Industry Choice and Product Language

Gerard Hoberg and Gordon Phillips*

October 4, 2012

ABSTRACT

We analyze the words that firms use to describe their products to examine in which industries firms operate. We find strong support for the proposition that when asset complementarities across product markets are strong as measured by industry product language overlaps, firms are more likely to operate across industries. More generally, multiple-industry firms avoid industries with more distinct language boundaries, as measured using language transitivity from rival to rival. Multiple-industry firms are also less likely to operate in industries with high within-industry word similarity and high economies of scale. These findings are consistent with firms choosing organizational form based on product market characteristics and asset complementarities.

*University of Maryland, and University of Southern California and National Bureau of Economic Research, respectively. Hoberg can be reached at ghoberg@rhsmith.umd.edu and Phillips can be reached at gordon.phillips@marshall.usc.edu. We thank Ricardo Alonso, Peter MacKay, Merih Sevilir, Albert Sheen, Denis Sosyura and seminar participants at the Western Finance Association meetings, the University of Chicago, City University of Hong Kong, Duisenberg School of Finance and Tinbergen Institute, Erasmus University, the Rotterdam School of Management, Rutgers University, Stanford University, Tilberg University, University of Illinois and the University of Mannheim for helpful comments. All errors are the authors alone. Copyright ©2011 by Gerard Hoberg and Gordon Phillips. All rights reserved.

Why do firms operate in multiple industries? This question has been viewed as particularly vexing given the discounts in conglomerate multiple-industry firm valuations documented by Lang and Stulz (1994) and Berger and Ofek (1995).[1] The literature has postulated both benefits (Stein (1997)) and costs (Scharfstein and Stein (2000)) of multiple-industry production taking the current industry structure as given. However, this literature has not addressed why multiple-industry firms choose some industry combinations and not others, and the role of product-market fundamentals in this choice.

Fundamentally, this choice is related to the trade-offs between specialization and coordination across industries. Theoretically, Becker and Murphy (1992) model how firms trade-off the costs of coordination across different tasks versus the gains to specialization in determining which tasks and products are grouped together. Our analysis examines firm industry choice directly by considering the extent to which firms in different industries share common product market language in the business descriptions of 10-Ks filed with the SEC. We build on these ideas to test central hypotheses related to theories of asset complementarities and the conditions under which multi-product firm production naturally arises.[2]

Our focus on analyzing the product words that firms use across and within industries is related to the theory of organizational languages in Crémer, Garicano, and Prat (2007). Crémer, Garicano and Prat focus on the key trade-off between facilitating internal communication and encouraging communication with other organizations. They conclude that distinct sets of technical words place a limit on firm scope. A broader scope allows for more synergies to be captured, but this has to be weighed against the cost of less precise communication in each unit. We find direct support for this link: when two product markets have a higher degree of language overlap, firms are more likely to jointly operate in these product markets in order

[1] This average discount has been shown to be related to self-selection by Campa and Kedia (2002), Graham, Lemmon, and Wolf (2002), and Villalonga (2004). Shin and Stulz (1998), Maksimovic and Phillips (2002), and Schoar (2002) examine ex post investment and productivity to understand the potential reasons for this discount. See Maksimovic and Phillips (2007) for a detailed survey.

[2] Panzar and Willig (1977), Teece (1980) and Panzar and Willig (1981)) provide an early analysis of economies of scope and multiple-industry production. For recent work on multi-product firms see Bernard, Redding, and Schott (2010)) and Goldberg, Khandelwal, Pavcnik, and Topalova (2010) for an analysis of changes to multiple-product firms in a developing country context.

to capture asset complementarities. Indeed we find that these firms exhibit greater product description growth, consistent with realized asset complementarities.

In our analysis, we first convert firm product text into a spatial representation of the product market following Hoberg and Phillips (2010a) (HP). In this framework, each *firm* has a product location in this space based on its product text that generates an informative mapping of likely competitors. A central innovation of this current article is to illustrate that *industries* also have locations in the product space, and relatedness analysis at the industry level can be used to examine theories of multi-product production. Our spatial framework thus allows an assessment of how similar industry languages are to each other, and which industries in the product market space are "between" any given pair of industries, providing unique measures of potential asset complementarities.[3]

Apple Inc. is an example of a firm which illustrates our key ideas. Its multiple-function products enable it to compete in multiple markets and offer differentiated products competing with cell phones, computers, and digital music - industries that are highly related today. Apple was successful in its decision to operate jointly in these industries and uses language that is used by single-industry focused firms in each of these markets. It has likely benefited from asset complementarities that are found across previous industry boundaries.

Although our main tests use a framework that relies on the validity of industry classifications, an additional innovation is that we also examine the links between product vocabulary and asset complementarities using a framework that is invariant to industry classifications. In particular, we consider the degree of transitivity in product language overlap among rival firms, and also among rivals of rival firms.[4] This concept of transitivity is related to the concept of industry boundaries or the

[3] "Between" industries are industries that are closer to each industry of a given industry pair than the industry pair is to each other based on product language similarity. We formally define this measure in the next section.

[4] This spatial representation does not impose transitivity on competitor networks. Similar to a Facebook circle of friends, each firm has its own set of competitors and competitors need not be overlapping with other firms competitors even within industry groups. This flexibility allows us to measure the degree to which a product market has strong boundaries (more transitivity indicates strong boundaries). SIC and NAICS industry groupings do not permit such an analysis because they mechanistically impose transitivity: if a firm A and firm B are competitors, and if firm B and firm C are competitors, then firms A and C are also competitors.

potential for product scope, as the ability to develop communication and language that can cross industry boundaries is essential in the realization of scope benefits.

We find that multiple-industry firms are more likely to operate in industry pairs that have high across-industry language overlap (asset complementarities), are vertically related, and when industry pairs have profitable, less contested opportunities between them. Multiple-industry firms are also *less* likely to operate within industries that exhibit high within-industry product similarity and high returns to scale. These findings are consistent with the existence of asset complementarities across industries, and with multiple-industry firms generating further asset complementarities from low cost entry into other profitable industries that lie between two industries in a given pair. These results remain robust after controlling for vertical integration, patent intensity and industry stability.

Our paper makes three main contributions. First, our paper examines in which industry combinations multiple-industry firms choose to operate based on industry product language. We find that asset complementarities, within-industry similarity, and the nature of industries lying between two industries can explain conglomerate industry choice. Second, we show how the fundamental industry characteristics consistent with asset complementarities and economies of scale differ in their effect on organizational form. Multiple-industry firms are more likely to operate across industries that are more likely to have high language overlap and less likely to operate in industries with high economies of scale. Third, we show evidence consistent with increases in product offerings by multiple industry firms when their respective industries exhibit high ex ante measures of language overlap consistent with the existence of asset complementarities. In all, our findings support theoretical links to organizational language, asset complementarities, and economies of scale. Our results also help to explain why so many firms continue to use the conglomerate structure despite potential negative effects on valuation as noted by past studies.

Our evidence is also consistent with the conclusion that multiple-industry production, as identified by the Compustat segment tapes, does not fit the historical view that multiple-industry firms operate unrelated business lines under one corporate headquarters, with diversification being the primary aim. Rather firms choose industry pairs in which to operate based on industry language overlaps and potential asset complementarities. For example, we find that roughly 69% of Compustat multiple-industry pairs are in industries that satisfy one of the two following conditions: (A) the language overlap of the pair is similarly as high as industry pairs in the same SIC-2, or (B) the industry pair is above the 90th percentile of vertical relatedness among all industry pairs. The magnitude of this finding suggests that studies aimed at explaining the behavior of diversified multiple-industry firms need more care in reducing the sample of Compustat multiple-industry firms to the much smaller subsample that plausibly has diversification as a primary motive.

Our paper proceeds as follows. In the next section, we present new measures of industry relatedness based on product language and we develop our key hypotheses. In Section II, we discuss our data, variables, and methods used to examine industry choice. Section III presents the results of our analysis of industry choice. Section IV presents our analysis of competitor firm product-market transitivity based on product language used by firms. Section V presents our analysis of subsequent product growth. Section VI concludes.

I Industry Fundamentals and Firm Organization

We ask whether there are certain fundamental industry characteristics - distinct from vertical relatedness - that make operating in two different industries valuable. The central hypothesis we examine is whether product market language overlap across industries and economies of scale impact which industries firms operate within and what types of firms operate across these industries. The foundation underlying why these factors should matter is that industry product language overlap captures the potential for product market synergies, and low cost entry into "between industries."

Our research foundation is related to the trade-off between specialization and coordination. Becker and Murphy (1992) model how firms trade-off the costs of coordinating workers across different tasks versus the gains to specialization across industries. In their analysis, specialization among complementary tasks links the division of labor to coordination costs, knowledge, and the extent of the market.

Workers invest in specialized knowledge until the costs of coordinating specialized workers outweigh the gains from specialization.[5] Our analysis captures the extent that different industries use different sets of specialized words as is theoretically modeled by Crémer, Garicano, and Prat (2007). A broader scope of language allows for more synergies to be captured, but at the cost of less precise communication within each unit. Our analysis examines the role of asset complementarities, as examined in the area of mergers by (Rhodes-Kropf and Robinson (2008) and Hoberg and Phillips (2010)).

Our focus on the potential for asset complementarities also relates to the proposition from Teece (1980) who writes "if economies of scope are based upon the common and recurrent use of proprietary knowhow or the common and recurrent use of a specialized and indivisible physical asset, then multiproduct enterprise (diversification) is an efficient way of organizing economic activity." Industry economies of scale, as Maksimovic and Phillips (2002)) emphasize, exert the opposite force as economies of scale increase the optimal size of a firm. Higher economies of scale reduce the incentive to produce across industry pairs as the relative advantage of operating within a single industry increases with economies of scale.

We discuss our key hypotheses through the lens of a spatial representation of the product market (see Hoberg and Phillips (2010a) for a discussion of the text-based product market space).[6] In this representation, all firms have a "location" on a high dimensional unit sphere that is determined by the overall vocabulary used in the given firm's 10-K business description.

We extend the previous firm-specific work of Hoberg and Phillips (2010a) by constructing new *industry* based measures of how groups of firms are related to each other. Thus the new measures in this paper capture how industries have a simple but highly informative representation in the product-market language space, which can be used to examine how industries relate to one another. Intuitively, an industry should be viewed as a cluster of firms in the product market space, and hence each

[5] The impact of communication on this trade-off has been theoretically studied by Bolton and Dewatripont (1994) and Alonso, Dessein, and Matouschek (2008).

[6] Note that the product market space is a full representation of the products that firms offer and the extent to which they are simple, and the space should not be interpreted as a geographic space.

industry has both a location, and also a degree to which it is spread-out in the product market space. Industries that are highly spread out have a high degree of within-industry product differentiation, for example, and likely offer industry participants additional protection from rivals.

The new fundamental measures of industries that are constructed in this paper allow us to assess how every pair of industries relates to one another and how products differ within industries. We first measure how close industries are in the product space using the extent of language overlap, *Across Industry Language Similarity (AILS)*, to capture potential asset complementarities. We also measure the extent of transitivity of competitor language within industry groupings, *TransComp*, how heterogeneous firms' products are within-industry, *Within Industry Language Similarity (WILS)*, and the extent to which other industries lie between the given industry pair in the product space, *Between Industries (BI).* We estimate economies of scale within industries, *Economies of Scale (Scale)*, using more traditional industry production functions.

We use these new industry relatedness measures from firm product text to test the following four hypotheses. These hypotheses are illustrated in Figures 1A to 1C, where each circle represents an industry in the product market space, and the size of the circle illustrates the degree to which the given industry is spread out (low within industry similarity).

[Insert Figure 1 Here]

H1: Asset complementarities and Across-Industry Language Similarity: Multiple-industry firms are more likely to produce in two industries that have overlapping product market language and higher potential for cross-industry asset complementarities. In addition, product offerings should expand when industries have higher asset complementarities.

The main idea underlying this hypothesis is that firms with more product language overlap are more likely to have assets that allow employees in each sector to engage in successful multi-product production across industries. Assets or resources that can be used in multiple industries increase the potential for product market synergies and additional product market offerings. H1 implies that the number of multi-product firms producing in a given industry pair should increase with cross-industry similarity. In addition, these firms should have higher ex post growth in product offerings. Figure 1A depicts industry X and Y as having a high degree of cross-industry similarity compared to other industry pairs, and H1 predicts that more multiple-product firms will choose to jointly operate in X and Y relative to other pairwise configurations.

H2: Economies of Scale: Multiple-industry firms are *less* likely to produce in industries that exhibit higher economies of scale.

This hypothesis comes from the relative gains from increasing production within an industry. If there are higher gains to production within an industry, versus in a new industry, a firm will have incentives to use any scarce resource such as managerial talent within an industry as in Maksimovic and Phillips (2002). Firms will not choose to move outside of an industry if the marginal product of within industry production is greater than the marginal product of outside industry production. If the industry has low relatedness with other industries, the firm may not expand at all, choosing to return excess resources to investors. The number of multiple-industry firms operating in a pair should thereby decrease with average pairwise economies of scale.

H3: Within-Industry Similarity: Multiple-industry firms are less likely to produce in industries that have high within-industry similarity, or industries with little potential for product differentiation.

This hypothesis comes from the proposition that industries with high similarity of product language are well defined, and the gains from specialization are large. We consider whether within industry similarity decreases the incentives for firms to operate in a particular industry as firms in industries with higher within-industry similarity are likely to have less unique products, and likely face more significant competition from their rivals due to the absence of product differentiation. If there are additional costs in setting up and operating firms with a multiple industry structure, such costs are likely to outweigh the smaller benefits of operating in industries with high within industry similarity. Figure 1B depicts industry X and Y as having a low degree of within-industry similarity compared to other industries, and hence firms residing in X and Y likely offer unique products, and H3 predicts that more multi-product firms will choose to jointly operate in X and Y relative to other pairwise configurations.

H4: Between-Industries: Multiple-industry firms are more likely to operate in an industry pair when the pair of industries has more high-value, less competitive industries, residing between the given pair.

The idea of this hypothesis is that producing in an industry pair that has high-value, less competitive industries between the industries may allow the multiple industry firm to more easily enter the between industries and produce products in these highly-valued concentrated product markets. Figure 1C depicts industry X and Y as having a third industry I_3 residing between them. If firms in I_3 are highly valued, then H4 predicts that multi-product firms will choose to operate in industries X and Y relative to other pairwise configurations.

II Data and Methodology

In this section we describe our sample of firms, the construction of key text-based variables used to examine where multiple-industry firms produce in the product space, and our identification of single-segment (also called pure-play) conglomerate competitors.

A The COMPUSTAT Industry Sample

We construct our COMPUSTAT sample using the industrial annual files to identify the universe of publicly traded firms, and the COMPUSTAT segment files to identify which firms are multiple-industry producers, and the industry of each segment. We define a conglomerate as a firm having operations in more than one SIC-3 industry in a given year. To identify segments operating under a conglomerate structure, we start with the segment files, which we clean to ensure we are identifying product-based segments instead of geographic segments. We keep conglomerate segments that are identified as business segments or operating segments. We only keep segments which report positive sales. We aggregate segment information into 3 digit SIC codes and only identify firms as multiple-industry firms when they report two or more three digit SIC codes. We identify 22,252 unique multiple-industry firm years from 1996 to 2008 (we limit our sample to these years due to required coverage of text-based variables), which have 62,058 unique conglomerate-segment-years. We also identify 56,491 unique pure play firm-years (firms with a single segment structure).

When we examine how multiple-industry firms change from year to year, we further require that a multi-industry structure exists in the previous year. This requirement reduces our sample to 18,589 unique conglomerate years having 53,126 segment-years. Because we use pure play firms to assess industry characteristics that might be relevant to the formation of multiple-industry firms, we also discard conglomerate observations if they have at least one segment operating in an industry for which there are no pure play benchmarks in our sample. We are left with 15,373 unique multiple-industry firm-years with 40,769 unique segment multiple-industry firm-years. This final sample covers 2,552 unique three digit SIC industry-years. As there are 13 years in our sample, this is roughly 196 industries per year.

We also consider a separate database of pairwise permutations of the SIC-3 industries in each year. We use this database to assess which industry pairs are most likely to be populated by multiple-industry firms that operate in the given pair of industries. This industry-pair-year database has 312,240 total industry pair x year observations (roughly 24,018 industry pair permutations per year).

B The Sample of 10-Ks

The methodology we use to extract 10-K text follows Hoberg and Phillips (2010a). The first step is to use web crawling and text parsing algorithms to construct a database of business descriptions from 10-K annual filings on the SEC Edgar website from 1996 to 2008. We search the Edgar database for filings that appear as "10-K," "10-K405," "10-KSB," or "10-KSB40." The business descriptions appear as Item 1 or Item 1A in most 10-Ks. The document is then processed using APL for text information and a company identifier, CIK.[7] Business descriptions are legally required to be accurate, as Item 101 of Regulation S-K requires firms to describe the significant products they offer, and these descriptions must be updated and representative of the current fiscal year of the 10-K.

C Word Vectors and Cosine Similarity

After we have the database of business descriptions we form word vectors for each firm based on the text in product descriptions of each firm. To construct each firm's word vector, we first omit common words that are used by more than 25% of all firms. Following Hoberg and Phillips (2010a), we further restrict our universe in each year to words that are either nouns or proper nouns.[8] Let M_t denote the number of such words. For a firm i in year t, we define its word vector $W_{i,t}$ as a binary M_t-vector, having the value one for a given element when firm i uses the given word in its year t 10-K business description. We then normalize each firm's word vector to unit length, resulting in the normalized word vector $N_{i,t}$.

Importantly, each firm is represented by a unique vector of length one in an M_t-dimensional space. Therefore, all firms reside on a M_t-dimensional unit sphere, and each firm has a known location. This spatial representation of the product space allows us to construct variables that more richly measure industry topography, for example, to identify other industries that lie between a given pair of industries.

The cosine similarity for any two word vectors $N_{i,t}$ and $N_{j,t}$ is their dot product $\langle N_{i,t} \cdot N_{j,t} \rangle$. Cosine similarities are bounded in the interval [0,+1] when both vectors are normalized to have unit length, and when they do not have negative elements, as will be the case for the quantities we consider here. If two firms have similar products, their dot product will tend towards 1.0 while dissimilarity moves the cosine similarity toward zero. We use the "cosine similarity" method because it is widely

[7] We thank the Wharton Research Data Service (WRDS) for providing us with an expanded historical mapping of SEC CIK to COMPUSTAT gvkey, as the base CIK variable in COMPUSTAT only contains the most recent link.

[8] We identify nouns using Webster.com as words that can be used in speech as a noun. We identify proper nouns as words that appear with the first letter capitalized at least 90% of the time in the corpus of all 10-K product descriptions. Previous results available from the authors did not impose this restriction to nouns. These results were qualitatively similar.

used in studies of information processing (see Sebastiani (2002) for a summary of methods). It measures the cosine of the angle between two word vectors on a unit sphere.

D Firm Restructuring over Time

We examine whether our spatial industry variables can explain how multiple-industry restructure over time, and we classify restructuring in three different ways. Because we consider the role of industry topography, the unit of observation for these variables is a pair of segments operating within a conglomerate. We define "New Segment Pairs" as when a given pair observed in a conglomerate in year t did not exist in the conglomerate in the previous year $t-1$. We then define "New Segment Pairs Likely Obtained through Growth" as pairs that did not exist in the conglomerate's structure in the previous year, and the conglomerate had fewer segments in year $t-1$ relative to year t. Finally, we define "New Segment Pairs Linked to SDC Acquisitions" as segment pairs that did not exist in the conglomerate's structure in the previous year, and the conglomerate was the target of an acquisition of at least ten percent of its assets between year $t-1$ and year t.

E Industry Variables

Our primary four industry variables are Across Industry Similarity (Asset Complementarities), Economies of Scale, Within Industry Similarity, and Fraction of Industries that are Between Industries. All but the Economies of Scale variable are direct functions of across- and within-industry language similarity calculations based on the firm-level relationships in Hoberg and Phillips (2010a). In this section, we discuss these variables and the additional industry variables we consider both as control variables and as variables of individual interest.

Because we seek to examine the industry pairs in which multiple-industry firms produce, to avoid any mechanistic relationships, we focus only on single-segment firms to calculate these industry relatedness variables. We then use the Compustat segment tapes to examine how observed conglomerate industry configurations relate to these text-based industry attributes computed from single-segment firms.

Because conglomerate segments are reported using SIC codes, our initial analysis relates to industry configurations and their incidence based on three-digit SIC code industry definitions. In later analysis, we relax this initial reliance on SIC-3 industry definitions and examine industry groupings using the fixed industry classifications (FIC) from Hoberg and Phillips (2010a), where firms are identified as competitors using text-based methods.

E.1 Text-based Industry Variables

Across Industry Language Similarity (AILS) of product market: This measure is based on industry product language overlap and captures the extent to which product descriptions of firms in two different industries use overlapping language. The AILS measure is meant to capture the similarities between the products that two industries produce and thus the potential for asset complementarities. Specifically, across industry similarity is the average textual cosine similarity of all pairwise permutations of the N_i and N_j firms in the two industries i and j, where textual similarity is based on word vectors from firm business descriptions (see Section II.C for a discussion of the cosine similarity method). Simply put, it captures the average proportion of product words that two randomly drawn firms from industry i and j will have in common.

Economies of Scale (Scale): This measure captures the gains to scale within an industry. This measure is captured by estimating a traditional Cobb-Douglas production function.[9] As with our measure of AILS, we estimate this measure for both traditional SIC industry groupings and the new text-based fixed industry classifications (FIC) of Hoberg and Phillips (2010a). We estimate the production function using firm-level data from Compustat. We estimate the production function using 10 years of lagged data for each firm in a given industry, with sales as the dependent variable. We include the following right-hand-side variables: net property plant and equipment for capital, the number of employees, the cost of goods sold and also firm

[9] We also estimate the industry economies of scale using a translog production function for robustness, and results are similar.

age. All variables are in natural logs, and variables except for age and the number of employees are deflated to 1987 real dollars using the wholesale price index. An industry's economies of scale variable is the sum of the coefficients on net property plant and equipment and the cost of goods sold.

Within Industry Language Similarity (WILS): This measure captures the product differentiation within an industry i. Within industry language similarity is the average cosine similarity of the business descriptions for all pairwise word permutations of the N_i firms in industry i (i.e., it is the degree of language overlap within an industry group).

Between Industries (BI): We use the across industry similarity measure (described above) to assess which other industries lie between any given industry pair. Specifically, a third industry is between two industries in a given industry pair if the third industry is closer in textual distance to each industry in the pair than the two industries in the pair are to each other.

The AILS measure discussed above is instrumental in computing the fraction of industries between a given pair. More formally, where $AILS_{i,j}$ denotes the across industry product language similarity of industries i and j, we define a third industry k as being *between* industries i and j if the following relationship holds.

$$AILS_{k,i} \leq AILS_{i,j} \qquad \text{AND} \qquad AILS_{k,j} \leq AILS_{i,j} \tag{1}$$

The fraction of industries between a given pair of industries i and j is therefore the number of industries k (excluding i and j) satisfying this condition divided by the total number of industries in the database in the given year (excluding i and j).

Transitivity of Competitors (TransComp): Transcomp is a measure of how weak a given product market's language boundaries are, and it is defined at the firm level using the basic TNIC correspondence from Hoberg and Phillips (2010a).[10] Transcomp is the fraction of a given focal firm's single segment rivals that are also TNIC rivals to the focal firm itself. Because TNIC links are direct estimates of language overlap, Transcomp measures the degree to which language overlap is transitive in a given product market. This variable by design lies in the interval [0,1]. Transcomp is a

[10] TNIC competitors are available for download at http://www.rhsmith.umd.edu/industrydata/industryclass.htm.

particularly stark measure of the potential for asset complementarities in a localized region of the product market space because it does not rely on the quality of the Compustat segment tapes and their potentially questionable SIC code designations. Markets with weak language boundaries, for example, offer more scope for multiple product firms to benefit from asset complementarities in neighboring markets.

E.2 Non-text-based industry control variables

Like Across Industry Language Similarity and the fraction of industries between a given pair, our first set of three additional control variables are a property of a pair of industries. These include a key control for industry-pair relevance, a measure of vertical relatedness, and a dummy identifying which industries are in the same two-digit SIC code. Because we aim to examine conglomerate incidence rates across industry pairs, controlling for industry pair relevance is important. For example, if multiple-industry firms were formed by randomly choosing among available pure play firms in the economy, then the incidence of conglomerate operating pairs would be related to the product of the fraction of firms residing in industries i and j. Therefore we define the Pair Likelihood if Random variable as the product (F_ixF_j), where F_i is the number of pure play firms in industry i divided by the number of pure play firms in the economy in the given year.

We consider the Input/Output tables to assess the degree to which a pair of industries is vertically related. The inclusion of this variable is motivated by studies examining vertically related industries and corporate policy and structure including Fan and Goyal (2006), Kedia, Ravid, and Pons (2008), and Ahern and Harford (2011). We consider the methodology described in Fan and Goyal (2006) to identify vertically related industries. Based on three-digit SIC industries, we use the “Use Table” of Benchmark Input-Output Accounts of the US Economy to compute, for each firm pairing, the fraction of inputs that flow between each pair.

Like within industry similarity and our economies of scale variable, we also consider two additional control variables that are a property of a single industry: patent applications and industry instability. We compute patent applications at the industry level as the fraction of total patents applied for by firms in the given industry (as a fraction of all patents applied for in the given year) scaled by the total assets of firms in the given industry in the given year. We multiply this quantity by ten thousand for convenience. We compute industry instability as the absolute value of the natural logarithm of the number of firms in the industry in year t divided by the number of firms in the same industry in year $t-1$. Industries with higher instability are experiencing changes in the industry's membership over time.

F Summary Statistics

Table I displays summary statistics for our conglomerate and pure play firms, and industry pair databases. Panel A shows that multiple-industry firms are generally larger than the pure play firms in terms of total value of the firm.

Panel B of the table compares randomly drawn pairs of SIC-3 industries to the SIC-3 industries comprising a conglomerate configuration. The panel shows that a randomly drawn pair of three digit SIC industries has 0.147 multiple-industry firms having segments operating in both industries of the given pair. Hence, the majority of randomly chosen industry pairs do not have multiple-industry firms operating in the pair. The average across-industry language similarity (AILS) of *random* pairs is 0.017, which closely matches the average firm similarity reported in Hoberg and Phillips (2010a). This quantity is nearly double for actual multiple-industry firms at 0.032, indicating that multiple-industry firms are far less diversified than previously thought. This conclusion is reinforced by comparing the fraction of all other industries lying between the given pair, which is 32.5% for random pairs, and just 9.7% for actual multiple-industry firms. Conglomerate industry pairs are in regions of the product space that are substantially closer together than randomly chosen industries. The average within-industry similarity, intuitively, is much higher at 0.086. This quantity is somewhat lower at 0.073 for actual multiple-industry firms.

[Insert Table I Here]

Table II displays the bivariate Pearson correlation coefficients for our key industry pair variables. The key variable we examine in the next section is the number of multiple-industry firms operating in a given pair. The first column of this table shows that this variable is positively related to across industry language similarity, and negatively related to within-industry similarity and the fraction of industries between a given pair. Although these univariate results hold for across- and within-industry similarity, multivariate results vary for the fraction of industries between variable (discussed later). This is related to the relatively high observed pairwise correlation of -69.1% between this variable and across industry similarity. Intuitively, industries that are further away likely have more industries residing between them. Our later results will show that multiple-industry firms are more likely to operate in industry pairs that have concentrated or high value industries residing in the product space between the given pair, but not when competitive or low value industries do.

Aside from the modest correlation between the between variable and the across industry language similarity variable, Table II shows that the other variables we consider have relatively low correlations. This fact, along with our very large database of 312,240 observations, indicates that multicollinearity is unlikely to be a concern in our analysis.

[**Insert Table II Here**]

Table III displays the mean values of our three key text variables for various conglomerate industry pairings. One observation is an industry pair permutation of an actual conglomerate. In Panel A, we find that multiple-industry firms populate industries with high across-industry similarity of 0.0304, which is 79% higher than the 0.017 of randomly chosen industry pairs. Hence, multiple industry firms are more likely to operate in industry pairs with higher levels of language overlap, consistent with their capturing asset complementarities. Multiple-industry firms also tend to populate industries with lower than average within-industry similarity, and industries having a lower than average number of other industries between them.

[**Insert Table III Here**]

In Panel B, we report results for smaller multiple-industry firms (two or three segments) compared to those of larger multiple-industry firms. The table suggests that larger multiple-industry firms tend to produce across a wider area of the product market space, as they have lower across industry similarity. They also tend to produce in industries with more industries between them, and industries that have higher within-industry similarity. In Panel C of Table III, we observe that most multiple-industry firms (30,525) are stable from one year to the next, although 3,259 of them reduce in size by one segment, and 600 multiple-industry firms reduce in size by two or more segments. Analogously, 4,741 firms increase in size by one segment, and 1,644 firms increase in size by two segments.

In Panel D, we observe that vertically related multiple-industry firms have average across industry similarities that are close to the average for all conglomerate pairs. However the panel also shows that across industry similarities are higher for industries having the same two digit SIC code pointing to relatedness of conglomerate chosen industry pairs. Both vertical industries and those in the same two-digit SIC code also have fewer than the average fraction of industries between them.

[**Insert Figure 2 Here**]

Figure 2 displays the large economic magnitude of the link between across-industry product language similarity and conglomerate firm industry choice. In particular, the solid line displays the distribution of across-industry product language similarity scores for randomly drawn industry pairs, and the dashed line displays this distribution for observed conglomerate firm industry pairs. The figure shows that the dashed line has a distribution that is (A) strongly shifted to the right relative to the solid line and (B) has a very large right tail as evidenced by the higher level of density on the right side of the figure and the large amount of mass to the right of 0.05. To put this shift in perspective, the median level of across industry similarity scores for conglomerate firm industry pairs resides at the 85.5th percentile among randomly drawn pairs.

III Firm Industry Choice

In this section we examine whether we can predict whether firms produce in particular industry pairs. We test whether potential asset complementarities measured through across-industry product language similarity, economies of scale, the fraction of industries between a particular industry pair, and the within-industry similarity matter for the number of multiple-industry firms producing in a particular industry pair. We also examine the impact of vertical relatedness using data from the input-output matrix.

Table IV presents OLS regressions where each observation is a pair of three digit SIC industries in a year derived from the set of all pairings of observed SIC-3 industries in the given year in the COMPUSTAT segment tapes. The dependent variable is the number of multiple-industry firms operating in the given industry pair. Put differently, it is the number of multiple-industry firms having segments in both industries associated with the given pair. Panel A displays results based on the entire sample of industry pairs. Panel B displays results for various subsamples that divide the overall sample based on the competitiveness or the valuations of industries lying between the industry pair.

[**Insert Table IV Here**]

Panel A shows that higher across-industry language similarity is associated with an increase in the number of multiple-industry firms producing in a particular industry, while average within-industry similarity decreases the multiple-industry firms producing in a particular industry. The table shows that multiple-industry firms tend to operate in more differentiated product markets, ie, those with low within industry similarity. Panel A also shows that the fraction of industries between a given pair also matters, and its sign also depends on the characteristics of the specific industries that lie between the pair.

Panels B and C show that when high value and concentrated industries are between, multiple-industry firms operate in the pair more often. The opposite is true for competitive low value industries. This result shows how industry boundaries can be crossed and redrawn presumably by using product market synergies to lower the cost of entry into previously concentrated product markets.

Table V examines how industry characteristics influence which industry pairs are added to multiple-industry firms in a given year. We consider raw segment additions for growing or stable multiple-industry firms, and we also consider the SDC mergers and acquisitions database. This allows us to separately consider segments likely added through growth, or those potentially acquired in a transaction. One observation is one pair of segments in an existing conglomerate in year t. We require the multiple-industry firm itself to exist in year t and year $t + 1$.

The dependent variable varies by Panel in Table V. The dependent variable in Panel A is the number of newly added conglomerate operating pairs. It is defined as the number of multiple-industry firms having new segments in both industries associated with a given pair in a given year (where the conglomerate did not have this segment in the previous year). In Panel B, we restrict attention to new segments in multiple-industry firms that previously had fewer segments in the previous year. Intuitively, these new segments were likely added through acquisition or organic investment. In Panel C, we restrict attention to new segments in firms that were the acquirer in an acquisition in the SDC database for a transaction amounting to at least ten percent of the firm's assets. The independent variables include various product market variables characterizing the industry pair.

[**Insert Table V Here**]

The results in Panel A of Table V show that segment pairs are likely to be added if across industry product language similarity and potential asset complementarities are high, and less likely when economies of scale are high. The panel also shows that the coefficient on the across-industry product similarity variable is highest when the industries between two industry pairs are highly concentrated and highly valued (and the lowest coefficient when the converse is true). This result is consistent with multiple-industry firms using complementary industry assets to extract product market synergies that allow them to lower the cost of entry into highly concentrated industries. We also see that multiple-industry firms are more likely to add segments when the fraction of industries between the conglomerate pair is high and the average within-industry similarity is low. These findings are present especially in concentrated and highly-valued industry pairs.

The results in Panels B and C further show that conglomerate segments are more likely to be added through growth or acquisition when concentrated and highly valued industries lie between the segment pairs. In particular, multiple-industry firms add such segments when the resulting industry pairs have high potential asset complementarities, low within-industry similarity, and a high fraction of industries lie between the industry pair. The results are broadly consistent with multiple-industry firms choosing to expand into industries with the potential for new differentiated products and related-industry synergy gains. These results also support the the theory of organizational languages in Crémer, Garicano, and Prat (2007), as multi-product firms appear to seek more asset complementarities across product markets when the product markets have more language overlap.

A Text based Industry Classifications

In this section, we replicate the multiple-industry firm choice analysis in Table IV using text-based industry classifications from Hoberg and Phillips (2010a). In particular, we focus on the Fixed Industry Classification with 300 industries (FIC-300), which is a set of 10-K based industries chosen to be roughly as granular as SIC-3. In order to implement this calculation, we first need to reassign each firm to a set of FIC-300 segments as a substitute for the SIC-3 segments indicated by Compsutat. This is achieved using the textual decomposition of each conglomerate firm into its respective segments from Hoberg and Phillips (2012). This decomposition generates a full set of single segment peers for each segment of each conglomerate, with associated weights that sum to one, and that best replicates the product offerings of the given conglomerate. For a conglomerate with N segments, we assign it to the N FIC-300 industries having the highest total weight in the Hoberg and Phillips (2012) decomposition. This methodology is parsimonious, and fully accounts for the documented improvements in conglomerate benchmarking illustrated in the paper. We refer readers to Hoberg and Phillips (2012) for details regarding the weighted conglomerate decomposition.

The main impetus for this analysis is to establish robustness using an alternative classification, but also to establish robustness using an industry classification based on text-based industry relatedness variables. We do not include this analysis as our primary analysis due to the fact that many variables are not as readily available using text-based classifications in this system, as text based classifications only become available starting in 1996. As a result of these limitations, our sample is restricted to 145,058 industry-pair-years rather than the 312,240 available in Table IV. Furthermore, we do not have measures of vertical relatedness in this setting, and variables requiring multiple years to compute such as our Economies of Scale variable are especially restrictive in limiting the size of the sample available using FIC-300 industries.

[**Insert Table VI Here**]

Table VI displays the results of this test using FIC-300 industries. The table shows that most of our key findings are highly robust to using FIC-300 instead of SIC-3 despite the smaller sample size. For example, multiple-industry firms are far more likely to operate in industry pairs with a high potential for asset complementarities (across-industry product language similarity), with a larger fraction of between industries, with more differentiated products (lower within industry similarity), and in industries that are more stable. However, two results differ from those in Table IV. First, multiple-industry firms are less likely to operate in high-patenting industries using FIC-300 industries, but are more likely to operate in high-patenting industries using SIC-3 industries. We find this result to be interesting especially given that FIC-300 industries are fully updated each year, whereas SIC-3 industries change little. Patenting is deeply related to innovation and product change, and hence we believe the FIC-300 results are likely more indicative of true conglomerate choice regarding patents. On the other hand, industry characteristics relating to asset complementarities and within industry similarity likely change less over time, and hence we see similar results for SIC-3 and FIC-300.

The second result that differs in Table VI and Table IV relates to our Economies of Scale variable, which is negative using SIC-3 and insignificant using FIC-300. We believe the reason for this difference is likely technical. The Economies of Scale variable requires a longer time series to properly estimate, and inadequate long-term FIC-300 data exists to make this possible.

IV Product Market Boundaries

In this section, we examine the robustness of our findings relating to across industry product language similarity and asset complementarities using a framework that does not rely on the Compustat segment database. This test is important for two reasons. First, Hoberg and Phillips (2010a) show that SIC-based classifications are inadequate to fully capture information about industry memberships.[11] Villalonga (2004) has also questioned the reliability of the Compustat segment database showing that the Compustat segment database does not capture multiple industry production. Second, we view the results regarding asset complementarities to be the primary contribution of the current article. Hence, re-examining the same predictions through a more refined framework can offer a highly discriminating test of robustness regarding our primary contribution.

Our alternative measure of the potential for asset complementarities is the degree of product market language overlap transitivity. This is a measure of how strong a given product market's language boundaries are. Markets with weak boundaries, for example, are likely susceptible to entry by firms in neighboring markets at relatively low cost due to asset complementarities. We define product market transitivity at the firm level using the basic TNIC correspondence from Hoberg and Phillips (2010a) and we compute the fraction of a given firm's rivals that also consider the given firm itself to be a rival. It is important to note that the TNIC industry classification relaxes the membership-transitivity restriction of SIC or NAICS based classifications, and firms that are rivals to firm B, which is a rival to firm A, might not consider firm A to be rivals. The product market language-transitivity variable by design thus lies in the interval [0,1], and Figure 3 displays the distribution of this variable for firms with more than one segment in the Compustat database, and separately for firms that have just one segment. It is also important to note that although we compute language transitivity for both multiple-industry firms and single-segment firms, we only use single segment firms as reference peers for the purposes of the calculation itself to maintain consistency with the rest of our study, and to ensure

[11] It is very telling that Apple is classified as a single-segment firm using Compustat segment database until 2007, five years after it introduced the iPod.

no mechanistic differences affect transitivity scores for multiple-industry firms.

Figure 3 shows a high degree of variability in the transitivity of product markets, and also that multiple-industry firms are fundamentally different from single segment firms regarding the degree of transitivity faced in their respective markets. In particular, single segment firms lie within a sharply bimodal distribution, and multiple-industry firms lie within a sharply unimodal distribution and generally have much lower transitivity when compared to single segment firms. Economically, we interpret this in terms of product market boundaries. We conclude that multiple-industry firms almost universally operate in product markets with weak boundaries, whereas single segment firms operate both in markets with weak boundaries, and in markets with stronger boundaries.

[**Insert Figure 3 Here**]

To build more intuition regarding product market transitivity, we next report the average product market transitivity and statistics regarding operating segments for firms in each Fama-French 48 Industry. A key focus is to examine how transitivity varies across intuitively identified industry groups. Column 3 reports the number of segments as the total count of segment year observations for our entire sample from 1996 to 2008. The final column reports the fraction of these operating segments that operate under a conglomerate structure (a firm with more than one segment). The figures in the final column may appear to be high, and so we remind readers that the fraction is based on segment counts and not firm counts, and the elevated figures reflect the fact that multiple-industry firms have multiple segments whereas single segment firms only have one.

[**Insert Table VII Here**]

The summary statistics in Table VII show that the strength of industry boundaries, measured using the degree of transitivity, varies widely across industry groupings. Some industries including beer exhibit very high industry transitivity as firms universally view each other as competitors. Other industries including construction and insurance have lower transitivity, indicating that competitors view different sub-groups of firms as primary rivals due to broader product offerings in these areas. These results would suggest that asset complementarities are likely more relevant in construction and insurance than in the beer industry. Analogously, asset complementarities are also likely more present in business services and retail compared to textiles. This is consistent with the emergence of broad retail empires such as Amazon.com, which likely benefit from asset complementarities. Indeed we confirm that Amazon.com has weak product market boundaries with a transitivity score that averages less than 20%. Relatedly, Apple also has a transitivity score close to 20% as well, providing evidence supportive of our earlier conjecture that Apple is a firm that likely benefits from strong asset complementarities.

Table VII also indicates the link between product market boundaries and reported multiple-industry firm segments from Compustat. For example, the table shows that segments in the construction and insurance industries are both more likely to be conglomerate segments when compared to most other industries. We next show this more formally, and is consistent with the predictions of asset complementarities as measured through weak industry boundaries. In contrast, although business services has a low degree of transitivity, its observed segments are not more likely than average to lie within multi-industry firms. This latter finding may appear puzzling at first, but it echoes the findings of other studies in this area including Hoberg and Phillips (2010a), which document that SIC codes are particularly weak and under-refined in the business services market. For example, Apple (prior to 2006) despite its iTunes digital music service and iPod was classified as a single segment firm just in the computer industry according to the Compustat segment tapes. This may seem surprising given that Apple's iPod line was announced by Apple on October 23, 2001, and released on November 10, 2001.

[**Insert Table VIII Here**]

Table VIII formally examines the association between industry transitivity and organizational form using three panels. Panel A examines highly transitive vs. low transitivity industries and shows that a higher fraction of competitors are multiple-industry firms in industries with low transitivity (60% versus 45%), and that this difference is large. Panel B examines this relationship across subsamples based on firm size and firm age, two variables that are also strongly linked to whether or not a firm is a conglomerate. Panel B shows that smaller, younger firms in highly transitive industries are especially less likely to be multiple-industry firms (just 24%). In contrast, segments are likely to be in multiple-industry firms if they are larger, older and in weakly transitive industries (78%). Panel B also shows that all three variables (size, age, and transitivity) are distinct and that each is separately economically important in explaining whether a firm is likely to be a conglomerate.

Panel C displays the results of logistic regressions, where one observation is one firm in one year, and the dependent variable is a dummy equal to one if the firm is a multiple-industry firm (defined as a firm having more than one segment in the Compustat tapes), and zero for a single segment firm. The independent variables include the degree to which the given firm is in transitive product market, and controls for firm age, size and profitability. The results show that multiple-industry firms are more likely to be in industries with weak boundaries (lower transitivity). Conglomerate multiple-industry firms are also more likely to be old, larger firms. Our finding that multiple-industry firms are producing in product markets with weaker product market boundaries is consistent with these firms choosing to operate in markets where asset complementarities are likely, and are also consistent with the theory of organizational languages in Crémer, Garicano, and Prat (2007).

We now examine whether these results hold in differences, and whether ex-ante changes in industry transitivity are linked to ex-post changes in conglomerate organization. In particular, we examine whether multiple-industry firms add or drop segments following changes in transitivity.

[Insert Table IX Here]

Table IX examines the logarithmic growth in the number of segments of the given conglomerate from year t to year t+1 as the dependent variable. All independent variables are measures of change in the given quantity over the three prior years from year t-3 to year t. In addition to three year changes in product market transitivity, we consider three year changes in R&D activity, CAPX activity, profitability and firm size.

Table IX shows that multiple-industry firms increase (decrease) the number of reported segments when transitivity decreases (increases). The results are consistent with multiple-industry firms responding to any weakening of product market boundaries by adding segments. Because weaker product market boundaries indicate that firms can more easily expand their scope, this finding is also consistent with firms reacting to changes in the potential for asset complementarities by changing their overall operating configuration.

This section offers a robustness check that is independent of the potentially unreliable SIC code links provided in the Compustat segment tapes. These results are instead based on a more direct measure of the potential for asset complementarities, and moreover, these results are based on a scope measure that is updated in each year (they are constructed from yearly firm 10-Ks). A stark test of this nature is prohibitively difficult using existing SIC or NAICS-based data because little cross-industry relatedness data is available, and moreover, these classifications are generally updated little over time.

V Growth of Product Offerings

Given our findings in earlier sections, we examine a finer prediction of H1 (asset complementarities) in this section. In particular, if multiple-industry firms indeed operate in some markets in order to act on potential asset complementarities, we should observe a positive link between potential asset complementarities and increases in firm product offerings over time. We thus examine if multiple-industry firms operating in industry combinations with greater across-industry product language similarity increase their product offerings over time. We consider the size of a firm's 10-K business description as a measure of the depth of a firm's product offerings in a given year. Because 10-ks are filed annually, we can assess the degree to which a firm increases its product offerings in a given year by examining the extent to which its business description grows from one year to the next. We can then examine whether this growth is related to ex-ante measures of asset complementarities.

[Insert Table X Here]

Table X presents the results of this test. The dependent variable is the firm's product description growth, defined as the natural logarithm of the number of words in the firm's business description in year t + 1 divided by the number of words in the firm's business description in year t. We consider the same explanatory variables as in Table IV, although we focus our attention on the across industry language similarity variable. Panel A displays results based on raw firm-level product description growth. Panel B displays results based on TNIC industry adjusted product description growth.

The results show that conglomerate product description growth is highly related to ex ante measures of asset complementarities as measured by across-industry product language similarity. The findings are consistent Hypothesis 1, which predicts that industry asset complementarities provide opportunities for multiple-industry firms to increase their product market offerings. The results are also consistent with the fundamental characteristics of asset complementarities as outlined by Teece (1980) and Panzar and Willig (1981).

VI Conclusions

We examine product language overlaps across industries using text-based computational analysis of firm business descriptions from 10-Ks filed with the SEC. We examine key hypotheses predicting in which industries multiple-industry firms are most likely to operate. We find that multiple-industry firms are more likely to operate in industry pairs with higher language overlap, in industry pairs that have highly valued product markets "between" them, and in industries with lower within-industry product similarity. These findings are consistent with firms using the multiple-industry structure to take advantage of asset complementarities and product synergies across markets.

We also examine firm entry into new industries through changes in the number of reported segments of multiple-industry firms. We find that multiple-industry firms are more likely to enter industries with high across-industry product language similarity, and are less likely to enter industries with high within-industry similarity and high economies of scale. These results are consistent with fundamental industry characteristics such as asset complementarities and economies of scale differing in their effect on organizational form.

We construct a more general test measuring the extent to which a product market has strong language boundaries. This test is based on the degree of transitivity of rival language overlaps. This approach relaxes the need to rely on the quality of the Compustat segment tapes, and the need to rely on a particular industry classification. Low levels of language overlap transitivity indicate strong industry language boundaries, and is consistent with a lower potential for asset complementarities and a reduced potential for scope. We find that multiple-industry firms are much more likely to operate in product markets with weak language boundaries, and that these results are economically large in magnitude. These results also affirm that our findings are robust when we use a framework that avoids potentially unreliable industry designations provided in the Compustat segment tapes.

Lastly, we present evidence consistent with increases in product offerings by multiple-industry firms when industries exhibit high ex ante measures of across-industry product language overlap, consistent with potential asset complementarities. In all, our findings support theoretical links to theories of organizational language and asset complementarities, economies of scale, and product market synergies. Our results also help to explain why so many firms continue to use the multiple-industry structure despite potential negative effects on valuation noted by past studies.

We conclude that multiple-industry firms choose which industries they operate within based on industry fundamentals. Our evidence is also consistent with the conclusion that multiple-industry production, as identified by the Compustat segment tapes, does not fit the historical view that multiple-industry conglomerate firms operate unrelated business lines under one corporate headquarters, with diversification being the primary aim.

References

Ahern, Kenneth, and Jarrad Harford, 2011, The importance of industry links in merger waves, University of Michigan and University of Washington Working Paper.

Alonso, Ricardo, Wouter Dessein, and Niko Matouschek, 2008, When does coordination require centralization?, *American Economic Review* 98, 145–79.

Becker, Gary S., and Kevin M. Murphy, 1992, The division of labor, coordination costs, and knowledge, *Quarterly Journal of Economics* 108, 1137–60.

Berger, Phillip, and Eli Ofek, 1995, Diversification's effect on firm value, *Journal of Financial Economics* 37, 39–65.

Bernard, Andrew, Stephen Redding, and Peter Schott, 2010, Multiple-product firms and product switching, *American Economic Review* 100, 70–97.

Bolton, Patrick, and Mathias Dewatripont, 1994, The firm as a communication network, *Quarterly Journal of Economics* 109, 809–39.

Campa, Jose, and Simi Kedia, 2002, Explaining the diversification discount, *Journal of Finance* 57, 1731–1762.

Crémer, Jacques, Luis Garicano, and Andrea Prat, 2007, Language and the theory of the firm, *Quarterly Journal of Economics* 122, 373–407.

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

Goldberg, P., N. Khandelwal, N. Pavcnik, and P. Topalova, 2010, Multi-product firms and product turnover in the developing world: Evidence from india, *Review of Economics and Statistics* 92, 1042–1049.

Graham, John, Michael Lemmon, and Jack Wolf, 2002, Does corporate diversification destroy value?, *Journal of Finance* 57, 695–720.

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

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

——— , 2012, The stock market, product uniqueness, and comovement of peer firms, Working Paper, University of Maryland and Southern California.

Kedia, Simi, Abraham Ravid, and Vicente Pons, 2008, Vertical mergers and the market valuation of the benefits of vertical integration, Rutgers Business School Working Paper.

Lang, Larry, and Rene Stulz, 1994, Tobin's q, corporate diversification, and firm performance, *Journal of Political Economy* 102, 1248–1280.

Maksimovic, Vojislav, and Gordon Phillips, 2002, Do conglomerate firms allocate resources inefficiently across industries? theory and evidence, *Journal of Finance* 57, 721–767.

——— , 2007, *Conglomerate Firms and Internal Capital Markets, Handbook of Corporate Finance: Empirical Corporate Finance* (North-Holland).

Panzar, J., and R. Willig, 1977, Economies of scale in multi-output production, *Quarterly Journal of Economics* 91, 481–93.

——— , 1981, Economies of scope, *American Economic Review* 71, 268–272.

Rhodes-Kropf, Matthew, and David Robinson, 2008, The market for mergers and the boundaries of the firm, *Journal of Finance* 63, 1169–1211.

Scharfstein, David, and Jeremy Stein, 2000, The dark side of internal capital markets: Segment rent seeking and inefficient investments, *Journal of Finance* 55, 2537–2564.

Schoar, Antoinette, 2002, The effect of diversification on firm productivity, *Journal of Finance* 62, 2379–2403.

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

Shin, Hyun-Han, and Rene M. Stulz, 1998, Are internal capital markets efficient?, *Quarterly Journal of Economics* 113, 531–552.

Stein, Jeremy, 1997, Internal capital markets and the competition for corporate resources, *Journal of Finance* 52, 111–133.

Teece, David J., 1980, Economies of scope and the scope of the enterprise, *Journal of Economic Behavior and Organization* 1, 223–247.

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

Table I: Summary Statistics

Summary statistics for firm value are reported for our sample of conglomerate and pure play firms (Panel A) for our sample from 1996 to 2008. Summary statistics for key variables of interest are reported for both multiple-industry and single-segment firms in Panel B. These variables are discussed in detail in Section II.E. Across industry language similarity (AILS), the fraction of industries between, and vertical relatedness are properties of an industry pair. AILS is the average pairwise similarity of firms in one of the industry in the pair with randomly drawn firms in the other industry. The fraction of industries between is the fraction of all other industries in the SIC-3 universe that lie between the given pair of industries, where betweenness is defined based on common vocabulary. Vertical relatedness is the degree of vertical relations based on the input-output tables. The remaining variables are the property of a single industry, and when tabulated over industry-pair observations, are simply the average of the characteristic for the two industries in the pair. Economies of scale is based on the estimation of a Cobb-Douglas production function over ten years, with sales being the dependent variable. Within industry similarity is the average pairwise similarity of randomly drawn firms in the given industry. Patent applications is at the industry level and is the fraction of total patents applied for by firms in the given industry. Industry instability is the absolute value of the logarithmic change in the number of firms in the given industry over the past year.

		Std.			
Variable	Mean	Dev.	Minimum	Median	Maximum
Panel A: Multiple-Industry (15,373 obs) and Pure-Play Firms (56,491 obs)					
Firm Value (Multi-Industry)	12430	48462	0.483	1228	1036340
Firm Value (Pure-Plays)	2450	18863	0.003	215.	1038648
Panel B: Industry Pairs (312,240 obs) and multiple-industry firms (15,373 obs)					
Number of multi-industry firms in Pair (Ind. Pairs)	0.147	0.855	0.0	0.0	57.0
Across Industry Language Simil. (Ind. Pairs)	0.017	0.010	0.000	0.014	0.169
Across Industry Language Simil. (multi-ind. firms)	0.032	0.019	0.000	0.025	0.138
Economies of Scale (Ind. Pairs)	0.701	0.434	0.000	0.942	1.326
Economies of Scale (multi-ind. firms)	0.000	0.012	0.000	0.000	0.720
Fraction of Industries Between Pair (Ind. Pairs)	0.325	0.257	0.000	0.267	0.992
Fraction of Industries Between Pair (multi-ind. firms)	0.097	0.133	0.000	0.042	0.992
Within Industry Language Simil. (Ind. Pairs)	0.086	0.038	0.000	0.081	0.433
Within Industry Language Simil. (multi-ind. firms)	0.073	0.030	0.010	0.066	0.188
Vertical Relatedness (Ind. Pairs)	0.003	0.014	0.000	0.000	0.536
Vertical Relatedness (multi-ind. firms)	0.027	0.066	0.000	0.006	0.536
Patent Applications (Ind. Pairs)	0.167	0.349	0.000	0.016	6.771
Patent Applications (multi-ind. firms)	0.375	0.495	0.000	0.183	4.544
Industry Instability (Ind. Pairs)	0.132	0.206	0.000	0.060	4.000
Industry Instability (multi-ind. firms)	0.457	0.194	0.000	0.446	1.600
Same 2-digit SIC Dummy (Ind. Pairs)	0.018	0.133	0.000	0.000	1.000
Same 2-digit SIC Dummy (multi-ind. firms)	0.228	0.371	0.000	0.000	1.000

Table II: Pearson Correlation Coefficients

Pearson Correlation Coefficients are reported for our sample of 312,240 observations of three digit SIC industry pairs from 1996 to 2008.

Row	Variable	Number of Operating Conglom. Pairs	Across Industry Language Similarity	Fraction of Industries Between	Within Industry Language Similarity	Industry insta-bility	Patent Applic-ations	Economies of Scale
			Correlation Coefficients					
(1)	Across Industry Similarity of Product Language	0.229						
(2)	Economies of Scale	-0.022	0.009					
(3)	Fraction of Industries Between Pair	-0.132	-0.691	0.003				
(4)	Within Industry Language Similarity	-0.044	0.184	0.059	-0.092			
(5)	Vertical Relatedness	0.200	0.165	0.002	-0.124	-0.049		
(6)	Patent Applications	0.038	-0.041	0.022	0.014	-0.140	0.024	
(7)	Industry Instability	-0.023	0.011	0.001	-0.018	0.008	-0.010	0.040

Table III: Conglomerate Multi-Industry Firm Summary

Summary statistics for various industry pairs from 1996 to 2008. Panel A compares observed multiple-industry pairs to randomly drawn industry pairs. Panel B displays observed multiple-industry pairs for firms of varying size. Panel C displays industry pairs for multi-industry firms that are growing, stable, or shrinking, as noted in the first column. Panel D displays conglomerate industry pairs for vertically integrated segments and for segments that are in the same two-digit SIC code.

Sub Sample	Across Industry Language Similarity	Within Industry Language Similarity	Fraction of Industries Between	# Obs.
Panel A: Overall				
All multi-industry firms	0.0296	0.0768	0.1293	40,769
Randomly Drawn SIC-3 Industries	0.0167	0.0862	0.3255	312,240
Panel B: By Conglomerate Size				
2 Segments	0.0341	0.0738	0.0867	6,365
3 Segments	0.0311	0.0750	0.1132	11,672
4-5 Segments	0.0289	0.0786	0.1366	15,794
6+ Segments	0.0247	0.0785	0.1790	6,938
Panel C: Shrinking, Stable, and Growing multi-industry firms				
Shrink by 2+ Segments	0.0268	0.0788	0.1490	600
Shrink by 1 Segment	0.0295	0.0779	0.1296	3,259
Stable Conglomerate	0.0301	0.0769	0.1260	30,525
Add 1 Segment	0.0282	0.0760	0.1414	4,741
Add 2+ Segments	0.0262	0.0739	0.1485	1,644
Panel D: Vertical and Same SIC-2 multi-industry firms				
Vertically Related Segments	0.0319	0.0717	0.0739	15,007
Same SIC-2 Segments	0.0471	0.0829	0.0291	8,015

Table IV: Where Multiple-Industry Firms Exist

OLS regressions with year fixed effects and standard errors clustered by year for our sample of 312,240 industry pairs from 1996 to 2008. One observation is one pair of three digit SIC industries in a year derived from the set of all permutations of feasible pairings. The dependent variable is the number of multiple-industry firms operating in the given industry pair. Panel A displays results based on the entire sample. Panels B and C display results for subsamples based on the competitiveness and valuations of industries lying between the given industry pair.

Row	Sample	Across Industry Language Similarity	Economies of Scale	Fraction Industries Between Pair	Within Industry Language Similarity	Vertical Relatedness	Patent Applications	Industry Instability	Pair Likelihood if Random	Same 2-digit SIC Code	# Obs. / RSQ
					Panel A: Full Sample						
(1)	All Industry Pairs	17.264	-0.237	0.120	-1.300	8.596	0.060	-0.069	0.947	0.082	312,240
		(19.06)	(-5.56)	(8.34)	(-11.55)	(6.96)	(4.44)	(-5.42)	(18.83)	(9.41)	0.130
					Panel B: Univariate Subsamples						
(2)	Concentrated Industry Pairs	27.334	-0.023	0.257	-0.969	3.649	0.048	-0.020	0.636	0.084	154,324
		(11.27)	(-1.22)	(6.73)	(-16.80)	(7.53)	(4.89)	(-1.77)	(8.61)	(6.60)	0.112
(3)	Competitive Industry Pairs	12.943	-0.355	-0.060	-1.569	7.921	0.066	-0.089	1.033	0.073	154,321
		(16.17)	(-5.60)	(-2.20)	(-9.62)	(7.59)	(3.26)	(-5.25)	(20.77)	(5.95)	0.107
(4)	High Firm Value Industry Pairs	21.295	-0.220	0.197	-1.251	5.648	0.021	-0.071	1.195	0.061	154,326
		(12.93)	(-5.49)	(5.57)	(-10.49)	(4.11)	(2.60)	(-5.43)	(19.55)	(6.07)	0.103
(5)	Low Firm Value Industry Pairs	11.673	-0.174	-0.005	-1.339	8.378	0.069	-0.048	0.736	0.116	154,319
		(14.86)	(-5.53)	(-0.74)	(-9.32)	(12.09)	(4.07)	(-3.05)	(12.46)	(5.87)	0.128
					Panel C: Bivariate Subsamples						
(6)	Concentrated + High Value	38.167	-0.032	0.425	-0.837	3.203	0.031	-0.009	0.778	0.065	65,904
		(5.96)	(-1.51)	(4.27)	(-13.36)	(4.14)	(3.55)	(-0.74)	(5.53)	(3.93)	0.114
(7)	Competitive + High Value	19.679	-0.339	0.153	-1.532	6.077	0.019	-0.101	1.284	0.060	88,422
		(11.95)	(-5.05)	(2.94)	(-8.87)	(3.88)	(1.25)	(-5.62)	(16.99)	(5.81)	0.101
(8)	Concentrated + Low Value	22.260	-0.014	0.158	-1.064	3.829	0.053	-0.027	0.593	0.109	88,420
		(8.97)	(-0.50)	(4.61)	(-11.71)	(7.29)	(3.80)	(-1.86)	(8.00)	(4.01)	0.118
(9)	Competitive + Low Value	8.824	-0.326	-0.268	-1.704	10.430	0.104	-0.076	0.803	0.120	65,899
		(8.95)	(-4.59)	(-13.60)	(-8.32)	(8.68)	(3.66)	(-3.25)	(14.00)	(4.61)	0.133

Table V: New Firm-Industry Segments

OLS regressions with year fixed effects and standard errors clustered by year. The dependent variable is the number of new multiple-industry segments in each three-digit SIC code pair in the given year. Panel A counts the number of new multiple-industry firms operating in both industries of an industry pair. Panel B restricts attention to new segments from multiple-industry firms that had fewer segments in the previous year. Panel C restricts attention to new segments of multiple-industry firms that were the acquirer in a transaction amounting to at least ten percent of the firm's assets.

		Across		Fraction	Within				Pair		
		Industry	Economies	Industries	Industry	Vertical	Patent	Industry	Likeli-	Same	
		Language	of	Between	Language	Relat-	Applic-	Inst-	hood if	2-digit	# Obs. /
Row	Sample	Similarity	Scale	Pair	Similarity	edness	ations	ability	Random	SIC Code	RSQ
					Panel A: Dep. Var = New Segment Pairs						
(1)	All Industry Pairs	2.440	-0.052	0.011	-0.253	0.746	0.008	-0.009	0.122	0.016	312,240
		(5.54)	(-4.15)	(3.60)	(-4.47)	(2.61)	(1.93)	(-3.32)	(4.72)	(3.90)	0.053
(2)	Concen. + High Value	5.969	-0.014	0.065	-0.153	0.523	0.005	-0.001	0.130	0.016	65,904
		(4.06)	(-1.94)	(3.07)	(-5.16)	(1.98)	(1.32)	(-0.59)	(2.62)	(2.48)	0.052
(3)	Concen. + Low Value	2.390	-0.077	-0.009	-0.334	0.527	0.004	-0.015	0.153	0.014	88,422
		(7.07)	(-4.32)	(-0.86)	(-4.43)	(2.04)	(1.06)	(-2.98)	(5.84)	(3.39)	0.048
(4)	Compet. + High Value	3.071	-0.010	0.022	-0.161	0.747	0.007	-0.006	0.075	0.018	88,420
		(4.14)	(-1.21)	(2.75)	(-4.58)	(2.91)	(1.82)	(-1.85)	(2.58)	(3.59)	0.039
(5)	Compet. + Low Value	1.448	-0.048	-0.038	-0.323	0.802	0.015	-0.007	0.108	0.022	65,899
		(3.18)	(-2.66)	(-3.81)	(-3.74)	(2.45)	(1.94)	(-1.15)	(3.66)	(2.81)	0.051
					Panel B: Dep. Var = New Segment Pairs Likely Obtained through Growth						
(6)	All Industry Pairs	2.018	-0.037	0.010	-0.202	0.606	0.007	-0.008	0.099	0.014	312,240
		(4.91)	(-3.72)	(4.18)	(-4.05)	(2.37)	(1.81)	(-3.53)	(4.32)	(3.65)	0.049
(7)	Concen. + High Value	4.201	-0.015	0.042	-0.117	0.419	0.006	-0.002	0.129	0.014	65,904
		(4.29)	(-2.44)	(3.32)	(-4.34)	(1.62)	(1.43)	(-1.03)	(2.62)	(2.27)	0.047
(8)	Concen. + Low Value	1.858	-0.052	-0.011	-0.269	0.469	0.004	-0.013	0.123	0.012	88,422
		(5.34)	(-3.52)	(-1.02)	(-4.09)	(2.14)	(1.19)	(-3.07)	(5.75)	(3.23)	0.044
(9)	Compet. + High Value	2.394	-0.008	0.017	-0.124	0.595	0.005	-0.005	0.061	0.015	88,420
		(3.62)	(-1.13)	(2.34)	(-4.39)	(2.76)	(1.59)	(-2.52)	(2.49)	(3.06)	0.033
(10)	Compet. + Low Value	1.283	-0.030	-0.030	-0.257	0.617	0.013	-0.006	0.086	0.020	65,899
		(3.18)	(-1.97)	(-3.22)	(-3.28)	(2.06)	(1.98)	(-1.38)	(3.10)	(2.72)	0.047
					Panel C: Dep. Var = New Segment Pairs Linked to SDC Acquisitions						
(11)	All Industry Pairs	0.242	-0.002	0.002	-0.018	0.072	0.001	-0.002	0.004	0.001	312,240
		(3.98)	(-1.44)	(2.16)	(-3.23)	(2.30)	(1.64)	(-2.02)	(1.77)	(2.43)	0.007
(12)	Concen. + High Value	0.602	-0.000	0.007	-0.008	0.036	0.001	-0.001	-0.001	0.001	65,904
		(2.42)	(-0.08)	(2.22)	(-3.17)	(1.25)	(1.58)	(-1.17)	(-0.37)	(1.80)	0.006
(13)	Concen. + Low Value	0.262	-0.002	0.001	-0.020	0.045	0.002	-0.002	0.006	0.001	88,422
		(4.21)	(-0.78)	(0.40)	(-2.94)	(1.50)	(1.86)	(-1.51)	(1.77)	(2.36)	0.007
(14)	Compet. + High Value	0.280	-0.002	0.002	-0.010	0.039	0.001	-0.001	0.003	0.001	88,420
		(2.43)	(-1.14)	(1.43)	(-3.30)	(2.19)	(1.35)	(-1.68)	(1.07)	(3.51)	0.004
(15)	Compet. + Low Value	0.176	-0.003	-0.003	-0.032	0.089	0.001	-0.003	0.005	0.001	65,899
		(2.17)	(-0.84)	(-0.89)	(-2.91)	(1.33)	(0.84)	(-2.24)	(2.02)	(1.49)	0.007

Table VI: Redefined Segments using text based classifications

OLS regressions with year fixed effects and standard errors clustered by year for our sample of 312,240 industry pairs from 1996 to 2008. One observation is one pair of three digit SIC industries in a year derived from the set of all permutations of feasible pairings. The dependent variable is the number of multiple-industry firms operating in the given industry pair. Panel A displays results based on the entire sample. Panels B and C display results for subsamples based on the competitiveness and valuations of industries lying between the given industry pair.

		Across		Fraction	Within			Pair			
		Industry	Economies	Industries	Industry	Patent	Industry	Likeli-	Vertical	Same	
		Language	of	Between	Language	Applic-	Inst-	hood if	Relat-	2-digit	# Obs. /
Row	Sample	Similarity	Scale	Pair	Similarity	ations	ability	Random	edness	SIC Code	RSQ
					Panel A: Where multiple-industry firms Exist (as in Table IV)						
(1)	All Industry Pairs	41.478	0.001	0.317	-0.654	-0.000	-0.195	0.136	N/A	N/A	145,058
		(22.93)	(0.02)	(9.42)	(-5.03)	(-3.00)	(-3.83)	(19.97)			0.084
					Panel B: New Conglomerate Segments (as in Table V)						
					Overall						
(2)	All Pairs	20.469	0.008	0.090	-0.465	-0.000	-0.120	0.074	N/A	N/A	145,058
		(15.54)	(0.19)	(4.54)	(-5.45)	(-0.47)	(-5.02)	(16.09)			0.080
					Segments Likely Obtained Through Growth						
(3)	All Pairs	4.422	-0.014	0.028	-0.041	-0.000	-0.025	0.015	N/A	N/A	145,058
		(7.03)	(-1.69)	(3.92)	(-2.18)	(-0.71)	(-4.92)	(8.75)			0.035
					Segments Likely Obtained Through Acquisition						
(4)	All Pairs	1.823	0.004	0.012	-0.032	-0.000	-0.010	0.005	N/A	N/A	145,058
		(8.28)	(0.59)	(2.86)	(-2.83)	(-0.41)	(-4.29)	(6.35)			0.016

Table VII: Product Market Transitivity by Industry

The table reports the average product market transitivity and statistics regarding operating segments for firms in each Fama-French 48 Industry. The number of segments is the total count of segment year observations for our entire sample from 1996 to 2008. The final column reports the fraction of these operating segments that operate under a conglomerate structure (a firm with more than one segment). Product Market transitivity is the fraction of peers of peers of a given firm that also consider the given firm itself to be a peer, as computed using the TNIC-3 industry classification.

Fama-French 48 Industry	Fraction Transitive	Total Segments	Fraction Conglomerate Segments
Aero	0.263	1190	0.882
Agric	0.551	1126	0.881
Autos	0.263	2953	0.783
Beer	0.879	564	0.670
BldMt	0.211	5843	0.896
Books	0.298	2244	0.851
Boxes	0.521	634	0.845
BusSv	0.105	18984	0.545
Chems	0.462	5438	0.878
Chips	0.327	8459	0.516
Clths	0.436	1723	0.577
Cnstr	0.211	2843	0.757
Coal	0.202	806	0.859
Comps	0.299	5319	0.466
Drugs	0.462	6195	0.264
ElcEq	0.172	2813	0.763
FabPr	0.123	1229	0.881
Food	0.551	2953	0.730
Fun	0.291	2850	0.669
Gold	0.323	846	0.363
Guns	0.263	630	0.908
Hlth	0.197	2298	0.560
Hshld	0.106	3440	0.757
Insur	0.196	7014	0.749
LabEq	0.282	3159	0.614
Mach	0.396	7943	0.814
Meals	0.620	2666	0.543
MedEq	0.197	4008	0.446
Mines	0.202	1511	0.763
Oil	0.202	7748	0.714
Other	0.381	2356	0.628
Paper	0.521	3078	0.850
PerSv	0.702	1728	0.638
RlEst	0.152	3480	0.875
Rtail	0.091	6962	0.537
Rubbr	0.348	2629	0.846
Ships	0.263	452	0.808
Smoke	0.343	294	0.738
Soda	0.538	694	0.731
Steel	0.401	3367	0.798
Telcm	0.737	6312	0.671
Toys	0.177	1569	0.662
Trans	0.521	4632	0.633
Txtls	0.914	1121	0.810
Whlsl	0.847	8531	0.769

Table VIII: Product Market Transitivity

Summary statistics and logistic regressions with year fixed effects and standard errors clustered by year for our sample of 40,330 Compustat firms from 1997 to 2008. Panel A and Panel B report summary statistics regarding the average fraction of firms are multiple-industry firms for various subsamples as noted. Panel C displays the results of logistic regressions for which the dependent variable is a dummy equal to one for a multiple-industry firm, and zero for a pure play firm. The independent variables include the degree to which firms are in transitive product markets, and controls for firm age, size and profitability. Product Market transitivity is the fraction of peers of peers of a given firm that also consider the given firm itself to be a peer, as computed using the TNIC-3 industry classification.

Sample	Transitivity Subsample	Fraction Multi-Industry	# Obs.
	Panel A: All Firms		
All Firms	Weakly Transitive	0.600	40,330
All Firms	Highly Transitive	0.451	40,333
	Panel B: Subsamples Based on Size and Age		
Small Young Firms Only	Weakly Transitive	0.416	11,674
Small Young Firms Only	Highly Transitive	0.243	14,800
Small Old Firms Only	Weakly Transitive	0.585	8,690
Small Old Firms Only	Highly Transitive	0.423	5,168
Large Young Firms Only	Weakly Transitive	0.583	6,601
Large Young Firms Only	Highly Transitive	0.463	7,317
Large Old Firms Only	Weakly Transitive	0.778	13,365
Large Old Firms Only	Highly Transitive	0.690	13,048

Row	Fraction Transitive	Log Sales	Log Firm Age	oi/ Sales	Obs. /RSQ
			Panel C: Logistic Regressions		
(1)	-1.753				80,663
	(-19.12)				0.052
(2)		0.372			80,663
		(24.05)			0.106
(3)			0.597		80,663
			(22.10)		0.094
(4)				0.046	80,663
				(0.85)	0.006
(5)	-1.645	0.294	0.396	-0.261	80,663
	(-16.92)	(17.46)	(14.54)	(-3.95)	0.170

Table IX: Product Market Transitivity and Changes to Conglomerate Competition

OLS regressions with year fixed effects and standard errors clustered by year for our sample of 11,754 multiple-industry firms from 2000 to 2008. The dependent variable is logarithmic growth in the number of segments of the given conglomerate from year t to year t+1. All independent variables are measures of change in the given quantity from year t-3 to year t. Product Market transitivity is the fraction of peers of peers of a given firm that also consider the given firm itself to be a peer, as computed using the TNIC-3 industry classification. We also consider three year changes in R&D activity, CAPX activity, profitability and firm size.

	Fraction	R&D/	CAPX/	oi/	Log	Document	Obs.
Row	Transitive	Sales	Sales	Sales	Assets	Length	/RSQ
(1)	-0.016						11,754
	(-2.14)						0.001
(2)		0.006					11,754
		(0.16)					0.001
(3)			-0.002				11,754
			(-0.15)				0.001
(4)				-0.006			11,754
				(-0.73)			0.001
(5)					0.006		11,754
					(2.28)		0.001
(6)						0.008	11,754
						(2.62)	0.002
(7)	-0.016	-0.003	-0.006	-0.011	0.005	0.006	11,754
	(-2.19)	(-0.06)	(-0.46)	(-0.99)	(2.07)	(2.16)	0.002

Table X: Product Description Growth

OLS regressions with year fixed effects and standard errors clustered by firm for our sample of 8,769 multiple-industry firms from 1997 to 2008. One observation is one conglomerate in one year. The dependent variable is the firm's product description growth, defined as the natural logarithm of the number of words in the firm's business description in year $t+1$ divided by the number of words in the firm's business description in year t. Panel A displays results based on raw firm-level product description growth. Panel B displays results based on TNIC industry adjusted product description growth.

	Across		Fraction	Within					Pair						
	Industry	Economies	Industries	Industry	Vertical	Patent	Industry	Same	Likeli-						
	Language	of	Between	Language	Relat-	Applic-	Inst-	2-digit	hood if	Document	R&D/	CAPX/	oi/	Log	# Obs.
Row	Similarity	Scale	Pair	Similarity	edness	ations	ability	SIC	Random	Length	Sales	Sales	Sales	Assets	RSQ
							Panel A: Product Description Growth								
(1)	0.723	0.014	0.013	0.180	0.001	0.000	-0.022	-0.011	0.001	-0.105	0.116	0.009	0.124	-0.001	8,769
	(3.01)	(0.22)	(0.53)	(1.49)	(0.01)	(1.13)	(-1.16)	(-1.16)	(1.16)	(-9.37)	(1.74)	(0.21)	(4.36)	(-0.40)	0.031
(2)	0.684								0.001	-0.106	0.140	0.007	0.125	-0.001	8,769
	(4.25)								(1.38)	(-10.39)	(2.37)	(0.16)	(4.39)	(-0.60)	0.030
							Panel B: Industry Adjusted Product Description Growth								
(3)	0.635	-0.030	-0.007	0.159	0.013	-0.000	-0.012	-0.006	0.001	-0.064	0.253	-0.008	0.105	-0.001	8,558
	(2.55)	(-0.45)	(-0.24)	(1.31)	(0.14)	(-0.45)	(-0.60)	(-0.64)	(1.03)	(-5.42)	(3.64)	(-0.18)	(3.60)	(-0.46)	0.009
(4)	0.788								0.001	-0.060	0.232	-0.005	0.106	-0.001	8,558
	(4.67)								(1.20)	(-5.66)	(3.80)	(-0.11)	(3.64)	(-0.35)	0.009

Figure 1: Visual depiction of industry organization hypotheses. Figure 1A depicts the concept of across industry similarity (potential asset complementarities). Industries X and Y have high levels of potential asset complementarities. Figure 1B depicts the concept of within industry language similarity (WILS). Industries with low levels of WILS occupy a larger volume of the product market space, and for example, industries X and Y have lower WILS as compared to I_1 or I_4. Figure 1C depicts the concept of between industries. Industry I_3 lies between industries X and Y.

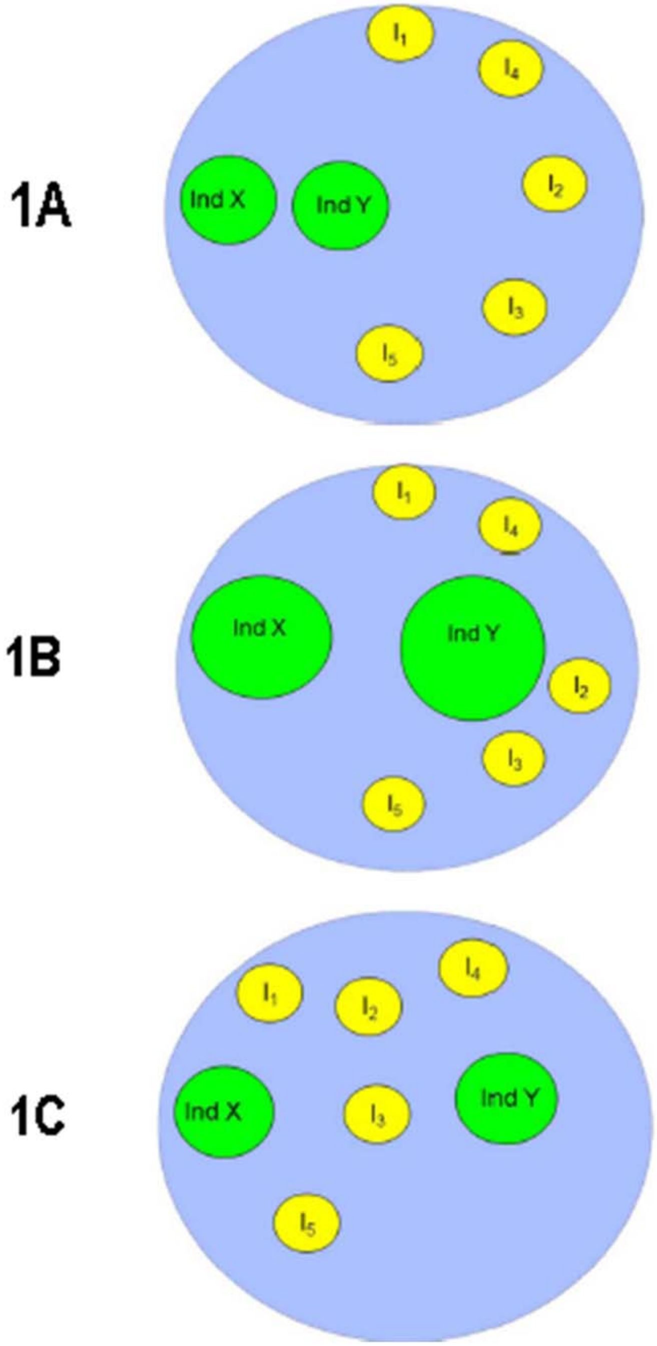


Figure 2: Distribution of Across Industry Language Similarity scores for randomly drawn industry pairs versus conglomerate industry pairs. Across industry similarity is the average pairwise 10-K textual similarity of firm pairs in each SIC-3 industry based on the text in each firm's business descriptions. The X-axis depicts the total fraction of all industry pairs with the given level of across industry language similarity and the Y-axis depicts levels of across industry language similarity ranging from zero to 0.05 (values above this level are grouped into the last datapoint to reduce the size of the graph. The gray line depicts the median across industry language similarity (0.023) for conglomerate industry pairs. This median is reached at the 85.5th percentile of across industry similarity for randomly drawn pairs. The large amount of mass on the RHS for conglomerate industry pairs indicates a thick right-tail of multiple-industry firms operating in industry pairs that are extremely similar relative to randomly drawn pairs.

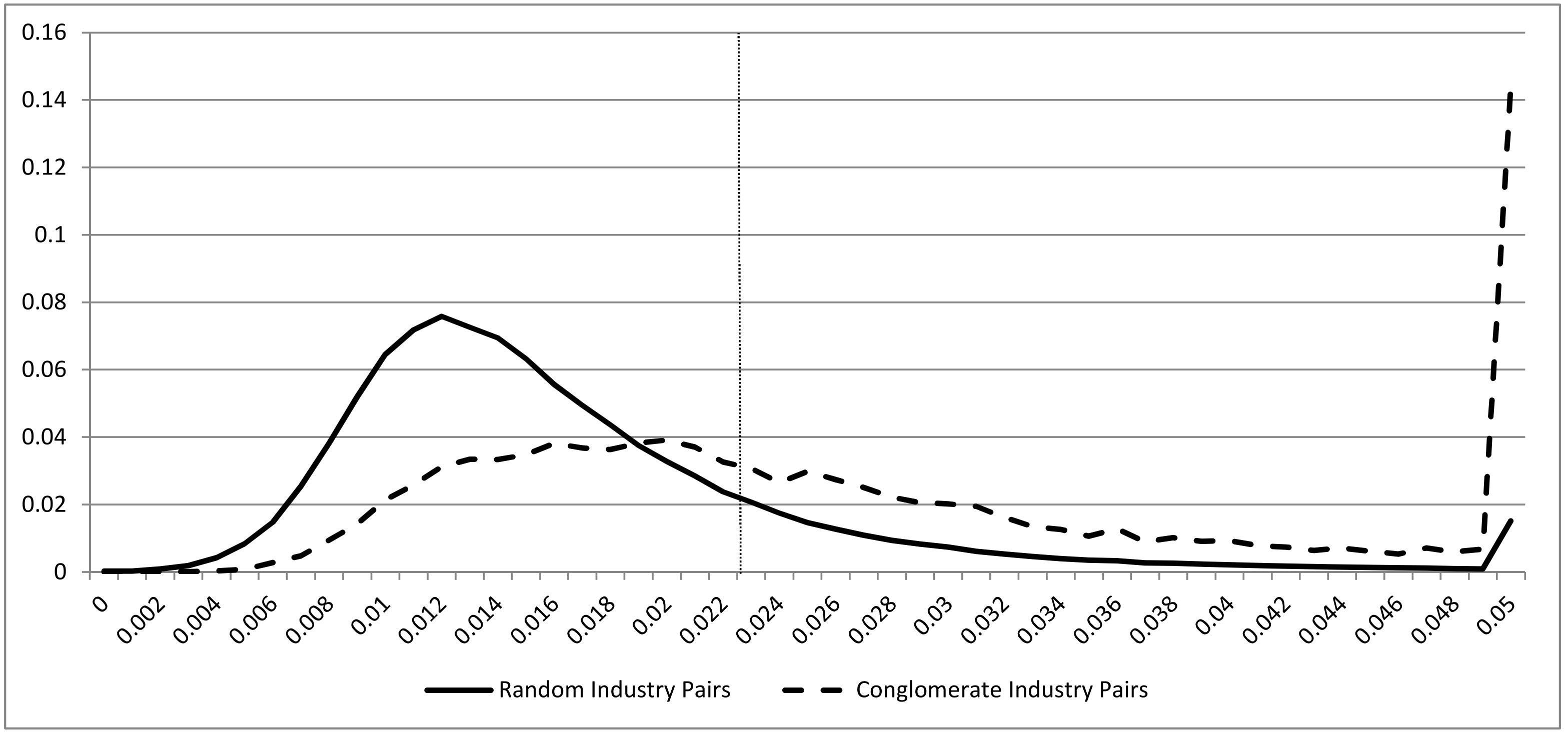
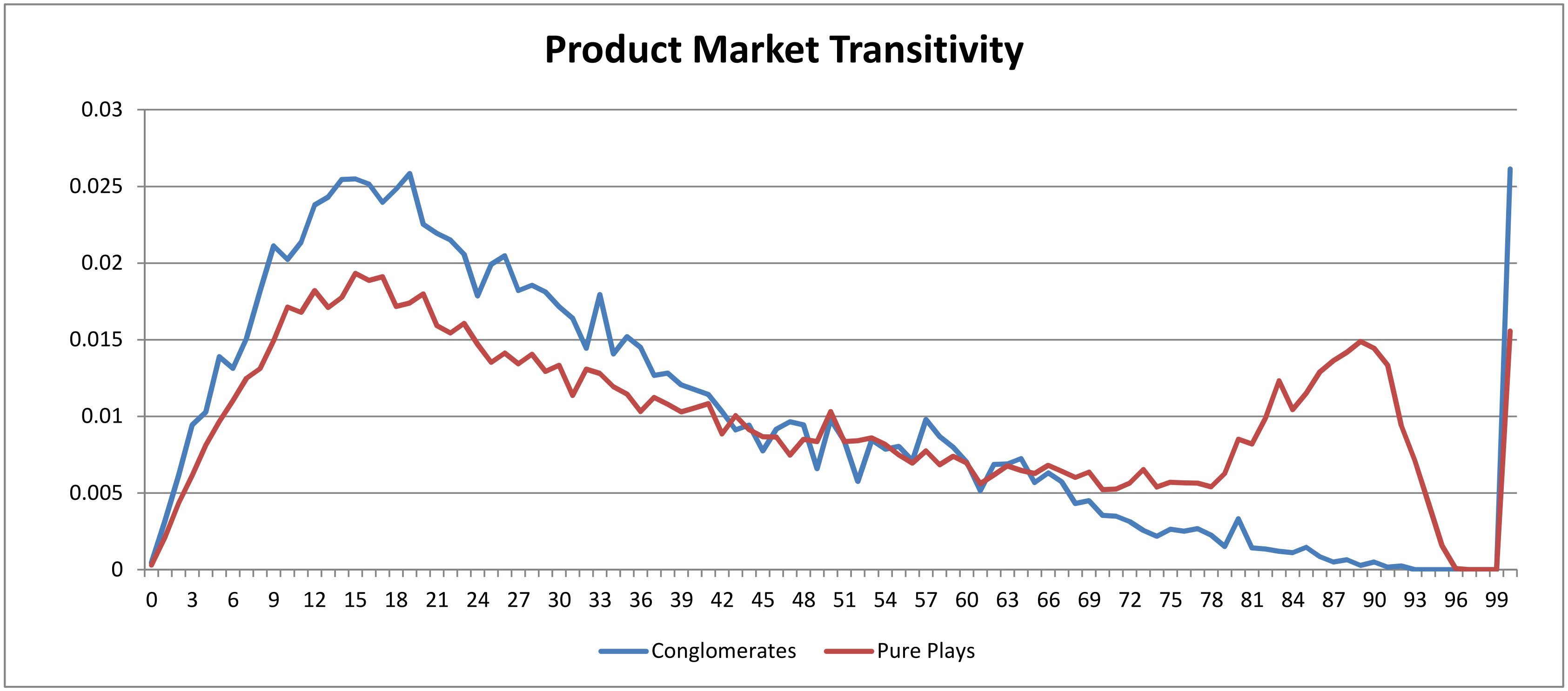


Figure 3: Density of product market transitivity for reported multiple-industry firms and pure play firms. Product market transitivity is the observed probability that firms A and C are rivals given that A and B are rivals and that B and C are rivals. Here a pair of firms is defined as being rivals if they are classified as such using the TNIC-3 industry classification. For the purposes of this figure, a conglomerate is defined as a firm that reports more than one segment in the Compustat Segment files, and a firm is defined as a pure play if it only reports one segment. The graph reports the probability density on the Y-axis, and the percentage level of transitivity (which is bound between zero and one hundred) on the X-axis.


Product Market Transitivity
Conglomerates
Pure Plays