0	0	0	0	0	0
	A	0	4	1	0	12	1	3	1	1	1
	B	0	1	420	6	27	20	1	1	7	9
	C	0	0	1	34	9	4	0	1	6	10
	D	0	12	29	8	1986	54	92	35	16	216
	E	0	3	60	5	55	549	17	7	14	77
	F	0	5	4	0	58	15	59	22	4	25
	G	0	0	2	0	38	3	27	96	9	21
	H	0	5	62	17	303	147	55	113	1226	484
	I	0	0	6	6	125	57	24	11	56	1007
		A	B	C	D	E	F	G	H	I	J

Figure 9: Matrix of Acquirer and Target SIC Codes

Law Firm	# Deals	Total Value (\$M)
Sullivan & Cromwell	421	\$2,206,201
Simpson Thacher & Bartlett	398	\$2,156,863
Wachtell, Lipton, R. & Katz	319	\$2,114,080
Skadden, Arps, S., M. & Flom	539	\$1,844,200
Davis Polk & Wardwell	290	\$1,777,311
Cleary Gottlieb S. & Hamilton	258	\$1,666,599
Weil, Gotshal & Manges	316	\$1,515,309
Latham & Watkins	446	\$1,348,833
Cravath, Swaine & Moore	268	\$1,315,508
Shearman & Sterling	263	\$1,140,714

Table 6: Top 10 Law Firms by Value of Deals, Buy Side

Law Firm	# Apport. Deals	Apport. Value (\$M)
Skadden, Arps, S., M. & Flom	326.5	\$825,338
Wachtell, Lipton, R. & Katz	178.8	\$794,602
Sullivan & Cromwell	214.5	\$756,676
Simpson Thacher & Bartlett	213.7	\$750,024
Davis Polk & Wardwell	153.9	\$570,017
Cleary Gottlieb S. & Hamilton	128.8	\$562,988
Latham & Watkins	275.9	\$544,244
Weil, Gotshal & Manges	180.2	\$490,403
Kirkland & Ellis	226.8	\$457,475
Cravath, Swaine & Moore	139.5	\$426,944

Table 7: Top 10 Law Firms by Apportioned Value of Deals, Buy Side

Law Firm	# Deals	Total Value (\$M)
Skadden, Arps, S., M. & Flom	608	\$2,845,414
Wachtell, Lipton, R. & Katz	333	\$2,311,804
Sullivan & Cromwell	491	\$2,296,923
Simpson Thacher & Bartlett	337	\$2,111,489
Cravath, Swaine & Moore	247	\$1,481,981
Davis Polk & Wardwell	304	\$1,424,276
Latham & Watkins	518	\$1,404,980
Shearman & Sterling	257	\$1,188,282
Fried, Frank, H., S. & Jacobson	210	\$1,151,410
Cleary Gottlieb S. & Hamilton	210	\$1,137,344

Table 8: Top 10 Law Firms by Value of Deals, Sell Side

Law Firm	# Apport. Deals	Apport. Value (\$M)
Skadden, Arps, S., M. & Flom	326.6	\$1,153,566
Wachtell, Lipton, R. & Katz	165.3	\$875,839
Sullivan & Cromwell	210.0	\$810,193
Simpson Thacher & Bartlett	159.8	\$733,628
Latham & Watkins	294.4	\$576,162
Cravath, Swaine & Moore	113.6	\$566,246
Davis Polk & Wardwell	138.5	\$498,225
Shearman & Sterling	114.2	\$445,019
Fried, Frank, H., S. & Jacobson	100.9	\$420,162
Cleary Gottlieb S. & Hamilton	91.8	\$399,916

Table 9: Top 10 Law Firms by Apportioned Value of Deals, Sell Side

References

- Afsharipour, Afra (2010). “Transforming the Allocation of Deal Risk Through Reverse Termination Fees”. In: *Vanderbilt Law Review* 63.
- Ayres, Ian and Robert Gertner (1989). “Filling Gaps in Incomplete Contracts: An Economic Theory of Default Rules”. In: *Yale Law Journal*.
- Burnett, Grace Maral (2019). “#MeToo Reps Becoming M&A Market Standard”. In: *Bloomberg Law*. URL: <https://perma.cc/8JNM-E67T>.
- Choi, Stephen, Robert Scott, and Mitu Gulati (2021). “Investigating the Contract Production Process”. In: *Capital Markets Law Journal*.
- Choi, Stephen J. et al. (2022). “Contract Production in M&A Markets”. In: *University of Pennsylvania Law Review* 171, p. 1881. URL: <https://heinonline.org/HOL/Page?handle=hein.journals/pnlr171&id=1915&div=&collection=>.
- Coase, R. H. (1937). “The Nature of the Firm”. In: *Economica* 4.16, pp. 386–405. DOI: 10.1111/j.1468-0335.1937.tb00002.x.
- Coates, John C. (2016). “Why Have M&A Contracts Grown? Evidence from Twenty Years of Deals.” In: *Harvard Law School John M. Olin Center Discussion Paper*.
- Davidoff, Steven M. (2007). “The Return of the Tender Offer”. In: *M&A LAW PROF BLOG*. URL: <https://perma.cc/3P6C-DT9U>.
- Efrati, Amir (2011). “Google Paid \$125 Million for Zagat”. In: URL: <https://www.wsj.com/articles/SB10001424053111904836104576560751396246430>.
- FactSet (n.d.). “FactSet Mergers”. In: (). URL: <https://www.factset.com/marketplace/catalog/product/factset-mergers>.
- Fisher, Daniel I. (2013). “DCGL Section 251(h): Top-Up Option No Longer Needed”. In: *AG DEAL DIARY*. URL: <https://perma.cc/V72Z-ZZPZ>.
- Fontenay, Elisabeth de (May 2020). *Windstream and Contract Opportunism*. Duke Law School Public Law & Legal Theory Series No. 2020-44. DOI: 10.2139/ssrn.3603752. URL: <https://ssrn.com/abstract=3603752>.
- Frankenreiter, Jens et al. (Feb. 2021). “Cleaning Corporate Governance”. In: *University of Pennsylvania Law Review* 170, p. 1. URL: <https://ssrn.com/abstract=3796628>.
- Greif, Avner (June 1, 1993). “Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders’ Coalition”. In: *The American Economic Review* 83.3, pp. 525–548. ISSN: 0002-8282. URL: <http://www.jstor.org/stable/2117532> (visited on 06/23/2015).
- Grossman, Sanford J. and Oliver D. Hart (Aug. 1986). “The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration”. In: *Journal of Political Economy* 94.4. Publisher: The University of Chicago Press, pp. 691–719. ISSN: 0022-3808. DOI: 10.1086/261404. URL: <https://www.journals.uchicago.edu/doi/abs/10.1086/261404> (visited on 02/17/2024).
- Gulati, Mitu and Robert E. Scott (2012). *The Three and a Half Minute Transaction: Boilerplate and the Limits of Contract Design*. University of Chicago Press. ISBN: 9780226924380.
- Hart, Oliver and John Moore (1988). “Incomplete Contracts and Renegotiation”. In: *Econometrica*.

- Hill, Claire A. (2001). “Why Contracts Are Written in Legalese Symposium: Theory Informs Business Practice”. In: *Chicago-Kent Law Review* 77.1, pp. 59–86. URL: <https://heinonline.org/HOL/P?h=hein.journals/chknt77&i=75> (visited on 03/12/2022).
- Hoffman, David (2014). “Whither Bespoke Procedure?” In: *University of Illinois Law Review*.
- Jennejohn, Matthew, Julian Nyarko, and Eric Talley (2022). “Contractual Evolution”. In: *University of Chicago Law Review* 89 (4). URL: <https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=6296&context=uclrev>.
- Kincaid, JP, RL Rogers, and BS Chissom (1975). “Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel”. In.
- Landa, Janet (1981). “A Theory of the Ethnically Homogenous Middleman Group: An Institutional Alternative to Contract Law”. In: *Journal of Legal Studies*.
- Li, Zehua, Neel Guha, and Julian Nyarko (2023). “Don’t Use a Cannon to Kill a Fly: An Efficient Cascading Pipeline for Long Documents”. In.
- Monson, Bryan (2015). “Note, The Modern MAC: Allocating Deal Risk in the Post-IBP v. Tyson World”. In: *Southern California Law Review*.
- Nyarko, Julian (Jan. 2021). “Stickiness and Incomplete Contracts”. In: *University of Chicago Law Review* 88.1, pp. 1–79. ISSN: 0041-9494. URL: <https://chicagounbound.uchicago.edu/uclrev/vol88/iss1/1>.
- Nyarko, Julian and Eric Talley (2022). “Corporate Climate: A Machine Learning Assessment of Climate Risk Disclosures”. In: *Business Law and the Transition to a Net Zero Economy*. Ed. by Andreas Engert et al. Chap. 1.
- Pandya, Sneha and Eric L. Talley (Jan. 2023). “Debt Textualism and Creditor-on-Creditor Violence: A Modest Plea to Keep the Faith”. In: DOI: 10.2139/ssrn.4317353. URL: <https://ssrn.com/abstract=4317353>.
- Quinn, Brian J. M. (2010). “Optionality in Merger Agreements”. In: *Delaware Journal of Corporate Law* 35 (3).
- Williamson, Oliver (1985). “The Economic Institutions of Capitalism”. In: *Free Press*.