The Information Costs of Organization: NLP Measures of Decision Making Capacity

M.J. Histen
College of Wooster

Coase's puzzle of economic organization asks why can't a larger firm do everything a smaller firm can do and more. Most studies underrepresent problems in decision making capacities which have been difficult to quantify. This study corroborates a more complete theory of firm using applications from natural language processing and information theory. Probit models regress an explanatory variable from textual analysis over annual reports against a discrete measure of organizational boundaries from financial disclosures. These estimates provide a cost calculus for organizing a marginal transaction internally or through markets, offering further implications for policy and managerial practice.

INTRODUCTION

When Coase (1937) studied the make versus buy decision, he formed the canonical problem of economic organization by proposing firms and markets as "alternative methods of coordinating production". Neither markets nor firms operate costlessly. They differ in structural ways and integration depends on the nature of the transaction. Williamson (1985) argued for asset specificity as most critical of transaction costs. If an asset's next best use is substantially less than the value of the current contract, the transaction is subject to expropriation by holdup. Vertical integration can assuage such hazards, and so asset specificity has seen the most testing in the empirical literature (Richman and Macher 2006). Results have been largely consistent with the theory with many unique measures of asset specificity being positively correlated with integration.

But Coase (1988) saw asset specificity as a weak foundation for the theoretical entirety of the firm. While Williamson proposed other critical dimensions of transactions costs, they hold considerably less empirical support. Of these, uncertainty has eluded rigorous treatment (Masten 2016). Uncertainty is defined as "unanticipated changes in circumstances surrounding an exchange" (Noordewier et al. 1990). Economic actors face limits on information processing capacity (Van Zandt 2000), causing coordination problems in developing a common language to contract on states of the world (Hart 1995). Organizing exchange internally requires firms to translate complex data from an uncertain environment into a profitable flow of outputs. The very idea of management serves this role (Langlois 2003). But management's capacities to respond to uncertainty operate under the constraint that collection, calculation, and communication are scarce (Arrow 1996). This prioritizes information costs as a significant component in transaction costs analysis in Coase's (1937) firm, in Knight's (1921) risk sharing, and in agency theories of the firm (Alchian and Demsetz 1972; Jensen and Meckling 1976).

Transaction costs stem from economic actors' limitations in processing information (Simon 1955). Acquiring information is expensive and so is communicating it. The boundary between firms and markets

forms as a least cost manner of conducting transactions along informational dimensions (Monteverde 1995). Yet empirical results have been uneven and contradictory in operationalizing capacities in response to uncertainty relative to organizational modes (see Richman and Macher (2006) for a survey). Quantifying decision-making capacity empirically has been notoriously difficult (Van Zandt 2000; Masten 2016). Here, I attempt an old problem with new techniques by measuring management's information capacities through textual artifacts. Through applications from natural language processing (NLP) and information theory, I measure relative differences in information sets as firms specialize in unique stocks according to their capabilities and industry. From a sample of 5000 firms, I regress this measure against an ordered integration variable compiled from a required footnote disclosure on operating stages. Both variables are derived systematically to minimize the author's estimates, which have been idiosyncratic and subjective in other studies (Richman and Macher 2006).

The larger the number of stages, the larger the number of different stocks of information, increasing the reliance on direction by management (Demsetz 1988). When the internal costs exceed some threshold, the stock is more cheaply housed across firms than internally. The findings here show where these dividing lines occur—quantifying Coase's initial description of management's diminishing returns as the countervailing force to one economy-sized firm. The probit estimates give a cost calculus for organizing a marginal transaction internally or through markets. The marginal effects estimate these thresholds to describe the information costs along each stage of integration, showing the cost of information increases at an increasing rate. These findings further support empirical approaches to transaction cost economics and text-based tools for organizational decision modeling. The distance between information bases has several implications for policy and managerial practice. Overall, the results and methodology point to several interesting avenues for future research.

The paper proceeds as follows. Section 2 presents the background context of organizations as information processors and related literature. Section 3 describes the data, the financial disclosure for the dependent variable, and the statistical measure of information, then carries out the empirical analysis. Section 4 interprets these results, and Section 5 gives concluding remarks.

THEOETICAL AND EMPIRICAL BACKGROUND

The modern firm as a going concern has long exceeded the value of its physical assets largely because the market assesses the value of the information base embedded within it (Arrow 1996). This information base depends not only on recipes of production or technical knowledge; it includes awareness of where complementary expertise lie and how to communicate it into the firm's decision making (Cohen and Levinthal 1990). Activating it requires dedicating nearly one-half of US workers (including managers) to information processing activities rather than production (Radner and Van Zandt 1992). These activities are distributed throughout the firm and the economy. All economic organization must solve the collocation of information and decision making either by moving the information to those with the decision rights, or the decision rights to those with the information (Jensen and Meckling 1976). Markets tend to operate by the latter, while firms solve the collocation problem by centralizing information through hierarchies. Paraphrasing Knight (1921), if "workers do, and managers tell them what to do," the firm must hire specialists to be in charge of processing information into actions taken by workers (Radner 1993). The movement from market to firm involves trading off the costs of centralizing information into instruction. The decision calculus over firm boundaries, then, conserves on information costs—costs changing with the scope of activities in the hierarchy.

Marshall identified the outcome of increasingly complex firms, noting "the development of the organism, whether social or physical, involves an increasing subdivision of functions between its separate parts on the one hand, and on the other a more intimate connection between them" (Marshall 1948). Nor can the design of such connective structures be disentangled from the distribution of expertise (Cohen and Levinthal 1990). The organization as a whole must have some degree of relevant overlapping background information. Fundamentally, effective communication within and across subdivisions consists of shared language and symbols (Dearborn and Simon 1958). Technical information from an R&D project must be

communicated in an understandable form to the accounting department. Management acts as this necessary interface; it specializes in economizing on the transmission of information (Arrow 1996). Management coordinates people, departments, and projects by processing information into a shared language. In public companies, a sufficient understanding of the firm's technical expertise must be communicated across all parties including capital providers and potential investors. Ultimately, management faces the extraordinary task of representing the information base embedded within the firm.

However, this information base is not management itself. No employees are permanently attached to a company, and the relevant information must overlap across time as well as people. The information base peculiar to a firm, then, is neither entirely in its physical assets nor its workforce inputs (Arrow 1996). It is partially a distinct structure reproducing itself through codifications of routines that form the regular patterns of the firm (Nelson and Winter 1982). These are the "invisible" assets accumulated through experience and refined by practice that shape a common language to communicate relevant information into actions across the firm and in perpetuity (Itami 1987). If the information base is a defining characteristic of the firm, processing it is a defining characteristic of management. They do so by interpreting dispersed expertise into shared codes that can communicate technical information all the way through to potential investors.

But there are limits on gathering, processing, and communicating information (Simon 1955), costs that reflect the basic reason why knowledge in society is dispersed (Hayek 1945). The ideal information base suggests an organizational trade-off between diversity and commonality of knowledge where expertise cannot be pushed so far as to undermine communication (Cohen and Levinthal 1990). Industries and firms are repositories of specialized codes that put information into work and they specialize in different stocks of information. As these codes become distinct, they form into distinct organizations. Airline firms specialize in different stocks of information than their manufacturers, and the formation of boundaries between them must be considered from the perspective of conserving on information costs (Demsetz 1988).

Such a perspective restates the canonical problem in organizational economics: where is the boundary between firms and markets as "alternative methods of coordinating production" (Coase 1937)? Asked another way, why can't a larger firm do everything a smaller firm can do and more (Knight 1921)? Supposing two stages of production are combined, if the acquired stage operates the same post-acquisition as pre-acquisition by replication, and the acquiring stage selectively intervenes when net gains can be ascribed to coordinated adaptations, then the combined firm can never do worse and will sometimes do better (Williamson 1985). Thus, the larger firm will realize greater value (Lewis 1983). This puzzle depends on replication and selective intervention assumptions. Williamson (1985) countered it by arguing that holding performance incentives constant between pre and post acquisition stages was "delusional". Grossman and Hart (1986) also emphasized differences between pre and post acquisition behaviors along investment incentives. In fact, most studies have equated the study of organizations with the study of incentive problems (Garicano 2000), focusing exclusively on a breakdown in the replication assumption.

These answers follow the footprints of Coase's insight into a world of positive exchange costs where some forms of governance perform better at organizing transactions than others. Though incentive problems correctly assume markets do not operate costlessly, they gloss over the heroic assumption embedded within selective information. By assuming the required information is freely in hand and the required calculations are costless to make, management is stripped of a meaningful role in interpreting and acting on information. This assumes away the limits on gathering, processing, and communicating information, which in effect assumes away nearly one-half of US workers. Even traditional price theory fails to consider the information costs associated with the selection of profit-maximizing quantities of outputs and inputs. The scope of the firm depends on theoretical and empirical research that more thoroughly model these information costs.

A rich body of literature attempts to theoretically model uncertainty and information processing through the bounded rationality of managers (Williamson 1985; Radner 1993; Garicano 2000). Operations research and management science literatures further examine decomposed decision designs for firms (see Rogers et al. (1991) for a survey). However, quantifying decision-making capacity empirically has been notoriously difficult (Van Zandt 2000; Masten 2016). Attempts to test this relationship have produced uneven and contradictory results in operationalizing this capacity relative to organizational modes (see Richman and

Macher (2006) for a survey). But NLP techniques tap into a new resource for measuring management's information capacities. Qualitative aspects of text contain rich information about the operations of firms but are challenging to represent. Textual analysis provides a channel to systematically account for and interpret such information. Through these methods, this study makes a new attempt at empirically substantiating information processing to provide a more complete theory of the firm.

EMPIRICAL INVESTIGATION AND RESULTS

Data

The notion of parsing text for patterns dates back at least to the 1300s when friars of the Dominican Order produced concordances of the Latin Vulgate with indexes of common phrases (The Catholic Encyclopedia 1908). More recently, the accounting and finance literature have given considerable attention to the annual 10-K filings with the Securities and Exchange Commission (SEC) (see Loughran and McDonald (2016) for a survey). Information plays a key role in a firm's operations and how markets assess its value. Much of the published work has focused on readability where researchers unpack the assimilation of information from text into asset prices (Li 2008; Guay et al. 2015; Lundholm et al. 2014). Many of these papers track a lexicon through documents to detect a sentiment or construct an index to correlate with stock prices (Henry 2008; Price et al. 2012; Loughran and McDonald 2015). For example, "non-GAAP" might signal an attempt by management to talk around financial performance (Black et al. 2017), which can then be correlated with future earnings. Authors have also applied similarity measures to detect document changes over time (Cohen et al. 2019). Monteverde (1995) makes an early attempt at quantifying text in the transaction costs literature. In his study, Monteverde proposes a measure on language use in semiconductor designers through reader judgment with a binary integration variable on a sample of 23 firms, providing support for the influence of information costs on firm boundaries.

I use the laboratory of firm annual 10-K filings to test a measure of management's information processing capacity from a sample of 5000 firms and regress it against an ordered integration variable compiled from audited financial statements. Both variables are derived systematically to minimize the author's estimates as follows. I pull all complete 10-K filings from the SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) website for the fiscal yearend 2016. I chose this date because of an updated reporting standard effective December 15, 2016. These audited filings are in HTML text format and contain an aggregation of all information submitted for each firm, such as exhibits, graphics, XBRL files, and more.

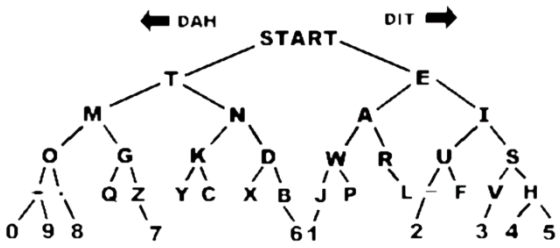
Similar to Cohen et al. (2019), I concentrate on the textual content of the document, specifically focusing on the Management Discussion and Analysis (MD&A) section. US public companies are required by the SEC to include a number of schedules including an MD&A, which serves as the major communications medium regarding information about the business. Management describes in detail production, procedures, performance, projections, and more—in short, its attempt to codify the firm's information base. I lemmatize the text to minimize ambiguity through Wordnet, a lexical data for English arranging subnets into hierarchies (Princeton University 2010). I also remove stop words with Python's NLTK standard English library plus an additional lexicon of generic balance sheet words corresponding to dates and written numbers, both included in Appendix 1. As is common in textual analysis, I rely on the bag-of-words technique which makes a critical assumption on independence to reduce the extraordinary dimensionality of a document (Loughran and McDonald 2016). Independence means the sequence of words are ignored and the analysis instead focuses on the distribution of terms. Because word counts often follow a power law distribution, certain words have particularly large impacts on the results. The level of uncertainty associated with a word based on its frequency in the library allows for a statistical measure to account for its information cost.

Information can be understood as a fundamental unit of communication, and this unit can be generalized and compared the same way as mass can be across physical objects. The bit, the common representation of this unit, gives a measure of "surprise". Information theory quantifies the difference in surprise to uncover an economy of communication. Given an alphabet (or card deck or dictionary or library), bits measure the

average number of yes-or-no questions needed to determine the exact symbol from that alphabet. Most symbols in the alphabet will not have the same frequency, so Shannon (1948) formalized the problem to minimize the number of questions needed to predict the draw. From there, codes are assigned to more frequent symbols to minimize transmitting costs.

Consider Figure 1, a graphic of the Morse code decision tree (Roomberg 2020). Samuel Morse, Joseph Henry, and Alfred Vail encoded English alphanumeric characters into dits and dahs to transmit them along an electric current and communicate natural language. Each pulse was costly, so Vail sought to minimize this by counting the frequency of English letters in the type-cases of a local newspaper in Morristown (Burns 2004). He then encoded characters according to this distribution. As a result, it takes less information to send "E" (one dit) than "H" (four dits). Symbols that are more surprising take more information to communicate. The price of a message depends on the amount of information needed to transmit it.

FIGURE 1
MORSE CODE DECISION TREE



Average bits measure uncertainty as a weighted height of the decision tree. Shannon named such a measure entropy (H) and generalized the formula to $H = -\sum_{i=1}^{n} p_i \times \log_2 p_i$. Entropy in effect quantifies uncertainty, disorderliness, or irregularity in a system which can be statistically compared across different stocks of data (Ekhosuehi and Osagiede 2012). Shannon entropy is the theoretical limit on the minimum number of bits needed to send a message. David Huffman (1952) developed a lossless optimal encoding algorithm as follows. Take the least two probability nodes, merge them by adding probabilities together, then take the next two smallest (including the merged one from before). Repeat until only a single node at the top remains and label the edges with zero or one in any order. Huffman coding applies shorter codes to more common symbols to minimize information costs, exactly as Vail attempted with the English alphabet in Morse code.

For example, consider two libraries and their frequencies, A and B respectively, as follows: $\{('a', 4), ('b', 2), ('c', 1), ('d', 1)\}$ and $\{('a', 2), ('b', 2), ('c', 2), ('d', 2)\}$. To minimize the number of bits needed to determine the exact letter, the best approach would be to ask questions that eliminate half the possibilities. With library A, an optimal machine would ask if the symbol is $\{'a'\}$; if not, is it $\{'b'\}$; if not, is it $\{'c'\}$; and thus fully span all possibilities. The average number of bits needed to determine a word would be the probability times the its distance down the decision tree; that is, the $P('a') \times 1 + P('b') \times 2 + P('c') \times 3 + P('d') \times 3 = 1.75$. It takes on average 1.75 bits to communicate a word in this library. A machine performing the same exercise on library B would yield 2 bits per letter. Because B is "more random", it takes more information to communicate a sequence. A machine for library A costs less information for the sequences of the same length.

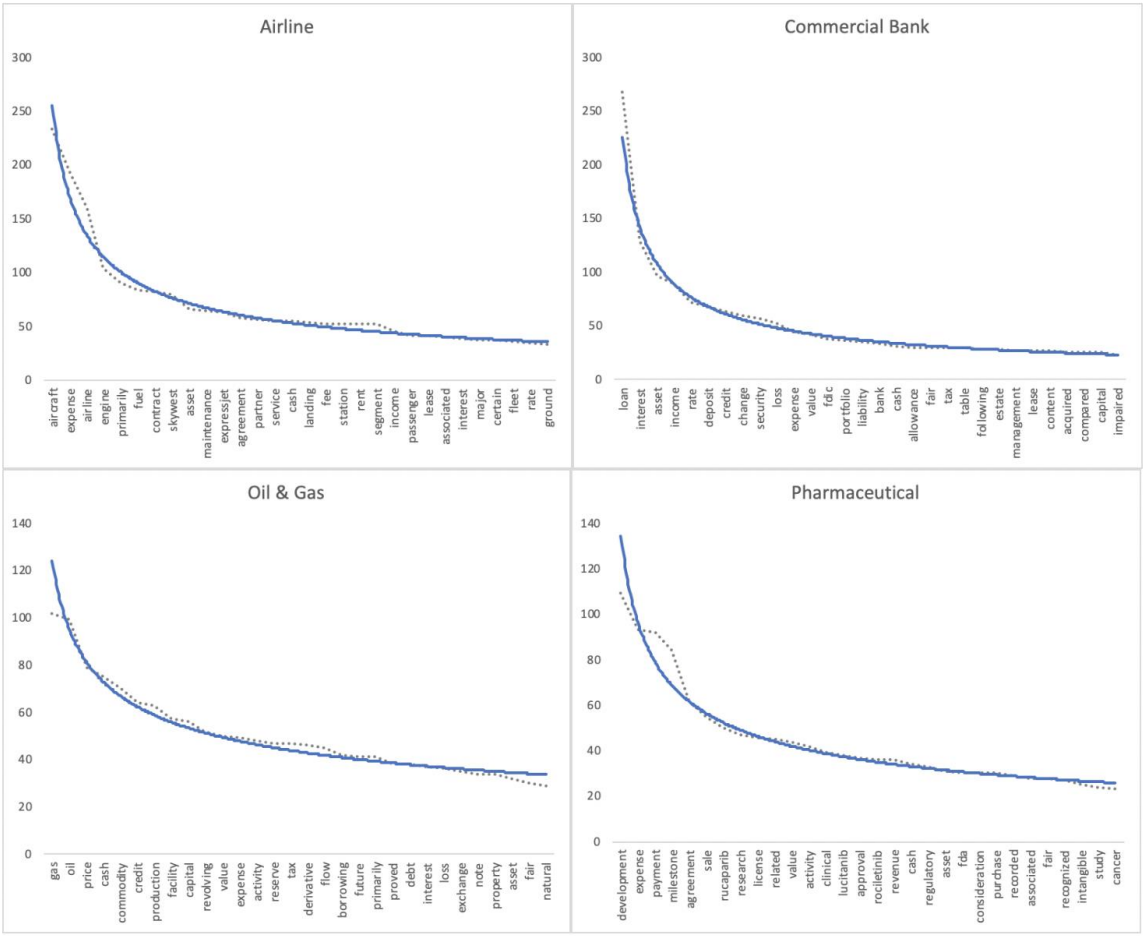
I apply such information measures to management discussion as an explanatory variable in firm boundaries. Some amount of redundancy in expertise is needed to communicate across subunits of the firm (Cohen and Levinthal 1990) which is precisely what information theory detects. Though none of the textual analysis papers cited above has directly leveraged an entropy measure, information theory has found many applications in economics (see Yang (2017) for a survey). Figure 2 displays the top thirty word counts for four firms in the sample. All disclosures are written in English, but the priority and frequency of word usage reflects a firm specific emphasis in stocks of information, and this organization of information can be quantified by their statistical variances. The information needed to communicate "aircraft" in an airline is

less costly than communicating it in an oil refinery firm—in precisely the same manner that "a" is less costly in library A than B. Recall that the power law distribution implies certain words can have significant impact.

The opposite case holds for "pipeline". Similarly, communicating "loan" in a bank is cheaper than other firms from the degree of overlapping background knowledge. Persistent discussion of terms in companies' annual reports, then, means very different things. An information theoretic measure compares such interpretations more generally. By measuring the differences, I can measure the price of management's information. That is, what are the relative differences in usage between such firms, and what happens when an aviation company owns a refinery? Comparison of these figures captures the magnitude of management's capacities in response to the scope activities within the firm.

FIGURE 2
SAMPLE FIRM WORD COUNTS

Dotted lines are raw data; solid lines are power law fits.



Airline firms specialize in different stocks of information than their suppliers. In fact, organizations have even been modeled as simply partitions of information sets (Garicano 2000). However, these stocks need to be considered along with the scope of operations to detect changes in the costs of information. I build an ordinal measure of organizational boundaries with a financial statement footnote disclosure for all companies filing with the SEC as required by ASC 280, segment reporting. The "reportable segments" disclosure requires management to aggregate operating segments with similar operating criteria based on the nature of the products or their production processes, class of customer, method of distribution, or

regulatory environment provided they are 10% of revenues, income, or assets (FASB, ASC 280-50-1). Operating segments are based on how management organizes their enterprise, defined as a business activity that earns revenues or expenses, has results reviewed by a chief operating decision maker, and has discrete financial information (FASB, ASC 280-50-1). That is, effective yearend 2016, reportable segments correspond to autonomous business units. These criteria align with Williamson's selective intervention thought experiment; they represent different material stages of operation in an overarching firm.

For example, in the data Delta Airlines, Inc. discloses two reportable segments: the airline segment, which provides air transportation for passengers and cargo, and the refinery segment, which includes pipelines and terminal assets to supply jet fuel. Alternatively, American Airlines Group, Inc. is managed as a single unit that provides air transportation for passengers and cargo. Lions Gate Entertainment Corp. reports three segments: motion picture production, television production, and media networks. About half of all firms consider themselves one segment (Table 1A and Table 1B below provide distribution data). The firm with the highest number of segments in the population, World Wrestling Entertainment, Inc., discloses nine segments ranging from media networks to live events to consumer merchandise like toys and attire.

Methodology

Firms allocate resources through instruction while markets allocate through prices. Management's job, therefore, is to convert information into instruction. The more complex these codes, the more costly it is to process information. Note that aggregation of information is not the right metaphor; instead the measure accounts for the costs of organizing information into firm specific codes. Industries and firms can be identified as repositories of specialized information put to work (Demsetz 1988). Management becomes more difficult with multiple, complex, or unfamiliar processes, demanding a greater share of their limited capacities and becoming more expensive to administer (Masten et al. 1991). When the costs exceed some threshold, the information stock is more cheaply housed across firms and exchanged through markets.

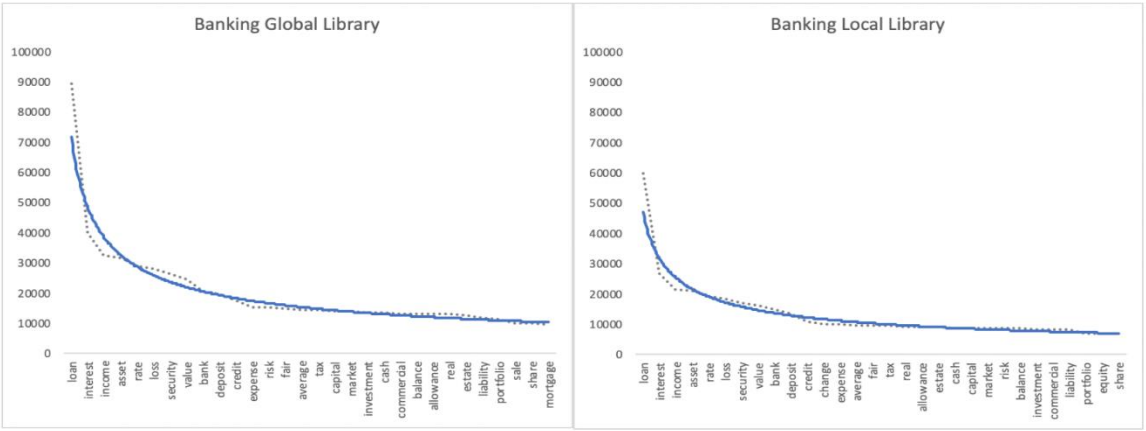
I measure the relationship between organizational scope as a discrete dependent variable and information costs as an explanatory variable through ordered probit models that control for competing explanations of integration. I take the dependent variable, organizational scope, from a financial statement footnote disclosure for all companies making filings with the SEC as required by ASC 280, segment reporting. Reportable segments correspond to autonomous business units in Williamson's selective intervention thought experiment; they represent different material stages of operation in an overarching firm.

I take the explanatory variable from a statistical measure of management information relative to the industry information stock. Each SEC filing must specify a Standard Industry Classification (SIC) identifier that describes the firm's primary industry of operation. I tally these into industry libraries by combining frequency distributions of all words by all firms in each industry. I actually construct two libraries for each industry: a global library and a local library. In the global library, I take the entire corpus of words by all firms in each SIC, encode it via the Huffman algorithm, then calculate an information cost per firm per the encoded global library. In the local library, the corpus is composed only of firms with a single reportable segment, then encoded, then an information cost is calculated for each firm per the encoded local library. Under this regime, not every firm contributes to the knowledge stock, so if a word doesn't appear in the library, it is assigned the maximum value in the tree. For example, if the SIC is "National commercial banks," the library is built only from firms that operate as national commercial banks alone. Firms that operate as national commercial banks and insurers, then, will have information related to their insurance practice more heavily penalized than in the global regime.

This second library therefore captures a more specific industry information stock. Figure 3 below provides examples for the banking global and local libraries, respectively, at the four digit SIC. The frequency distributions form the cumulative library from every firm in the classifier (an aggregation of the examples from Figure 2 by industry). These are encoded into an optimized decision tree by the Huffman algorithm (similar to letter organization in the Morse code diagram from Figure 1). Each MD&A is then run through the tree to determine an entropic weight as an estimate of the firm's specific information base.

FIGURE 3
GLOBAL AND LOCAL INDUSTRY LIBRARIES, RESPECTIVELY

Dotted lines are raw data; solid lines are power law fits.



SIC codes have a top down structure beginning with general characteristics and narrowing to more specific industries. This means two digit codes represent major industries while four digit codes describe more specialized business groupings. I adjust the granularity of industry by moving through four digit, three digit, two digit, and letter categorizations of SICs in the series of regressions below. For example, banks are categorized in the following regime: 4-digit: {(6021, "National commercial banks"), (6022, "State commercial banks), (6153, "Short-Term Business Credit Institutions"), (6162, "Mortgage Bankers & Loan Correspondents")}; 3-digit: {(602, "Commercial banks"), (615, "Business Credit Institutions), (616, "Mortgage Bankers")}; 2-digit: {(60, "Depository institutions"), (61, "Non-depository institutions"}; and letter category: {(H, "Finance, insurance, & real estate")}. Because four digit SICs are the most precise, some industries have few firms. To ensure a robust explanatory variable, I only include industries with at least 50 firms with no firm comprising more than 5% of the library. This implies a stricter condition at the local library level since only one segment firms are included, of which there are less. The full tables by SIC with library word counts are included in Appendix 2.

Table 1A presents summary data on the dependent variable by SIC for the global library, and Table 1B presents summary data for the local library. Sample size increases as SIC digit decreases because the industry category broadens to include more firms. Firms with five or more reportable segments were combined into the last bucket to estimate the model due to their infrequency, and in some probit runs, firms with four or more reportable segments were combined.

TABLE 1A
SUMMARY DATA ON GLOBAL DEPENDENT VARIABLE

	4 SIC		3 SIC		2 SIC		L SIC	
Stage	Count	%	Count	%	Count	%	Count	%
1	1394	73%	1657	71%	2395	61%	2918	57%
2	219	12%	322	14%	675	17%	941	19%
3	183	10%	233	10%	504	13%	724	14%
4	102	5%	135	6%	339	9%	500	10%

TABLE 1B
SUMMARY DATA ON LOCAL DEPENDENT VARIABLE

	4 SIC		3 SIC		2 SIC		L SIC	
Stage	Count	%	Count	%	Count	%	Count	%
1	1390	76%	1618	74%	2247	65%	2900	58%
2	193	11%	264	12%	542	16%	932	19%
3	155	8%	192	9%	415	12%	710	14%
4	86	5%	108	5%	269	8%	480	10%

To test the generality of the relationship between information costs and firm boundaries firms, I draw a significant sample of SEC filing firms for the fiscal 2016 yearend. The SEC maintains the EDGAR site for all filed financial data sets by quarter from January 2009 and beyond. I collected all filing firms by central index key (CIK) from 2016 and pulled their annual statements for balance sheet data, reportable segments disclosures, SICs, and text from the MD&A section. Because I wish to test organizational choices among successful firms, I take an operational definition of "success" as material financial condition. I only include firms with more than $100K in assets or liabilities for a focused sample. Though this drops about 500 observations, most are shell companies or firms with minimal filing information. For the 5000+ remaining in the population, I run a series of ordered probit models with organizational choice as the dependent variable and the information cost measure as the explanatory variable. This is the typical structure in the literature and I argue for the direction of the relationship through Williamson's (1996) remediableness criterion. An extant mode of organization can be presumed to be efficient, or in the words of the Chicago school, "what exists is ultimately the best guide to what should exist" (Eisner 2017). My claim is that the coefficients on information costs measure efficient firms, and their organizational choices correspond to efficient boundary conditions.

I include the following variables to control for other possible explanations for integration. Firm size is a common control in studies predicting levels of integration in the transaction cost framework (Pisano 1990; Saussier 2000). I control for it by total assets. A natural assumption is that larger firms have more integration given their growth strategies and leverageable assets. Surprisingly, firm size was not correlated with information cost measures. Many large firms are so specialized as to operate as a single or few segments. Following Loughran and McDonald (2014), I also use MD&A word count as a control to account for some firms and industries with MD&A sections spanning tens of thousands of words (e.g., depository institutions). I use a simple leverage ratio to control for asset specificity as an account for the proportion of assets relative to firm value (Henisz 2000). Stronger measures of asset specificity and industry heterogeneity are controlled for through industry fixed effects.

An ordered probit model is appropriate because there is a natural ordering of the dependent variable (from less integrated to more) but the values only reflect a ranking. Note that a significant underlying assumption of an ordered probit regression is that the relationship between each pair of outcome groups is the same. This proportional odds assumption allows for one model to describe the relationship. A Brant test for all probit runs confirms the proportional odds assumption was not violated. However, there are not enough degrees of freedom to conduct the test for every SIC granularity with four stages of integration, so these results are excluded. Moreover, probit regressions are commonly used to model this problem and provide comparable interpretations of the results with other studies (Monteverde 1995; Kalnins and Mayer 2004; Loughran and McDonald 2016). To ensure model specification, I conduct a number of robustness checks, including an ordered logit and a substituted explanatory variable through cosine similarity, another comparison tool for textual artifacts. All robustness checks support the regression results below and available by request.

Results

The results of six probit runs are summarized in Table 2A under the global library and support my argument that firm boundaries congeal around management's information processing capacities. The information measure exhibits a statistically significant association with organizational structure after controlling for other factors. Moreover, the control variables perform as the theory suggests, further corroborating the model. Because probit models differ by a scale factor, the magnitude of the effects must be interpreted through marginal analysis. Recall that the dependent variable refers to the number of reportable segments in the firm and the explanatory variable tracks the information costs measure of its MD&A.

TABLE 2A
GLOBAL PROBIT RESULTS

	Dependent Variable Regime					
	3 Stage Boundary				4 Stage Boundary	
SIC digits	4	3	2	L	2	L
Information	2.1299*	1.9279*	1.3933*	0.8872*	1.4614*	0.9385*
	(0.3519)	(0.2868)	(0.2140)	(0.1669)	(0.2074)	(0.1616)*
Log Size	0.5012*	0.4654*	0.1455*	0.4954*	0.1449*	0.4876*
	(0.0382)	(0.0316)	(0.0067)	(0.0188)	(0.0064)	(0.0181)
Log Leverage	0.3526*	0.3283*	0.1103*	0.3399*	0.1091*	0.3227*
	(0.1039)	(0.0820)	(0.0167)	(0.0474)	(0.0165)	(0.0466)
Log Word count	0.0494*	0.0772*	0.0858*	0.3064*	0.0879	0.3121*
	(0.1188)	(0.1022)	(0.0223)	(0.0630)	(0.0216)	(0.0610)
Intercept 1	11.2903	10.3804	9.8081	8.2625	10.0820	8.4026
	(1.1688)	(0.9766)	(0.7452)	(0.5751)	(0.7210)	(0.5553)
Intercept 2	11.8115	10.9594	10.4433	8.8977	10.7167	9.0373
	(1.1711)	0.9787	(0.7468)	(0.5765)	(0.7226)	(0.5566)
Intercept 3	n/a	n/a	n/a	n/a	11.4259	9.7427
					(0.7249)	(0.5584)
Industry Controls	yes	yes	yes	yes	yes	yes
No of obs	1,898	2,347	3,913	5,083	3,913	5,083
Pseudo R^2	0.167	0.152	0.160	0.141	0.143	0.125
** Indicates significance at .01 level. Standard errors in parentheses.*						

Table 2B displays marginal effects for the explanatory variables which I interpret as the limits of management's processing capacity per level of integration. Nonintegrated firms have lean information costs; management efficiently communicates core topics of the industry's knowledge stock. For all probit runs, as integration increases, the cost of information increases at an increasing rate. Given the remediableness argument above and examining the four digit SIC, successful two stage integration shows management's information costs increase by 16% on average, and successful two to three stage integration shows management's information costs increase 40% on average. Examining the letter digit SIC, successful two stage integration shows management's information costs increase 6% on average, successful two to three stage integration shows management's information costs increase 10% on average; and successful

three to four stage integration shows management's information costs increase 14% on average. The marginal effects are smaller as SIC increases because the industry stock of knowledge expands. A 16% information cost increase in "National commercial banks" knowledge stock cannot be directly compared to a 6% increase in "Finance, insurance, & real estate" knowledge stock. To simply compare word counts, the former library holds 1.5 million versus 12.5 million in the latter.

TABLE 2B
GLOBAL PROBIT MARGINAL EFFECTS FOR INFORMATION VARIABLE

	Dependent Variable Regime					
	3 Stage Boundary Mfx				4 Stage Boundary Mfx	
SIC digits	4	3	2	L	2	L
Pr(1)	-0.5562*	-0.5377*	-0.4228*	-0.2829*	-0.4455*	-0.3007*
	(0.0896)	(0.0780)	(0.0641)*	(0.0529)	(0.0623)	(0..0515)
Pr(2)	0.1613*	0.1639*	0.1009*	0.05627	0.1068*	0.0604*
	(0.0275)	(0.0250)	(0.0158)	(0.0110)	(0.0156)	(0.0108)
Pr(3)	0.3950*	0.3739)	0.3219*	0.2266*	0.1534*	0.1049*
	(0.0649)	(0.0553)	(0.0490)	(0.0423)	(0.0222)	(0.0182)
Pr(4)	n/a	n/a	n/a	n/a	0.1852*	0.1354*
					(0.0268)	(0.0234)
** Indicates significance at .01 level. Standard errors in parentheses.*						

These results become significantly stronger in the local library. Recall this regime includes firms that exist in a single SIC when constructing the library, implying a more specific industry knowledge stock. Table 3A summarizes the results of six probit runs under the local library and offers even stronger support for my argument. The information measure is again statistically significant and the control variables perform as theory suggests.

TABLE 3A
LOCAL PROBIT RESULTS

	Dependent Variable Regime					
	3 Stage Boundary				4 Stage Boundary	
SIC digits	4	3	2	L	2	L
Information	4.1854*	3.8301*	3.9248*	2.9256*	3.8360*	2.8345*
	(0.3348)	(0.2799)	(0.2078)	(0.1512)	(0.1984)	(0.1449)
Log Size	0.4965*	0.4727*	0.4898*	0.4987*	0.4757*	0.4848*
	(0.0404)	(0.0340)	(0.0240)	(0.0187)	(0.0229)	(0.0179)
Log Leverage	0.3094*	0.3252*	0.3241*	0.3736*	0.3120*	0.3519*
	(.1055)	(0.0865)	(0.0575)	(0.0467)	(0.0563)	(0.0458)
Log Word count	-0.1054	-0.0255	0.1170	0.1753*	0.1246	0.1818*
	(0.1247)	(.1094)	(0.0800)	(0.0638)	(0.0771)	(0.0614)
Intercept 1	17.3398	16.2218	17.1958	14.4487	16.8200	14.0775

	(1.1288)	(0.9637)	(0.7170)	(0.5238)	(0.6811)	(0.4979)
Intercept 2	17.8850	16.8032	17.8387	15.1232	17.4610	14.7502
	(1.1333)	(0.9677)	(0.7206)	(0.5266)	(0.6848)	(0.5008)
Intercept 3	n/a	n/a	n/a	n/a	18.2001	15.4863
					(0.6890)	(0.5038)
Industry Controls	yes	yes	yes	yes	yes	yes
No of obs	1,824	2,182	3,473	5,022	3,473	5,022
Pseudo R^2	0.209	0.190	0.203	0.174	0.180	0.152
** Indicates significance at .01 level. Standard errors in parentheses.*						

TABLE 3B
LOCAL PROBIT MARGINAL EFFECTS FOR INFORMATION VARIABLE

	Dependent Variable Regime					
	3 Stage Boundary Mfx				4 Stage Boundary Mfx	
SIC digits	4	3	2	L	2	L
Pr(1)	-0.9689*	-0.9488*	-1.0818*	-0.8871*	-1.0711*	-0.8692*
	(0.0694)	(0.0622)	(0.0504)	(0.0420)	(0.0491)	(.0410)
Pr(2)	0.3044*	0.3087*	0.2828*	0.1914*	0.2838*	0.1901*
	(0.0286)	(0.0258)	(0.0174)	(0.0118)	(0.0174)	(0.0117)
Pr(3)	0.6644*	0.6402	0.7990*	0.6957*	0.3697*	0.3005*
	(0.0524)	(0.0461)	(0.0399)	(0.0338)	(0.0222)	(0.0167)
Pr(4)	n/a	n/a	n/a	n/a	0.4175*	0.3786*
					(0.0257)	(0.0211)
** Indicates significance at .01 level. Standard errors in parentheses.*						

The marginal effects in Table 3B capture the magnitude of the effects. Once more, nonintegrated firms have lean information costs; management efficiently communicates core topics in the industry. For all probit runs, as integration increases, the cost of information increases at an increasing rate. This increase is even sharper than under the global library. Examining the four digit SIC, integration from one to two stages shows management's information costs increasing by 30% on average, and integration from two to three stages shows management's information costs increase 66% on average. Examining the letter digit SIC, successful two stage integration shows management's information costs increasing by 20% on average; successful two to three stage integration shows management's information costs increasing by 30% on average; and successful three to four stage integration shows management's information costs increasing by 38% on average.

Note that these marginal coefficients are further affected by industry. The most dramatic effects for variations in information stocks occur for firms in the transportation and manufacturing industries. Particularly in manufacturing, firms have enormous variety in technology and production which involves highly specific information relevant to only that activity. Increasing the scope of activities in such a firm, then, requires dedicating substantial administrative resources. However, because disclosures do not categorize the operating segments integrated into, I cannot draw more general conclusions than identifying which industries place the highest strain on management capacities.

DISCUSSION

By proposing markets and firms as substitutes, Coase's (1937) problem of economic organization evokes two questions. First, why are there firms when (at least in principle) all economic activity can be organized through markets? There's no reason assembly line workers in a factory can't buy and sell their semi assembled components as they move along production lines. Coase recognized that there is a cost to using the price mechanism, and that firms exist to supersede these costs. It follows to ask why do firms not expand indefinitely? If by organizing the firm can eliminate certain costs, why are there any market transactions at all? The outcome of this process is the selective intervention thought experiment. Coase originally argued rather vaguely for diminishing returns to management, but since then authors have emphasized mostly incentive problems (Garicano 2000).

The problem of economic organization has spurred an entire literature of answers. This paper leverages information's role in the puzzle with emphasis on the selective intervention piece. Though its role has been explored theoretically, information processing stories have struggled with empirical support. Previous answers have relied on idiosyncratic measures with contradictory results (Richman and Macher 2006). With new tools from NLP, I contribute a unique study to the traditional transaction cost story. I apply fairly common econometric techniques on novel data to corroborate a more complete theory of the firm. The explanatory variable quantifies the firm's information base embedded within textual artifacts to draw conclusions about organizational boundaries through a measure collected from financial disclosures.

Information in ordinary economic theory plays a crucial role in the firm (Arrow 1996). Production requires physical resources and information about how to combine them (Garicano 2000) but that information is costly to produce, maintain, and use (Demsetz 1988). Ultimately, there is a tradeoff between the benefits from collocating information against the costs of assimilating and processing it, and this comparative assessment drives the boundary of the firm. The organization as a whole must share some level of background information (Cohen and Levinthal 1990). But as expertise becomes more specialized, more resources must be dedicated to communication systems to move relevant information between subunits. Management serves as this interface by interpreting and structuring patterns from technical expertise through to potential investors. They represent the firm's information base by codifying it into a common language and the MD&A in particular documents this communication.

These results provide a quantitative interpretation on the upper bounds of management as interface—Coase's (1937) "diminishing returns to management". Though many authors have proposed more explicit answers (for examples, refer to Williamson (1985) for disadvantages of bureaucratic decision making, Simon (1947) for exponentially increasing communication, or Radner (1993) on larger teams making worse decisions), the study here places information processing costs as the primary effect. Larger planning problems lead to increasing management costs (Kikuchi et al. 2018), an idea echoed by Hayek (1945) on dispersed knowledge and the costs of acting on it. The marginal effects above show information costs increase at an increasing rate with each additional stage of integration, corroborating Coase with explicit thresholds on the boundary decision.

Interpreting information through the methodology here offers several applications in managerial practice. Regarding acquisition strategy, analysts can expand their quantitative analyses over financial data to include textual artifacts. NLP techniques immensely scale up the interpretability of textual information. In particular, textual similarities between business units can anticipate communication frictions and other redundancies from combining operations. Because communication systems are entangled with the distribution of expertise (Cohen and Levinthal 1990), weighing or optimizing the effects through NLP helps evaluate the costs of added administrative burdens. These decisions have industry specific considerations, too. Such analyses are equally applicable to business units within the firm. Examining information costs by subdivision can reveal where communication costs are highest which might otherwise be invisible. This insight highlights where interface efforts could be improved, or where disintegration decisions or other misaligned synergies might be considered. Applied more generally, NLP techniques over competitors or industry data could place reasonable bounds on management and establish market standards. Furthermore, this approach supports both rotational programs within firms and R&D in broad domains. Both activities

contribute to the generalized background understanding in a firm, improving communication between subunits.

The results further suggest a story about gains left on the table from firms being under-integrated. Again, the excess of market capitalization over book value of assets reflects the premium placed on a firm's information base. The operation of a market costs something and forming an organization supersedes some of those costs, but it is always possible to revert to the market if internal organization fails to do the task more cheaply. Though firms see large increases in their processing costs from absorbing additional operating stages, they still outperform the market. This implies information cannot be easily traded across boundaries (Bresnahan and Levin 2012) and there is substantial reward in getting around that (at least enough to make 66% increases in information costs worthwhile). This is consistent with research showing diversified forms are often more productive than single-segment firms (Schoar 2002). By contrast, the contrapositive to high transaction costs motivating internal organization is that markets can only exist when transaction costs are sufficiently low. Currently, transaction costs seem rather high for distributing information processing across firms. But NLP techniques might reduce contracting costs to improve the information flow in exchanges. For example, information measures over potential trading partners might improve complementarities by maximizing distances between their respective information bases, allowing parties to take more advantage of specialization (or minimizing them to detect coordination advantages). Like Coase's telephone, the ultimate effect on firm boundary will depend if NLP techniques benefit external or internal coordination more. Given the richer textual information available within firms versus across firms, I suspect managerial technique will benefit in the short run.

These methods open up several avenues for future research as well. A statistical quantification of textual information is a needed ingredient in many analyses. For example, these measures relative to environmental change could reveal how a firm's information set becomes more or less specialized in reaction to Schumpeterian change. The specificity of information stocks currently has unknown sensitivities to innovation. Further corroborating this relationship with labor mobility speaks even louder to innovation over time. Reframing the analysis here by focusing on performance measures instead of taking the status quo as an efficient outcome would improve precision on integration decisions. Industry cross sectional or relative firm comparisons could offer dollar value estimates on changes in organizational structures, as could time series or difference in difference econometrics over firm mergers pre and post acquisition. Similarly, framing the problem in industrial organization terms instead of organizational economics could shed light on information and market structure effects. Research designs teasing out significant differences in information sets between stages of production might reveal the strategic encapsulation of certain knowledge. Ultimately, the analysis here supported by further research offers many implications for both policy and managerial practice.

However, in weighing these conclusions, several caveats are in order. First, there are many competing explanations of the boundary of the firm. The emphasis on information measures here could be coordinated more closely with these alternatives. The regressions must also be interpreted with some caution regarding the employed variables. The construct of information costs offers a direct statistical measure for empirical investigation, but relies on necessary assumptions for text analysis. While statistical relationships surely matter in natural language, their representation in the bag of words technique superficially collapses the much deeper problem of language comprehension. The use of segment reporting provides a consistent measure for the selective intervener's stages through footnote disclosure criteria. Though reportable segments include transfer pricing in their criteria, the measure does not distinguish carefully between types of integration, nor does the study here consider the industry integrated into (instead forming only a ranking). Lastly, industry classifications through SICs have been inconsistent in their standardizations (Jacobs and O'Neil 2003). In defense of these critiques, these assumptions are largely acted on but not made by the author, which enhances the replicability of the study.

CONCLUDING REMARKS

The price of modern firms above their physical assets reflects the value of the information base embedded within it. In fact, organizations have even been modeled as simply partitions of information sets (Garicano 2000). Production requires physical resources and information about how to combine them but acquiring information is expensive and so is communicating it. The firm must hire specialists to be in charge of economizing on the transmission of information (Arrow 1996). Management serves as the interface by interpreting and structuring patterns from technical expertise through to potential investors.

But there is a tradeoff between the benefits from collocating information against the costs of assimilating and processing it. The organization as a whole must share some level of background information (Cohen and Levinthal 1990). As expertise becomes more specialized, more resources must be dedicated to communication systems to move relevant information between subunits. Industries and firms are repositories of specialized codes that put information into work and they specialize in different stocks of information. As these codes become distinct, they form into distinct organizations. Airline firms specialize in different stocks of information than their suppliers, and the formation of boundaries between them must be considered from the perspective of conserving on information costs (Demsetz 1988). The larger the number of stages, the larger the number of different stocks of knowledge, increasing the reliance on direction by management (Demsetz 1988). When the internal costs exceed some threshold, the stock is more cheaply housed across firms than internally. The boundary between firms and markets minimizes transaction costs along informational dimensions (Monteverde 1995).

But measuring information processing in organizational forms has been challenging (Masten 2016). Though its role has been explored theoretically, information processing stories have struggled with empirical support (Richman and Macher 2006). I develop a measure through textual analysis to systematically account for and interpret firm specific information. Qualitative aspects of text contain rich information about the operations of firms but are difficult to represent. Using NLP techniques, I quantify the firm's information base embedded within textual artifacts with applications from information theory. I measure sources of difference in management information relative to the industry to describe an economy of communication. I regress this explanatory variable against a novel measure of organizational boundaries collected from financial disclosures to corroborate a more complete theory of the firm. Both variables are derived systematically to minimize the author's estimates, which have been idiosyncratic and subjective in other studies (Richman and Macher 2006).

The results demonstrate how firm boundaries congeal around management's information processing capacities. The probit estimates provide a cost calculus for organizing a marginal transaction internally or through markets. For all probit runs, as integration increases, the cost of information increases at an increasing rate. These results provide a quantitative interpretation to Coase's (1937) diminishing returns to management—in fact, they show where dividing lines occur. The marginal effects estimate these thresholds to describe the information costs along each stage of integration.

Furthermore, these findings validate empirical approaches to transaction cost economics and text-based tools for organizational decision modeling. Modeling information through the methodology here offers several implications for managerial practice. NLP techniques immensely scale up the interpretability of textual information and the distance between information bases can be leveraged in a variety of ways. In particular, performing textual similarities between business units before acquisitions can anticipate communication frictions and other redundancies from combining operations. Alternatively, examining information costs by subdivision could highlight where interface efforts could be improved, or where disintegration decisions might be considered. NLP techniques show promise in contracting exchanges, too. For example, information measures over potential trading partners could maximize differences between information bases to take advantage of specialization or minimize them to act on coordination advantages. More generally, the results point to a story about gains left on the table from firm integration. Overall, the results and methodology here supported by further research offer several applications for policy and managerial practice.

REFERENCES

Alchian, A., & Demsetz, H. (1972) Production , Information Costs, and Economic Organization, *American Economic Review, 62*(5), 777–795.

Arrow, K. (1996). The Economics of Information: An Exposition. *Empirica*, *23*, 119-128.

Black, E.L., Christensen, T.E., Joo, T.T., & Schmardebeck, R. (2017). The Relation Between Earnings Management and Non-GAAP Reporting. *Contemporary Accounting Research*, *34*(2), 750–782.

Bresnahan, T., & Levin, J. (2012). Vertical integration and market structure. *NBER Working Papers 17889*, National Bureau of Economic Research, Inc.

Burns, R.W. (2004). *Communications: An International History of the Formative Years*. Institution of Electrical Engineers, London.

The Catholic Encyclopedia (Vol. 4.). (1908). New York, NY: Robert Appleton Company.

Coase, R.H. (1937). The Nature of the Firm. *Econometrica*, *4*(16), 386–405.

Coase, R.H. (1988). The Nature of the Firm, 1. Origin, 2. Meaning, 3. Influence. *Journal of Law, Economics, and Organization, 4*(1), 3–47.

Cohen, W., & Levinthal, D. (1990). Absorptive Capacity: A New Perspective on Learning and Innovation. *Administrative Science Quarterly*, *35*(1), 128-152.

Cohen, L., Malloy, C.J., & Nguyen, Q. (2019). Lazy Prices. *Academic Research Colloquium for Financial Planning and Related Disciplines*, SSRN 1658471.

Dearborn, R., & Simon, H. (1958).Selective perception in executives. *Sociometry*, *21*, 140-144.

Demsetz, H. (1988). The Theory of the Firm Revisited. *Journal of Law, Economics, and Organization*, *4*(1), 141–161.

Eisner, M.A. (2017). *Antitrust and the Triumph of Economics: Institutions, Expertise, and Policy Change*. University of North Carolina Press, Chapel Hill.

Ekhosuehi, V.U., & Osagiede, A. (2012). The Entropy-Theoretic Stability Index for Manpower Systems. *International Journal of Operations Research*, *9*, 120–128.

FASB (Financial Accounting Standards Board). (May 2014; effective December 15, 2016). Accounting Standards Update (ASU) No. 2014–09. Accounting standards codification.

Garicano, L. (2000). Hierarchies and the Organization of Knowledge in Production. Jou*rnal of Political Economy*, *108*(5), 874-904.

Grossman, S.J., & Hart, O.D. (1986). The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration. *Journal of Political Economy*, *94*(3), 691–719.

Guay, W., Samuels, D., & Taylor, D. (2015). *Guiding Through the Fog: Financial Statement Complexity and Voluntary Disclosure*. Working paper, University of Pennsylvania.

Hart, O. (1995). *Firms, Contracts, and Financial Structure*. Oxford, UK: Clarendon Press.

Hayek, F.A. (1945). The Use of Knowledge in Society. *American Economic Review*, *35*(4), 519–530.

Henisz, W.J. (2000) The Institutional Environment for Economic Growth. *Economics and Politics*, *12*(1), 1–31.

Henry, E. (2008). Are Investors Influenced by How Earnings Press Releases Are Written? *Journal of Business Communication*, *45*, 363–407.

Huffman, D. (1952). A Method for the Construction of Minimum-Redundancy Codes. *Proceedings of the IRE, 40*(9), 1098–1101.

Itami, H. (1987) *Mobilizing Invisible Assets*. Cambridge, MA: Harvard University Press

Jacobs, J., & O'Neil, C. (2003). On the Reliability (or Otherwise) of SIC Codes. *European Business Review*, 15(3), 164–169.

Jensen, M., & Meckling, W. (1976). Theory of the Firm: Managerial Behavior, Agency Costs, and Capital Structure. *Journal of Financial Economics*, *3*(4), 305-360.

Kalnins, A., & Mayer, K.J. (2004). Relationships and Hybrid Contracts: An Analysis of Contract Choice in Information Technology. *Journal of Law, Economics, and Organization*, *20*(1), 207–229.

Kikuchi, T., Nishimura, K., & Stachurski, J. (2018). Span of Control, Transaction Costs, and the Structure of Production Chains. *Theoretical Economics, 13*, 729–760.
Knight, F.H. (1921). *Risk, Uncertainty, and Profit*, New York, NY: Houghton Mifflin.
Langlois, R.N. (2003). The Vanishing Hand: The Changing Dynamics of Industrial Capitalism. *Industrial and Corporate Change, 12*(2), 351–385.
Lewis, T.R. (1983). Preemption, Divestiture, and Forward Contracting in a Market Dominated by a Single Firm. *American Economic Review, 73*(5), 1092–101.
Li, F. (2008). Annual Report Readability, Current Earnings, and Earnings Persistence. *Journal of Accounting and Economics, 45*(2), 221–47.
Loughran, T., & McDonald, B. (2014) "Measuring Readability in Financial Disclosures. *Journal of Finance, 69*(4), 1643–71.
Loughran, T., & McDonald, B. (2015). The Use of Word Lists in Textual Analysis. *Journal of Behavioral Finance, 16*(1), 1–11.
Loughran, T., & McDonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research, 54*(4), 1187–1230.
Lundholm, R.J., Rogo, R., & Zhang, J. (2014). Restoring the Tower of Babel: How Foreign Firms Communicate with US Investors. *The Accounting Review, 89*, 1453–85.
Marshall, A. (1948) *Principles of Economics* (8th ed.). New York, NY: Macmillan.
Masten, S., Meehan J., Jr., & Snyder, E. (1991). The Costs of Organization. *Journal of Law, Economics, & Organization, 7*(1), 1-25.
Masten, S. (2016). *Transaction Cost Economics*. Elgar Research Reviews in Economics, Cheltenham, UK: Edward Elgar Publishing.
Monteverde, K. (1995). Technical Dialog as an Incentive for Vertical Integration in the Semiconductor Industry. *Management Science, 41*(10), 1624–1638.
Nelson, R., & Winter, S. (1982). *An Evolutionary Theory of Economic Change*. Cambridge, MA: Belknap Press.
Noordewier, T.G., John, G., & Nevin, J.R. (1990). Performance Outcomes of Purchasing Arrangement in Industrial Buy-Vendor Relationships. *Journal of Marketing, 54*, 80–93.
Pisano, G.P. (1990). The R&D Boundaries of the Firm: An Empirical Analysis. *Administrative Science Quarterly, 35*(1), 153–176.
Price, S.M., Doran, J.S., Peterson, D.R., & Bliss, B. A. (2012). Earnings Conference Calls and Stock Returns: The Incremental Informativeness of Textual Tone. *Journal of Banking and Finance, 36*(4), 992–1011.
Princeton University. (2010). About WordNet. WordNet. Princeton University.
Radner, R., & Van Zandt, T. (1992). Information Processing in Firms and Returns to Scale. *Annales d'Economie et de Statistique, 25*(26): 265-298.
Radner, R. (1993). The Organization of Decentralized Information Processing. *Econometrica, 61*(5), 1109–1146
Richman, B.D., & Macher, J.T. (2006). *Transaction Cost Economics: An Assessment of Empirical Research in the Social Sciences*. Duke Law School Legal Studies, Paper No. 115.
Rogers, D.F., Plante, R.D., Wong, R.T., & Evans, J.R. (1991). Aggregation and Disaggregation Techniques and Methodology in Optimization. *Operations Research, 39*, 553–582.
Roomberg, R. (Received 2020). Learn Morse Code. Retrieved from www.learnmorsecode.com.
Saussier, S. (2000). Transaction costs and contractual incompleteness: The case of Électricité de France. *Journal of Economic Behavior and Organization, 42*(2),189–206.
Schoar, A. (2002). Effects of Corporate Diversification on Productivity. *Journal of Finance, 57*(7), 2379-2403.
Shannon, C. (1948). A Mathematical Theory of Communication. The Bell System Technical Journal, *27*, 379–423.
Simon, H. (1947). *Administrative Behavior*. New York, NY: Macmillan.

Simon, H. (1955). A Behavioral Model of Rational Choice. *Quarterly Journal of Economics*, *69*(1), 99–118.
Van Zandt, T. (2000). Real-Time Decentralized Information Processing as a Model of Organizations with Boundedly Rational Agents. *Review of Economic Studies*, *66*(3), 633–58.
Williamson, O.E. (1985). *The Economic Institutions of Capitalism*. New York, NY: The Free Press.
Williamson, O.E. (1996). *The Mechanisms of Governance*. New York, NY: Oxford University Press.
Yang, J. (2017). Information Theoretic Approaches in Economics. *Journal of Economic Surveys, 32*(3), 940–960.

APPENDIX 1

Stop Word Libraries

A1 LIST 1
PYTHON'S NLTK STANDARD ENGLISH STOP WORDS

{'can', 'was', 'he', 'needn', 'her', 'now', 'itself', "won't", 'herself', 'than', 'me', 'for', 'under', 'their', 'an', 'whom', 'before', 'haven', 'same', 'until', 'of', 'yours', 'isn', 'off', "haven't", 'were', 'that', 'further', 'having', 'does', 'my', "that'll", 'ourselves', 'no', 'why', "isn't", "don't", 'all', "you're", 'wasn', 'hers', 'over', 'very', 'few', 'will', 'him', 'myself', "doesn't", "hasn't", 'do', 'there', 'from', 'against', 'which', 'being', 'a', 'am', 'yourselves', 'out', 'is', 'and', "you've", 'doing', 'at', 'mightn', 'in', 'as', 'more', 'not', 'hadn', 'only', 't', 'most', 'm', 'by', 'i', 'into', "hadn't", 'y', 'any', 'our', 'so', 'about', 'such', 'down', "you'd", 'have', "aren't", "you'll", 'on', 'then', 'aren', 'those', 'but', 'how', 'ain', 'd', 'what', 'to', "couldn't", 'shouldn', "shouldn't", 'where', 'or', 'couldn', 'o', 'ma', 'too', 'up', 'them', "should've", "mightn't", 's', 'hasn', 'through', 'she', 'its', 'don', 'you', 'wouldn', 'other', 'been', "didn't", 'who', 'this', 'above', "needn't", 'll', 'are', 'his', 'your', 'because', 'shan', 'weren', 'himself', 'the', 'both', "wasn't", 'here', 'below', 'each', 'should', "mustn't", 'with', 'between', 're', 'it', 'when', 'mustn', 've', 'nor', 'won', 'has', 'after', 'just', 'yourself', 'did', 'during', "it's", 'once', 'theirs', 'doesn', 'again', 'some', "she's", 'they', 'had', 'if', 'while', 'be', 'didn', "shan't", 'we', 'these', "wouldn't", 'ours', "weren't", 'themselves', 'own'}

A1 LIST 2
ADDITIONAL BALANCE SHEET STOP WORDS

{'company', 'business', 'financial', 'net', 'gross', 'total', 'due', 'year', 'end', 'annual', 'increase', 'decrease', 'incline', 'decline', 'prior', 'fiscal', 'hundred', 'thousand', 'million', 'billion', 'statement', 'accounting', 'january', 'february', 'march', 'april', 'may', 'june', 'july', 'august', 'september', 'october', 'november', 'december', 'one', 'two', 'three', 'four', 'five', 'six', 'seven', 'eight', 'nine', 'ten', 'eleven', 'twelve', 'thirteen', 'fourteen', 'fifteen', 'sixteen', 'seventeen', 'eighteen', 'nineteen', 'twenty', 'thirty', 'forty', 'fifty', 'sixty', 'seventy', 'eighty', 'ninety'}

APPENDIX 2

Summary Data By Sic Code (Sans Stop Words)

A2 TABLE 1
GLOBAL DATA

4 SIC			3 SIC			2 SIC			L SIC		
SIC	Firms	Words (000s)	SIC	Firms	Words (000s)	SIC	Firms	Words (000s)	SIC	Firms	Words (000s)
2834	380	2,093	283	491	2,668	28	627	3,640	D	1,855	11,168
6798	294	2,316	602	339	4,237	73	500	3,161	H	1,258	12,457
6022	223	2,667	679	325	2,474	60	445	5,160	I	804	5,196
1311	162	1,282	737	271	1,708	67	352	2,551	E	395	3,501
7372	139	864	384	181	968	38	280	1,569	B	317	2,198
6021	113	1,555	131	162	1,282	36	267	1,497	G	260	1,666
3841	108	597	367	132	745	13	223	1,682	F	133	874
7389	105	764	738	124	857	49	179	1,932	C	61	460
3674	86	547	603	102	898	35	166	1,005		5,083	37,519
6035	81	632	382	83	494	63	141	1,925			
2836	78	410	581	69	387	62	139	1,370			
6331	68	944	633	68	944	48	114	820			
5812	61	355		2,347	17,662	37	95	629			
	1,898	15,024				20	88	546			
						50	76	447			
						61	76	848			
						80	76	640			
						58	69	387			
							3,913	29,807			

A2 TABLE 2
LOCAL DATA

4 SIC			3 SIC			2 SIC			L SIC		
SIC	Firms	Words (000s)	SIC	Firms	Words (000s)	SIC	Firms	Words (000s)	SIC	Firms	Words (000s)
2834	380	2,093	283	491	2,668	28	627	3,640	D	1,855	11,168
6798	294	2,316	602	339	4,237	73	500	3,161	H	1,258	12,457
6022	223	2,667	679	325	2,474	60	445	5,160	I	804	5,196
1311	162	1,282	737	271	1,708	67	352	2,551	E	395	3,501
7372	139	864	384	181	968	38	280	1,569	B	317	2,198
6021	113	1,555	131	162	1,282	36	267	1,497	G	260	1,666
3841	108	597	367	132	745	13	223	1,682	F	133	874
7389	105	764	738	124	857	49	179	1,932		5,022	37,060
3674	86	547	603	102	898	35	166	1,005			
6035	81	632	622	55	307	62	139	1,370			
2836	78	410		2,182	16,144	48	114	820			
6221	55	307				65	94	499			
	1,824	14,034				87	87	542			
							3,473	25,428			