The Incentives for Vertical Mergers and Vertical Integration

Laurent Frésard, Gerard Hoberg and Gordon Phillips*

March 10, 2013

ABSTRACT

We examine the incentives for firms to vertically integrate through vertical mergers and production. We develop a new firm-specific measure of vertical integration using 10-K text to identify the extent a firm's products span vertically related product markets. We find that firms in high R&D industries are less likely to vertically integrate or engage in vertical mergers, and are more likely to initiate customer or supplier relationships outside of the firm. These findings are consistent with firms with unrealized innovation avoiding integration to maintain ex-ante incentives to make relationship specific investments and maintain residual rights of control as in Grossman and Hart (1986). In contrast, firms in high patenting industries with stable product markets are more likely to vertically integrate consistent with control rights being obtained by commercializing firms to reduce potential ex-post holdup.

*University of Maryland, University of Maryland, and University of Southern California and National Bureau of Economic Research, respectively. Frésard can be reached at lfresard@rhsmith.umd.edu, Hoberg can be reached at ghoberg@rhsmith.umd.edu and Phillips can be reached at gordon.phillips@marshall.usc.edu. All errors are the authors alone. Copyright ©2013 by Laurent Frésard, Gerard Hoberg and Gordon Phillips. All rights reserved.

The scope of firm boundaries and whether to organize transactions within the firm (integration) or by using external purchasing is a major topic in both finance and economics. Williamson (1971), Williamson (1979) and Klein, Crawford, and Alchian (1978) pioneered this area through their theory of transaction cost economics and ex-post holdup given contractual incompleteness. Firms choose the organizational form that minimizes transaction costs and ex-post holdup. Grossman and Hart (1986) in their property rights theory of the firm show that ex-ante investment effects also matter. They show that ex-ante incentives to invest in relationship specific assets are reduced through vertical integration giving residual rights of control to the controlling firm.

We examine how the distinction between incentives for ex-ante investment in relationship specific assets, and the potential for ex-post holdup, influence the extent to which a firm is vertically integrated and whether firms engage in vertical mergers. Our focus is on both vertical integration and vertical mergers. Our analysis of the incentives for relationship specific investment, ex-post holdup and vertical mergers is distinct from other motives for mergers including neoclassical theories, agency theories and horizontal theories of mergers.[1]

We analyze vertical integration and vertical mergers using new firm-specific measures of vertical integration and relatedness constructed using text-based analysis of firm product descriptions filed with the Securities and Exchange Commission. We focus on how these firm product vocabularies relate to commodity descriptions from the BEA input-output tables. We then measure how vertically integrated firms are, how their vertical integration changes over time, and the extent to which mergers represent vertical transactions. Our study is thus able to analyze vertical integration beyond the single industry focus of studies in industrial organization.[2]

We show that the distinction between unrealized innovation through R&D and

[1]See Maksimovic and Phillips (2001), Jovanovic and Rousseau (2002), and Harford (2005) for neoclassical and q theories and Morck, Shleifer, and Vishny (1990) for an agency motivation for mergers, and Phillips and Zhdanov (2013) for a recent horizontal theory of mergers.

[2]These early studies include for instance Monteverde and Teece (1982) focusing on automobile manufacturing, Masten (1984) focusing on airplane manufacturing, and Joskow (1987) focusing on coal markets. An extensive review of the empirical literature on vertical integration can be found in Lafontaine and Slade (2007) and Bresnahan and Levin (2012).

realized innovation through patents is important in the propensity to vertically integrate through changes in own-production and vertical mergers. We find that firms in high R&D industries are less likely to vertically integrate through either own production or vertical mergers. These results are consistent with R&D being a measure of technological firm-specific investments that are non-contractible and non-verifiable as in Grossman and Hart (1986) and Hart and Moore (1990).[3] Our findings are consistent with firms avoiding integration and remaining separate to maintain ex-ante incentives to invest in relationship specific assets through R&D and to maintain residual rights of control as in Grossman and Hart (1986).

These results also extend to external purchases from major customers and suppliers. We find that firms are more likely to remain separate and have arms length customer-supplier relationships when they are in industries with significant R&D activity. These findings match the predictions that, for unrealized innovation, firms need to not only avoid integration, but also need major relationships to promote strong incentives that ensure adequate investment in technologies that can be used by other firms (relationship specific investment). These results are further consistent with Cen, Dasgupta, and Sen (2011) and Johnson, Karpoff, and Yi (2012), who suggest that mergers can threaten relationship specific investments when firms have explicitly identified customer - supplier links.

In contrast, we find that firms in industries with high patenting rates and thus with more realized innovation with strong property rights are more likely to vertically integrate both through own-production and vertical mergers. In these high patenting industries, owners of the realized innovations have more legally enforceable residual rights of control, and are thus more likely to merge because the residual rights of control are most efficiently assigned to the firm that will use and commercialize the patented technology. This distinction between high R&D and patents is nontrivial. High R&D does not necessarily lead to high patenting rates. As reported by Cohen, Nelson, and Walsh (2000) in their survey of 1478 R&D labs, high R&D in many cases

[3] Acemoglu (1996) also considers how search market frictions can make technological investments partner-specific, thus creating relationship specific assets. In addition Allen and Phillips (2000) and Kale and Shahrur (2007) show that R&D increases with firm interaction with alliance partners, and customer and supplier contracts.

does not lead to patents given concerns about appropriabilty. Their survey points to the ability of others to work around patents using information conveyed by the patent application causing managers not to patent in many industries.[4]

A recent example of this realized innovation effect is Apple Computer's 2008 purchase of a chip maker, PA Semi, for $278 million. This firm specialized in making fast, low-power processors, now used in the iPad. Apple also purchased Intrinsity, an Austin TX low power processor manufacturer, in April 2010, now used in the iPhone. In these cases, the need for additional ex-ante commitment to relationship specific investment is lower as the products have already been patented and developed. However, a more pressing issue is that Apple faced the risk of ex-post holdup as they commercialized, produced and marketed the new iPad and iPhone. Buying the firms to gain control of their realized innovations removes the possibility of ex-post hold up, and eliminates the potential sale of related products to other competing manufactures.[5]

We also consider the role of supply chain maturity. Our motivation relies on the fact that the decision to vertically integrate, or to engage in vertical mergers, should relate to the characteristics of the supply chain itself. For example, assuming the benefits of integration derive from ongoing operations, such decisions should be more profitable when the supply chain is expected to be stable. An unstable supply chain can reduce the horizon during which benefits are realized. Because reorganization typically entails a high level of fixed costs, firms in unstable supply chains should be less willing to vertically integrate as the shorter horizon over which gains are realized may not be adequate to cover the high fixed costs.

We find strong empirical support for the proposed positive link between maturity and vertical integration. In particular, both vertical integration and vertical mergers, are positively related to maturity and stability as captured by older firm age, lower

[4]Cohen, Nelson and Walsh report "firms typically protect the profits due to invention with a range of mechanisms, including patents, secrecy, lead time advantages and the use of complementary marketing and manufacturing capabilities. Of these mechanisms, however, patents tend to be the least emphasized by firms in the majority of manufacturing industries, and secrecy and lead time tend to be emphasized most heavily.

[5]A NY Times article (April 28, 2010) supports this conclusion as it claims the purchases were made for use in Apple's products, to increase the products' attractiveness, and to keep the technology out of competitors' hands.

M/B ratio, higher PP& E/assets, and lower product market fluidity. These results suggest that firms avoid integration during earlier phases of their development, but later consider integration as growth options are later fully exercised and industries become more stable.

We demonstrate that our new text-based measures of vertical relatedness have significant advantages over existing measures based on NAICS codes. By linking the vocabulary of firms' business descriptions to that of the commodities in the input-output tables, we are able to identify vertical relatedness between firms in a more direct way compared to existing measures. Our focus on vertical links extends the work of Hoberg and Phillips (2010), who examine horizontal links between firms and their impact on horizontal merger synergies.

We find three key results that validate the new text based measures as valid measures of vertical integration. First, we examine the correlation between accounts receivables and accounts payables for firms predicted to be vertically related using the new text-based method and the existing NAICS-based method, and find that these correlations are an order of magnitude more negative for the text based method than for the NAICS-based method. We show that this negative link strongly supports the conclusion that the links we identify are based on vertical orientation.

Second, unlike existing measures of vertical relatedness (based on NAICS codes), the new measure does not pick up a disproportionate number of links to financial firms (which are less relevant for assessing the theoretical links we hypothesize).

Third, the new measures do not have to rely on the quality of the Compustat segment tapes, nor the quality of the NAICS classification, to measure vertical integration.[6]

Overall our paper reconciles some of the tension between the ex-post hold up literature of Klein, Crawford, and Alchian (1978) and Williamson (1979), and the ex-ante incentive effect of assigning residual rights of control of Grossman and Hart (1986). We show that firms in industries with high unrealized innovation are less likely to integrate either through own production or vertical mergers, while firms

[6]The Compustat segment reporting has been shown to have reporting biases as documented by Hyland (1999) and Villalonga (2004).

in industries with significant realized patents and thus stronger property rights are more likely to integrate to avoid ex-post holdup.

Our paper adds to the literature on vertical integration and mergers by Fan and Goyal (2006) and the recent paper by Ahern and Harford (2012) on how economic shocks affect mergers. Fan and Goyal examine the stock market reactions of vertical mergers where vertical integration is identified through links between SIC codes and the input-output tables. They do not study how relationship specific assets and ex-post holdup affect vertical integration. Ahern and Harford examine the extent to which shocks in the supply chain cause firms to vertically integrate through mergers. Also relevant, Acemoglu, Johnson, and Mitton (2009) indicate that contracting costs and financial development are important drivers of vertical integration in a large sample covering 93 countries.

Our focus is on examining the role of relationship specific investment through examining unrealized and realized innovation, the role of supply chain maturity, and how these issues relate to vertical integration and mergers. Our findings also complement those of Atalay, Hortascu, and Syversen (2012), who suggest that vertical integration promotes the transfer of intangible inputs such as R&D capital or know-how.

Our paper adds to papers that examine how mergers interact with customer and supplier relationships. Fee and Thomas (2004) and Shahrur (2005) investigate the effects of horizontal mergers on firms with explicitly identified customer and supplier relationships. Our results also complement the recent evidence on mergers and innovation by Bena and Li (2013) who examine the impact of mergers on ex-post innovation, and Phillips and Zhdanov (2013) who examine how anticipated mergers affect R&D. Our paper focuses on the ex-ante incentives to vertically integrate, and examines how R&D and patents impact these incentives and shape firm boundaries.

The remainder of this paper is organized as follows. Section II presents the data and develops our new measure of firm-level vertical integration. Section III validates our text-based vertical relatedness network. Section IV summarizes the characteristics and statistical properties of vertical integration and vertical mergers.

Section V examines the determinants of vertical integration and vertical mergers, and Section VI concludes.

II Data and Methodology

We draw from multiple data sources to create our sample and our key variables: 10-K business descriptions, COMPUSTAT, the SEC Edgar database, the Bureau of Economic Analysis (BEA) Input-Output (IO) tables, data on mergers from Securities Data Corporation, and data on announcement returns from CRSP. We present each data source in turn, and detail the construction of our new text-based variables.

A Data from 10-K Business Descriptions

We start with the COMPUSTAT sample of firm-years from 1996 to 2008 with sales of at least $1 million and positive assets. We then use the Edgar database to extract text in the Business Description section of annual firm 10-Ks. We thus require that a given observation has a 10-K filed on the SEC Edgar website with a machine readable business description. The methodology we use to extract and process 10-K text follows Hoberg and Phillips (2010a). The first step is to use web crawling and text parsing algorithms to construct a database of business descriptions from 10-K annual filings on the SEC Edgar website from 1996 to 2008. We search the Edgar database for filings that appear as "10-K," "10-K405," "10-KSB," or "10-KSB40." The business descriptions appear as Item 1 or Item 1A in most 10-Ks. The document is then processed using APL to extract the business description text and a company identifier, CIK.[7] Business descriptions are legally required to be accurate, as Item 101 of Regulation S-K requires firms to describe the significant products they offer, and these descriptions must be updated and representative of the current fiscal year of the 10-K. There are 74,379 firm-years in the Compustat/Edgar universe.

[7] We use the SEC Analytics database from the Wharton Research Data Service (WRDS) to obtain a historical mapping of SEC CIK to COMPUSTAT gvkey, as the base CIK variable in COMPUSTAT only contains the most recent link.

B Data from the Input-Output Tables

Consistent with existing studies (e.g. Fan and Goyal (2006)), we use data from the BEA as the foundation to define vertical relatedness. The IO tables account for the dollar flows between all producers and purchasers in the U.S. economy (including households, the government, and foreign buyers of U.S. exports). Relevant to our analysis, these tables are based on two primitive concepts: 'commodity' outputs (defined by the Commodity IO Code), and producing 'industries' (defined by the Industry IO Code). A commodity is any good or service that is produced. The 'Make' table reports the dollar value of each commodity produced by a given industry, and in the 2002 IO tables, there are 424 distinct commodities and 426 industries in the Make table. An industry can produce more than one commodity.[8]

The 'Use' table reports the dollar value of each commodity that is purchased by each industry or by final end-user.[9] There are 431 commodities in the Use table purchased by 439 industries or final end-users.[10] The 'Detailed Item Output' table provides the textual description (Item Description) of each commodity as well as sub-commodities (Item IO code) and their dollar value. We record and utilize this verbal information to construct our measure of vertical relatedness. We describe three data structures we compute from the IO Tables, which we refer to later in this section: (1) Commodity-to-commodity (upstream to downstream) correspondence matrix (V), (2) Commodity-to-word correspondence matrix (CW), and (3) Commodity-to-'exit' (supply chain) correspondence matrix (E).

In addition the the numerical values in the BEA data, we use an often overlooked resource: the detailed verbal description for each commodity, which is critical to our identification of vertically related firms. The 'Detailed Item Output' table decom-

[8] In the 2002 IO edition, the average (median) number of commodities produced per industry is 18 (13). The output of an industry tends to be rather concentrated: The average (median) commodity concentration ratio per industry is 0.78 (0.81).

[9] While costs are reported in both purchaser price and producer costs, we use producers' prices.

[10] There are seven commodities in the Use table that are not in the Make table: non-comparable imports, used and second hand goods, rest of the world adjustment to final uses, compensation of employees, indirect business tax and nontax liability, and other value added. There are thirteen 'industries' in the Use table that are not in the Make table. These correspond to 'end users' and include personal consumption expenditures, private fixed investment, change in private inventories, exports and imports, and federal and state government expenditures.

poses each commodity into sub-commodities, and lists each commodity's respective vocabulary (its verbal description) and its total production.[11] Each verbal description uses between 1 to 25 distinct words (the average is 8) that summarize the nature of the good or service defined as each commodity.[12] For each commodity, we extract these verbal descriptions to identify commodity words. We follow the conventions in Hoberg and Phillips (2010a) regarding stop words, and we only include nouns and proper nouns, and we discard all words that appear in more than 25% of all commodity descriptions. We identify 10,660 unique commodity words. We then define the commodity-word correspondence matrix (CW) as a three-column matrix having the first column being a commodity, the second column being an associated word and the third being its weight (relative to the commodity's total output).

Given the textual description in Detailed Item Output table relates to commodities (and not industries), we focus on the intensity of vertical relatedness between pairs of commodities and we construct the sparse square matrix V based on the extent to which a given commodity is vertically linked to (upstream or downstream) another commodity in the supply chain. To do this, from the Make Table, we create $SHARE$, an $I \times C$ matrix (Industry $\times$ Commodity) that contains the percentage of commodity c produced by a given industry i. The USE matrix is a $C \times I$ matrix that records the dollar value of industry i's purchase of commodity c as an input. The $CFLOW$ matrix is then given by $USE \times SHARE$, and is the $C \times C$ matrix of dollar flows from an upstream commodity c to a downstream commodity d. Similar to Fan and Goyal (2006), we define the $SUPP$ matrix as $CFLOW$ divided by the total production of the downstream commodity d. $SUPP$ records the fraction of commodity c that is used as an input to produce commodity d. Similarly, the matrix $CUST$ is given by $CFLOW$ divided by the total production of the upstream commodities c and records the fraction of commodity c's total production that is used to produce commodity d. The V matrix is then defined as the average of $SUPP$ and $CUST$. A larger element in V indicates a stronger vertical relationship between commodities

[11] There are 5,459 sub-commodities in the 'Detailed Item Output' table.

[12] For instance, the commodity 'Footwear Manufacturing' (IO Commodity Code #316100) has 15 subcategories, such as #316211 described as 'rubber and plastics footwear', or #316212 described as 'house slippers'.

c and d.[13] Note that V is sparse (i.e., most commodities are not vertically related in the supply chain) and is non-symmetric as it features distinct downstream ($V_{c,d}$) and upstream ($V_{d,c}$) directions.

Finally, we create an exit correspondence matrix E to account for the production that flows out of the U.S. supply chain. To do so, we use the industries that are present in the Use table but *not* in the Make table. These correspond to 'final users' and include personal consumption expenditures; private fixed investment; change in private inventories; exports; import; federal government defense consumption, and investment; federal government non-defense consumption, and investment; state and local government education consumption, and investment; and state and local government other consumption and investment. E is a one-column matrix where one row represents a commodity and the column-vector contains the fraction of each commodity's output that flows to users outside the U.S. supply chain.

C Text-based Vertical Relatedness

Our contribution is identifying vertical relatedness between firms by jointly using the vocabulary in firm 10-Ks and the vocabulary defining the BEA IO commodities. We link each firm in our Compustat/Edgar universe to the IO commodities using textual similarity analysis between the given firm's business description vocabulary and the textual description of each BEA commodity. Because vertical relatedness is observed from BEA data at the IO commodity level (see description of the matrix V above), we can score every pair of firms i and j based on the extent to which they are upstream or downstream by (1) mapping i's and j's text to the subset of observed IO commodities it provides, and (2) then determining i and j's vertical relatedness using the relatedness matrix V.

A first step is to identify the universe of product market vocabulary that will be used to compute textual mappings from each firm to the set of commodities. We limit attention to words that appear both in the IO commodities definitions (10,660 words), and in the Hoberg and Phillips (2010a) post-processed universe (60,507 nouns

[13] Alternatively, we consider in unreported tests the maximum between $SUPP$ and $CUST$, $SUPP$, or $CUST$ alone to define vertical relatedness. Our results are robust.

and proper nouns appearing in 10-K business descriptions excluding any "common words" appearing in more than 25% of all 10-K business decriptions). There are 5,997 such words in this intersection. To ensure we do not score broad words or potentially horizontally contaminated words as vertical, we further reduce this universe by discarding broad words from this universe if they are in the bottom tercile based on the fraction of times each given word coappears with its industry peer words (based on IO commodities) in observed 10-K filings.[14] Henceforth, we refer to this targeted vocabulary as the "intersection vocabulary". Define M as the number of firms in our database in a given year. To account for the fact that some business decription words for each firm do not appear in the IO vocabularies, we additionally compute an M-vector F as the fraction of firm i's business description vocabulary that is present in the intersection vocabulary. We use the vector F to adjust our measures of vertical integration later in this section.

Our next step is to compute the 'firm to IO commodity correspondence matrix' B. This matrix has dimension $M \times C$, where C is the number of IO commodities, and M is the number of firms. An entry $B_{m,c}$ (row m, column c) is simply the cosine similarity of the text in the given IO commodity c, and the text in firm m's business description using only vocabulary from both that are in the "intersection vocabulary". Hence, this calculation is based only on non-common nouns and proper nouns as in Hoberg and Phillips (2010a). We use the cosine similarity method because it naturally controls for document length and is a well-established method for computing document similarities (see Sebastiani (2002)). The cosine similarity is simply the normalized dot product (see Hoberg and Phillips (2010a)) of the word-distribution vectors of the two vocabularies (documents) being compared. Cosine similarities are bounded in [0,1], and a value close to one indicates that firm i's product market vocabulary is a near exact match relative to IO commodity c's product market vocabulary. Hence, the matrix B indicates which IO commodity a given firm's products is most similar to. Most elements of the matrix B are zero or close to zero.

We then measure the extent to which firm i is upstream relative to firm j as

[14] This approach is akin to the notion of broad versus local vocabularies as discussed in Hoberg and Phillips (2010), who also focus on the local subset of words after discarding the tercile of most broad words.

follows:

$$UP_{ij} = F[i]F[j] \;\; [B \cdot V \cdot B']_{i,j}. \tag{1}$$

The triple product $(B \cdot V \cdot B')$ is an $M \times M$ matrix of unadjusted upstream-to-downstream links between all firms i to firms j. Note that direction is important, and this matrix is not symmetric. Upstream relatedness of i to j is thus the i'th row and j'th column of this matrix, which we then adjust by the extent to which firm i and firm j vocabularies load on the IO vocabulary (as measured by $F[i]$ and $F[j]$, the i'th and j'th element of the M-vector F, respectively). Firm-pairs receiving the highest scores for vertical relatedness are those having vocabulary that maps most strongly to IO commodities that are vertically related according to the matrix V (constructed only using BEA relatedness data), and those having vocabularies that overlap non-trivially with the vocabularies that are present in the IO commodity dictionary according to the matrix B. Thus, firm i is located upstream from firm j when i's business description is strongly associated with commodities that are used to produce commodities whose description resembles firm j's products description.

Downstream relatedness is simply the mirror image of upstream relatedness, $DOWN_{ij} = UP_{ji}$. By repeating this procedure for every year in our sample (1996-2008), the matrices UP and $DOWN$ provide a (time-varying) network representation of vertical links among individual firms.

D Measuring Vertical Integration and Supply Chain Exit

Using the quantities developed above, we can directly compute the extent to which a given firm is 'vertically integrated' based on the diagonal elements of the triple product as follows:

$$VI_i = F[i]^2 \;\; [B \cdot V \cdot B']_{i,i}. \tag{2}$$

With this measure, a firm is more vertically integrated when its own 10-K business description contains words that are vertically related (upstream or downstream) to other words also in its own business description. Intuitively, when a firm's business description contains more word pairs that are vertically related, the firm likely offers products or services at different stages of the supply chain. The above measure of

vertical integration is a property of a single firm, whereas upstream or downstream links are a property of a firm-pair.

Note that our text-based definition of vertical integration does not directly measure whether a firm owns and operates upstream and downstream assets, or whether vertical structure is used to actually facilitate movements of physical goods along a production chain within the firm. As highlighted by Atalay, Hortascu, and Syversen (2012), most vertical structures are not primarily concerned with the internal transfer of physical goods but instead mediate efficient transfers of intangible inputs within firms.[15] In the same vein, our text-based measures thus focus on the goods and services that firms provide to the market, and hence capture the potential for transfer of both tangible and intangible inputs along the supply chain.

In addition, we use the IO vocabulary to characterize whether a firm supplies products or services that are related to commodities that exit the U.S. supply chain. Using the exit correspondence matrix E defined above, we can quantify the degree to which a firm is at the end of the supply chain. We compute this quantity as follows:

$$\text{End Users}_i = F[i] \;\; [B \cdot E]_i. \tag{3}$$

A larger value indicates that a firm's product description loads highly on IO commodities that flow to end users. We compute analogous measures of the extent to which the firm sells to retail, the government, or export buyers by replacing the matrix E with the fraction sold to each sub-group instead of the amount sold to all end users.

E Historical NAICS-based Vertical Relatedness Network

We later compare the properties of our text-based vertical network to those of the NAICS-based measure used in previous research. We now describe this historical NAICS-based approach. A critical difference is that the NAICS based network is computed using the BEA industry space, and not the BEA commodity space. This is by necessity given that links to NAICS are at the level of BEA industries. Avoiding

[15] Using comprehensive data on ownership structure, production, and shipment patterns throughout the U.S. economy, they report that upstream units ship only small shares of their outputs to own-firm downstream units.

the need to link to BEA industries is one of many reasons why the textual vertical network offers many advantages. More generally, the compounding of imperfections in BEA industries, NAICS industries, and the Compustat segment tapes, for example, may result in potential horizontal contaminations, especially when firms are in markets that do not cleanly map to NAICS industries.

To compute the NAICS-based network, we use methods that parallel those discussed above for the BEA commoditity space (matrix V), but we instead focus on the BEA industry space and construct a matrix Z. We first compute the BEA industry (not BEA commodity) matrix $IFLOW$ as $SHARE \times USE$, which is the dollar flow from industry i to industry j. We then obtain $ISUPP$ and $ICUST$ by dividing $IFLOW$ by the total production of industry j and i respectively (using parallel notation as was used to describe the construction of V). The matrix Z is simply the average between $ICUST$ and $ISUPP$.

F Granularity

Following common practice in the literature (e.g. Fan and Goyal (2006)), we use two thresholds to identify meaningful relatedness using the NAICS-based vertical network: 1% and 5%. This ensures that our results can be compared to those in the existing literature. A given industry i is then upstream (downstream) relative to industry j when Z_{ij} (Z_{ji}) is larger than this 1% or 5% threshold. Finally, to determine vertical relatedness between individual firms, we use the correspondence table between IO and NAICS to map these values from each firm (from our Compustat/Edgar universe) to a BEA IO industry. It is noteworthy that, based on this historical construct, all the firms in a given NAICS industry are classified as upstream or downstream to all the firms in another NAICS industry if these two industries have vertical relatedness larger than 1% or 5%. We label these two vertical networks as 'NAICS-1%' and 'NAICS-5%', respectively.

We will report in the next section that the NAICS-1% and NAICS-5% networks have granularity levels of 9.48% and 1.37% respectively. That is, 9.48% of all possible firm-pairs (in the Compustat/Edgar universe) are deemed to be vertically related

when using the NAICS-1% network. This network is thus very coarse. At 1.37% granularity, the NAICS-5% network is quite fine. To ensure relevant comparisons between our textual networks and the NAICS networks, we choose two granularity levels for our text-based network that resemble those of NAICS-based networks : 10% and 1%. These two versions of the text-based vertical network thus define as vertically related the top 10% and top 1% most vertically related firm-pairs as indicated by the aforementioned text scores. We label these networks as 'Text-10%' and 'Text-1%'.

G Discussion

Numerous existing studies have used the IO tables to identify vertical relatedness between pairs of Compustat firms. However, these studies do not use verbal links, and instead were required to (1) determine the intensity of vertical relatedness between pairs of IO industries (from the IO tables), and (2) assign each Compustat firm to one IO industry using a correspondence between IO industries and SIC or NAICS codes.[16] This second step is potentially problematic for four reasons. First, the researcher has to rely on the quality of the link between IO industries and NAICS codes. Second, the researcher has to rely upon the individual quality of NAICS codes as a classification, which has been challenged by studies including Hoberg and Phillips (2010a). Third, the researcher faces difficulty to properly assign conglomerate firms into one particular NAICS industry. Four, the researcher has little flexibility regarding choice of granularity, as NAICS are only classified at a small number of arbitrary levels (which may be too broad and may also, as our later results will show, capture horizontal links rather than vertical links).

Our text-based approach does not have the limitations of the NAICS classification. In particular, a firm's upstream and downstream links are customized and unique to its own business description. In contrast, when using NAICS-based methods, any two firms in the same NAICS code will have the same set of upstream and downstream firms. The non-transitive nature of our classification allows us to

[16] We use the 2002 version of the IO table throughout the paper. Because the BEA provides a correspondence between IO and NAICS industries, we focus on the NAICS classification instead of SIC.

accommodate the true intra-industry heterogeneity in product offerings observed in Hoberg and Phillips (2010a), who document that relaxing industry transitivity results in dramatic improvements in a researcher's ability to explain a host of financial variables. Moreover, our text-based approach allows us to compute firm-level vertical integration as a continuous variable and over time. This allows us to examine how vertical integration changes over time, its determinants, and how it relates to merger transactions, and the initiation of customer-supplier relationships. Finally, our text-based approach provides the flexibility to select any granularity level. By defining upstream (downstream) firms as those having a minimum level of upstream (downstream) textual relatedness UP_{ij} ($DOWN_{ij}$), the researcher can calibrate the relatedness network to have any desired granularity.

III Network Statistics and Validation

A Vertical Network Statistics

We compare the properties of five key relatedness networks in this section: Text-10%, Text-1%, NAICS-1%, NAICS-5%, and the TNIC network developed by Hoberg and Phillips (2010a). The first four networks are intended to capture vertical relatedness, and the TNIC network is intended to capture horizontal relatedness. For each network, Panel A of Table I presents various statistics relating to granularity, the degree of overlap, and the degree of focus on financial firms.

[Insert Table I Here]

The first row of Panel A presents the level of granularity of each network. The NAICS-1% and NAICS-5% networks display granularity levels of 9.48% and 1.37% respectively. By construction, the 'Text-10%' and 'Text-1%' vertical networks have 10% granularity and 1% granularity, respectively. Finally, the TNIC network, which is designed to match the granularity of SIC-3, has a granularity of 2.33%. This puts the Text-1% network nicely in perspective, and suggests that this network has slightly less than half the relatedness pairs as does SIC-3.

Reassuringly, the second to fourth rows in Panel A shows that the four vertical networks exhibit low overlap with the horizontal TNIC network, or with the horizon-

tal SIC and NAICS networks. Hence, we conclude that none of the networks have a material direct amount of contamination from known horizontal links. Despite this, the fifth and sixth rows of Panel A illustrate that the vertical networks are quite different in other ways. Only 10.59% of firm-pairs in the NAICS-1% network are also present in the Text-10% network. Similarly, only 1.10% of firm-pairs are in both the Text-1% and NAICS-5% networks.

The last three rows of Panel A are perhaps most the most informative. The eighth row reports the fraction of firm-pairs that includes at least one financial firm (SIC code ranging between 6000 and 6999). The results show that the presence of financial firms is quite low in our text-based vertical network, as it amounts to only 6.78% and 1.01% of vertically linked firms. In contrast, financial firms account for a surprisingly large 48.44% and 34.31% of firm-pairs in the NAICS-based vertical networks. These latter statistics illustrate that each network's propensity to score financial firms as vertically related or not is a first-order dimension upon which these networks disagree. When we discard financial firms, the overlap between our text-based network and NAICS-based network roughly doubles (e.g. 19.90% of non-financial firm-pairs in the NAICS-1% network are also present in the Text-10% network). As theories of vertical relatedness and integration often focus on non-financial concepts such as relationship specific investment and ownership of assets, these results support the use of the text-based vertical network as providing more balance across more theoretically-relevant industries.

B Validating Vertical Orientation

The tests in the previous section illustrate many relevant statistical properties of the vertical networks that we consider. A more direct question is: can we directly show that the relatedness that these networks capture is indeed vertical relatedness? This is an important question, as firms can be related in many ways, including horizontal links, partial horizontal links, or other more sophisticated horizontally oriented links including asset complementarities. A unique feature of vertical relatedness is that it uniquely relates to a supply chain. We construct a test of the extent to which any proposed vertical relatedness network is indeed vertical based on the extent to

which accounts receivable (AR) and accounts payable (AP) respond to shocks in a way that is consistent with adjacency along a supply chain (as opposed to horizontal links, which have parallel positioning, and are not adjacent along a supply chain).

[Insert Figure 1 Here]

To illustrate the implications of supply chain adjacency, consider Figure 1. Suppose firm 5 has a large positive industry shock to its accounts receivable. Because it is horizontally positioned relative to the given firm 5, firm 2 should also experience a positive shock to its accounts receivable. In contrast, firms 3 and 6, because they are downstream relative to firm 5, should experience a large positive shock to their accounts payable. Firms 1 and 4, being upstream, are less likely to experience large shocks in this example as they are one step separated from the shock. If a researcher forms a network trying to portray vertical relationships in this space, we would expect the peers they identify to have adjacent shocks in their ARs and APs if the links are indeed vertical. However, if the researcher mistakenly identifies horizontal links and deems them to be vertical, we would expect the peers they identify to have parallel (not adjacent) shocks in their AR and AP.

To operationalize these predictions in our setting, we consider trade-credit shocks among firm pairs. When AR increases for a supplier, one should expect an adjacent increase in the AP of its customers. We first compute for each firm-year $\triangle AR$ as $\frac{AR_t - AR_{t-1}}{AR_t + AR_{t-1}}$ and $\triangle AP$ as $\frac{AP_t - AP_{t-1}}{AP_t + AP_{t-1}}$.[17] Critical to our examination, we then compute the difference ($\triangle AR - \triangle AP$), which we label $TOPSIDE_{i,t}$ (we choose this label because a high value indicates that firm i is experiencing a larger shock on its "top-side" ARs than on its bottom-side APs).

Intuitively, firms that are horizontally related will tend to experience trade-credit shocks that are positively correlated: top-sided or bottom sided at the same time. In contrast, firms that are vertically related will experience *negatively* correlated trade-credit shocks, and one firm in the vertical pair will be top-sided while the other will be bottom-sided. To measure firm pairwise $TOPSIDE$ correlations for a given network, we estimate the following regression, where one observation is one firm-pair

[17] By construction, $\triangle AR$ and $\triangle AP$ can take value between +1 and -1 and are thus not influenced by outliers.

that is a member of a given network:

$$TOPSIDE_{i,t} = \alpha + \gamma \cdot TOPSIDE_{j,t} + \eta_t + \epsilon_{i,t} \tag{4}$$

The subscript i corresponds to an upstream firm and j to a downstream firm indicated by the given network being tested. We account for time variation in aggregate trade credit shocks (e.g. macroeconomic conditions) by including year fixed effects (η). In more refined tests, we then focus on sub-samples of firm-pair observations where (1) $|\triangle AR_{i,t}| > |\triangle AP_{i,t}|$, or (2) $|\triangle AR_{j,t}| < |\triangle AP_{j,t}|$). The former condition focuses on positive shocks to the AR of upstream firms, while the latter focuses on positive shocks to the AP of downstream firms. The prediction is that the coefficient η should be positive for horizontal networks, and negative for vertical networks.

Panel B of Table I displays the results. As expected, the estimate for η is significantly positive for the TNIC network, which is a horizontal network. In contrast, η is systematically negative for the four vertical networks we consider. However, the estimates of η is far more negative, and are also statistically different from zero, only for our text-based networks. Not surprisingly, results are strongest of all for the text-1% network (the t-statistics ranges between 3.19 and 4.55), where the likelihood of contamination due to excessive breadth is minimized. None of the estimates of η are significant for the NAICS-based vertical network, and the coefficient estimates for η are an order of magnitude smaller.

Overall, these results suggest that our text-based vertical networks are far more likely to capture true vertical links as compared to the historical NAICS-based approach, which exhibits either only very weak vertical links or a fair amount of horizontal contamination.

IV Summary Statistics: Vertical Integration and Vertical Mergers

We now use our text-based approach to analyze firms' degree of vertical integration. We start by describing the properties of our text-based measure of vertical integration (defined above) or sample of mergers and then in the next section examine the

economic determinants of vertical integration and vertical mergers.

A Description and Evolution of Vertical Integration

Table II presents summary statistics for our text-based measure of vertical integration and the main variables used in the analysis. We describe these variables and their construction in the Appendix. The sample covers the period 1996-2008 and exclude regulated utilities and financial firms. We further require observations to have non-missing values for each variable.[18] We have 48,208 firm-year observations.

[Insert Table II and Figure 2 Here]

Statistics for the text-based variables are reported in Panel A. The table indicates a substantial heterogeneity in the degree of vertical integration among firms. The average and median value of vertical integration are 0.791 and 0.431 respectively, but the maximum is 15.402. This skewness is largely confirmed in Figure 2 that displays the empirical distribution of vertical integration. While most firms exhibit a low potential for integration, a non-negligible fraction features business descriptions that contain many words that are strongly vertically related. This conforms to the intuition that most firms in the economy are in fact not vertically integrated, and the firms situated on the right tail of the distribution are likely to be relatively smaller set of firms that are vertically integrated.

Figure 3 displays the evolution of vertical integration over time, and we note a clear trend away from integration over time. The average firm is thus less vertically integrated in recent years than in the late 1990s. Figure 4 shows that vertical integration is heterogeneous across sectors. Based on the Fama and French 48 industries, excluding financials (which we omit from our analysis), sectors such as Health, Drugs, and Telecommunication appear to be essentially non-integrated. This conforms to the intuition of our major results: these sectors are known for higher levels of research and development expenditures, and hence would avoid integration due to the need to ex-ante incentivize investment in relationship specific investment. In contrast, firms active in the Steel, Aerospace, or Electrical Equipment Sectors exhibit

[18] The only exception is Fluidity (see Hoberg, Phillips, and Prabhala (2012)) that can only be measured from 1997 onwards.

a larger degree of vertical integration. These results conform to the mirror image of the above regarding RSI, and additionally conform to a second major finding in our study: vertical integration is more common in more mature industries that have fewer growth options and less fluid product markets.

[Insert Figure 3 and 4 Here]

Panel A of Table II reports additional statistics regarding supply chain exit (these measures identify the extent to which a firm is located at the end of the supply chain, as it sells goods to entities outside the supply chain of US firms). These measures have a simple interpretation as the fraction of output that flows to each end user. The average value for Retail is 0.36, which indicates that roughly 36% of firms' products are linked to IO commodities that are purchased by retail consumers. Similarly, about 3.2% of firms' products are purchased by the U.S. Government, and about 8.7% leave the U.S. supply chain as exports. This latter figure is in line with the survey paper by Bernard, Jensen, Redding, and Schott (2007), who document that about 4% of all U.S. firms were exporters in 2000. This number rises to 18% for manufacturing firms.

Overall, the average total fraction of output that flows to 'End users' is 0.484, with substantial variation in cross section. This variable captures the extent to which a given firm is located near the end of the supply chain. Due to the magnitude of this quantity, we view it as important to control for a firm's proximity to the end of the supply chain in our later analysis of vertical integration, as a firm located near the end of the supply chain has fewer options regarding integration, especially on the downstream side because end users are located downstream. Panel B reports summary statistics for our key control variables. They are generally in line with related studies (see Hoberg and Phillips (2010)).

[Insert Table III Here]

To complete the description of our new measure of vertical integration, Table III presents average summary statistics of our key variables across quartiles of vertical integration. Several interesting patterns emerge. We first remark, as expected, that vertically integrated firms are generally further from the end of the supply chain, and sell less output to retail and the government. However, they are more focused

on exports. This may be related in part to the broader finding we report that firms become more vertically integrated when their product markets are more mature and less fluid. Vertical integration may also be more attractive to exporting firms as a form of risk management, as such firms face greater risk due to international factors, and may see benefits from risk reduction on the domestic aspects of their supply chains.

We also find that vertically integrated firms operate in more stable product markets (they have lower Fluidity as developed by Hoberg, Phillips, and Prabhala (2012)) and also more concentrated markets (based on the TNIC HHI from Hoberg and Phillips (2010)). Moreover, consistent with Atalay, Hortascu, and Syversen (2012), vertically integrated firms appear to be larger, older, and more capital-intensive firms, with lower growth opportunities. Interestingly, integrated firms spend relatively less on R&D than non-integrated firms. Yet, based on the U.S. Patent and Trademark Office data (the NBER data archive), vertically integrated firms own larger portfolios of patents. The differences are quite substantial: firms in the bottom quartile own 4.288 patents on average compared to 10.287 in the top quartile. These univariate results summarize a key finding in the paper, which we document more formally later. Vertical integration is negatively correlated with R&D, likely due to the need to incentivize relationship specific investment ex-ante. However, the realization of successful innovation (patents) marks a time of increased firm maturity, and the end of this cycle as RSI is not needed ex-post. Hence, we observe increased integration at this time.

As expected, vertically integrated firms also have more Compustat segments, where segments are identified from the NAICS-based Compustat segment tapes. For completeness and comparison, we also report an alternative measure of vertical integration that uses the non-text NAICS based vertical network and the Compustat segment tapes: VI(Segment). Following Acemoglu, Johnson, and Mitton (2009), we use the matrix of industry relatedness Z (defined above) to compute the average vertical relatedness across a firm's existing segments. VI(Segment) is assigned a larger value when the segments that a given firm operates in share stronger vertical relations. Although this measure can only be obtained for firms that report more

than *one* segment (which represents less than 33% of observations in our sample), it is positively correlated with our text-based measure (the correlation coefficient is 0.20).

[Insert Table IV Here]

To further illustrate our text-based measure of vertical integration, Table IV displays the 30 most vertically integrated firms in 2008. The most integrated firm is Parker Hannifin (www.parker.com), a leading diversified manufacturer of motion and control technologies and systems, which is active in areas such as Aerospace, Climate Control, Hydraulics, Pneumatics or Process Control. This firm's product offerings have many channels of vertical relatedness. For example, it sells hoses, manifolds, gauges, and also generators heaters, and dryers. The second most integrated firm is Eaton (www.eaton.com), a global company offering a wide range of products for the Automotive, Eletrical, Plastics and Brakes sectors. In particular, this firm offers automotive and truck products and parts at various stages of aggregation: hydraulics, motors, pumps, valves, transmissions, connectors, and electronics. Overall, a review of these most vertically integrated firms suggests a high degree of actual vertical potential.

Interestingly, among these 30 most integrated firms, eight report only a single segment in Compustat despite the fact that their websites indicate product offerings with strong vertical potential. An example of a single segment firm is Franklin Electric. Regarding its product offerings in the area of fueling systems, the firm offers products in the following vertically related areas: fuel dispensing, piping and containment, service station hardware, and even transport. While highly integrated, these firms also generally rank rather low on the measure of integration based on Compustat segments.

[Insert Table V Here]

Table V presents transition matrices examining future vertical integration as a function of initial vertical integration. Panel A displays one-year transition probabilities, and Panels B and C examine three-year and six-year transition probabilities respectively (based on non-overlapping time periods). The rows indicate the initial vertical integration quintile of a firm and the columns indicate which integration

quintile the firm is in t years later. Overall, the results across all panels show that vertical integration is highly persistent. Vertically integrated firms are very likely to be vertically integrated ex-post at one, three, and six year horizons. Interestingly, the persistence is somewhat lower for firms that have medium levels of vertical integration compared to firms that are highly integrated or highly dis-integrated. For example, a firm in the lowest or highest quintile remains in this category 85.8% and 87.8% of the time, respectively. In contrast, firms in the intermediate quintiles are more likely to migrate to a category with lower level of vertical integration over time. For example, in Panel C, firms in quintile 4 move to quintile 3 with a 27.5% probability, while firms in quintile 2 move to quintile 3 with a 20% likelihood. These findings are consistent with Figure 3 and indicate a tendency towards less vertical integration over time.

B Vertical Mergers

Our sample of corporate merger and acquisition transaction is from the Securities Data Corporation SDC Platinum database. We consider all announced and completed U.S. acquisitions with announcement dates between January 1, 1996 and December 31, 2008 that are coded as a merger, an acquisition of majority interest, or an acquisition of assets. To be able to distinguish between vertical and non-vertical transactions, we further require that both the acquirer and the target firm are publicly traded with available Compustat data, and a 10-K that is available on the SEC Edgar website.

Table VI displays the summary statistics describing the composition of our transactions sample. Panel A shows that our sample consists of 5,929 transactions when we limit attention to publicly traded acquirers and targets, and when we exclude transactions that involve financial firms (SIC codes between 6000 and 6999), which we exclude from our regression analysis following convention in the literature. Panel B and Panel C reports how many of these transactions are deemed to be vertical using the Text-10% network and the NAICS-1% network, respectively.

[Insert Table VI Here]

Given that both the Text-10% and NAICS-1% vertical networks are chosen to have similar granularity levels, it is perhaps surprising that the networks disagree rather strongly regarding the fraction of transactions that are vertically related. For our primary sample excluding financials, we observe that 23% are vertically related using the Text-10% network. Using the NAICS-1% network, which is similarly as granular, we observe that just 9.7% are vertically related. This difference, which arises despite roughly the same granularity across the two networks, is economically very large. This difference suggests that the accumulated noise associated with less direct vertical relatedness measures greatly reduces the ability to identify more direct vertically related transactions.

C Wealth Effects in Vertical Mergers

We next consider the stock market reaction to vertical and non-vertical mergers. Table VII displays the average abnormal announcement returns (in percent) of combined acquirers and targets (Panel A), acquirers (Panel B), targets (Panel C) in vertical and non-vertical transactions. Vertical transactions are identified using our new text-based vertical network. We present these results to compare with previous results based on SIC codes and to see if results are stronger with our new text-based vertical network.

[Insert Table VII Here]

Overall the results are strong. Excess stock returns for the acquirer and the combined target and acquirer are positive and are significantly larger for vertical than for non-vertical transactions. Although target returns are very positive, results are not significantly different for vertical and non-vertical deals for targets. Overall, our results show similar patterns relative to those reported in Fan and Goyal (2006).

Our results also help to explain the difference in results regarding value creation in Fan and Goyal (2006) (vertical mergers have positive announcement returns, sample ends in 1996) versus Kedia, Ravid, and Pons (2011) (vertical mergers have negative announcement returns out of sample after 1996). Our sample begins in 1996, and we find that vertical mergers have positive announcement returns using our text-

based measures of vertical relatedness. This suggests that part of the disagreement in this existing literature is likely related to the imperfections of the NAICS-based measures, which likely became more stark as the technology era changed the nature of the product market during this period.

In all, these results shed light on an important debate regarding the value that is created in vertical transactions, and also support to the conclusion that the text-based methods capture vertical integration well.

V The Determinants of Vertical Integration and Vertical Mergers

In this section, we examine the economic determinants of vertical integration and vertical mergers. We are particularly interested in examining the link between the extent of integration, innovation activities, firm maturity, and product market variables.

A Vertical Integration

We estimate panel data regressions where the dependent variable is our text-based measure of vertical integration. To account for the large degree of persistence in this variable, as well as to control for unobserved firm characteristics, we include firm fixed effects. Thus, our regressions focus on how firms *modify* their degree of vertical integration over time, as a function of changes in observable characteristics. We also include year fixed effects to isolate the apparent trend towards dis-integration in our sample, and we cluster standard errors by firm to further account for the persistent nature of vertical integration (see Petersen (2009)). To facilitate the economic interpretation of the results, we standardize the independent variables so that they have unit standard deviation.

[Insert Table VIII Here]

Table VIII presents the regression results. For the main right hand side variables of interest (in particular industry R&D/Sale and #Patents), we consider *industry* characteristics (using the average across TNIC industries) instead of own-firm vari-

ables. This choice is dictated by two important considerations. First, focusing on industry variables lessens endogeneity concerns. For example, the firm (e.g. its senior management) directly chooses the degree of vertical integration (our dependent variable), but has little choice regarding his or her industry's overall level of investment (e.g. R&D activities), especially holding firm fixed effects constant. Second, the theoretical incentives to vertically integrate should be mostly driven by the peculiarities of product markets as a whole (e.g. R&D intensity), rather than firm-specific attributes. For instance, as in Acemoglu, Johnson, and Mitton (2009), the incentive to invest in relationship-specific ventures, or the extent of holdup problems, are primarily determined by the specific product or service being exchanged between firms, and hence this is better identified using industry variables.

The first column of Table VIII displays the central results for our text-based vertical network. Our first main result is that firms operating in industries with high levels of R&D are less vertically integrated. This main result supports our primary hypothesis that firms will avoid integration when ex-ante incentives to invest in relationship specific assets are high. Importantly, this main result does not obtain for the non-text NAICS-based vertical network as seen in column (3). This indicates that more direct identification of vertical relatedness is crucial for this inference, as horizontal contamination or contamination due to asset complementarities, as suggested by our earlier tests in Panel B of Table I, would likely imply a more positive relationship.

Our second major finding is that firms with more patents are more likely to vertically integrate. This is consistent with the ex-post realization of successful innovation alleviating the ex-ante need to incentivize RSI. Hence firms with successful innovation do not face RSI incentives and are also more mature, and hence they are more likely to integrate to reduce the threat of ex-post holdup. Column (2) reveals that our results are also robust when we account for the interaction between R&D and patents. The R&D and patent variables reliably have distinct and opposite effects on firms' propensity to vertically integrate.[19]

[19] Arguably, some industries with high level of R&D activities also display high level of patenting. Yet, in our sample, the correlation between R&D and patenting activity is low. This correlation is 0.15 across industries, and 0.044 across firms.

The third major finding of the table is that firms are more likely to integrate when they are more mature and more stable. In particular, integration is negatively related to fluidity, positively related to PPE/assets, negatively related to the market to book ratio, and positively related to size (log assets). The strong link to firm maturity is likely related to the irreversibility of integration, and firms will be more willing to commit to integration when product markets are more stable, growth options are less important, and hence they are less likely to need to dis-integrate later due to changes in the product market. This issue of irreversibility also relates to the high fixed costs of integration, as integration is likely only profitable if gains are likely to remain stable over a suitably long horizon to amortize the fixed costs.

Finally, Table VIII also shows the results that firms closer to the end of the supply chain (End Users) are less likely to integrate, and firms having an observed conglomerate structure (# Segments) are more likely to integrate.[20]

It is also important to contrast column (1) (text-based vertical integration) and column (3) (non-text NAICS-based vertical integration) of Table VIII. Using the NAICS-based approach, we find no link between vertical integration and R&D and patenting. We also find no link to product market competition, fluidity, or firm size. Remarkably, we even do not observe the more mechanistic and expected link between end users (end of the supply chain) and vertical integration. We would expect that firms near the end of the supply chain have fewer opportunities to integrate as they can only integrate in one direction, whereas firms in the middle of the supply chain have two directions to choose from. The only results we find in column (3) are a strong but mechanistic link to the number of segments, and negative links to firm age and PPE/assets. Hence, not only are results for the NAICS-based measures generally very weak, but these latter results actually conflict with the maturity hypothesis. These results underscore the critical nature of more directly identifying measures of vertical integration.

Overall, the results for column (1) versus column (3) suggest that the cumulative

[20]In unreported tables, we show that the results presented in Table VIII remain unchanged if use lagged independent variables, if we consider the natural logarithm of vertical integration to account for its skewness, if we estimate the model without including firm fixed effects, and if we exclude variables associated with our other findings.

noise required to construct vertical links using (A) mappings from BEA commodities to BEA industries, (B) additional mappings from BEA industries to NAICS industries, and (C) further reliance on the potentially Compustat segment tapes reduces the clarity of the NAICS-based vertical network compared to the text-based network (which does not rely on any industry mappings and does not rely on the quality of the Compustat segment tapes).

B Vertical Mergers and Market Relationships

To better understand the economic forces that drive vertical integration, we perform two complementary analyses. First, we study the determinants of vertical mergers, as vertical transactions represent a direct way firms can modify their degree of integration. Second, we focus on the initiations of major customer-supplier relationships. Arguably, the use of direct relationships (in this case without integration) is a partial substitute for how firms can manage their needs along the supply chain. Examining the determinants of both types of change (vertical acquisitions and new customer/supplier relationships) provides a more complete picture of how vertical integration arises over time, as we can compare integration to separate production.

B.1 The Determinants of Vertical Mergers

We consider logistic regressions to examine which factors explain firms' decision to merge with vertically related firms. The dependent variables are binary variables indicating whether a given firm is an acquirer or a target in vertical or non-vertical transactions in a given year. For consistency, we consider the same set of explanatory variables as in Table VIII and consider lagged values. We also cluster standard errors at the industry and year level, and we consider both the text-based network (Table IX) and the NAICS based network (Table X) when assessing which transactions are in fact vertically related.

Table IX displays the results of these tests when we use our text-based network to identify vertical transactions. We find strong differences in the types of firms that participate in vertical and non-vertical mergers. Columns (1) to (4) show strong

support for our hypothesis regarding the negative link between industry R&D and vertical integration. In particular, column (1) shows that industry R&D is negatively and significantly related to the probability of being an acquirer in a vertical transaction. In contrast, firms in these same R&D intensive industries are in fact more likely to be acquirers in non-vertical transactions (second column). Analogously, targets in non-vertical transactions are also more likely to be in R&D intensive industries (column (4)). In vertical transactions, however, column (3) reveals that the level of R&D is not a significant predictor of transaction incidence. Overall, firms requiring incentives for ex-ante relationship specific investment (i.e. firms in R&D intensive industries) are less likely to merge vertically. This is consistent with our earlier findings that these same firms are also less likely to vertically integrate in general.

[Insert Table IX Here]

We next focus on the level of patenting activity. Consistent with our earlier findings that ex-post successful innovation indicates maturity and lessens the need to incentivize relationship specific investment ex-ante, we find that vertical mergers are more likely to take place between firms in industries with more patenting activity (columns (1) and (3)). The opposite is true for non-vertical mergers, where firms in low patenting industries are less likely to participate in restructuring transactions. In columns (5) to (8) we consider the interaction between R&D and patenting activity. The results confirm again the distinct role played by R&D and patenting activity in explaining the incidence of vertical transactions.

Table IX also shows broad support for our hypothesis that maturity is an important positive determinant of vertical integration. Column (3) indicates that firms with lower market-to-book ratios, and those in capital intensive and stable industries (i.e. low fluidity), are more likely to be targets of vertical mergers. In contrast, targets in non-vertical mergers are more likely to be in fluid product markets in which capital intensity is lower. Across specifications, we also observe that older firms are more likely to participate in vertical acquisitions, while younger firms are more likely to be involved in non-vertical deals. These findings are consistent with the following interpretation of the U-shaped relationship between firm age and restructuring activity noted in Arikan and Stulz (2011). Younger firms engage in non-vertical

transactions likely to capitalize on asset complementarities, and older more mature firms experience an increase in acquisition activity as their focus turns to vertical acquisitions.

Table IX also reveals that vertical transactions are not related to the degree of product market competition (measured with the TNIC HHI). Hence, our earlier finding that firms in more competitive markets are more likely to integrate is not driven by vertical transactions, and instead is thus more likely due to organically-driven integration. We conclude that the observed link between competition and integration is likely due to risk management concerns in competitive markets, and not due to market power hypotheses (which would be more salient in transactional data given the direct effects transactions have on market structure). Analogous to our earlier findings, we also find the expected result that firms closer to the end of the supply chain are less likely to participate in vertical mergers.

The results are less decisive if we instead focus on the NAICS-based network to differentiate between vertical and non-vertical transactions (Table X). While firms in R&D intensive industries do appear to participate in non-vertical transactions, there is no evidence that relationship specific investment through R&D intensity is related to vertical mergers. This difference might be due to horizontal contamination or contamination relating to asset complementarities, as discussed earlier. Furthermore, patent intensity is only related to the incidence of vertical targets and not vertical acquirers. We also find only mixed results supporting the maturity hypothesis when we consider the NAICS-based network. In particular, vertical targets are positively related to firm age and negatively related to the M/B ratio as expected, but these results switch sign when we consider vertical acquirers.

[Insert Table X Here]

Overall, when we consider the text-based vertical network, our results are broadly in line with our earlier results regarding vertical integration and the problems with the NAICS-based network to measure vertical integration. Specifically, if the NAICS-based network is an unreliable way to measure vertical integration due to non-vertical links, especially asset complementarities, we would likely expect a positive link to R&D given the direct link between R&D and technological asset complementari-

ties. Our receivables and payables validation test reported earlier, coupled with this finding, support this conclusion.

B.2 Product-Market Relationships

We now focus on major customer and supplier relationships with business partners. Because such market transactions occur *outside* firm boundaries, they represent substitutes for vertical integration. For example, a manager needing input (e.g. a good or an service) can either generate it on her own through a vertically integrated structure, or can buy from a supplier without integration.

We rely on two different data sources to identify the intensity of major customer or supplier relationships. First, as in Fee and Thomas (2004), we consider the Compustat Segment files. Public firms must report important customers that provide at least 10% of overall revenues, and in most cases they report the identity (i.e. the name in abbreviated or truncated form) of the customer. Following Fee and Thomas (2004), we use a text-based matching algorithm to identify the customer firms. Focusing on (non-financial) firms in our sample, we identify 7,503 customer-supplier pairs involving 3,147 unique suppliers and 1,913 customers over the period 1996-2008.

We also consider Capital IQ's "Key Developments" database as an alternative measure of new customer-supplier relationships. This database contains corporate announcements from over 20,000 sources including public news announcements, company press releases, regulatory filings, call transcripts, and company websites.[21] We concentrate exclusively on announcements classified as 'client announcements' for which both partners have available links to the Compustat and the SEC Edgar database. Because coverage starts in 2002, we restrict attention to announcements occurring between 2002 and 2008. Our sample consists of 2,586 (non-financial) pairs of clients.

[21] Capital IQ filters this data to eliminate duplicates, extraneous information, identify the companies involved, and to categorize each announcement into event categories. Categories include M&A rumors, new business expansion, changes in corporate guidance, delayed filing, or new clients. Note that the matching of Capital IQ announcing firms and the Compustat database is perfect as both datasets are obtained from the same data provider.

Using both sources, we then study how R&D intensity, the number of patents, and firm maturity relates to *newly initiated* market relationships. Based on the Compustat Segment data, we define an initiation as an instance in which a firm pair first appears as a customer-supplier pair in a given year. With the Capital IQ data, we define an initiation as an instance of a firm pair relationship newly appearing in a given year. The Capital IQ approach has the limitation that a relationship is only visible in the data when the firm announces its business relationship.

Table XI displays regressions in which the dependent variable is an indicator variable that is one when the firm initiated a significant reported customer or supplier relationship. We include similar R&D and patenting variables and the other variables as in our vertical merger tables. As before, the right-hand-side variables are lagged, and the standard errors are clustered at the industry and year level.

[Insert Table XI Here]

Table XI reports results from logistic regressions examining the likelihood of initiating a market relationship. Several important results obtain. First, firms are more willing to have customer-supplier links in industries with high R&D intensity. We find a positive and significant relationship between R&D intensity and the likelihood of reporting new major customer-supplier relationships. The cross term reveals that this relationship is strengthened in high R&D industries that also have a significant numbers of patents consistent with the conclusion that firms that are in industries that continue to conduct R&D and patent needing the stronger incentives for relationship specific investment that come from separate ownership. These results hold across both data sources (Compustat segment-based data results in column (1) and Capital IQ-based Announcement-based in column (2)).

Overall these findings are consistent with firms in high R&D product markets substituting toward more customer-supplier relationships and away from vertical integration. These findings also provide especially strong support for the Grossman and Hart (1986) and the Hart and Moore (1990) theories of relationship specific investment. In particular, firms needing greater ex-ante incentives for relationship specific investment (R&D firms) remain non-integrated to maintain RSI incentives. These firms instead initiate new major customer-supplier relationships, which in fact

promote these RSI incentives.

VI Conclusions

Our paper studies when firms choose to vertically integrate and transact within the firm. We focus on both changes to overall firm-specific vertical integration, and also vertical mergers. We examine how the distinction between incentives for ex-ante relationship specific investment and the potential for ex-post holdup both influence vertical integration and vertical transactions. We also examine the role of supply chain maturity, as vertical integration is more likely when the supply chain is stable enough to support long-term gains from operational synergies.

We measure vertical integration at the firm level across multiple industries using computational text analysis of firm product descriptions and how they relate to product words from the input-output tables. This new measure captures how vertically integrated firms are, how their vertical integration changes over time and the extent mergers represent vertical transactions.

We show that the distinction between unrealized innovation through R&D and realized innovation through patents affects the propensity to vertically integrate through own production and also through vertical mergers. Firms in high R&D industries are less likely to vertically integrate through either channel. This finding is consistent with firms remaining separate to maintain ex-ante incentives to invest in relationship specific assets and maintain residual rights of control. These firms thus have high incentives to invest in technologies that can be used by relationship partners.

In contrast, firms in industries with high patenting rates and thus high realized innovation and strong property rights are more likely to vertically integrate both through own production and vertical mergers. In high patenting industries, owners have more legally enforceable residual rights of control and are more likely to merge as the residual rights of control can be obtained by commercializing firms to prevent ex-post holdup..

We also find strong empirical support for firm maturity impacting vertical integration. Both vertical integration and vertical mergers are positively related to a myriad of variables with links to maturity and stability: older firm age, lower M/B ratio, higher PP& E/assets, and lower product market fluidity. These results suggest that firms are more likely to integrate as industries mature and a stable industry outlook is present. This is consistent with firm's needing industries to be stable for longer periods so that they can justify the additional costs of integration.

Overall our paper reconciles some of the tension between the ex-post hold up literature of Klein, Crawford and Alchain (1979) and Williamson (1979), and the ex-ante incentive effects of assigning residual rights of control as in Grossman and Hart (1986). We show that firms in industries with high levels of unrealized innovation are less likely to integrate either through own production or vertical mergers, while firms in industries with significant realized innovation and stronger property rights are more likely to integrate.

Appendix

In this appendix, we describe the control variables used in this study (see Panel B of Table II). We report Compustat items in parenthesis when it applies. All ratios are winsorized at 1% in each tail.

- *Fluidity* is computed as the Product Market Fluidity variable in Hoberg, Phillips, and Prabhala (2012)). This variable measures the extent to which rivals in a given firm's product market are experiencing changes in their product offerings.
- *HHI* measures the degree of concentration (of sales) within TNIC industries. We compute HHI as the TNIC HHI in Hoberg and Phillips (2010a) using the text-based TNIC horizontal industry network.
- *R&D/sales* is equal to research & development expenses (XRD) scaled by the level of sales (SALE). This variable is set to zero when R&D is missing.
- *#Patents* is the number of (granted) patents a firm owns. Patents data are obtained from the US Patent and Trademark Office data (the NBER Patent data archive).
- *PPE/assets* is equal to the level of property, plant and equipment (PPENT) divided by total assets (AT).
- *Log(assets)* is the natural logarithm of the firm assets (AT).
- *Log(age)* is the natural logarithm of one plus the firm age. Age is computed as the current year minus the firm's founding date. When we cannot identify a firm's founding date, we use its listing vintage (based on the first year the firm appears in the Compustat database).
- *#Segments* is the number of operating segments observed for the given firm in the Compustat segment database. We measure operating segments based on the NAICS classification.
- *MB* is the firm's Market-to-Book ratio. It is computed as total assets (TA) minus common equity (CEQ) plus the market value of equity ((CSHO×PRCC_F)) divided by total assets.

- *VI(Segment)* measures firm-level vertical integration based on Compustat Segments. It is computed as the average vertical relatedness across a firm's distinct NAICS segments. Vertical relatedness is based on the matrix Z (defined in Section II.G) that relies on the 2002 BEA Input-Output table.

References

Acemoglu, Daron, 1996, A microfoundation for social increasing returns in human capital accumulation, *Quarterly Journal of Economics* 111, 779–804.

——— , Simon Johnson, and Todd Mitton, 2009, Determinants of vertical integration: Financial development and contracting costs, *Journal of Finance* 64, 1251–1290.

Ahern, Kenneth, and Jarrad Harford, 2012, The importance of industry links in merger waves, *Journal of Finance (forthcoming)* University of Michigan and University of Washington Working Paper.

Allen, Jeffrey W., and Gordon M. Phillips, 2000, Corporate equity ownership, strategic alliances, and product market relationships, *Journal of Finance* 55, 2791–2815.

Arikan, Asli, and Rene M. Stulz, 2011, Corporate acquisitions, diversification, and the firm's lifecycle, Ohio State University Working Paper.

Atalay, Enghin, Ali Hortascu, and Chad Syversen, 2012, Why do firms own production chains?, *Working Paper*.

Bena, Jan, and Kai Li, 2013, Corporate innovations and mergers and acquisitions, *Journal of Finance (forthcoming)*.

Bernard, Andrew B., Bradford J. Jensen, Stefan J. Redding, and Peter K. Schott, 2007, Firms in international trade, *NBER Woking Paper 13054*.

Bresnahan, Timothy, and Jonathan Levin, 2012, Vertical integration and market structure, Working Paper.

Cen, Ling, Sudipto Dasgupta, and Rik Sen, 2011, Discipline or disruption? stakeholder relationships and the effect of takehover threats, .

Cohen, Wesley M., Richard R. Nelson, and John P. Walsh, 2000, Protecting their intellectual assets: Appropriability conditions and why us manufacturing firms patent (or not), *National Bureau of Economic Research* 7552, 1–50.

Fan, Joseph, and Vidhan Goyal, 2006, On the patterns and wealth effects of vertical mergers, *Journal of Business* 79, 877–902.

Fee, Edward C., and Shawn Thomas, 2004, Sources of gains in horizontal mergers: Evidence from customers, suppliers, and rival firms, *Journal of Financial Economics* 74, 423–460.

Grossman, Sanford J., and Oliver D. Hart, 1986, The cost and benefits of ownership: A theory of vertical and lateral integration, *Journal of Political Economy* 94, 691–719.

Harford, Jarrad, 2005, What drives merger waves?, *Journal of Financial Economics* 77, 529–560.

Hart, Oliver, and John Moore, 1990, Property rights and the nature of the firm, *Journal of Political Economy* 98, 1119–1158.

Hoberg, Gerard, and Gordon Phillips, 2010, Product market synergies in mergers and acquisitions: A text based analysis, *Review of Financial Studies* 23, 3773–3811.

——— , 2010a, Text-based network industry classifications and endogenous product differentiation, University of Maryland Working Paper.

——— , and Nagpurnanand Prabhala, 2012, Product market threats, payout, and financial flexibility, *Journal of Finance (forthcoming)*.

Hyland, David, 1999, Why firms diversify: An empirical investigation, Working Paper.

Johnson, William C., Jonathan M. Karpoff, and Sangho Yi, 2012, Why do ipo firms have takeover defenses?, .

Joskow, Paul L., 1987, Contract duration and relationship specific investments, *American Economic Review* 77, 168–175.

Jovanovic, Boyan, and P Rousseau, 2002, The q-theory of mergers, *American Economic Review* 92, 198–204.

Kale, Jayant R., and Husayn K. Shahrur, 2007, Corporate capital structure and the characteristics of suppliers and customers, *Journal of Financial Economics* 83, 321–365.

Kedia, Simi, Abraham Ravid, and Vicente Pons, 2011, When do vertical mergers create value?, *Financial Management* 40, 845–877.

Klein, Benjamin, Robert G. Crawford, and Armen A. Alchian, 1978, Vertical integration, appropriable rents, and the competitive contracting process, *Journal of Law and Economics* 21, 297–326.

Lafontaine, Francine, and Margaret Slade, 2007, Vertical integration and firm boudaries: The evidence, *Journal of Economic Literature* 45, 629–685.

Maksimovic, Vojislav, and Gordon Phillips, 2001, The market for corporate assets: Who engages in mergers and asset sales and are there efficiency gains?, *Journal of Finance* 56, 2019–2065.

Masten, Scott E., 1984, The organization of production: Evidence from the aerospace industry, *Journal of Law and Economics* 27, 403–417.

Monteverde, Kirk, and David J. Teece, 1982, Supplier switching costs and vertical integration in automobil industry, *Bell Journal of Economics* 13, 206–213.

Morck, Randall, Andrei Shleifer, and Robert Vishny, 1990, Do managerial motives drive bad acquisitions, *Journal of Finance* 45, 31–48.

Petersen, Mitchell, 2009, Estimating standard errors in finance panel data sets: Comparing approaches, *Review of Financial Studies* 22, 435–480.

Phillips, Gordon M., and Alexei Zhdanov, 2013, R&d and the incentives from merger and acquisition activity, *Review of Financial Studies* 34-78, 189–238.

Sebastiani, Fabrizio, 2002, Machine learning in automated text categorization, *ACMCS* 34, 1–47.

Shahrur, Husayn, 2005, Industry structure and horizontal takeover: Analysis of wealth effects on rivals, suppliers, and corporate customers, *Journal of Financial Economics* 76, 61–98.

Villalonga, Belen, 2004, Does diversification cause the diversification discount, *Financial Management* 33, 5–27.

Williamson, Olivier E., 1971, The vertical integration of production: Market failure consideration, *American Economic Review* 61, 112–123.

———, 1979, Transaction-cost economics: The governance of contractual relations, *Journal of Law and Economics* 22, 233–261.

Table I: Vertical Network Summary Statistics

This table displays characteristics of our new Text-based vertical network and the existing NAICS-based vertical network. Our sample is based on annual firm observations from 1996 to 2008. We consider five networks: Text-10% and Text-1% networks correspond to vertical networks based on textual analysis (see Section II) set at a 10% and respectively 1% granularity level, NAICS-1% and NAICS-5% correspond to vertical networks based in the 2002 BEA Input-Output Table with relatedness cutoffs of 1% and 5% respectively, and TNIC corresponds to the horizontal Text-based Network Industry Classification developed by Hoberg and Phillips (2010). Panel A presents the level of granularity of each network, their overlap, the fraction of firm-pairs in each network sharing the same NAICS or SIC industries, as well as the fraction of firm-pairs containing a financial firm (SIC code between 6000 and 6999). Panel B presents validity tests for vertical relatedness based on the correlation of trade credit shocks among vertically related firms (as outlined in Section III.). The coefficient γ is obtained from OLS regressions of trade credit shocks of upstream firms on trade credit shocks of downstream firms. We report t-statistic below the coefficients.

Network:	Text-10%	Text-1%	NAICS-1%	NAICS-5%	TNIC
Panel A: Granularity and Overlap					
Granularity	10%	1%	9.48%	1.37%	2.33%
% of pairs in TNIC	1.09%	2.30%	2.67%	2.89%	100%
% of pairs in the same SIC	0.64%	0.91%	0.35%	0.20%	38.10%
% of pairs in the same NAICS	0.48%	0.56%	0.30%	0.18%	38.11%
% of pairs in the same SIC or NAICS	0.96%	0.74%	0.35%	0.20%	41.24%
% of pairs in Text-10%	100%	100%	10.59%	12.90%	4.80%
% of pairs in Text-1%	10%	100%	1.10%	1.12%	1.00%
% of pairs that include a financial firm	6.79%	1.01%	48.44%	34.31%	58.72%
% of (no fin.) pairs in Text-10%	100%	100%	19.90%	19.29%	11.63%
% of (no fin.) pairs in Text-1%	10%	100%	2.14%	1.71%	2.44%
Panel B: Validity Test (Trade Credit Shocks)					
γ (unconditional)	-0.0003	-0.0024	-0.0001	-0.0001	0.0071
(t-statistic)	(2.02)	(4.55)	(0.93)	(0.03)	(15.91)
γ (if $\mid \triangle AR_{i,t} \mid > \mid \triangle AP_{i,t} \mid$)	-0.0002	-0.0028	-0.0002	-0.0007	0.0071
(t-statistic)	(0.85)	(3.50)	(0.90)	(0.92)	(12.37)
γ (if $\mid \triangle AR_{j,t} \mid < \mid \triangle AP_{j,t} \mid$)	-0.0004	-0.0022	-0.0001	-0.0005	0.054
(t-statistic)	(1.96)	(3.19)	(0.051)	(0.06)	(8.38)

Table II: Summary Statistics

This table displays summary statistics for the main variables used in the analysis. Our sample is based on annual firm observations from 1996 to 2008. Observations are required to be in COMPUSTAT and our 10-K database. All variables are defined in the Appendix. Panel A contains variables that are constructed from textual analysis (firm 10-Ks and the text description of the 2002 BEA Input-Output Table). Panel B contains variables that are constructed from COMPUSTAT, CRSP, and the NBER Patent dataset.

Variable	Mean	St. Dev	Min	p25	p50	p75	Max	#Obs.
Panel A: Data from Text Analysis								
VI	0.791	1.015	0.000	0.185	0.431	0.994	15.402	48,208
Retail	0.364	0.078	0.053	0.309	0.358	0.411	0.910	48,208
Government	0.032	0.018	0.000	0.020	0.029	0.040	0.513	48,208
Export	0.087	0.021	0.000	0.073	0.088	0.101	0.272	48,208
End users	0.484	0.073	0.100	0.435	0.484	0.529	0.919	48,208
Panel B: Data from Existing Literature								
Fluidity	6.785	3.280	0.329	4.330	6.263	8.718	24.113	43,776
HHI	0.258	0.224	0.014	0.103	0.176	0.332	1.000	48,208
R&D/sales	0.061	0.123	0.000	0.000	0.001	0.068	0.669	48,208
#Patents	8.022	65.158	0.000	0.000	0.000	1.000	3621.917	48,208
PPE/assets	0.260	0.218	0.000	0.091	0.193	0.366	0.888	48,208
log(assets)	5.787	1.844	2.529	4.388	5.598	6.983	10.881	48,208
log(age)	2.975	1.087	0.000	2.303	2.996	3.714	5.037	48,208
#Segments (NAICS)	1.590	1.067	1.000	1.000	1.000	2.000	12.000	48,208
MB	2.010	1.469	0.624	1.120	1.516	2.286	8.351	48,208
VI(Segment)	0.015	0.043	0.000	0.000	0.000	0.007	0.639	48,208

Table III: Summary Statistics Across Vertical Integration Quartiles

This table displays summary statistics for the main variables used in the analysis by (annual) quartiles of vertical integration (VI). Our sample is based on annual firm observations from 1996 to 2008. Observations are required to be in COMPUSTAT and our 10-K database. All variables are defined in the Appendix. Panel A contains variables that are constructed from textual analysis (firm 10-Ks and the text description of the 2002 Input-Output Tables). Panel B contains variables that are constructed from COMPUSTAT, CRSP, and the NBER Patent dataset.

Variables	Quartile 1 (Low VI)	Quartile 2	Quartile 3	Quartile 4 (High VI)
Panel A: Data from Text Analysis				
VI	0.096	0.297	0.668	2.102
Retail	0.400	0.385	0.360	0.310
Government	0.041	0.032	0.029	0.026
Export	0.082	0.084	0.088	0.095
End users	0.524	0.501	0.477	0.432
Panel B: Data from Existing Literature				
Fluidity	8.107	7.706	6.372	4.959
HHI	0.223	0.248	0.280	0.282
R&D/sales	0.111	0.065	0.047	0.023
#Patents	4.288	7.152	10.358	10.287
PPE/assets	0.189	0.250	0.286	0.315
log(assets)	5.488	5.629	5.808	6.222
log(age)	2.669	2.764	3.039	3.428
#Segments (NAICS)	1.319	1.445	1.589	2.007
MB	2.492	2.110	1.873	1.567
VI(Segment)	0.005	0.009	0.016	0.028

Table IV: Examples of Vertically Integrated firms: Top 30 in 2008

This table displays to 30 most vertically integrated firms in 2008 based on our text-based measure of vertical integration (VI). For each firm, the table also presents the number of segments (from COMPUSTAT), the location on the empirical distribution of VI, as well as the location on the empirical distribution on the NAICS-based measure of vertical integration (VI (NAICS).

Company	Rank	#Segments	Perc.(VI)	Perc.(VI (NAICS))
PARKER-HANNIFIN CORP	1	2	1.000	0.000
EATON CORP	2	5	0.999	0.966
EMERSON ELECTRIC CO	3	6	0.999	0.991
HANDY & HARMAN LTD	4	5	0.999	0.969
DOVER CORP	5	4	0.998	0.641
FRANKLIN ELECTRIC CO INC	6	1	0.998	0.717
SONOCO PRODUCTS CO	7	3	0.997	0.000
MATTHEWS INTL CORP -CL A	8	6	0.997	0.884
P & F INDUSTRIES -CL A	9	2	0.997	0.760
CLARCOR INC	10	3	0.996	0.000
LEGGETT & PLATT INC	11	3	0.996	0.710
AMPCO-PITTSBURGH CORP	12	2	0.996	0.681
CARLISLE COS INC	13	6	0.995	0.962
SYNALLOY CORP	14	2	0.995	0.000
SCHNITZER STEEL INDS -CL A	15	3	0.995	0.000
HARBIN ELECTRIC INC	16	1	0.994	0.717
SUPREME INDUSTRIES INC	17	1	0.994	0.851
EASTERN CO	18	2	0.993	0.000
MYERS INDUSTRIES INC	19	1	0.993	0.000
LYDALL INC	20	4	0.993	0.775
TRINITY INDUSTRIES	21	5	0.992	0.545
PULSE ELECTRONICS CORP	22	4	0.992	0.858
GIBRALTAR INDUSTRIES INC	23	1	0.992	0.000
HAWK CORP	24	1	0.991	0.000
COMMERCIAL VEHICLE GROUP INC	25	1	0.991	0.000
ROCK-TENN CO	26	3	0.991	0.664
ENPRO INDUSTRIES INC	27	3	0.990	0.992
GORMAN-RUPP CO	28	1	0.990	0.681
PARK OHIO HOLDINGS CORP	29	3	0.989	0.724
GREIF INC -CL A	30	5	0.989	0.000

Table V: Transition Probabilities for Vertical Integration

This table presents transition matrices examining ex-post text-based vertical integration (VI) for firms with varying levels of initial VI. Our sample is based on annual firm observations from 1996 to 2008. We examine the one, three and six year transition probabilities, respectively. In Panel A we examine the distribution of one year transition probabilities. In Panel B, we divide our sample into four non-overlapping three year intervals. We then examine how VI changes for firms from one interval to the next. For the six year text, we divide our sample into two six years non-overlapping intervals and repeat the same test. In each case, we assign firms to quintiles in each ex-ante interval based on the level on ex-ante VI. Holding breakpoints fixed for each interval, we then group observations in quintiles based on their ex-post level of VI. The results is a 5x5 grid, in which we compute the empirical distribution of transitions.

Sample	Quintile 1 (Least VI)	Quintile 2	Quintile 3	Quintile 4	Quintile 5 (Most VI)	# Obs.
			Panel A: One-year Transition Probabilities			
Quintile 1	0.858	0.130	0.009	0.002	0.001	7359
Quintile 2	0.151	0.678	0.158	0.011	0.002	7601
Quintile 3	0.016	0.174	0.671	0.135	0.005	7641
Quintile 4	0.004	0.014	0.153	0.723	0.105	7639
Quintile 5	0.001	0.002	0.006	0.114	0.878	7568
			Panel B: Three-year Transition Probabilities			
Quintile 1	0.785	0.176	0.027	0.008	0.004	4342
Quintile 2	0.229	0.536	0.199	0.034	0.003	4704
Quintile 3	0.041	0.231	0.523	0.189	0.016	4784
Quintile 4	0.012	0.041	0.226	0.573	0.148	4818
Quintile 5	0.002	0.005	0.022	0.172	0.799	4783
			Panel C: Six-year Transition Probabilities			
Quintile 1	0.737	0.198	0.038	0.017	0.010	1988
Quintile 2	0.297	0.437	0.200	0.057	0.009	2353
Quintile 3	0.072	0.270	0.434	0.194	0.030	2457
Quintile 4	0.023	0.072	0.275	0.473	0.156	2660
Quintile 5	0.003	0.013	0.045	0.213	0.726	2701

Table VI: Vertical Mergers - Sample Description

This table presents the sample of M&A transactions. We look separately at transactions that involve two non-financial firms (column (1)), and transactions that include financial firms (column (2)). Financial firms are firms with (4-digit) SIC code between 6000 and 6999. Panel A includes all transactions where the acquirer or the target are public firms included in COMPUSTAT. Panel B includes all transactions where acquirers and targets are public firms included in COMPUSTAT. Panel C includes all transactions where the participating firms (acquirer and target) are included in the text-based vertical network (Text-10%). Panel D includes all transactions where the participating firms (acquirer and target) are included in the NAICS-based vertical network (NAICS-1%)

Sample	Excluding Financials	Including Financials
	(1)	(2)
Panel A: Transactions with identified (public) acquirers and targets		
# Transactions	5,929	7,848
# Mergers	2,087	2,918
# Acquisitions	3,842	4,930
Panel B: Vertical Transactions - Text-based network		
# Transactions	1,369	1,493
Fraction of Vertical Transactions	23%	19%
# Upstream Transactions	696	755
# Downstream Transactions	673	738
Panel C: Vertical Transactions - NAICS-based network		
# Transactions	577	856
Fraction of Vertical Transactions	9.7%	10.9%
# Upstream Transactions	268	400
# Downstram Transactions	309	456

Table VII: Wealth Effects in Vertical and Non-Vertical Mergers

This table displays the average abnormal announcement returns (in percent) of combined acquirers and targets (Panel A), acquirers (Panel B), targets (Panel C) in vertical and non-vertical transactions. Vertical transactions are identified using our new Text-based vertical network (Text-10%). The sample is from 1996 to 2008. Announcement returns are computed over various windows: CAR(-10,0) is computed between day $t = -10$ to day $t = 0$ (where $t = 0$ is the announcement date); CAR(0) is computed at $t = 0$; and CAR(-1,1) is computed between day $t = -1$ to day $t = +1$. The abnormal returns results after subtracting the return of the CRSP value-weighted market index over each event window. The combined abnormal return is obtained as the weighted sum of the acquirer and target abnormal returns, where the weights correspond to their relative market value (evaluated at day $t = -10$)). The last column reports the t-statitics corresponding to tests of difference between vertical and non-vertical transactions [(i)-(ii)].

Deal type:	All Transactions	Vertical (i)	Non-Vertical (ii)	(i)-(ii)
		Panel A: Combined Acquirers and Targets		
CAR(-10,0)	0.60%	1.00%	0.47%	(2.15)
CAR(0)	0.41%	0.71%	0.31%	(3.69)
CAR(-1,1)	0.75%	1.10%	0.64%	(2.78)
		Panel B: Acquirers		
CAR(-10,0)	0.22%	0.48%	0.13%	(1.18)
CAR(0)	0.18%	0.47%	0.08%	(2.84)
CAR(-1,1)	0.46%	0.88%	0.32%	(2.60)
		Panel C: Targets		
CAR(-10,0)	7.19%	6.97%	7.27%	(0.45)
CAR(0)	5.21%	5.24%	5.20%	(0.01)
CAR(-1,1)	7.83%	7.47%	7.96%	(0.80)

Table VIII: The Determinants of Vertical Integration

This table examines the determinants of vertical integration. OLS Regressions are based on annual firm observations from 1996 to 2008. The dependent variable is a measure of vertical integration (as defined in Section II.D). A higher value implies that the firm is more vertically integrated. In columns (1) and (2), it is the text-based measure of vertical integration. In columns (3) and (4), it is the NAICS-based measure. One observation is one firm in one year. All the variables are defined in the Appendix. TNIC(X) refers to the industry (TNIC) average of variable X. We include a dummy variables that equals one for missing values of Fluidity (that starts only in 1997). The independent variables are standardized to have a standard deviation of one prior to fitting regressions to permit more intuitive comparisions and interpretation. All specifications include firm and year fixed effects. Standard errors are clustered by firm and correct for heteroskedasticity. We report t-statistics in parentheses.

Dependent Variable:	VI	VI	VI	VI
Measure:	(Text-based)		(NAICS-based)	
	(1)	(2)	(3)	(4)
TNIC(R&D/sale)	-0.024	-0.023	0.000	0.000
	(5.49)	(5.23)	(1.10)	(0.86)
TNIC(#Patents)	0.008	0.010	0.000	0.000
	(3.15)	(1.98)	(1.43)	(0.37)
TNIC(R&D/sale) × TNIC(#Patents)		-0.002		0.002
		(0.54)		(1.17)
TNIC(Fluidity)	-0.012	-0.012	0.000	0.000
	(2.08)	(2.06)	(0.99)	(0.97)
TNIC(PPE/asset)	0.028	0.028	-0.001	-0.001
	(3.64)	(3.63)	(2.25)	(2.22)
TNIC(HHI)	-0.079	-0.079	0.000	0.000
	(16.45)	(16.45)	(0.46)	(0.47)
End Users	-0.070	-0.070	0.000	0.000
	(13.19)	(13.18)	(0.27)	(0.27)
#Segments (NAICS)	0.029	0.029	0.013	0.013
	(3.32)	(3.32)	(15.24)	(15.25)
log(Assets)	0.054	0.054	-0.001	-0.001
	(6.53)	(6.53)	(1.64)	(1.64)
log(Age)	0.024	0.024	-0.001	-0.001
	(2.87)	(2.86)	(2.23)	(2.23)
MB	-0.006	-0.006	0.000	0.000
	(3.21)	(3.22)	(1.16)	(1.17)
#Obs.	48,208	48,208	48,208	48,208
Adjusted R2	0.878	0.878	0.823	0.823

Table IX: The Determinants of Vertical Mergers - Text-based

This table examines the determinants of vertical and non-vertical mergers. We display results from logistic regression where the dependent variable is a dummy indicating whether the given firm is an acquirer or a target in a vertical or non-vertical merger or an aquisition of assets in a given year. One observation is one firm in one year. Vertical transactions are identified using our new Text-based vertical network (Text-10%). The sample is from 1996 to 2008. All the independent variables are lagged by one year and are defined in the Appendix. All estimations include year fixed effects. TNIC(X) refers to the industry (TNIC) average of variable X. We include a dummy variables that equals one for missing values of Fluidity (that starts only in 1997). Standard errors are clustered by industry and year. We report t-statistics in parentheses.

Dependent variable:	Acquirer?		Target?		Acquirer?		Target?	
Deal type:	Vertical	Non-Vertical	Vertical	Non-Vertical	Vertical	Non-Vertical	Vertical	Non-Vertical
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
TNIC(R&D/sale)	-1.458	1.017	-0.287	1.472	-1.565	0.869	-0.502	1.602
	(3.09)	(5.21)	(0.63)	(6.80)	(3.31)	(3.93)	(1.08)	(6.71)
TNIC(#Patents)	0.004	-0.003	0.004	-0.004	0.004	-0.005	0.003	-0.002
	(4.91)	(1.90)	(4.63)	(2.24)	(3.35)	(2.19)	(2.74)	(0.84)
TNIC(R&D/sale) × TNIC(#Patents)					0.010	0.021	0.020	-0.020
					(0.72)	(1.66)	(1.70)	(1.18)
TNIC(Fluidity)	-0.019	0.046	-0.030	0.057	-0.019	0.046	-0.029	0.057
	(1.34)	(5.66)	(2.21)	(6.61)	(1.33)	(5.66)	(2.18)	(6.64)
TNIC(PPE/asset)	-0.089	-0.559	0.093	-0.717	-0.088	-0.561	0.100	-0.714
	(0.42)	(2.15)	(0.43)	(3.63)	(0.41)	(2.17)	(0.46)	(3.62)
TNIC(HHI)	-0.182	-0.080	0.092	-0.338	-0.180	-0.072	0.097	-0.347
	(1.05)	(0.74)	(0.52)	(2.89)	(1.04)	(0.66)	(0.55)	(2.98)
End Users	-6.062	0.130	-5.601	1.042	-6.066	0.120	-5.609	1.052
	(13.99)	(0.42)	(11.93)	(3.14)	(13.98)	(0.38)	(11.94)	(3.18)
#Segments (NAICS)	0.263	0.025	0.217	0.066	0.264	0.027	0.218	0.063
	(11.73)	(1.39)	(9.44)	(3.41)	(11.79)	(1.52)	(9.54)	(3.26)
log(Assets)	0.351	0.432	0.385	0.307	0.350	0.432	0.384	0.307
	(18.07)	(32.49)	(17.72)	(22.06)	(17.92)	(32.42)	(17.65)	(22.11)
log(Age)	0.046	-0.079	0.153	-0.048	0.046	-0.079	0.154	-0.047
	(1.21)	(3.37)	(3.88)	(2.00)	(1.22)	(3.37)	(3.90)	(2.00)
MB	0.007	0.110	-0.170	-0.028	0.007	0.109	-0.172	-0.027
	(0.24)	(8.80)	(4.41)	(1.86)	(0.23)	(8.74)	(4.43)	(1.81)
#Obs.	37,873	37,873	37,873	37,873	37,873	37,873	37,873	37,873

Table X: The Determinants of Vertical Mergers - NAICS-based

This table examines the determinants of vertical and non-vertical mergers. We display results from logistic regression where the dependent variable is a dummy indicating whether the given firm is an acquirer or a target in a vertical or non-vertical merger or an aquisition of assets in a given year. One observation is one firm in one year. Vertical transactions are identified using our new NAICS-based vertical network (NAICS-1%). The sample is from 1996 to 2008. All the independent variables are lagged by one year and are defined in the Appendix. All estimations include year fixed effects. TNIC(X) refers to the industry (TNIC) average of variable X. We include a dummy variables that equals one for missing values of Fluidity (that starts only in 1997). Standard errors are clustered by industry and year. We report t-statistics in parentheses.

Dependent variable:	Acquirer?		Target?		Acquirer?		Target?	
Deal type:	Vertical	Non-Vertical	Vertical	Non-Vertical	Vertical	Non-Vertical	Vertical	Non-Vertical
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
TNIC(R&D/sale)	-0.708	0.838	0.258	1.369	-1.214	0.901	0.050	1.607
	(1.18)	(4.57)	(0.51)	(6.78)	(1.96)	(4.46)	(0.09)	(7.40)
TNIC(#Patents)	0.001	0.001	0.002	0.000	-0.004	0.001	0.001	0.003
	(0.63)	(1.09)	(2.63)	(0.37)	(1.50)	(1.38)	(0.56)	(2.73)
TNIC(R&D/sale) × TNIC(#Patents)					0.055	-0.008	0.024	-0.032
					(3.31)	(0.77)	(1.34)	(2.95)
TNIC(Fluidity)	0.022	0.031	0.037	0.037	0.023	0.031	0.038	0.036
	(1.19)	(4.13)	(1.87)	(4.77)	(1.25)	(4.13)	(1.89)	(4.78)
TNIC(PPE/asset)	0.554	-0.516	0.410	-0.540	0.539	-0.517	0.415	-0.550
	(1.25)	(2.75)	(1.09)	(3.55)	(1.22)	(2.76)	(1.10)	(3.61)
TNIC(HHI)	-0.638	-0.030	-0.209	-0.218	-0.623	-0.034	-0.197	-0.237
	(2.03)	(0.30)	(0.78)	(2.09)	(1.99)	(0.35)	(0.74)	(2.28)
End Users	-1.157	-1.668	1.015	-1.210	-1.205	-1.668	1.010	-1.213
	(1.71)	(5.87)	(1.51)	(4.11)	(1.77)	(5.87)	(1.50)	(4.12)
#Segments (NAICS)	0.099	0.130	0.104	0.147	0.104	0.129	0.106	0.144
	(3.15)	(8.23)	(3.01)	(9.11)	(3.31)	(8.14)	(3.06)	(8.90)
log(Assets)	0.368	0.426	0.310	0.333	0.367	0.426	0.309	0.335
	(12.59)	(34.78)	(9.87)	(26.88)	(12.54)	(34.54)	(9.80)	(27.09)
log(Age)	-0.099	-0.031	0.108	0.002	-0.098	-0.031	0.108	0.002
	(1.98)	(1.45)	(1.92)	(0.08)	(1.97)	(1.45)	(1.93)	(0.09)
MB	0.059	0.101	-0.098	-0.037	0.058	0.102	-0.099	-0.036
	(1.70)	(8.11)	(2.27)	(2.56)	(1.67)	(8.14)	(2.29)	(2.49)
#Obs.	37,873	37,873	37,873	37,873	37,873	37,873	37,873	37,873

Table XI: The Determinants of Market Relationships

This table examines the determinants of market relationships. We display marginal effects of logistic regressions where the dependent variable identifies the existence of market relationships. In columns (1) and (2) the dependent variable is a dummy indicating whether a given firm appears in a pair of customer-supplier links based on Compustat in a given year. The sample is from 1996 to 2008. In columns (3) and (4) the dependent variable is a dummy indicating whether a given firm appears in a pair of customer-supplier links based on Capital IQ new client announcements database in a given year (as defined in Section **??**.**??**). The sample is from 2002 to 2008. All the independent variables are lagged by one year and are defined in the Appendix. TNIC(X) refers to the industry (TNIC) average of variable X. Standard errors are clustered by industry and year. We report t-statistics in parentheses.

Sample:	Comp. segments	CIQ Clients	Comp. segments	CIQ Clients
Dependent Variable:	Initiation?	Initiation?	Initiation?	Initiation?
	(1)	(2)	(3)	(4)
TNIC(R&D/sale)	1.749	1.123	1.280	0.582
	(9.02)	(3.28)	(5.55)	(1.28)
TNIC(#Patents)	0.004	0.005	-0.001	-0.001
	(4.15)	(3.55)	(0.81)	(0.35)
TNIC(R&D/sale) × TNIC(#Patents)			0.056	0.060
			(5.28)	(2.92)
TNIC(Fluidity)	-0.010	-0.014	-0.010	-0.013
	(1.24)	(1.06)	(1.14)	(0.96)
TNIC(PPE/asset)	-0.009	-0.214	-0.010	-0.206
	(0.05)	(0.69)	(0.05)	(0.67)
TNIC(HHI)	-0.153	-0.436	-0.122	-0.379
	(1.60)	(2.24)	(1.30)	(1.94)
End Users	0.145	5.637	0.126	5.617
	(0.51)	(11.76)	(0.44)	(11.74)
#Segments (NAICS)	-0.041	-0.213	-0.034	-0.206
	(2.40)	(7.27)	(2.02)	(7.05)
log(Assets)	0.342	1.204	0.341	1.205
	(27.53)	(47.47)	(27.58)	(47.24)
log(Age)	0.020	0.043	0.022	0.047
	(1.09)	(1.38)	(1.17)	(1.52)
MB	0.069	0.137	0.070	0.136
	(5.92)	(5.56)	(6.07)	(5.52)
#Obs.	37,873	23,736	37,873	23,736

Figure 1: Payable and Receivables on the Supply Chain.

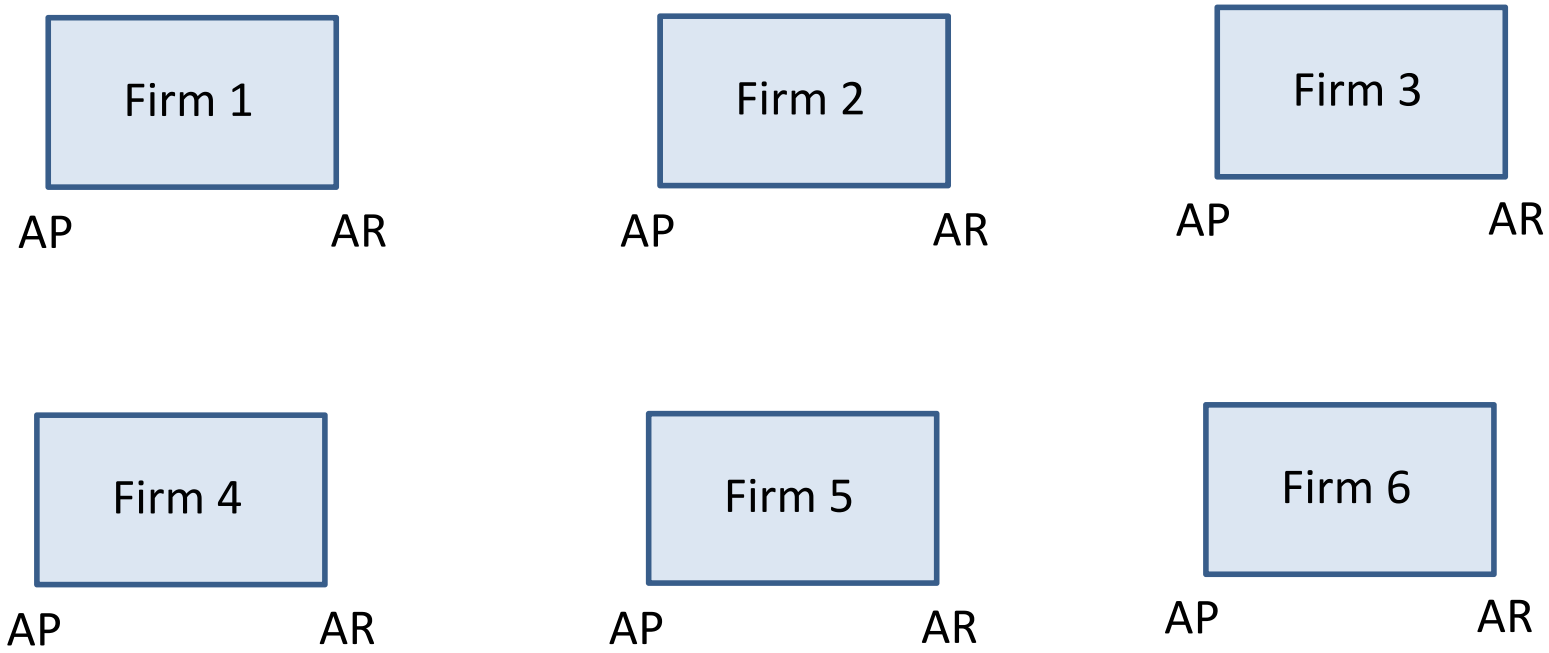

Figure 2: Empirical Distribution of (Text-based) Vertical Integration. This figure shows the the empirical distribution of our text-based measure of vertical integration (VI) as defined in Section II. The sample period is 1996-2008.

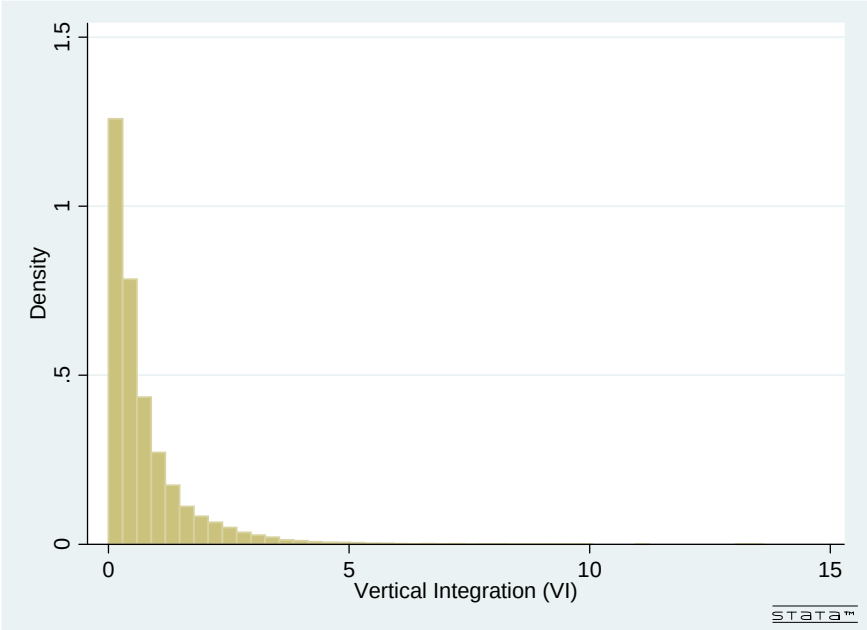

Figure 3: Evolution of (Text-based) Vertical Integration over time. This figure shows the time evolution of our text-based measure of vertical integration (VI) as defined in Section II. We report the annual average of VI across all sample firms. The sample period is 1996-2008.
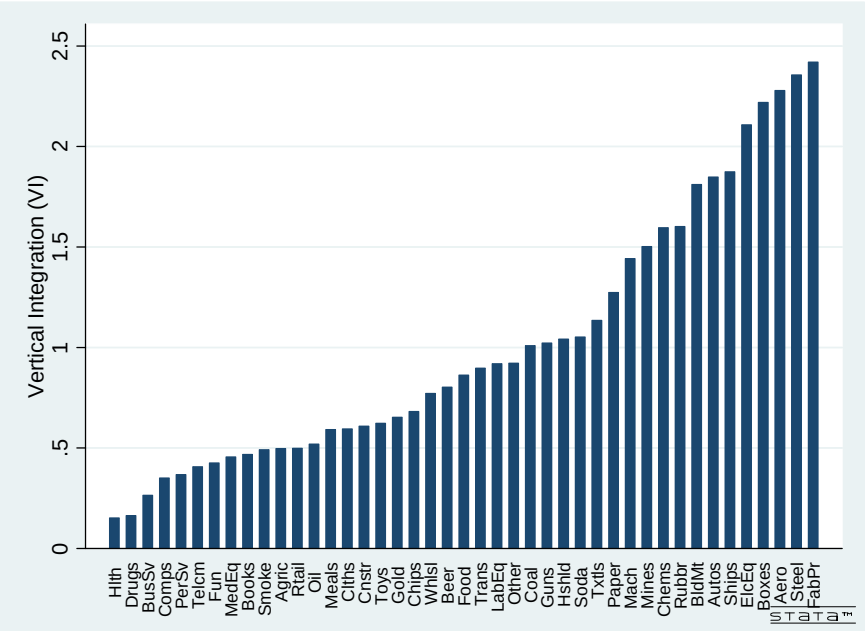
Figure 4: Text-based Vertical Integration across Fama and French 48 sectors. This figure shows our text-based measure of vertical integration (VI) as defined in Section II across sectors. We report the average VI across all firms in a given sector (over all years). The sample period is 1996-2008.