Innovation Activities and the Incentives for Vertical Acquisitions and Integration

Laurent Frésard, Gerard Hoberg and Gordon Phillips*

October 22, 2014

ABSTRACT

We examine the incentives for firms to vertically integrate through acquisitions and production. We develop a new firm-specific measure of vertical integration using 10-K product text to identify the extent to which 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 become targets in vertical acquisitions. These findings are consistent with firms with unrealized innovation avoiding integration to maintain *ex ante* incentives to invest in intangible assets and maintain residual rights of control as in Grossman and Hart (1986). In contrast, firms in high patenting industries with mature product markets are more likely to be vertically integrated consistent with control rights being obtained by firms to facilitate commercialization of already realized innovation.

*University of Maryland, University of Southern California, 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 gerard.hoberg@marshall.usc.edu and Phillips can be reached at gordon.phillips@marshall.usc.edu. We thank Ken Ahern, Jean-Noel Barrot, Thomas Bates, Giacinta Cestone, Oliver Hart, Thomas Hellmann, Sébastien Michenaud, Ivo Welch and seminar participants at the 2014 American Finance Association Meeting, Arizona State University, Carnegie Mellon, Dartmouth College, Humboldt University, IFN Stockholm, Harvard-MIT Organizational Economics joint seminar, Tsinghua University, UBC, UCLA, the University of Alberta, the Universidad de los Andes, the University of Maryland, the University of Washington, and VU Amsterdam for helpful comments. All errors are the authors alone.

The scope of firm boundaries and whether to organize transactions within the firm (integration) or by using external purchasing is of major interest in understanding why firms exist. 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 control rights are key to understanding firm boundaries and their influence on ex ante investment. They show that ex ante incentives for a firm to invest in relationship-specific assets are reduced through vertical integration – for the firm that gives up its residual rights of control to the other contracting firm. Holmstrom and Milgrom (1991) and Holmstrom and Milgrom (1994) also emphasize the role of incentives in firm structure.[1]

In this paper, we show that the costs and benefits of integration are related to the stage of development of intangible assets. Atalay, Hortascu, and Syverson (2014) report that actual physical shipments between vertically related plants are infrequent. Our paper is consistent with this being explained at least in part by the link between integration and intangibles. Our paper is also the first to document a strong role for the stage of innovation. We examine unrealized innovation in the form of R&D, and realized innovation characterized by patents. We find that this distinction plays an important role in explaining firm boundaries, which we capture using new firm-specific measures of integration and vertical 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 Bureau of Economic Analysis (BEA) input-output tables to measure vertical relatedness. We measure how vertically integrated firms are by looking at whether the products (and services) they offer span vertically related markets. Similarly, we identify the extent to which mergers and acquisitions represent vertical transactions

[1] Gibbons (2005) summarizes the large literature and highlights that the costs and benefits of vertical integration – and hence why integration matters – largely depend on transactions costs, opportunistic behaviors, contractual incompleteness, and the specificity of the assets involved in transactions.

by examining if participating firms' products are vertically linked.

This novel approach enables us to analyze vertical integration dynamically, to relate firms' organization structure to the stage of their innovation activities, and to examine differences both broadly across industries and narrowly through within-firm variation.[2] Using a sample of almost 7,000 publicly-traded firms over the 1996-2008 period, we find strong evidence that firms in R&D intensive industries (unrealized innovation) are less likely to be vertically integrated. In sharp contrast, we find that firms in industries characterized by high patenting intensity, and thus with more realized innovation are more likely to be vertically organized.

These results are economically large: In our baseline specification, firms' vertical integration decreases by 10% in response to a one-standard deviation increase in R&D intensity, and increases by 7% following a one-standard deviation increase in patenting intensity. Moreover, our results hold across several specifications. In particular, they hold in within-firm specifications, indicating that firms modify their degree of integration over time when the R&D and patenting profile of their industry changes.

The distinction between unrealized and realized innovation also matters in vertical acquisitions. We estimate that firms operating in R&D intensive industries are less likely to be acquired in vertical transactions. In contrast, firms in patent intensive industries are more likely to be targeted in vertical deals. In addition, we show that vertical acquisitions tend to occur at times where target firms have accumulated more patents, in contrast to non-vertical acquisitions which typically take place after a slower accumulation of patents. Our framework for how vertical acquisitions are affected by the development of intangible assets is distinct from other theoretical motives for acquisitions including neoclassical theories, agency theories and horizontal theories.[3]

[2] This is in contrast to many single-industry studies in industrial organization. Earlier studies include Monteverde and Teece (1982) focusing on automobile manufacturing, Masten (1984) focusing on airplane manufacturing, and Joskow (1987) focusing on coal markets. Reviews of the empirical literature can be found in Lafontaine and Slade (2007) and Bresnahan and Levin (2012).

[3] 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 acquisitions, and Phillips and Zhdanov (2013) for a recent horizontal theory of acquisitions.

The observed contrast between R&D and patent intensity in explaining patterns of vertical integration and acquisitions is consistent with the predictions of a simple incomplete contracting model in which the decision of firms to integrate vertically depends on the stage of development of the specific asset exchanged in the transaction. We present such a model in the next section. The model predicts that when the asset is still in the form of R&D (unrealized innovation) firms optimally will not integrate.[4] Firms avoid integration and remain separate to maintain ex ante incentives for the upstream firm to further invest in developing the potential related product extensions, and to maintain residual rights of control. In contrast, when the product is more fully developed and the product and its features are protected by a patent (realized innovation), the owner of the realized innovation has more legally enforceable residual rights of control. Incentives for ongoing separate innovation decrease while incentives for the downstream firm to commercialize and integrate the product with additional product extensions or related products increase with integration between the downstream and upstream firm. At that time, integration optimally allocates the residual rights of control to the downstream firm that will use the realized innovation and commercialize it.

This distinction between high R&D and patents is nontrivial. High R&D does not necessarily lead to high patenting rates. There are also cross-sectional differences in patenting across industries. As reported by Cohen, Nelson, and Walsh (2000) in their survey of 1,478 R&D labs, high R&D may not lead to patents due to concerns about appropriability. 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.

Two recent examples of the effects we document are Microsoft's recent purchases of Skype and Nokia. Skype specialized in making VoIP phone and video calls over the Internet. After purchasing Skype, Microsoft integrated Skype into Windows and also into Windows phones. Microsoft's purchase of Nokia in 2013 is also a good example.

[4] Acemoglu (1996) argue that technological investments are partner-specific, thus creating relationship-specific assets that are difficult to contract on. 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, consistent with the needed R&D incentives to develop relationship-specific assets.

One insider indicated that the deal between the two companies would help to bring the "hardware closer to the operating system and achieve a tighter integration." Buying firms to gain control of their realized innovations facilitates commercialization either through reduced ex post hold-up or increased commercialization incentives.[5] Note that the acquirer who vertically integrates need only pay the reservation price of the seller.

An industry that exemplifies our findings is the network equipment industry, which includes Cisco, Broadcom, Citrix, Juniper, Novell, Sycamore, and Utstarcom. During our sample, and using our measures, we find that firms in this industry jointly experienced (A) levels of R&D that peaked and began to decline, (B) levels of patenting activity that rose four to five fold, and (C) levels of vertical integration that also rose four to five fold. The conversion of unrealized innovation into realized patented innovation reduced the incentives for relationship-specific investment by these firms, and also increased the incentives to vertically integrate in order to transfer decision rights to the party commercializing the patents.[6]

Our paper is related to the recent paper by Atalay, Hortascu, and Syverson (2014). They document that vertical shipments of goods within manufacturing across plants is not common. Using comprehensive data on ownership structure, production, and shipment patterns throughout the U.S. economy, Atalay, Hortascu, and Syverson (2014) report that upstream units ship only small shares of their outputs to own-firm downstream plants. However, their data does not capture the use of vertically related intangible assets within firms. The authors suggest that intangible assets may drive vertical integration. Our paper provides direct evidence that intangibles are indeed related to vertical integration. Our results are consistent with vertical structures being less concerned with the internal transfer of physical goods across plants but instead mediating efficient transfers of intangible inputs within firms, as in the example of post-purchase integration of Microsoft and Skype.

We also consider the role of supply chain stability and maturity. For example,

[5] See http://www.businessinsider.com/why-microsoft-bought-skype-an-insider-explains-2011-5.

[6] The 2014 IBISWORLD industry report on the Telecommunication Networking Equipment Manufacturing confirms the trend towards more integration in this market. Players in this industry seek to offer "end-to-end" and "all-in-one" solutions.

assuming the benefits of integration derive from ongoing operations, the decision to integrate should be more profitable when the supply chain is more mature and 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 duration of gains may not be adequate to cover the high fixed costs. We find empirical support for the proposed positive link between maturity and vertical integration. In particular, both vertical integration and vertical transactions are positively related to maturity as captured by firm age, lower M/B ratio, and more tangible assets (higher PP&E/assets).

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 as in Grossman and Hart (1986). We show that firms in industries with high unrealized innovation are less likely to vertically integrate, while firms in industries with significant realized patents and thus stronger property rights are more likely to integrate to facilitate commercialization of the realized innovation.

Our paper adds to the literature on vertical integration and acquisitions by Fan and Goyal (2006), Kedia, Ravid, and Pons (2011), Ahern (2012) and Ahern and Harford (2013). Fan and Goyal examine the stock market reactions of vertical deals where vertical integration is identified through links between SIC codes and the Input-Output tables. Kedia et al. show that vertical mergers create more value in imperfectly competitive markets. Ahern shows that division of stock-market gains in mergers is determined in part by customer or supplier relations and their relative bargaining power. 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 analysis also adds to papers that examine how acquisitions 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 further 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.

Our last contribution is methodological. We demonstrate that our new text-based firm-specific measures of vertical relatedness have significant advantages over existing measures based on coarser industry definitions, such as 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 more directly. For example, the use of NAICS or SIC classifications linked to BEA data to define vertical relatedness is problematic because NAICS and SIC classifications are based on production processes and not products themselves.[7] Our new measure of vertical relationships does not rely on the quality of the Compustat segment tapes, nor the quality of the NAICS classification.[8] Our focus on vertical links extends the work of Hoberg and Phillips (2014), who examine horizontal links between firms and their impact on horizontal merger synergies using 10-K text.

The remainder of this paper is organized as follows. Section II develops a simple model of vertical integration. Section III presents the data and develops our new measure of vertical relatedness between firms. Section IV summarizes our new measure on vertical integration and examines its economic determinants. Section IV examines vertical transactions. Section VI concludes.

[7] See http://www.naics.com/info.htm. The Census Department states "NAICS was developed to classify units according to their production function. NAICS results in industries that group units undertaking similar activities using similar resources but does not necessarily group all similar products or outputs."

[8] The Compustat segment reporting has been shown to have reporting biases as documented by Hyland (1999) and Villalonga (2004). Note that measures of vertical integration using SIC or NAICS codes relies heavily on the quality of segment reporting.

II A Simple Model of Integration

In this section we develop a simple incomplete contracting model of the timing of integration. Our model predicts that the decision of firms to integrate vertically depends on the stage of the product development. In the spirit of Grossman and Hart (1986), we assume that R&D is non-contractible and non-verifiable, as are commercialization and integration expenditures by the downstream firm. If the upstream firm is acquired by the downstream firm it thus has no incentives to further invest in R&D to further develop additional product extensions. Incentives for ongoing separate innovation thus decrease with integration, while the incentives of the downstream firm to commercialize the product and integrate any product extensions are higher under integration when it owns the product and its extensions. Thus, our model predicts that firms integrate after a sufficiently large product innovation is realized, such that further innovations have a decreasing effect on the product price.

There are two firms in the economy, an upstream supplier and a downstream firm. At each time period t, they are cooperating to produce a product, with a base price P_t^b. The final price P_t, which we define later, further depends on the level of commercialization and product integration operations chosen by the downstream firm. The upstream supplier can do an x_t amount of R&D research that can result in new patentable features and extensions that can increase the price consumers are willing to pay for the product with a cost $k_t = c(x_t) = Sx_t^g$, and the producer can do a y_t amount of commercialization and integration activities that can also boost the price of the product with a cost of $m_t = c(y_t) = Ry_t^h$. We assume that both $g > 1$ and $h > 1$ so that costs are convex. The discount rate is r.

The base price P_t^b takes a value in the set $\{P_0, P_1, \dots P_N\}$, with $P_s < P_{s+1}$ $(0 \leq s \leq N-1)$ and $P_{s+1} - P_s < P_s - P_{s-1}$ $(0 \leq s \leq N-1)$. A success in R&D research at time t results in new features and product enhancements. These product enhancements result in a patent, and boost the base price from P_s to P_{s+1} $(0 \leq s \leq N-1)$. Additional product features and extensions have a positive but decreasing effect on prices. We use X_t to denote the result of R&D research which is realized and observed by both parties at the end of time period t, such that

$X_t = 1$ corresponding to a success and $X_t = 0$ a failure. The probability of success is determined by the R&D expenditures $p(X_t = 1) = x_t$.

To keep the model relatively straightforward, we assume that the increase in price resulting from commercialization and integration expenditures is deterministic, and it increases the base price P_t^b by an amount of y_t if the firms are separate , and by an amount of $\rho(y_t)$ if the firms are integrated. Both the level and the marginal product of the commercialization and integration expenditures are higher in integration, such that $\rho(y_t) > y_t$ and $\rho'(y_t) > 1$.[9]

The timing of the actions is as follows. At each time period t, given the outcome of the R&D success state and patent at the end of last time period X_{t-1}, we have

1. Downstream producer decides whether to integrate and negotiates with the upstream supplier if it decides to integrate. Let I_t stands for the integration decision, 1 for integration, and 0 for separation.

2. R&D expenditures x_t and commercialization expenditures y_t are decided by firms as ex-ante investments.[10]

3. Renegotiation occurs if separated.

4. By the end of the period, the R&D success is realized, so that at the beginning of next period $t + 1$, both firms observe the value of X_t.

The realization of R&D (whether patent is granted) is the key to determine when the firms will integrate. Since we make the assumption that X_t is realized at the end of each period, the final price is equal to $P_t = P_t^b(1 + y_t)$ under separation and $P_t = P_t^b(1 + \rho(y_t))$ under integration, with the base price $P_t^b = P_N$ if the last period

[9] The assumption that the marginal product of commercialization expenditures is higher in integration can arise from the supplier not cooperating fully (withholding some information or selling related products to other firms) with the downstream firm if separate. We do not model the specific reason for the marginal product of commercialization expenditures being higher under integration.

[10] We could equivalently consider the case where the upstream firm buys the downstream firm. This would occur if the downstream firm does the R&D and the upstream firm customizes the product features before supplying the product. We make the assumption that the downstream firm buys the upstream as the case we consider as it is more plausible than the upstream buying the downstream, which we see infrequently in reality. Furthermore, as along as there are separate skills (R&D and commercialization or customization), one set of incentives would be lost in integration as in Grossman and Hart (1986) for both cases post integration.

base price is $P^b_{t-1} = P_N$ or $P^b_t = P_s + (P_{s+1} - P_s)X_{t-1}$ if the last period base price is $P^b_{t-1} = P_s < P_N$. Note that the base price is a contingent variable given the last-period R&D result X_{t-1}.

The bargaining power of the upstream supplier is always α (and the downstream producer $1-\alpha$) in both ex-ante negotiation for integration and ex-post renegotiation for splitting total surplus. The model and timing of events are summarized in Figure 1.

We model integration as a real option that when exercised is costly to reverse. Thus, firms do not integrate until the marginal benefits of staying separate decrease for the upstream firm.

We now present a sequence of propositions that we prove in Appendix 1.

Proposition 1

R&D expenditures are higher under separation, while commercialization and product integration expenditures are higher under integration.

Proposition 2

If $P^b_t = P_N$, then both firms prefer to integrate. Hence, where V is the value function, we have $V(P_N) = V(P_N; I = 1) > V(P_N; I = 0)$.

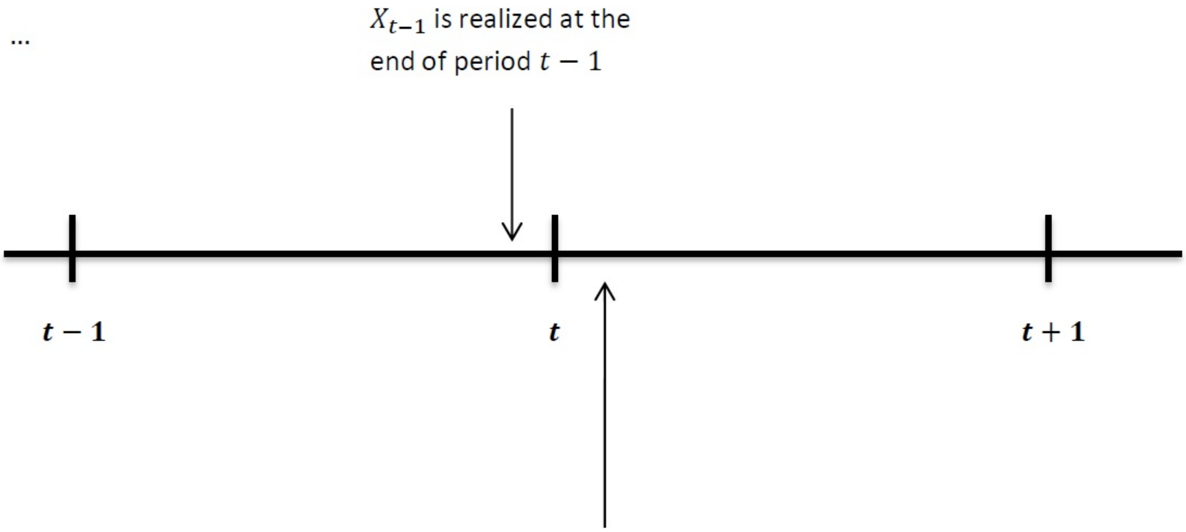
Our next proposition, Proposition 3, gives our key result, which predicts when it is optimal for the firms to integrate. It is the key proposition that we empirically test.

Proposition 3

There exists a state s^* such that $V(P_s) = V(P_s; I = 1) \geq V(P_s; I = 0)$ for any $s \geq s^*$, and $V(P_s) = V(P_s; I = 1) < V(P_s; I = 0)$ for any $s < s^*$. The state s^* would then be the triggering state for integration.

This proposition is illustrated in Figure 2.

Figure 1:



At the beginning of period t

Actions:

(1) Producer decides I_t given X_{t-1}
(2) Choose x_t and y_t given I_t
(3) Renegotiation

Prices:

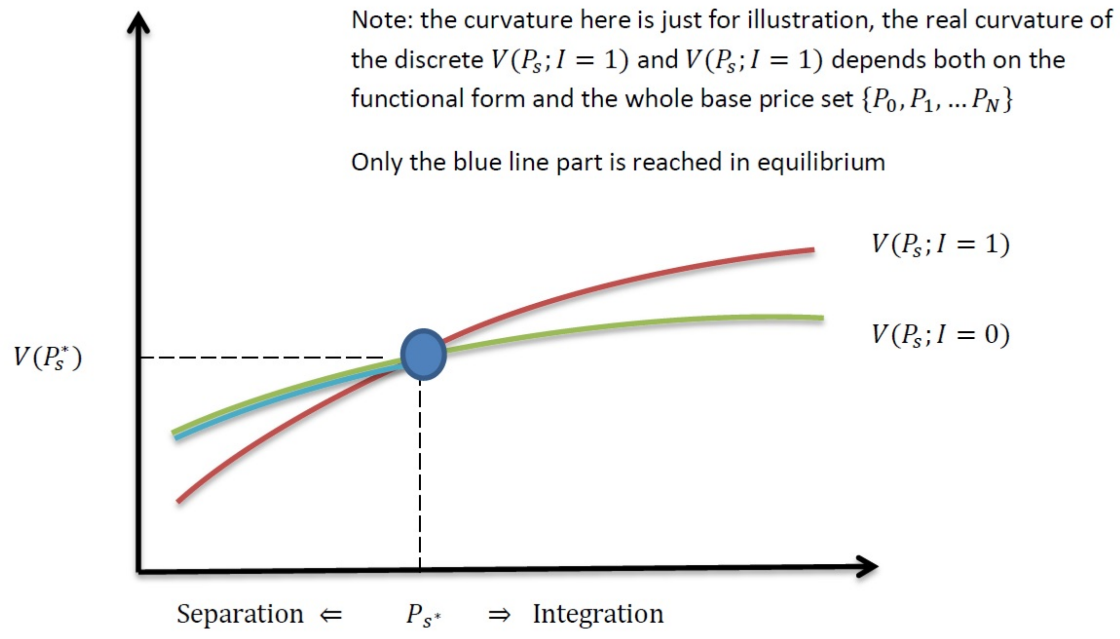
$$P_t = \begin{cases} P_t^b(1+y_t) & \text{if } I_t = 0 \\ P_t^b\big(1+\rho(y_t)\big) & \text{if } I_t = 1 \end{cases}$$

$$P_t^b = \begin{cases} P_s + (P_{s+1} - P_s)X_{t-1} & \text{if } P_{t-1}^b = P_s \text{ with } 0 \le s < N \\ P_N & \text{if } P_{t-1}^b = P_N \end{cases}$$

Payoffs:

(1) $TS_t = P_t - Sx_t^g - Ry_t^h$
(2) The split of the sales depends on α
(3) Supplier's profit is αTS_t, and producer's profit is $(1-\alpha)TS_t$

Figure 2:


Note: the curvature here is just for illustration, the real curvature of the discrete $V(P_s; I = 1)$ and $V(P_s; I = 1)$ depends both on the functional form and the whole base price set $\{P_0, P_1, \ldots P_N\}$

Only the blue line part is reached in equilibrium

III Data and Methodology

We draw from multiple data sources to create our sample and our key variables: 10-K business descriptions, the Input-Output (IO) tables from the Bureau of Economic Analysis (BEA), COMPUSTAT, the SEC Edgar database, data on mergers and acquisitions 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 (2014). 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.[11] 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.

B Data from the Input-Output Tables

We use both commodity text and numerical data from the the Input-Output (IO) tables from the BEA. 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.[12]

The 'Use' table reports the dollar value of each commodity that is purchased by each industry or by final end-user.[13] There are 431 commodities in the Use table purchased by 439 industries or final end-users.[14] The 'Detailed Item Output' table

[11] 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.

[12] 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).

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

[14] 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

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: (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 decomposes each commodity (i.e. each IO Commodity Code) into sub-commodities (labeled by 'IO Item Code'). For each sub-commodity the BEA details a respective vocabulary (a verbal description) and provides the dollar value of its total production. A commodity's total production as reported in the Input-Output Table is thus the sum of the production of its sub-commodities.[15] Each sub-commodities' verbal description uses between 1 to 25 distinct words (the average is 8) that summarize the nature of the good or service provided.[16] Table II contains an example of product text for the BEA 'photographic and photocopying equipment' commodity. We extract the verbal descriptions of sub-commodities to form the sets of words associated with each commodity. We label these sets 'commodity words'.

[Insert Table II Here]

We do not include all words in our analysis. We start with the convention in Hoberg and Phillips (2014) and we only consider nouns and proper nouns. We then apply four additional screens to ensure our identification of vertical links is conservative. First, because commodity vocabularies identify a stand-alone product

'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.

[15] There are 5,459 sub-commodities in the 'Detailed Item Output' table associated with 427 commodities. The average number of sub-commodities per commodity is 12, the minimum is 1 and the maximum is 154.

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

market, we discard from the commodity vocabularies any expressions that indicate a vertical relation such as 'used in', 'made for' or 'sold to'. Second, we remove any expressions that indicate exceptions to what is sold in the given product market (e.g, we drop phrases beginning with 'except' or 'excluding'). Third, we discard common words that appear in large numbers of commodity vocabularies, as such words are not discriminating.[17]

Finally, we remove from each commodity vocabulary, any words that do not frequently co-appear with other words in the given commodity vocabulary. This further ensures that horizontal links or links capturing horizontal asset complementarities are not mislabeled as vertical links. We identify these broad words by examining the fraction of times each word in a given IO commodity co-appears with other words in the same IO commodity when the given word appears in a 10-K business description. We use all 10-Ks from 1997 to compute this fraction for each word in each commodity, and we then discard words in the bottom tercile (the broad words). For example, if there are 21 words in an IO commodity description, we would discard 7 of the 21 words using this method.[18] In all, we are left with 7,735 remaining commodity words that uniquely identify vertically related product markets in tight word clusters.

The information from the 'Detailed Item Output' table also enables us to determine the economic importance of each word for a given commodity's output. To obtain it, we compute the *relative* economic contribution of a given sub-commodity (ω) as the dollar value of its production relative to its commodity's total production. Each word in the sub-commodity's textual description is then assigned the same ω. Because a word can appear in the text of several sub-commodities within a commodity, we sum its ω's by commodity. Hence, a given commodity word is economically more important if it is used in the text of sub-commodities that account for a larger share of the commodity's output. We then define the commodity-word correspondence matrix (CW) as a three-column matrix having the first column being a given commodity, the second column being an associated commodity word, and the third

[17] There are 250 such words including accessories, air, attachment, commercial, component. See Appendix 2 for a full list.

[18] This approach (and our use of terciles) is based on Hoberg and Phillips (2010), who also focus on the local subset of words after discarding the tercile of most broad words.

being its economic importance.

Because the textual description in the 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 (upstream or downstream) to another commodity in the supply chain. To do this, from the Make Table, we create $SHARE$, an $I \times C$ matrix (Industry × 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 c and d.[19] 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

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

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

We identify 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 by computing the similarity between the given firm's business description and the textual description of each BEA commodity. Because vertical relatedness is observed from BEA 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 IO commodities it provides, and (2) determining i and j's vertical relatedness using the relatedness matrix V.

When computing all textual similarities, we limit attention to words that appear in the Hoberg and Phillips (2014) post-processed universe. We also note that we only use text from 10-Ks to identify the product market each firm operates in (vertically links between vocabularies are then identified using BEA data as discussed above). Although uncommon, a firm will sometimes mention its customers or suppliers in its 10-K. For example, a coal manufacturer might mention in passing that its products are "sold to" the steel industry. To ensure that our firm-product market vectors are not contaminated by such vertical links, we identify and remove any mentions of customers and suppliers using 81 phrases, which we list in Appendix 3. Although we feel this step is important, we also note that our results are robust if we exclude this step.

Ultimately, we represent both firm vocabularies and the commodity vocabularies from BEA as vectors with length 60,507, which is the number of nouns and proper nouns appearing in 10-K business descriptions in Hoberg and Phillips (2014). Each element of these vectors corresponds to a single word. If a given firm or commodity does not use a given word, the corresponding element in its vector will be set to zero. By representing BEA commodities and firm vocabularies as vectors in the same space, we are able to assess firm and commodity relatedness using cosine similarities.

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. In this cosine similarity calculation, commodity word vector weights are assigned based on words' economic importance from the CW matrix (see above), and firm word vectors are equal weighted following the TNIC construction used in Hoberg and Phillips (2014). This calculation is also based only on non-common nouns and proper nouns as in Hoberg and Phillips (2014). 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 (2014)) 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 follows:

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

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. 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 other commodities whose description resembles firm j's product 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 NAICS-based Vertical Relatedness

Given we are proposing a new way to compute vertical relatedness between firms, we compare the properties of our text-based vertical network to those of the NAICS-based measure used in previous research, which we describe now. One critical difference is that the NAICS-based vertical network is computed using the BEA industry space, and not the BEA commodity space. This is by necessity because the links to NAICS are at the level of BEA industries. Avoiding 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 and NAICS industries for example, may result in potential horizontal contaminations, especially when firms are in markets that do not cleanly map to NAICS industries. In particular, the Census Department states “NAICS was developed to classify units according to their production function. NAICS results in industries that group units undertaking similar activities using similar resources but does not necessarily group all similar products or outputs.”

To compute the NAICS-based network, we use methods that parallel those discussed above for the BEA commodity space (matrix V), but we focus on the BEA industry space and construct an analogous matrix Z. We first compute the BEA industry 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 $ISUPP$.

Following common practice in the literature (see for example Fan and Goyal (2006)), we use two numerical 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 to a BEA IO industry.

We label these two vertical networks as 'NAICS-10%' and 'NAICS-1%', respectively, as the NAICS-10% and NAICS-1% 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-10% network. This network is thus very coarse. At 1.37% granularity, the NAICS-1% network is quite fine. To ensure relevant comparisons between our textual networks and the NAICS networks, we choose two similar granularity levels for our text-based network that resemble those of the 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 'Vertical Text-10%' and 'Vertical Text-1%'.

Note that this discussion highlights a second key difference between the NAICS-based method and our text-based network method. The NAICS-based method imposes the same similarity on all firms within an industry based on the industry-wide degree of vertical relatedness. When using NAICS-based methods, any two firms in the same NAICS code will thus have exactly the same set of upstream and downstream peers. In our text-based method, a firm's upstream and downstream links are customized and unique to its own business description. The non-transitive nature of our classification allows us to accommodate informative intra-industry heterogeneity in product offerings.

E Vertical Network Statistics

We compare the properties of five key relatedness networks: Vertical Text-10%, Vertical Text-1%, NAICS-10%, NAICS-1%, and the TNIC-3 network developed by Hoberg and Phillips (2014). The first four networks are intended to capture vertical relatedness, and the TNIC-3 network is calibrated to be as granular as are three-digit SIC industries, and is intended to capture horizontal relatedness. For each network, Panel A of Table III presents various statistics relating to granularity, the degree of overlap, and the degree of focus on financial firms.

[Insert Table III Here]

The first row of Panel A presents the level of granularity of each network. The NAICS-10% and NAICS-1% networks display granularity levels of 9.48% and 1.37% respectively. By construction, the 'Vertical Text-10%' and 'Vertical Text-1%' vertical networks have 10% granularity and 1% granularity, respectively. Finally, the TNIC-3 network, which is designed to match the granularity of SIC-3, has a granularity of 2.33%. This puts the Vertical 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 show that the four vertical networks exhibit low overlap with the horizontal TNIC-3 network, or with the horizontal 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 10.48% of firm-pairs in the NAICS-10% network are also present in the Vertical Text-10% network. Similarly, only 1.21% of firm-pairs are in both the Vertical Text-1% and NAICS-1% networks.

The last three rows of Panel A are perhaps 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 9.20% and 1.80% 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-10% network are also present in the Vertical 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.

F Validation of the Text-Based Approach

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 show that the relatedness that these networks capture is 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 relates to a supply chain. We construct a test of the extent to which any proposed vertical relatedness network is 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 being consistent with horizontal links).

Intuitively, our test is based on how firms that are related vertically versus horizontally should respond to trade-credit shocks. Firm pairs that are vertically related will experience *negatively* correlated shocks in accounts payable versus accounts receivable due to their supply chain adjacency as in a recent paper by Kalemli-Ozcan, Kim, Shin, Sorensen, and Yesiltas (2014). For example, a shock to an upstream industry's *receivables* should be associated with a similar shock to the downstream industry's *payables*. In contrast, firms that are horizontally related should experience trade-credit shocks in either accounts payable or accounts receivable that are positively correlated. We define trade credit as accounts payable minus accounts receivable for each firm. We then regress changes in trade credit of upstream firms on the changes in the trade credit of downstream firms. The complete details and results of this test, which support the conclusion that our text-based approach strongly measures vertical links, are presented in Appendix 4.

IV Vertical Integration

We now use our text-based approach to analyze firms' degree of vertical integration and its association with firms' innovation activities. We explain how we measure integration within a given firm, describe the properties of our new measure, and examine its economic determinants.

A Text-based Integration

Using the quantities developed above, we can directly compute the extent to which a given firm is 'vertically integrated'. This is simply given by a firm's vertical relatedness with itself ($UP_{i,i}$ or interchangeably $DOWN_{i,i}$). Using our notation from Section III, we have:

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

With this measure, a firm is more vertically integrated (i.e. higher value for VI_i) 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, this occurs when a firm offers products or services at different stages of a supply chain.[20]

We complement and validate the measure VI by identifying directly whether firms mention that they are vertically integrated in their 10-Ks. We specifically search for the terms 'vertical integration' and 'vertically integrated' in each 10-K. We exclude cases where firms indicate they are not integrated or lack integration. We create a dummy variable $VI(search)$ that equals one when a firm states explicitly that it is vertically integrated in a given year. Because this measure is not derived from word

[20] Unfortunately, data limitation prevents us to determine the economic importance of each product from firms' product description. Hence, while VI is a novel measure that uniquely captures firm-level vertical integration based on the evolution of its product description, it cannot account for each firm products' importance.

vectors, it enables us to validate our text-based measure, and provides us with an alternative assessment of vertical integration.

In addition, we 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 in Section III, we can quantify the degree to which a firm is at the end of the supply chain. We compute this quantity as follows:

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

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.

B Descriptive Statistics

Table IV presents summary statistics for our measures of vertical integration and the main variables we use in the analysis. We describe all variables in Appendix 5. Our sample covers the period 1996-2008 and excludes regulated utilities and financial firms. We further require observations to have non-missing values for each variable. We have 45,198 firm-year observations corresponding to 6,924 distinct firms.

[Insert Table IV and Figure 3 Here]

Statistics for the text-based variables are reported in Panel A. The table indicates that the degree of vertical integration among firms is quite heterogeneous. The average and median value of vertical integration (VI) are 0.012 and 0.008 respectively, and the maximum is 0.116. This skewness is largely confirmed in Figure 3 that displays the empirical distribution of VI. While most firms exhibit a low potential for integration, a non-negligible fraction feature business descriptions that contain many words that are strongly vertically related. These are the firms whose business descriptions have large cosine similarity scores with vocabularies of one or more adjacent vertically linked BEA commodities. The lower means and medians conform 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 the relatively smaller set of firms that are vertically integrated.

[Insert Figure 4 Here]

Figure 4 displays the evolution of vertical integration over time, and we note a clear trend away from integration. This pattern obtains both using equally-weighted or sales-weighted averages, indicating that the average firm is less vertically integrated in recent years than in the late 1990s.

Panel A of Table IV reports fraction of output that flows to various end-users at the end of the supply chain. The table shows that roughly 35% of firms' products are linked to IO commodities that are purchased by retail consumers. About 3.1% flows to the U.S. Government, and 8.6% leaves the U.S. supply chain as exports. This latter figure is in line with Bernard, Jensen, Redding, and Schott (2007), who document that 4% of 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 0.470.[21] Panel B reports summary statistics for our key control variables. They are generally in line with related studies (see Hoberg and Phillips (2010) for example).

[Insert Table V Here]

To complete the description of our new measure of vertical integration, Table V presents average summary statistics of our key variables across quartiles of vertical integration. Several interesting patterns emerge. Reassuringly, we find that firms that mention being vertically integrated ($VI(search) = 1$) display a high level of integration according to our measure. While only 3.1% of firms describe themselves as vertically integrated in the lowest quartile, 14.4% do so in the highest quartile. We also note, 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, and consistent with the fact that a large fraction of US trade of goods takes place intra-firm (e.g. Zeile (1997) or Antras (2003)), vertically integrated firms are more

[21] 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 when we examine vertical integration, as a firm located near the end of the supply chain has fewer options regarding integration, especially on the downstream side.

focused on exports.[22]

Notably and consistent with the prediction of the model in Section II, integrated firms spend less on research and development than non-integrated firms. Yet, based on the U.S. Patent and Trademark Office data (the NBER data archive), vertically integrated firms appear to own larger portfolios of patents. These univariate results summarize the key finding in the paper, which we document more formally later. Vertical integration is negatively correlated with R&D expenses, likely due to the need to incentivize relationship-specific investment ex ante. However, patent intensity is positively related to integration. The realization of successful innovation – measured using the grant of patents – lowers the marginal product of investment in additional product features but increases the marginal product of commercialization and integration expenditures, leading firms to integrate vertically.

This result is illustrated by the evolution of the networking equipment industry over our sample period. Firms active in this industry include Cisco, Broadcom, Citrix, Juniper, Novell, Sycamore, and Utstarcom. As highlighted by Figure 5 these firms became significantly more vertically integrated over time. Their average level of integration (VI) rose four to five fold in twelve years. These firms jointly experienced (A) levels of R&D that peaked in 2002 and then began to sharply decline, and (B) levels of patenting activity that increased four to five fold starting in 2001. The observed dynamics are broadly consistent with the idea that the conversion of unrealized innovation into realized patented innovation increased the incentives to vertically integrate as the importance of ownership and control rights shifts from the smaller innovative firms to the larger commercializing firms.

[Insert Figure 5 Here]

We also find that vertically integrated firms operate in more concentrated markets (based on the TNIC HHI from Hoberg and Phillips (2014)). Moreover, consistent with Atalay, Hortascu, and Syverson (2014), vertically integrated firms appear to be larger ($log(Assets)$), older ($log(Age)$), and more capital-intensive firms ($PPE/assets$), with lower growth opportunities (measured using market-to-book ra-

[22] Further consistent with this idea, the Internet Appendix (Table IA.1) presents evidence that our measure of vertical integration is positively correlated with the intensity of related-party trade, measured using aggregate industry data from the U.S. Census Bureau.

tios MB). Vertically integrated firms also have more business segments ($\#Segments$), 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 (less than 33% of observations in our sample), it is positively correlated with our text-based measure (the correlation coefficient is 0.20).

To further illustrate our text-based measure of vertical integration, Table VI displays the 30 most vertically integrated firms in 2008. The most integrated firm is Handy & Harman (www.handyharman.com) a diversified global company offering a wide range of products serving the construction, electronics, telecommunication, home appliance and transportation sectors. In particular, this firm offers precious metals, tubing, engineered materials, electronics and meat cutting products. The second 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. Overall, a review of these vertically integrated firms suggests a high degree of actual vertical relatedness.

[Insert Table VI Here]

Interestingly, among these 30 most integrated firms, 13 report only a single segment in Compustat despite the fact that their websites indicate product offerings with strong vertical potential. An example of such a single segment reporting 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. Even though they are highly integrated, firms in this category rank rather low on existing non-text measures of integration based on Compustat segments.

[Insert Table VII Here]

Table VII 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 81.5% and 86.3% 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. These findings are consistent with Figure 4 and indicate a tendency towards less vertical integration over time.

C What Drives Vertical Integration?

To examine what factors are related to firm's degree of vertical integration, we estimate panel data regressions where the dependent variable is our measure of vertical integration (VI). We focus on within-industry and within-firm specifications and include year fixed effects in all specifications to isolate the apparent trend towards dis-integration in our sample. 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 results. Baseline estimates for our text-based measure (VI) are reported in columns (1) to (2). For the main right hand side variables of interest (in particular *R&D/Sale* and *#Patents/assets*), we consider industry characteristics (using the equally-weighted average across TNIC-3 industries) instead of own-firm variables.[23] This choice is dictated by two considerations. First, focusing on industry measures lessens endogeneity concerns as they apply to both integrated and non-integrated firms (see Acemoglu, Aghion, Griffith, and Zilbotti (2010)). Indeed, while the firm directly chooses its degree of vertical integration, it has little choice regarding its industry's overall level of R&D or patenting activities. Second, the theoretical incentives to vertically integrate should be mostly driven by the peculiarities of the product markets, rather than firm-specific attributes. For instance, as in Acemoglu, Johnson, and Mitton (2009), the incentive to invest in intangibles are primarily determined by the specific product or service being exchanged between firms. Hence this is better captured using industry variables.[24]

The estimates from Table VIII largely confirm the univariate evidence: firms operating in industries with high levels of R&D are less vertically integrated. This result supports our primary hypothesis that firms avoid integration when ex ante incentives to invest in intangible assets through R&D are needed to be high. This result obtains both within industries (when we include industry fixed effects in column (1)) and within firms (with firm fixed effects in column (2)). This latter result is important as it indicates that firms modify their degree of vertical integration over time as a function of changes in their industry R&D profile. Economically, the negative link between R&D and integration is substantial: A one standard deviation increase in R&D intensity is associated with a 10% decrease in our text-based measure of integration in the within-industry specification, and with a 1.7% decrease in the within-firm specification.

In sharp contrast, the coefficients on *#Patents/assets* are positive and significant. All else equal, firms operating in industries with higher patenting intensity are more likely to be vertically integrated. Again, the economic magnitude of our estimates is large: Integration increases by 6.8% (1.9%) following a one standard deviation increase in patenting intensity in the within-industry (within-firm) specification. This

[23] We note, however, that our conclusion is unchanged if we use own-firm R&D and patent variables instead of industry-level variables. The results are presented in the Internet Appendix (IA.2).

[24] We are conservative in our inferences and cluster standard errors by industry and year to capture potential serial autocorrelation within industries.

result consistent with the ex post realization of successful innovation alleviating the ex ante need to incentivize investments in relation-specific assets. Firms with successful innovation face less incentives for relation-specificity and are more likely to integrate to reduce the threat of ex post holdup.

The estimates also indicate that firms are more likely to be integrated when they are more mature. In particular, integration is positively related to capital intensity and size in all specifications. It is also positively related to firms' age and negatively related to Market-to-Book in within-industry specifications. The 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 mature, growth options are less important, and hence they are less likely to need to disintegrate 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. Notably, firms that are closer to the end of the supply chain ($End\ Users$) are less likely to integrate, whereas firms having an observed conglomerate structure ($\#Segments$) are more likely to operate in a vertically integrated structure.

Overall, our analysis provides strong and robust evidence that R&D and patenting activities have opposite effects on vertical integration.[25] Yet, as R&D and patenting activities are positively related, some industries with high levels of R&D activities also display high levels of patenting.[26] To account for this possibility, we introduce an interaction term between them and report the results in columns (3) and (4). The coefficient on this interaction is negative, while patenting remains positive and R&D remains negative. The results confirm that unrealized and realized innovation

[25] In additional tests that we present in the Internet Appendix for brevity, we show that the results hold when we use lagged values of the independent variables (Table IA.3), when we use sales-weighted industry measures instead of equally-weighted measures (Table IA.4), when we focus solely on industry R&D and patent intensity and exclude the additional control variables (Table IA.5), when we include industry R&D and patent intensity separately (Table IA.6), and when we consider the natural logarithm of vertical integration to account for skewness (Table IA.7). We also consider the NAICS-based measure of vertical integration ($VI(Segment)$) and report the results in the Internet Appendix (Table IA.8). Overall, the results are much weaker, probably reflecting the difficulty to measure the dynamics of firm-level vertical integration by combining Compustat segments and vertical links at NAICS-industry level.

[26] In our sample, R&D and patenting activity are not perfectly correlated. This correlation is 0.33 across firms, and 0.58 across industries.

have opposite effects on firms' propensity to vertically integrate. The results with industry fixed effects are statistically and economically stronger but the results with firm fixed effects remain significant. Our interpretation of this negative interaction is that the ongoing R&D effect dominates and that innovation incentives remain important in these industries. Thus, the higher incentives from separate ownership dominate over any ex post holdup effect that encourages integration.

Note that the distinction between R&D and patents is non-trivial. High R&D does not necessarily lead to high patenting rates because the rents from some type of inventions are better appropriated without the use of patents. For instance, using the Carnegie Mellon Survey (CMS) on industrial R&D in the manufacturing sector, Cohen, Nelson, and Walsh (2000) report large differences across industries in the use of patents to protect inventions. In many cases, inventors prefer to keep their innovation secret to limit the risk that rivals use the information disclosed in the patent application process to work around the patents.

[Insert Table IX Here]

To further understand the distinct roles of R&D and patents on vertical integration, we use data from the CMS to separate (manufacturing) industries into two groups based on the importance of secrecy to protect innovation ('Low' and 'High' secrecy).[27] We then investigate whether and how the effects of R&D and patents on vertical integration differ across low and high secrecy industries. Table IX presents the results. Consistent with the key importance of property rights, we observe that patenting intensity is positively and significantly associated with vertical integration in low secrecy industries where patents are effective at protecting innovation. In contrast, vertical integration is largely unrelated to patenting intensity in high secrecy industries, where the rights of control of the realized innovation are less legally enforceable. These results are consistent with the idea that unrealized and realized innovation are distinctly related to firms' propensity to vertically integrate due to opposite investment incentives.

[27] The survey was performed with 1,478 R&D units or laboratories in 1994 and covers 34 manufacturing industries defined at the SIC 3-digit level. We measure the importance of secrecy using the average of two variables capturing the importance of secrecy for product innovation and for process innovation. We then assign firms into the low secrecy group if their SIC-3 industry is below the sample median, and to the high secrecy group if their SIC-2 industry is above the sample median.

V Vertical Acquisitions

To provide additional evidence on the economic forces that drive vertical integration, we study the determinants of vertical acquisitions, as these transactions represent a direct way firms can modify their degree of integration. To assess theoretical predictions in the literature (e.g., Grossman and Hart (1986)), we concentrate on targets (the sellers of assets) as they are the party that loses control rights due to the transaction. The trade-off between ex ante investment incentives and ex post hold-up should be important for target firms. We thus examine how R&D and patenting activity are related to the likelihood of being a target in vertical or non-vertical transactions.

A Transactions Sample

Our sample of transactions is from the Securities Data Corporation SDC Platinum database. We consider all announced and completed U.S. transactions 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. As we are interested in situations where the ownership of assets changes hands, we only consider acquisitions that lead acquirers to own majority stakes. 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 X displays the summary statistics describing the composition of our transactions sample. Panel A shows that the sample consists of 3,460 transactions when we limit attention to publicly traded acquirers and targets, and when we exclude transactions that involve financial firms and utilities (SIC codes between 6000 and 6999 and between 4000 and 4999), which we exclude from our regression analysis following convention in the literature. Panel A further reports how many of these transactions are classified as vertical using the Vertical Text-10% network and the NAICS-10% network, respectively.

[Insert Table X Here]

Given that both the Vertical Text-10% and NAICS-10% vertical networks are chosen to have similar granularity levels, it is perhaps surprising that the networks disagree in magnitude regarding the fraction of transactions that are vertically related. For our primary sample excluding financials, we observe that 39% are vertically related using the Vertical Text-10% network. Using the NAICS-10% network, which is similarly as granular, we observe that just 13% are vertically related. For any network with a granularity of 10%, if transactions are random, we expect to see 10% of transactions belonging to this network. The fact that we find 39% is strong evidence that many transactions occur between vertically related parties. The far lower figure for NAICS-10%, which arises despite the fact that the two networks have roughly the same granularity, is economically very large. This difference suggests that the accumulated noise associated with less direct vertical relatedness measures based on NAICS greatly reduces the ability to identify more direct vertically related transactions. We also note that with both networks, vertical deals are almost evenly split between upstream and downstream transactions.

We also find that transactions classified as vertical are followed by an increase in our firm-level measure of vertical integration (VI). Using the Vertical Text-10% network, acquirers in vertical transactions experience an increase of 6% in VI from one year prior to one year after the acquisition. In contrast, acquirers in non-vertical transactions experience a decrease of 0.70% in VI. In contrast, when we use the NAICS-10% network to identify vertical mergers, vertical acquirers see a negligible increase of of 0.30% in VI.

Panel B of Table X displays the average abnormal announcement returns (in percent) of combined acquirers and targets in vertical and non-vertical transactions. We present these results mainly to compare with previous research (based on either SIC or NAICS codes). Confirming existing evidence, the combined returns across all transactions are positive and range from 0.49% to 0.94%. Notably, when vertical transactions are identified using our text-based measure, the combined returns appear to be larger in vertical transactions than in non-vertical transactions. This is in line with the idea that vertical deals are value-creating on average (as in Fan and Goyal (2006)). Yet, the differences in announcement return between vertical and non-vertical transactions are never significant when vertical relatedness is identified using the NAICS-based network.

B Profile of Targets in Vertical Transactions

Table XI presents the R&D and patenting profile of targets in vertical and non-vertical deals. We focus on all transactions and we use our text-based network (10%) to identify vertical deals. We consider both industry- (i.e. TNIC-3) and firm-level measures of R&D and patenting activity. In Panel A, we observe a large difference between between targets in vertical and non-vertical deals. When compared to benchmark firms that never participate in any takeover transactions over the sample period (labeled non-merging firms), vertical targets exhibit lower levels of R&D and hold more patents. In contrast, targets in non-vertical deals are more R&D intense, and display lower patenting intensity.

[Insert Table XI Here]

To formally test these differences, we account for the fact that targets in vertical and non-vertical deals can differ on dimensions other than their R&D and patent profile. In Panels B and C, each actual target (vertical and non-vertical) is directly compared to a matched target with similar characteristics. For every transaction, we select matches from the subset of firms that did not participate in any transaction over the three years that precede the transaction. Matched targets are the nearest neighbors from a propensity score estimation. In panel B, we obtain matched targets based on industry (defined using the Fixed Industry Classification (FIC) of Hoberg and Phillips (2014))) and size. In panel C, we obtain matched targets based on FIC industries, size, age, Market-to-book ratio, PPE/Assets, the fraction of End Users from the BEA data, and the number of segments.

The results in Table XI are consistent with high R&D firms remaining separate to preserve strong ex ante incentives to create new innovation consistent with Grossman and Hart (1986). In contrast target firms have higher patenting activity consistent with acquirers buying high patent firms to reduce ex post hold-up that may occur when firms are attempting to commercialize the innovation.

[Insert Figure 6 Here]

These patterns are confirmed in Figure 6 when we look at the average patenting activity and R&D of target firms *prior* to their acquisition. Strikingly, vertical acquisitions tend to occur after targets experience a period of increased patenting activity (either measured with *log(1+#Patents)* or *#Patents/assets).* The realization of successful innovation (i.e. the grant of patents) marks a time of increased firm maturity, and may indicate the end of the innovation cycle. As the marginal product of additional R&D investment declines, we observe increased integration at this time. The mirror image appears true for non-vertical acquisitions, which tend to cluster after a period of lower patenting activity. Although the dynamics are less clear-cut, Figure 6 confirms that there are large differences in R&D intensity between the firms that are acquired in vertical and non-vertical deals. Non-vertical targets have much higher R&D intensity than vertical targets.

C Multivariate Analysis

We estimate logistic regressions to examine the likelihood of becoming a target. The dependent variable is an indicator variable indicating whether a given firm is a target in a vertical or a non-vertical transaction, as noted in the column headers, in a given year. We consider our text-based network when identifying which transactions are in fact vertically related. For consistency, we consider the same set of explanatory variables as in Table VIII. We cluster standard errors at the industry (using the FIC data from Hoberg and Phillips (2014)) and year level.

Table XII displays the results of these regressions. We find strong differences between the types of firms targeted in vertical and non-vertical deals. In particular, column (1) indicates that even after we control for other factors, firms in high R&D industries are less likely to be a target in a vertical transaction. In contrast, column (2) shows that firms in these same R&D intensive industries are in fact more likely to be targets in non-vertical transactions. This is consistent with Phillips and Zhdanov (2013), who document that horizontally related high R&D firms are likely to integrate to internalize R&D competition.

[Insert Table XII Here]

We next focus on the level of patenting activity. Consistent with our earlier findings that ex post successful innovation indicates maturity and lowers the returns from separate R&D investment, we find that vertical targets are more likely to be firms in industries with more patenting activity. The opposite is true for non-vertical acquisitions, where firms in high patenting industries are less likely to be acquired in restructuring transactions. In columns (3) to (4) we consider the interaction between R&D and patenting activity. The negative coefficient on this interaction that is only significant for non-vertical transactions further confirms the differences across these types of transactions. The positive coefficient for industry patenting activity also remains robust for vertical acquisitions.

Table XII also supports our hypothesis that maturity is an important positive determinant of vertical transactions. For instance, column (1) indicates that firms with lower market-to-book ratios and older firms are more likely to be targets of vertical deals. In contrast, targets in non-vertical deals are more likely to be young and are in less capital intensive industries. These findings are consistent with the following interpretation of the U-shaped relationship between firm maturity and restructuring activity noted in Arikan and Stulz (2011). Younger firms engage in non-vertical transactions likely to capitalize on asset complementarities, while more mature firms increase their acquisition activity as their focus turns to vertical acquisitions.

Table XII also reveals that vertical transactions are not significantly related to the degree of product market competition (measured with the 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 likely due to organically-driven integration. Therefore, the observed link between competition and integration is more likely driven by risk management concerns in competitive markets, and not due to the 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.

Overall, our results are broadly consistent with our earlier findings. The distinction between realized and unrealized innovation are key characteristics that determine who participates in vertical and non-vertical transactions.

VI Conclusions

Our paper examines vertical integration and the boundaries of the firm. We examine when firms choose to vertically integrate and transact within the firm and how integration is related to innovative activity. We examine changes to overall firm-specific vertical integration and vertical acquisitions. We examine how the distinction between incentives for intangible investments through R&D and the potential for ex post holdup both influence vertical integration and the likelihood of vertical transactions. We also examine the role of maturity, as vertical integration is more likely when the supply chain is mature enough to support long-term gains from operational synergies.

We measure vertical integration at the firm-level across multiple industries using computational linguistics 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 acquisitions 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. Firms in high R&D industries are less likely to vertically integrate through own-production and vertical acquisitions. These findings are consistent with firms remaining separate to maintain ex ante incentives to invest in intangible capital and to maintain residual rights of control, as in the property rights theory of Grossman, Hart and Moore.

In contrast, firms in industries with high patenting rates and thus high realized innovation and strong property rights are more likely to be vertically integrated and vertically integrate through acquisitions. 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. These results reconcile 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 also find empirical support for firm maturity impacting vertical integration. Both vertical integration and vertical mergers are positively related to variables that capture firm maturity. Firms that are older, more capital intensive, and with lower market to book values are more likely to vertically integrate and be the target of vertical acquisitions. These findings are consistent with integration taking place when product markets are mature, so that firms have time to recover additional fixed costs of integration.

Appendix 1: Model Proofs

In this appendix, we first show how the integration decision can be viewed as a real option. We then present the proofs of the four propositions from the text, along with a lemma needed for the proofs.

Optimal Timing of Integration As A Real Option

First note that after integration $I_t^* = 1$, integration stays forever ($I_{t+\tau}^* = 1$ for any $\tau > 0$). Here, the R&D investment by the supplier $x_{t+\tau}^* = 0$ for any $\tau >= 0$ since it is non-contractible. For the producer's investment in integration, we have $y_t^* = \text{argmax}_{y_t}[P_s(1+\rho(y_t)) - Ry_t^h]$ in each period if the base price in integration has been improved to P_s. Denote the maximized value and the perpetuity value by:

$$\begin{aligned} v(P_s; I=1) &= P_s(1+\rho(y_t^*)) - Ry_t^{*h} \\ V(P_s; I=1) &= \max_{y_t}[P_s(1+\rho(y_t)) - Ry_t^h] + \frac{V(P_s; I=1)}{1+r} \\ &= v(P_s; I=1)[1 + \frac{1}{1+r} + \frac{1}{(1+r)^2} + \dots] = \frac{1+r}{r} v(P_s; I=1) \end{aligned}$$

The optimal y_t^* thus depends on P_s in the following way

$$P_s \rho'(y_t^*) = Rhy_t^{*h-1}$$

A second observation is that the only state variable for the value function is the base price P_t^b, which is assumed to be equal to P_s ($s < N$) at time t. Therefore, we can define continuation value of separation and the value function $V(P_s)$ recursively as follows for $s < N$

$$V(P_s; I=0) \;=\; \max_{\{x_t, y_t\}} \underbrace{[P_s(1+y_t) - Ry_t^h - Sx_t^g]}_{\text{time } t \text{ profit}} + \underbrace{\frac{1}{1+r}[x_t V(P_{s+1}) + (1-x_t)V(P_s)]}_{\text{expected future profit}}$$

For $s = N$ any additional R&D expenditures thus cannot increase the base price

Table I:

$P_t^b = P_s$ $(s < N)$	$V(P_s)$	x^*	y^*
Integration $I = 1$	$V(P_s; I = 1) > V(P_s; I = 0)$	0	$P_s \rho'(y^*) = Rhy^{*h-1}$
Separation $I = 0$	$V(P_s; I = 1) < V(P_s; I = 0)$	$\frac{V(P_{s+1}) - V(P_s)}{1+r} = Sgx^{*g-1}$	$P_s = Rhy^{*h-1}$

anymore so:

$$V(P_N; I = 0) \;=\; \max_{y_t} P_N(1 + y_t) - Ry_t^h + \frac{V(P_N)}{1+r}$$

The optimal y_t^* and x_t^* also depend on P_s:

$$\begin{aligned} P_s &= Rhy_t^{*h-1} \\ \frac{V(P_{s+1}) - V(P_s)}{1+r} &= Sgx_t^{*g-1} \end{aligned}$$

The value function is thus:

$$V(P_s) \;=\; \max\{V(P_s; I = 1), V(P_s; I = 0)\}$$

Assuming for now that we solved the value function (which we do later in the appendix), the optimal decisions in each state can be summarized in Table I (above).

We now prove the propositions that we gave earlier in the paper.

Proposition 1

R&D expenditures are higher in separation, while commercialization and product integration expenditures are higher in integration.

Proof:

In integration we have $x^* = 0$. In separation we must have $x^* > 0$, otherwise assuming $x^* = 0$, by definition $V(P_s) = V(P_s; I = 0) = \max_y P_s(1+y) - Ry^h + \frac{V(P_s)}{1+r}$. So we solve that $V(P_s) = \frac{1+r}{r}[\max_y P_s(1+y) - Ry^h] < \frac{1+r}{r}v(P_s; I = 1) = V(P_s; I = 1)$, which gives a contradiction. So as long as separation is chosen, $x^* > 0$, which from the FOC, we can derive that $V(P_{s+1}) > V(P_s) = V(P_s; I = 0)$ (if separation is chosen when the base price last period is P_s).

Proposition 2

If $P_t^b = P_N$, then both firms prefer to integrate so $V(P_N) = V(P_N; I = 1) > V(P_N; I = 0)$.

Proof:

Assuming separation is chosen, then $V(P_N) = V(P_N; I = 0) = \max_y[P_s(1 + y) - Ry^h] + \frac{V(P_N)}{1+r}$. So we can solve that $V(P_N) = V(P_N; I = 0) = \frac{1+r}{r}\max_y[P_N(1 + y) - Ry^h] < \frac{1+r}{r}\max_y[P_N(1 + \rho(y)) - Ry^h] = V(P_N; I = 1)$, which is a contradiction. Therefore, we must have $V(P_N) = V(P_N; I = 1) > V(P_N; I = 0)$.

Lemma 1

Value function $V(P_s)$ is increasing in P_s.

Proof:

First note that the value of integration $V(P_s; I = 1)$ is always increasing in P_s. By the Envelope theorem, we have that $\frac{\partial V}{\partial P_s}(P_s; I = 1) = \frac{1+r}{r}(1 + \rho(y^*)) > 0$. Now just analyze by cases, if separation is chosen, given base price P_s, by the proof in Proposition 1 we know that $V(P_{s+1}) > V(P_s)$; otherwise integration is chosen, then $V(P_s) = V(P_s; I = 1) < V(P_{s+1}; I = 1) \leq V(P_{s+1})$. So in both cases, value function is increasing in base price. Also, we could see directly that $V(P_s; I = 0) < V(P_{s+1}; I = 0)$ since $V(P_s; I = 0)$ is increasing in P_s, $V(P_s)$, and $V(P_{s+1})$.

Solution of $V(P_s)$ by Backward Induction

Integration is a real option, and the base price is the only state variable. The series of value functions $\{V(P_0), V(P_1), \ldots V(P_s)\}$ is solved by backward induction.

- $P_b = P_N$: we know that $V(P_N) = V(P_N; I = 1) = \frac{1+r}{r}v(P_N; I = 1)$ which can be solved directly

- $P_b = P_{N-1}$: note that the value of integration is pre-determined as $V(P_{N-1}; I = 1) = \frac{1+r}{r}v(P_{N-1}; I = 1)$, so if $V(P_{N-1}) = V(P_{N-1}; I = 0)$ then it must be true that $V(P_{N-1}; I = 0)$ is the solution solving the following equation on M and M must be greater than $V(P_{N-1})$

$$M = \max_{\{x,y\}}[P_{N-1}(1 + y) - Ry^h] + [\frac{1}{1 + r}(xM + (1 - x)V(P_N)) - Sx^g]$$

- $P_b = P_s$: by now $V(P_{s+1})$ is known. Again, solve the following equation on M, then $V(P_s) = \max\{V(P_s; I = 1), M\}$

$$M = \max_{\{x,y\}}[P_s(1 + y) - Ry^h] + [\frac{1}{1 + r}(xM + (1 - x)V(P_{s+1})) - Sx^g]$$

The above is a valid solution as long as the integration decision is monotonic in s, in other words, there is a triggering state s^* such that separation is chosen whenever $s < s^*$ and integration is chosen when $s \geq s^*$

Assumption 2

The increase in price P_s decreases with each successive innovation such that the series of value functions solved using the above method satisfies this condition: $V(P_{s+2}) - V(P_{s+1}) < V(P_{s+1}) - V(P_s)$.

The following Proposition then claims that there exists a triggering state s^*, so the series of value functions solved using the backward induction is the true solution.

Note that under this assumption, Lemma 1 holds, and the marginal benefit of R&D expenditures which equals $\frac{V(P_{s+1})-V(P_s)}{1+r}$ in separation is decreasing in base price, and the optimal level of R&D expenditures in separation is also decreasing. Also note that even though the function $V(P; I = 1)$ is convex in P, we could make the increment in base price so small such that conditions in Assumption 3 hold.

Proposition 3

There exists a state s^* such that $V(P_s) = V(P_s; I = 1) \geq V(P_s; I = 0)$ for any $s \geq s^*$, and $V(P_s) = V(P_s; I = 1) < V(P_s; I = 0)$ for any $s < s^*$. The state s^* would then be the triggering state for integration.

Proof:

We only need to prove that there does not exist a state s such that integration is chosen with base price P_s while separation is chosen with base price P_{s+1}.

In state s, we have

$$\begin{aligned} V(P_s; I = 1) &= \max_y [P_s(1 + \rho(y)) - Ry^h] + \frac{V(P_s; I = 1)}{1 + r} \\ V(P_s; I = 0) &= \max_y [P_s(1 + y) - Ry^h] + \frac{V(P_s)}{1 + r} + \max_x [\frac{V(P_{s+1}) - V(P_s)}{1 + r} x - Sx^g] \end{aligned}$$

Integration is chosen in state s meaning $V(P_s) = V(P_s; I = 1)$ and

$$\underbrace{\max_y [P_s(1 + \rho(y)) - Ry^h] - \max_y [P_s(1 + y) - Ry^h]}_{\substack{\text{Increments in TS by commercialization expenditures in integration} \\ \text{if the integration as a real option is exercised right now}}} \quad > \quad \underbrace{\max_x [\frac{V(P_{s+1}) - V(P_s)}{1 + r} x -}_{\substack{\text{Increments in TS by R\&D} \\ \text{Continuation value in separ}}}$$

First note that $\max_x [\frac{V(P_{s+1}) - V(P_s)}{1+r} x - Sx^g]$ is always non-negative so we must have $\max_y [P_s(1 + \rho(y)) - Ry^h] - \max_y [P_s(1 + y) - Ry^h] > 0$.

The difference $\max_y [P_s(1 + \rho(y)) - Ry^h] - \max_y [P_s(1 + y) - Ry^h]$ is a function of P_s and the derivative with respect to P_s is $\rho(y^1) - y^0$ (by the Envelope Theorem), with y^1 and y^0 the optimum under integration and separation. Note that from the FOC we have $P_s = Rh(y^0)^{h-1}$ and $P_s = Rh(y^1)^{h-1}$, since $\rho'(y^1) > 1$ and $h > 1$ we must have $y^1 > y^0$ (commercialization expenditures are larger in integration) and thus $\rho(y^1) > \rho(y^0) > y^0$. So the difference is increasing in P_s so that

$$\begin{aligned} &\max_y [P_{s+1}(1 + \rho(y)) - Ry^h] - \max_y [P_{s+1}(1 + y) - Ry^h] \quad > \\ &\quad \max_y [P_s(1 + \rho(y)) - Ry^h] - \max_y [P_s(1 + y) - Ry^h] \end{aligned}$$

The net benefit of R&D expenditures $\max_x [\frac{V(P_{s+1}) - V(P_s)}{1+r} x - Sx^g]$, however, is decreasing in P_s because the increments in the value function $V(P_{s+1}) - V(P_s)$, by assumption, are decreasing in P_s, so must have

$$\max_x [\frac{V(P_{s+2}) - V(P_{s+1})}{1 + r} x - Sx^g] \quad < \quad \max_x [\frac{V(P_{s+1}) - V(P_s)}{1 + r} x - Sx^g]$$

Combining the three inequalities, we have that

$$\max_y[P_{s+1}(1+\rho(y)) - Ry^h] - \max_y[P_{s+1}(1+y) - Ry^h] \quad > \quad \max_x[\frac{V(P_{s+2}) - V(P_{s+1})}{1+r}x - Sx^g]$$

So the exercise value (exercise the option of integration) is greater than the continuation value (in separation) in state $s+1$. From this it is easy to see that $V(P_{s+1}; I = 1) > V(P_{s+1}; I = 0)$ since otherwise $V(P_{s+1}; I = 0)$ would be equal to $\frac{1+r}{r}\left\{\max_y[P_{s+1}(1+y) - Ry^h] + \max_x[\frac{V(P_{s+2})-V(P_{s+1})}{1+r}x - Sx^g]\right\}$, which is less than $V(P_{s+1}; I = 1)$ which is equal to $\frac{1+r}{r}\max_y[P_{s+1}(1+\rho(y)) - Ry^h]$.

Therefore, if in state s, integration is chosen, then in state $s+1$, integration will be chosen too. By induction, all states after s will be under integration. Given the fact that the two firms start as separated, and in the final state N they must choose integration, there must exist a triggering state s^* such that integration is chosen in states $s \geq s^*$ and separation is chosen in states $s < s^*$. In other words, s^* would be the state in which the real option of integration is exercised in equilibrium.

Note that some states of the world could not be reached in equilibrium. For example, for any $s > s^*$, the base price P_s would never appear in equilibrium since the two firms have integrated at state s^*. So the total surplus $V(P_{s^*})$ would be the highest one reached in equilibrium, which is also the final value in integration.

Appendix 2: Excluded BEA Words

Because they are used in a large number of commodities, we exclude the following words from the BEA commodity vocabulary we use to compute vertical relatedness: accessories, accessory, air, airs, attachment, attachments, commercial, commercials, component, components, consumer, consumers, development, developments, equipment, exempt, expense, expenses, ga, gas, industrial, industrials, net, part, parts, processing, product, products, purchased, purchase, receipt, receipts, research, researches, sale, sales, service, services, system, systems, unit, units, work, works, tax, taxes, oil, repair, repairs, aids, aid, air, apparatuses, apparatus, applications, application, assemblies, assembly, attachments, attachment, automatic, auxiliary, bars, bar, bases, base, blocks, block, bodies, body, bulk, business, businesses, byproducts, byproduct, cares, care, centers, center, collections, collection, combinations, combination, commercials, commercial, completes, complete, components, component, consumers, consumer, consumption, contracts, contract, controls, control, covers, cover, customs, custom, customers, customer, cuts, cut, developments, development, directly, distributions, distribution, domestic, dries, dry, equipments, equipment, establishments, establishment, exempt, expenses, expense, facilities, facility, fees, fee, fields, field, finished, finish, finishings, finishing, gas, generals, general, greater, hands, hand, handling, high, hot, individuals, individual, industrials, industrial, industries, industry, installations, installation, lights, light, lines, line, maintenances, maintenance, managements, management, manmade, manufactured, manufacture, materials, material, naturals, natural, nets, net, offices, office, only, open, operated, operate, organizations, organization, others, other, pads, pad, paid, pay, parts, part, permanent, portable, powers, power, processing, products, product, productions, production, public, purchased, purchase, purposes, purpose, receipts, receipt, reclassified, reclassify, repairs, repair, researches, research, sales, sale, self, services, service, sets, set, shares, share, shipped, similar, singles, single, sizes, size, small, soft, specials, special, stocks, stock, storages, storage, supplies, supply, supports, support, surfaces, surface, systems, system, taxes, tax, taxable, technical, this, trades, trade, transfers, transfer, types, type, units, unit, used, without, work, works.

Appendix 3: 10-K Phrase Exclusions

Because we use 10-K text only to identify a firm's own-product market location (vertically related vocabulary is identified using BEA data), we exclude any part of a sentence that follows any of the following 81 phrases: buy, buys, sells its, are sold, buying, products for, for sale, for their, used in, used by, used as, used for, used with, used primarily, used mainly, used commonly, primarily used, mainly used, commonly used, for use, uses, utilized, serve, serving, serves, sold to, sold primarily, sold mainly, sold commonly, designed for, supply of, supply for, supplier to, supplied to, service to, purchase, purchaser, purchasers, customer, customers, user, users, for application, equipment for, equipment to, equipment by, product for, product to, product by, solution for, solution to, solution by, component for, component to, component by, application for, application to, application by, system for, system to, system by, equipments for, equipment for, equipment to, equipments to, equipments by, products for, products to, products by, solutions for, solutions to, solutions by, components for, components to, components by, applications for, applications to, applications by, systems for, systems to, systems by.

Appendix 4: Validation test of the new Vertical Integration Measure

To operationalize our tests of whether our text based measure captures vertical integration, 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 $\frac{AP_t - AP_{t-1}}{AP_t + AP_{t-1}}$.[28] 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).

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

To measure firm pairwise $TOPSIDE$ correlations for a given network, we estimate the following regression, where one observation is one firm-pair 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) $\mid \triangle AR_{i,t}| > \mid \triangle AP_{i,t}|$, or (2) $\mid \triangle AR_{j,t}| < \mid \triangle AP_{j,t}|$). 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 γ should be positive for horizontal networks, and negative for vertical networks. We present the results of the these tests in Panel B of Table III.

The results in Panel B of Table III show that γ is systematically negative for the vertical networks we construct. However, the estimates of γ are 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-statistic ranges from 3.19 to 4.55), where the likelihood of contamination due to breadth is minimized. None of the estimates of γ are significant for the NAICS-based vertical network, and the coefficient estimates are an order of magnitude smaller. In the last column we can see that the estimates of γ for the TNIC-3 horizontal network are significantly positive, as is predicted for horizontal relationships.

The results of these tests show that horizontally related firms experience positively correlated responses in accounts payable and accounts receivable, whereas our vertically related firm pairs experience negatively correlated responses. These results provide a strong validation test of our new measure of vertical linkages. These tests further illustrate that our measures of vertical relatedness statistically capture vertical integration and information about vertical links, whereas NAICS-based measures of vertical integration are likely contaminated.

Appendix 5: Variable Descriptions

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

- *VI* measures the degree to which a firm offers products and services that are vertically related based on our new text-based approach to measure vertical relatedness (as defined in Section IV.A)).
- *VI(search)* id a dummy variable that indicates whether a given firm mentions that it is vertically integrated in its 10-K report in a given year ((as defined in Section IV.A)).
- *Retail* measures the degree to which a firm-year's products flow to retail customers (as defined in Section IV.A)).
- *Government* measures the degree to which a firm-year's products flow to the government (as defined in Section IV.A)).
- *Export* measures the degree to which a firm-year's products are exported (as defined in Section IV.A)).
- *End users* is the sum of *Retail*, *Government*, and *Export*. It measures degree to which a firm-year's products exit the supply chain.
- *HHI* measures the degree of concentration (of sales) within TNIC-3 industries. We compute HHI as the TNIC HHI in Hoberg and Phillips (2014) using the text-based TNIC-3 horizontal industry network.
- *Industry 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. We average this firm level measure over all firms in a given industry.
- *Industry #Patents/assets* is the number of (granted) patents a firm owns scaled by the level of assets (AT). We average this firm level measure over all firms in a given industry. Patents data are obtained from the US Patent and Trademark Office data (the NBER Patent data archive). Patent data are available until 2006. We assume patents for 2007 and 2008 remain at the level of 2006 in our tests.[29]
- *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).

[29] All the results in the paper hold if we instead restrict our sample to the period 1996-2006 to have complete information on patents.

- *#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 III.D) 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.

——— , Phillipe Aghion, Rachel Griffith, and Fabrizio Zilbotti, 2010, Vertical integration and technology: Theory and evidence, *Journal of the European Economic Association* pp. 989–1033.

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

Ahern, Kenneth, 2012, Bargaining power and industry dependence in mergers, *JFE* 103, 530–550.

——— , and Jarrad Harford, 2013, 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.

Antras, Pol, 2003, Firms, contracts, and firm structure, *Quarterly Journal of Economics* pp. 1375–1418.

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 Syverson, 2014, Vertical integration and input flows, *forthcoming American Economic Review* pp. 1120–1148.

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.

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.

Gibbons, Robert, 2005, Four formal(izable) theories of the firm?, *Journal of Economic Behavior and Organization* 58, 200–245.

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.

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

——— , 2014, Text-based network industry classifications and endogenous product differentiation, University of Maryland Working Paper.

Holmstrom, Bengt, and Paul Milgrom, 1991, Multi-task principal-agent problems: Incentive contracts, asset ownership and job design, *Journal of Law, Economics and Organization* 7, 24–52.

——— , 1994, The firm as an incentive system, *American Economic Review* 84, 972–991.

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

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.

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.

Zeile, William, 1997, U.s. intrafirm trade in goods, *Survey of Current Business* pp. 23–38.

Table II: BEA vocabulary example: Photographic and Photocopying Equipment

Description of Commodity Sub-Category	Value of Production ($Mil.)
Still cameras (hand-type cameras, process cameras for photoengraving and photolithography, and other still cameras)	266.1
Projectors	72.4
Still picture commercial-type processing equipment for film	40.5
All other still picture equipment, parts, attachments, and accessories	266.5
Photocopying equipment, including diffusion transfer, dye transfer, electrostatic, light and heat sensitive types, etc.	592.4
Microfilming, blueprinting, and white-printing equipment	20.7
Motion picture equipment (all sizes 8mm and greater)	149.0
Projection screens (for motion picture and/or still projection)	204.9
Motion picture processing equipment	23.0

Note: This table provides an example of the BEA commodity 'photographic and photocopying equipment' (IO Commodity Code #333315). The table displays its sub-commodities and their associated product text, along with the value of production for each sub-commodity.

Table III: Vertical Network Summary Statistics

Network:	Vertical Text-10%	Vertical Text-1%	NAICS-10%	NAICS-1%	TNIC-3
		Panel A: Granularity and Overlap			
Granularity	10%	1%	9.48%	1.37%	2.33%
% of pairs in TNIC-3	1.33%	2.39%	2.67%	2.89%	100%
% of pairs in the same SIC	0.74%	1.03%	0.35%	0.20%	38.10%
% of pairs in the same NAICS	0.59%	0.66%	0.30%	0.18%	38.11%
% of pairs in the same SIC or NAICS	0.81%	1.09%	0.35%	0.20%	41.24%
% of pairs in Vertical Text-10%	100%	100%	10.48%	13.18%	6.15%
% of pairs in Vertical Text-1%	10%	100%	1.18%	1.21%	1.09%
% of pairs that include a financial firm	9.20%	1.80%	48.44%	34.31%	58.72%
% of (no fin.) pairs in Vertical Text-10%	100%	100%	19.90%	19.29%	11.63%
% of (no fin.) pairs in Vertical Text-1%	10%	100%	2.14%	1.71%	2.44%
		Panel B: Validity Test (Trade Credit Shocks)			
γ (unconditional)	-0.0006	-0.0024	-0.0001	-0.0001	0.0071
(t-statistic)	(-3.37)	(4.57)	(-0.93)	(-0.03)	(15.91)
γ (if $\mid \triangle AR_{i,t} \mid > \mid \triangle AP_{i,t} \mid$)	-0.0006	-0.0030	-0.0002	-0.0007	0.0071
(t-statistic)	(-2.40)	(-3.71)	(-0.90)	(-0.92)	(12.37)
γ (if $\mid \triangle AR_{j,t} \mid < \mid \triangle AP_{j,t} \mid$)	-0.0006	-0.0027	-0.0001	-0.0005	0.054
(t-statistic)	(-2.47)	(-3.83)	(-0.051)	(-0.06)	(8.37)

Note: This table displays various characteristics for five networks: Vertical Text-10% and Vertical Text-1% vertical networks, NAICS-10% and NAICS-1% vertical networks, and the TNIC-3 horizontal network. The coefficient γ in Panel B is obtained from OLS regressions of trade credit shocks of upstream firms on trade credit shocks of downstream firms (See Appendix 1). We report t-statistic below the coefficients.

Table IV: Summary Statistics

Variable	Mean	St. Dev	Min	p25	p50	p75	Max	#Obs.
Panel A: Data from Text Analysis								
VI	0.012	0.011	0.000	0.004	0.008	0.016	0.116	45,198
VI(search)	0.079	0.270	0.000	0.000	0.000	0.000	1.000	45,198
Retail	0.352	0.081	0.001	0.297	0.347	0.401	0.859	45,198
Government	0.031	0.021	0.000	0.018	0.026	0.038	0.971	45,198
Export	0.086	0.025	0.001	0.069	0.085	0.102	0.301	45,198
End users	0.470	0.076	0.086	0.421	0.470	0.516	0.973	45,198
Panel B: Data from Existing Literature								
R&D/sales	0.059	0.121	0.000	0.000	0.000	0.062	0.669	45,198
#Patents/assets	0.007	0.020	0.000	0.000	0.000	0.002	0.112	45,198
log(1+#Patents)	0.600	1.176	0.000	0.000	0.000	0.693	8.195	45,198
PPE/assets	0.263	0.221	0.000	0.089	0.195	0.374	0.888	45,198
HHI	0.259	0.225	0.014	0.103	0.177	0.334	1.000	45,198
log(assets)	5.684	1.784	2.529	4.330	5.505	6.839	10.881	45,198
log(age)	2.953	1.060	0.000	2.303	2.996	3.664	5.037	45,198
#Segments	1.550	0.999	1.000	1.000	1.000	2.000	12.000	45,198
MB	1.980	1.444	0.624	1.110	1.497	2.247	8.351	45,198
VI(Segment)	0.013	0.037	0.000	0.000	0.000	0.006	0.639	45,198

Note: This table displays summary statistics for the variables used in the analysis. All variables are defined in Appendix 2.

Table V: Summary Statistics Across Vertical Integration Quartiles

Variables	Quartile 1 (Low VI)	Quartile 2	Quartile 3	Quartile 4 (High VI)
	Panel A: Data from Text Analysis			
VI	0.002	0.006	0.011	0.028
VI(search)	0.031	0.053	0.089	0.144
Retail	0.396	0.366	0.346	0.301
Government	0.039	0.031	0.027	0.025
Export	0.079	0.084	0.087	0.095
End users	0.515	0.482	0.462	0.422
	Panel B: Data from Existing Literature			
R&D/sales	0.098	0.062	0.047	0.027
#Patents/assets	0.006	0.008	0.008	0.007
log(1+#Patents)	0.420	0.504	0.640	0.835
PPE/assets	0.199	0.247	0.285	0.320
HHI	0.231	0.258	0.272	0.274
log(assets)	5.355	5.485	5.771	6.123
log(age)	2.736	2.788	2.970	3.318
#Segments	1.318	1.422	1.572	1.890
MB	2.355	2.080	1.880	1.606
VI(Segment)	0.006	0.009	0.013	0.024

Note: This table displays summary statistics by (annually sorted) quartiles of vertical integration (VI). All variables are defined in Appendix 2.

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

Company	Rank	#Segments	VI	Perc.(VI)	Perc.($VI(Segment)$)
HANDY & HARMAN LTD	1	5	0.091	1	0.969
PARKER-HANNIFIN CORP	2	2	0.079	0.999	0.000
EATON CORP	3	5	0.076	0.999	0.966
EMERSON ELECTRIC CO	4	6	0.074	0.999	0.991
FRANKLIN ELECTRIC CO INC	5	1	0.073	0.998	0.717
COMMERCIAL VEHICLE GROUP INC	6	1	0.069	0.998	0.000
ROCKWOOD HOLDINGS INC	7	5	0.069	0.997	0.959
SCHNITZER STEEL INDS -CL A	8	3	0.064	0.997	0.000
LEGGETT & PLATT INC	9	3	0.062	0.997	0.710
DOVER CORP	10	4	0.058	0.996	0.641
SIFCO INDUSTRIES	11	2	0.055	0.996	0.994
MYERS INDUSTRIES INC	12	1	0.053	0.996	0.000
AMPCO-PITTSBURGH CORP	13	2	0.053	0.995	0.681
SONOCO PRODUCTS CO	14	3	0.052	0.995	0.000
LKQ CORP	15	1	0.052	0.995	0.000
P & F INDUSTRIES -CL A	16	2	0.052	0.994	0.760
BERKSHIRE HATHAWAY	17	9	0.051	0.994	0.000
PRECISION CASTPARTS CORP	18	2	0.051	0.993	0.790
MATTHEWS INTL CORP -CL A	19	6	0.051	0.993	0.884
RELIANCE STEEL & ALUMINUM CO	20	1	0.050	0.993	0.000
CARLISLE COS INC	21	6	0.050	0.992	0.962
UNVL STAINLESS & ALLOY PRODS	22	1	0.050	0.992	0.000
AMERICAN AXLE & MFG HOLDINGS	23	1	0.049	0.992	0.000
ENCORE WIRE CORP	24	1	0.049	0.991	0.000
HAWK CORP	25	1	0.049	0.991	0.000
KANSAS CITY SOUTHERN	26	1	0.049	0.991	0.000
AMERICAN ELECTRIC TECH INC	27	3	0.049	0.990	0.885
DREW INDUSTRIES INC	28	1	0.049	0.990	0.000
CHINA PRECISION STEEL INC	29	1	0.048	0.989	0.000
COLEMAN CABLE INC	30	1	0.048	0.989	0.000

Note: The table displays the 30 most vertically integrated firms in 2008 based on our text-based measure of vertical integration (VI). The table also presents the number of Compustat segments, the VI score, the firm's percentile VI ranking, and the firm's percentile $VI(Segment)$ ranking.

Table VII: Transition Probabilities for Vertical Integration

Sample	Quintile 1 (Least VI)	Quintile 2	Quintile 3	Quintile 4	Quintile 5 (Most VI)	# Obs.
			Panel A: One-year Transition Probabilities			
Quintile 1	0.815	0.153	0.025	0.006	0.001	7,262
Quintile 2	0.150	0.647	0.174	0.026	0.004	7,491
Quintile 3	0.028	0.164	0.625	0.168	0.014	7,542
Quintile 4	0.006	0.027	0.160	0.685	0.121	7,657
Quintile 5	0.002	0.003	0.012	0.120	0.863	7,845
			Panel B: Three-year Transition Probabilities			
Quintile 1	0.692	0.230	0.054	0.018	0.006	4,680
Quintile 2	0.214	0.484	0.234	0.055	0.013	4,987
Quintile 3	0.059	0.207	0.460	0.237	0.037	5,079
Quintile 4	0.013	0.056	0.213	0.534	0.183	5,364
Quintile 5	0.006	0.011	0.028	0.175	0.780	5,705
			Panel C: Six-year Transition Probabilities			
Quintile 1	0.619	0.254	0.079	0.031	0.017	2,423
Quintile 2	0.250	0.396	0.231	0.098	0.026	2,603
Quintile 3	0.075	0.218	0.381	0.258	0.067	2,716
Quintile 4	0.022	0.087	0.235	0.447	0.209	3,086
Quintile 5	0.009	0.022	0.050	0.200	0.719	3,338

Note: The table presents transition matrices examining ex post text-based vertical integration (VI) for firms with varying levels of initial VI. We display one, three and six year transition probabilities, respectively. In each case, we assign firms to quintiles in each ex ante interval based on the level of ex ante VI. Holding breakpoints fixed for each interval, we then group observations in quintiles based on their ex post level of VI.

Table VIII: The Determinants of Vertical Integration

Dep. Variable:	**(Text-based) VI**			
	Baseline		Interaction	
	(1)	(2)	(3)	(4)
Ind.(R&D/sales)	-0.100	-0.017	-0.081	-0.012
	(-13.91)	(-3.24)	(-9.35)	(-1.90)
Ind.(#Patents/assets)	0.068	0.019	0.084	0.025
	(9.49)	(3.99)	(9.25)	(3.82)
Ind.(R&D/sales) × Ind.(#Patents/assets)			-0.014	-0.005
			(-3.35)	(-1.78)
Ind.(PPE/assets)	0.022	0.014	0.023	0.014
	(1.68)	(1.50)	(1.81)	(1.48)
HHI	-0.107	-0.055	-0.106	-0.055
	(-15.98)	(-11.77)	(-15.75)	(-11.66)
End User	-0.24	-0.142	-0.24	-0.142
	(-31.18)	(-21.56)	(-31.22)	(-21.54)
#Segment (NAICS)	0.131	0.041	0.131	0.041
	(22.98)	(6.83)	(22.99)	(6.84)
log(Assets)	0.051	0.124	0.051	0.124
	(10.47)	(11.35)	(10.49)	(11.25)
log(Age)	0.021	0.014	0.02	0.014
	(4.38)	(1.37)	(4.32)	(1.35)
MB	-0.016	0.005	-0.016	0.005
	(-5.13)	(1.93)	(-5.15)	(1.95)
Industry Fixed Effects	Yes	No	Yes	No
Firm Fixed Effects	No	Yes	No	Yes
#obs.	45,198	45,198	45,198	45,198
Adj. R^2	0.526	0.855	0.527	0.855

Note: The dependent variable is vertical integration VI. All independent variables are defined in Appendix 2. The independent variables are standardized for convenience. Standard errors are clustered by industry and year and we report t-statistics in parentheses.

Table IX: The Determinants of Vertical Integration: Secrecy vs. Patents

Dep. Variable:	(Text-based) VI			
	Low	High	Low	High
	(1)	(2)	(3)	(4)
Ind.(R&D/sales)	-0.141	-0.104	-0.036	-0.004
	(-9.26)	(-6.65)	(-2.89)	(-0.33)
Ind.(#Patents/assets)	0.042	0.002	0.023	0.013
	(3.04)	(0.15)	(2.56)	(1.62)
Ind.(PPE/assets)	-0.027	0.077	0.017	-0.041
	(-0.94)	(3.39)	(0.72)	(-1.77)
HHI	-0.173	-0.165	-0.069	-0.089
	(-10.55)	(-12.53)	(-6.37)	(-7.35)
End User	-0.25	-0.311	-0.227	-0.179
	(-14.81)	(-21.81)	(-11.96)	(-9.80)
#Segment (NAICS)	0.102	0.155	0.051	0.068
	(7.47)	(11.39)	(3.17)	(4.04)
log(Assets)	0.036	0.057	0.114	0.238
	(3.14)	(5.94)	(4.58)	(7.43)
log(Age)	0.002	-0.012	-0.008	0.012
	(0.11)	(-1.04)	(-0.25)	(0.41)
MB	-0.001	-0.046	0.001	0.004
	(-0.11)	(-6.17)	(0.11)	(0.66)
Industry Fixed Effects	Yes	Yes	No	No
Firm Fixed Effects	No	No	Yes	Yes
#obs.	9,409	9,333	9,409	9,333
Adj. R^2	0.570	0.509	0.857	0.815

Note: The dependent variable is vertical integration VI. All independent variables are defined in Appendix 2. The 'Low' and 'High' group correspond to industries where the importance of secrecy (as opposed to patents) for protecting innovation is below and respectively above the sample median as defined in the text. The sample is limited to 34 manufacturing industries. The independent variables are standardized for convenience. Standard errors are clustered by industry and year and we report t-statistics in parentheses.

Table X: Mergers and Acquisitions - Sample Description

Measure:	**All**	**Text-Based**		**NAICS-based**	
Deal type:		Vertical	Non-Vertical	Vertical	Non-Vertical
		Panel A: Sample Description			
# Transactions	3,460	1,368	2,092	460	3,000
% Vertical (Non-Vertical)		39.54%	60.46%	13.29%	86.71%
# Upstream		687		199	
# Downstream		681		261	
		Panel B: Combined Acquirers and Targets Returns			
CAR(0)	0.49%	0.65%	0.38%^{a}	0.46%	0.49%
CAR(-1,1)	0.86%	0.97%	0.79%^{a}	0.55%	0.91%
# Transactions	3,256	1,301	1,995	427	2,829

Note: Panel A displays statistics for vertical and non-vertical transactions (non-financial firms only). A transaction is vertical if the acquirer and target are pairs in the Vertical Text-10% network or the NAICS-10% network. Panel B displays the average cumulated abnormal announcement returns (CARs) of combined acquirers and targets. We include the superscript ^{a} when the difference in CARs between vertical and non-vertical transactions is significant at the 5% level.

Table XI: Vertical Transactions - Deal-level Analysis

Variable:	Ind.(R&D/ sales)	R&D/ sales	Ind.(#Patents/ Assets)	#Patents/ Assets
	Panel A: Whole Sample			
(i) Vert. Targets	0.0555	0.0424	0.0076	0.0091
(ii) Non-Vert. Targets	0.1262	0.0813	0.0077	0.0069
(iii) Non-Merging Firms	0.0905	0.0622	0.0075	0.0082
t-statistic [(i)-(ii)]	(-16.67)	(-9.32)	(-0.30)	(3.36)
t-statistic [(i)-(iii)]	(-9.20)	(-5.34)	(0.27)	(1.45)
t-statistic [(ii)-(iii)]	(11.03)	(6.07)	(0.74)	(-2.42)
	Panel B: Matched Targets I			
(i) Vert. Targets	0.0555	0.0424	0.0076	0.0091
(ii) Matched Vert. Targets	0.0953	0.0592	0.0072	0.0065
t-statistic [(i)-(ii)]	(-9.64)	(-4.01)	(0.94)	(3.44)
(i) Non-Vert. Targets	0.1262	0.0813	0.0077	0.0069
(ii) Matched Non-Vert. Targets	0.1073	0.0696	0.0079	0.0072
t-statistic [(i)-(ii)]	(4.15)	(2.66)	(-0.62)	(-0.52)
	Panel C: Matched Targets II			
(i) Vert. Targets	0.0555	0.0424	0.0076	0.0091
(ii) Matched Vert. Targets	0.0802	0.0477	0.0061	0.0048
t-statistic [(i)-(ii)]	(-6.63)	(-1.36)	(3.81)	(6.39)
(i) Non-Vert. Targets	0.1262	0.0813	0.0077	0.0069
(ii) Matched Non-Vert. Targets	0.0931	0.0584	0.0072	0.0062
t-statistic [(i)-(ii)]	(7.67)	(5.62)	(1.62)	(1.21)

Note: Transactions are defined as vertical when the acquirer and target are in pairs in the Vertical Text-10% network. In Panel A, we compare targets of vertical and non-vertical deals, and non-merging firms. In Panel B, each target is compared to a "matched" non-merging target using a propensity score model based on industry , size, and year. In Panel C, the propensity score is based on industry, size, age, Market-to-Book, PPE/Assets, End Users, # of NAICS Segments, and year. We report *t*-statistics corresponding to tests of mean differences.

Table XII: The Determinants of Vertical Target Acquisitions

Dep. Variable:	**Prob(Target)**			
Deal type:	Vertical	Non-Vertical	Vertical	Non-Vertical
	(1)	(2)	(3)	(4)
Ind.(R&D/sales)	-0.312	0.340	-0.248	0.505
	(-5.66)	(10.92)	(-3.03)	(11.24)
Ind.(#Patents/asse)	0.338	-0.177	0.373	0.016
	(11.41)	(-4.55)	(10.85)	(0.330)
Ind.(R&D/sales) × Ind.(#Patents/assets)			-0.043	-0.127
			(-1.26)	(-4.93)
Ind.(PPE/assets)	-0.023	-0.188	-0.016	-0.135
	(-0.68)	(-3.61)	(-0.46)	(-2.55)
HHI	-0.100	-0.182	-0.100	-0.167
	(-2.57)	(-5.51)	(-2.55)	(-4.99)
End User	-0.348	0.187	-0.346	0.208
	(-10.35)	(6.73)	(-10.31)	(7.48)
#Segment (NAICS)	0.171	-0.013	0.171	-0.012
	(7.87)	(-0.51)	(7.85)	(-0.46)
log(Assets)	0.660	0.431	0.660	0.434
	(17.64)	(14.74)	(17.67)	(14.93)
log(Age)	0.180	-0.031	0.181	-0.026
	(4.95)	(-1.18)	(4.97)	(-1.00)
MB	-0.269	-0.071	-0.271	-0.073
	(-5.18)	(-2.44)	(-5.22)	(-2.54)
#obs.	45,198	45,198	45,198	45,198
Pseudo R^2	0.116	0.045	0.116	0.048

Note: The dependent variable in the logistic models is a dummy indicating whether the given firm is a target in a vertical or non-vertical transaction in a given year. Vertical transactions are identified using the Vertical Text-10% network. All independent variables are defined in Appendix 2. The independent variables are standardized for convenience. Standard errors are clustered by industry and year and we report t-statistics in parentheses.

Figure 3: 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 IV.A. The sample period is 1996-2008.

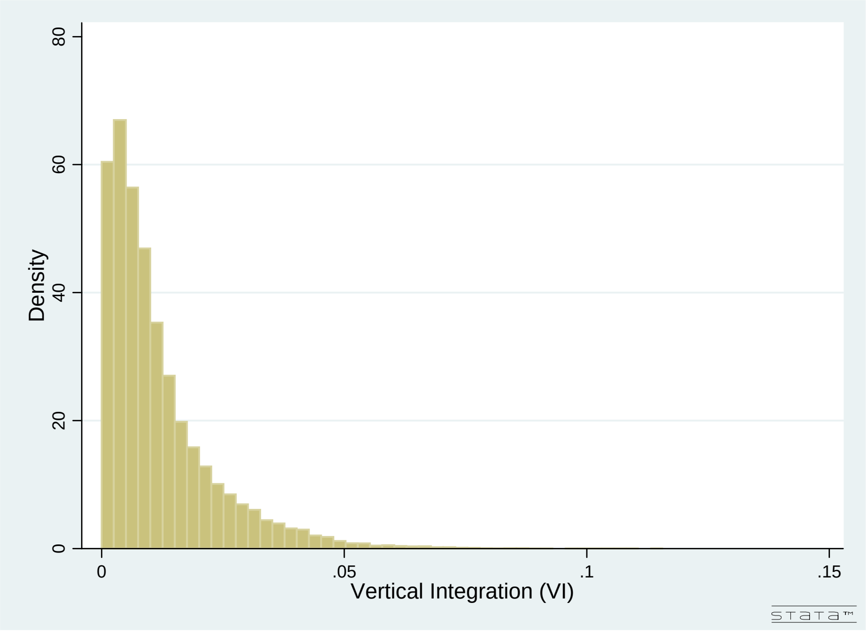


Figure 4: Evolution of sample-wide average (text-based) Vertical Integration over time. Vertical integration (VI) is defined in Section IV.A. The solid blue line is the annual equal-weighted average VI. The dashed red line is the corresponding sales-weighted average.

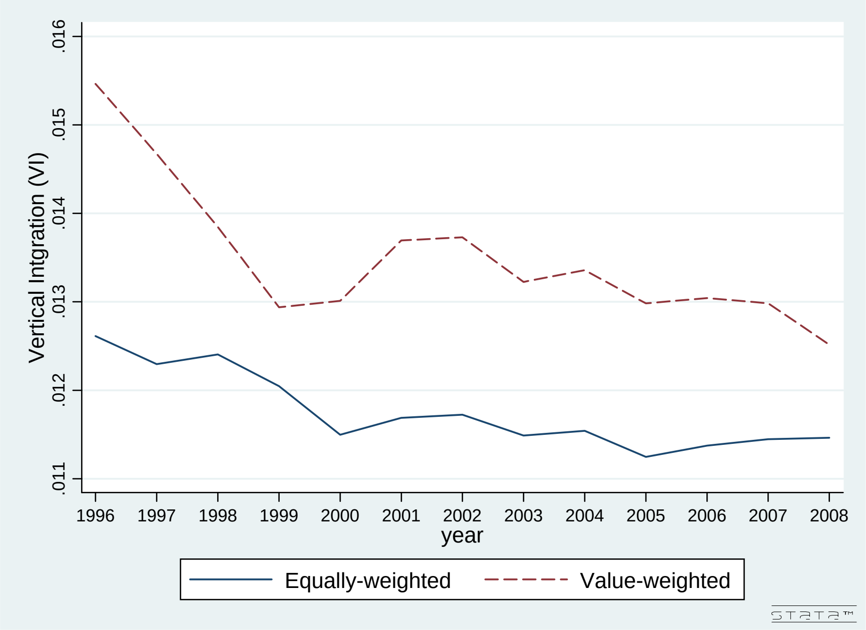


Figure 5: An Example: the Network Equipment Industry. The figure plots the evolution of text-based vertical integration (VI), patenting activity (log(#patents) and #patents/assets) and R&D activity (R&D/sales) for seven representative firms in the network equipment industry: Cisco, Broadcom, Citrix, Juniper, Novell, Sycamore, and Utstarcom.

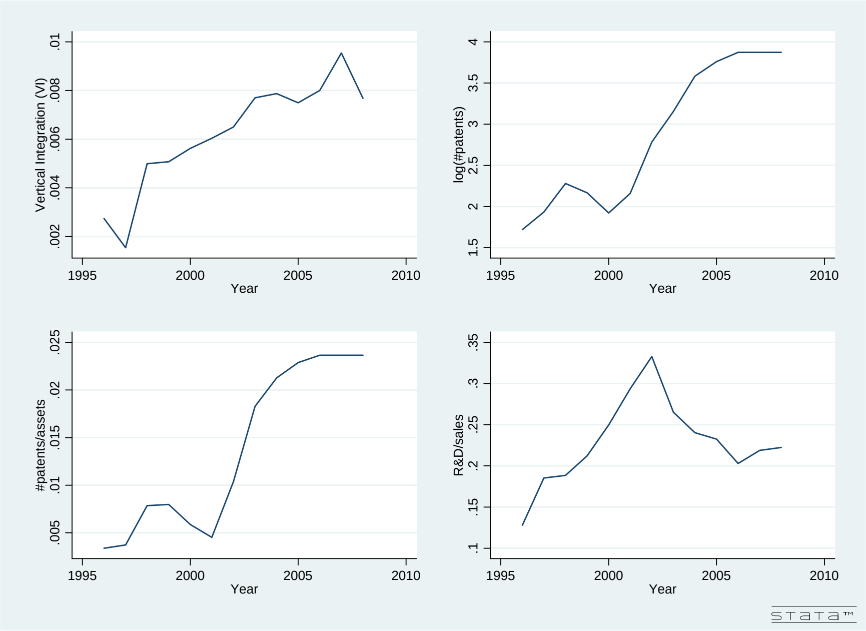
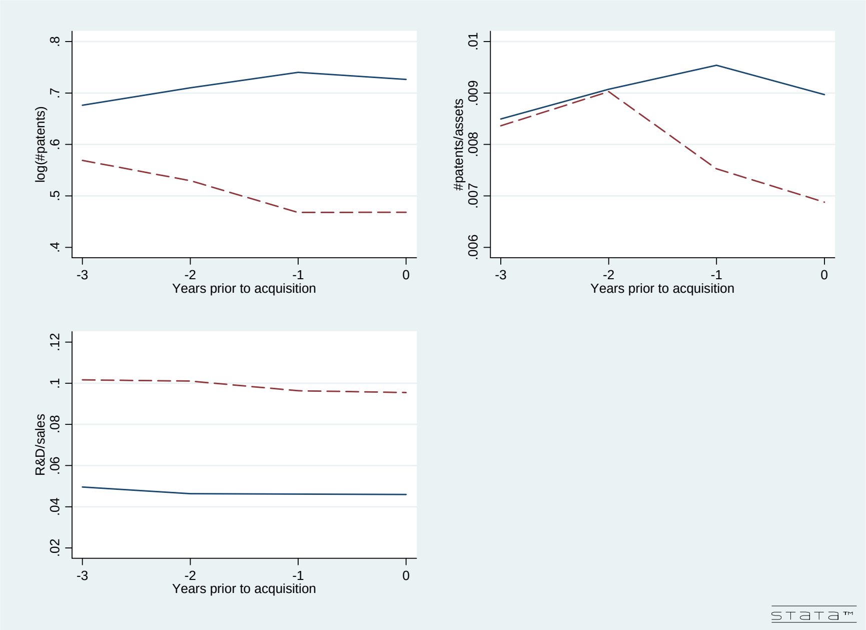


Figure 6: R&D and Patents prior to Acquisitions. The figure shows the average R&D (lower panel) and patenting activity (upper panel) of firms that are targets in vertical and non-vertical acquisitions prior to the acquisition. Solid lines represent vertical transactions identified using the Vertical Text-10% network. Dashed lines represent non-vertical transactions.


log(#patents)
Years prior to acquisition
#patents/assets
Years prior to acquisition
R&D/sales
Years prior to acquisition
stata™