Conglomerate Industry Spanning

Gerard Hoberg and Gordon Phillips[*]

April 28, 2011

ABSTRACT

We use text-based analysis of business descriptions from 10-Ks filed with the SEC to examine in which industries conglomerates are most likely operate and to understand conglomerate valuations. We find that conglomerates are most likely to operate in industry pairs that are closer together in the product space and in industry pairs that have profitable opportunities "between" them. Examining cross-sectional conglomerate valuations, we find that conglomerates that are more difficult to reconstruct using text-analysis of firm pure plays tend to trade at modest premia. The conglomerates that are most easy to replicate trade at small discounts relative to matched pure-play firms. These findings are consistent with conglomerate firms generating product synergies when producing in related profitable industries.

[*] University of Maryland and University of Maryland and National Bureau of Economic Research respectively. Hoberg can be reached at ghoberg@rhsmith.umd.edu and Phillips can be reached at gphillips@rhsmith.umd.edu. We thank seminar participants at Stanford University and University of Illinois for helpful comments. All errors are the authors alone.

Why do firms operate in multiple industries? This question has been the focus of much research that takes the industries that firms operate within as given and examines outcomes such as valuations and investment decisions. Existing explanations for multiple industry production based on investment decisions include advantages of internal capital markets (Stein (1997)), agency problems (Lang and Stulz (1994) Berger and Ofek (1995), Denis, Denis, and Sarin (1997), Scharfstein and Stein (2000)), conglomerate learning about ability (Matsusaka (2001)) and managerial talent that can be used in different industries (Maksimovic and Phillips (2002)). This literature does not examine why conglomerate firms span some industries and not others.

We take a different approach in this study. We analyze in which industries conglomerate firms are likely to produce using fundamental product market characteristics. We examine whether conglomerate firms may create value through operating in related industries that allow for synergies and new products to be created. The idea we examine is simple and is related to that of Rhodes-Kropf and Robinson (2008) who examine the importance of asset complementarities in mergers. While Robinson and Rhodes-Kropf examine firm complementarities, we examine potential industry complementarities. We ask whether certain industry characteristics - distinct from vertical relatedness - make operating in two different industries valuable? Are industries related in certain ways that make it likely that firms will find it optimal to produce in multiple industries? Apple Computer is an example of a firm that now produces in cell phones, computers, and digital music - industries that are today very related to each other. It faces some firms that operate individually in each of these industries but more firms today are attempting to operate in these related industries.

We use text-based analysis of conglomerate and pure play business descriptions from 10-Ks filed with the SEC to examine in which industries conglomerates are most likely operate and to understand cross-sectional conglomerate valuation. Following Hoberg and Phillips (2010a), we convert firm product text into a spatial representation of the product market. In this framework, each firm, and each industry, has a product location based on words that allows measurement of how close industry product markets are to each other. Our framework also allows an assessment of which industries in product market space are between any given pair of industries, where between industries are industries that are closer to each industry of a given industry pair than they are to each other.

In our primary analysis, we examine within and cross industry similarity measures using text based analysis. We also examine the relatedness of other industries to SIC industry pairs and whether the type of other industries that are close to or “between” industry pairs impacts whether firms choose to produce in particular industries. We control for other measures of relatedness including vertical integration in assessing the impact of our text-based measures of relatedness.

We find that conglomerate pairs are more likely to operate in industry pairs that are closer together in the product space, industry pairs that have profitable opportunities surrounding them, and in industries with lower within industry product similarity. Conglomerate firms are also less likely to produce in industries that span competitive, low-value industries. These findings are consistent with product synergies, where conglomerates producing in two related profitable industries may be able to enter profitable industry niches.

The second part of our study reexamines conglomerate firm valuation by redefining the benchmark single-segment firms used to reconstruct and value conglomerates. We use text-based analysis to redefine conglomerate pure-play benchmarks to better understand the cross-sectional variation in conglomerate valuation premia and discounts. Our focus is on the cross-sectional variation in the valuation - not the average discount or premium of conglomerate firms - a topic that has been extensively studied previously.[1] Our methodology identifies pure-play firms that are most similar, based on product text descriptions, to conglomerate firms. We use these pure-play firms and the product text distance they are from conglomerate firms to understand cross-sectional differences in conglomerate valuation. We examine both

[1] Wernerfelt and Montgomery (1988), Lang and Stulz (1994), Berger and Ofek (1995), Comment and Jarrell (1995), Servaes (1996), Lins and Servaes (1999), Rajan, Servaes, and Zingales (2000) and Lamont and Polk (2002) find evidence of a diversification discount. Laeven and Levine (2007) find discount in financial conglomerates. However this average discount has been shown to be driven by self-selection by Campa and Kedia (2002), Graham, Lemmon, and Wolf (2002), and Villalonga (2004b) and by data problems by Villalonga (2004a) and merger accounting by Custodio (2010). See Maksimovic and Phillips (2007) for a detailed survey.

simple text-based identifications of the pure plays that use product vocabulary that best explains the vocabulary of the conglomerate, and enhanced benchmarks that additionally seek to match the conglomerate along five key accounting dimensions including profitability and growth.

We find that on average conglomerates do not trade at a discount relative to text-matched single segment firms. However, this average effect masks important cross-sectional variation. We find that conglomerates that are more difficult to reconstruct using pure plays tend to trade at modest premia and those conglomerates that are most easy to replicate trade at small discounts. These findings are consistent with conglomerate firms producing in related industries that have product synergies and in industries that are also more difficult to enter for competing single-segment firms.

The contributions of our paper are two-fold. Our first main contribution is to use word similarity from firm product descriptions to understand in which industries conglomerate firms choose to produce. We calculate several unique relatedness measures across and within industries. We calculate within industry similarity to measure how different firms are within SIC classifications to capture potential product differentiation. We also calculate pair-wise industry relatedness to measure the similarity of industries to each other. The final measure we calculate is the number of industries "between" two industries, where an industry is between two others if it is closer in similarity to each of the two industries individually than they are to each other. This last measure captures potential competition between two industry segments and also the potential for multiple industry firms to introduce new products at low cost in those product spaces.

The second main contribution of our paper is to help understand cross-sectional conglomerate valuation. We use text-based analysis to better form a matched set of pure-play competitors so that each firm and each segment has its own set of competitors. This new "network" centric method of viewing competition draws on firm centric notion of competition analogous to a Facebook circle of friends. In order to find competitor firms, we use the similarity of firms to each reported segment of the conglomerate firm and then weight these firms by how similar they are to the overall multiple industry firm. We weight these firms differentially so that firms closer to a given firm receive a higher weight based on text-based distances. Both the set of weights and segment allocations provide information about the competitive structure a conglomerate faces in its respective segments.

Our measure of how difficult a conglomerate is to replicate using pure play firms is a more accurate measure of a conglomerate's overall competitive position. Using this method we can construct comparable valuations and then can assess whether conglomerate firms with fewer direct close competitors have higher values and whether conglomerate firms that span high-value concentrated industries have different valuations that those that span competitive industries with positions that are easier to replicate.

We add to the existing literature in two central ways. First, we examine which industries conglomerates are more likely to operate within and the characteristics of these and surrounding industries. Previous research has examined investment decisions of conglomerate firms including Lang and Stulz (1994), Rajan, Servaes, and Zingales (2000), Scharfstein and Stein (2000) and Maksimovic and Phillips (2002), but, with the exception of Maksimovic and Phillips (2008) and Hann, Ogneva, and Ozbas (2011), do not examine how industry characteristics affect production decisions. Maksimovic and Phillips examine how long-run industry characteristics affect acquisition decisions by conglomerate and single-segment firms, but do not examine in which industry pairs conglomerate firms choose to produce. Hann, Ogneva, and Ozbas (2011) show that producing in different industries lowers a firm's cost of capital consistent with a coinsurance effect.

Second, with respect to conglomerate valuation, many authors have examined whether diversified firms trade at an average discount relative to single-segment competitors. While we do show that the average discount disappears by finding better matched pure play firms, this is not our main contribution. Our major contribution relative to this literature is to show that cross-sectionally the discount varies by industry characteristics and the uniqueness of conglomerate firm characteristics.

Our paper proceeds as follows. In the next section we discuss our new measures of industry relatedness and spanning and how our paper provides tests of the potential for product market synergies by focusing on within and cross-industry similarity. In Section 2, we present our methodology and how we calculate our new text based measures of within and cross-industry similarity. Section 3 contains the results of our tests of how industry relatedness and spanning affect conglomerate production. Section 4 analyzes cross-sectional conglomerate valuation and Section 5 concludes.

I Industry Relatedness and Spanning

We ask whether there are certain industry characteristics - distinct from vertical relatedness - that make operating in two different industries valuable? The central hypothesis we examine in this paper is whether the potential for product market synergies and industry competition influence in which industries conglomerate firms choose to produce. Our analysis also examines the competition within an industry and in industries that are close to, or between, industry pairs to see if conglomerates are more likely to produce in industry pairs that have the potential for asset complementarities and synergies. The idea is similar to that of Rhodes-Kropf and Robinson (2008) and Hoberg and Phillips (2010b) who examine whether asset complementarities and synergies are important for mergers.

We also address whether the industries an industry pair ”spans” influence whether a conglomerate firm produces within that industry pair. Industry spanning is the idea that there may be in industry that is more similar to both industries of a particular industry pair that the two industries of the pair are to each other. The idea we test is whether a conglomerate is more likely to produce in a particular industry pair, if that pair spans other high valued, less competitive industries. Producing in that industry pair may allow the multiple industry pair to more easily produce products for the high-valued concentrated industry.

We generate industry pair characteristics using text-based analysis of business descriptions from 10-Ks filed with the SEC. We then examine use these industry characteristics to understand in which industries conglomerates are most likely operate and to understand cross-sectional conglomerate valuation. We discuss the way we gather and process these 10-K product descriptions in the next section. In this section we introduce conceptually the variables we will measure to capture how industries are related to each other. We consider these new measures in addition to more typical industry-relatedness measures of vertical integration.

We construct three new variables that allow us to assess how every pair of industries relates to one another. In particular, we measure how far apart industries are in the product space, how heterogeneous their products are within-industry, and the extent to which other industries lie between the given industry pair in the product space. We use our database of pure play firms - firms with no industry segments from the COMPUSTAT segment tapes - to construct these measures, as this permits a more straight-forward interpretation of the measures. Because our COMPUSTAT conglomerate database reports industry membership using SIC codes, we define an industry for the purposes of this discussion as a three digit SIC code.[2]

The first industry relatedness variable we calculate is *Across Industry Similarity.* This measures how close industry i and industry j are in the product space. Suppose industry i has N_i pure play firms, and industry j has N_j firms. Across industry similarity is the average cosine similarity of all pair permutations using word vectors from a firm's product descriptions. We discuss this cosine similarity measure fully described in the next section. Each pair includes one firm from industry i and one from industry j. Industries located closer together likely share asset complementarities or economies of scope.

Second, we measure *Within Industry Similarity.* Suppose industry i has N_i pure play firms. Within industry similarity is the average cosine similarity for all pair permutations of the N_i firms. Firms in industries with higher within industry similarity likely have less unique products, and face more significant competition from their rivals due to the absence of product differentiation. For an industry pair, within industry similarity as the average within industry similarity of industry i and j.

Third, we also measure the fraction of other industries between a pair of indus-

[2] Later in this paper we relax this definition to explore conglomerate valuations. Due to the high dimensionality of the industry spanning tests we construct here, we do not consider alternative industry definitions here as using firm-specific industry definitions (as in Hoberg and Phillips (2010a)) would render our calculations infeasible due to the exponential growth in industry trio permutations.

tries i and j, which we label *Between Industries.* Because all firms, and hence all industries, have a unique location on a high dimensional unit sphere, we can assess whether other industries lie in the space between a given pair. This novel measure can be used to examine whether conglomerates benefit from business opportunities lying between their segments, perhaps through asset complementarities. The Across Industry Similarity measure discussed above, which is available for any pair of industries, is instrumental in computing the fraction of industries between a given pair. Where $AIS_{i,j}$ denotes the Across Industry Similarity of industries i and j, we define a third industry k as being *between* industries i and j if the following relationship holds.

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

The Fraction of Industries Between a given pair of industries i and j is therefore the number of industries (excluding i and j) satisfying this condition divided by the number of industries in the database in the given year (excluding i and j). We also consider a dummy variable identifying industry pairs for which no other industries lie between them.

From these measures we test the following three hypotheses:

H1: Cross-Industry Similarity: Conglomerate firms are more likely to produce in two industries that have high cross-industry similarity and thus are easier to manage with more potential synergies.

We test this hypothesis by examining the number of conglomerate firms that operate in each pair-wise segment and examine whether the number of conglomerate firms increase in pair-wise industry similarity. We also predict that conglomerate benchmark valuation will increase the harder the conglomerate firm is to replicate with pure-play firms.

H2: Within-Industry Similarity: Conglomerate firms are less likely to produce in industries that have high within industry similarity and those with high competition.

We test this hypothesis by examining whether the number of conglomerate firms decrease when the industry pairs have high within industry similarity and high competition.

H3: Between-Industry Spanning: Conglomerate firms are more likely in an industry pair when the industries spanned by the pair-wise combination include high-value, less competitive industries.

We examine the fraction of industries that are between each pair-wise combination of industries and test wither conglomerate firms producing in a particular pair-wise combination increase when the industries spanned or between these industries are highly valued and less competitive.

II Data and Methodology

In this section we briefly describe our conglomerate database, our database of pure-play (non-conglomerate) firms, and the construction of key text-based variables used to examine where conglomerates produce in the product space.

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 segment files to identify which firms are conglomerates, 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 conglomerate firms when they report two or more three digit SIC codes. We identify 22,252 unique conglomerate 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 conglomerates change from year to year, we further require that a conglomerate exist 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 conglomerates, 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 conglomerate firm-years with 40,769 unique segment conglomerate 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 conglomerates that span 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) and Hoberg and Phillips (2010b). 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.[3] 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.

[3] 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.

C Word Vectors and Cosine Similarity

We employ methods used in Hoberg and Phillips (2010a) and Hoberg and Phillips (2010b) to construct word vectors and measure similarity. The first step is to form word vectors for each firm based on the 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 5% of all firms. We then consider the universe of all product words in the 10-K universe in each year. 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 used in studies of information processing (see Sebastiani (2002) for a summary of methods). It measures the angle between two word vectors on a unit sphere.

D Control Variables and Vertical Integration

In addition to our three new industry similarity and relatedness variables, we include control variables for industry size, vertical relatedness, and a dummy identifying which industries are in the same two-digit SIC code. As we aim to examine conglomerate incidence rates across industry pairs, controlling for industry size is important.

For example, if conglomerates formed by randomly choosing among available pure play firms in the economy, then the incidence of conglomerate spanning 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 $(F_i x F_j)$, 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 whether conglomerates tend to span vertically related industry pairs. The inclusion of this control 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.

Finally, we consider a dummy variable set equal to one if a given pair of three digit SIC industries lies in the same two-digit SIC industry.

E Conglomerate Restructuring

We examine whether our industry spatial variables can explain how conglomerates restructure, and we classify restructuring in four 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 "Segment Pair Disappears" as a dummy equal to one if the given pair does not exist in the conglomerate's structure in the following year. We then define "Segment Pair Likely Sold or Closed" as a dummy equal to one if the given pair does not exist in the conglomerate's structure in the following year, and the conglomerate has fewer segments in year $t + 1$ relative to year t. We define "Segment Pair Likely Reclassified" as a dummy equal to one if the given pair does not exist in the conglomerate's structure in the following year, and the conglomerate has at least as many segments in year $t + 1$ relative to year t. Finally, we define "Segment Pair Like Sold Off" as a dummy equal to one if the given pair does not exist in the conglomerate's structure in the following year, and the conglomerate was the target of an acquisition of at least ten percent of its assets in year $t + 1$.

F Summary Statistics

Table I displays summary statistics for our conglomerate and pure play firm, and industry pair databases. Panel A shows that we our conglomerates are generally larger than our pure play firms in terms of total value of the firm, and they also generally operate in markets that are more concentrated, as measured by their VIC-7.06 HHI.

Panel B of the table shows that a randomly drawn pair of three digit SIC industries has 0.147 conglomerates having segments operating in both industries of the given pair. Hence, the majority of randomly chosen industries do not have conglomerates spanning them. The average across industry similarity is 0.017, which closely matches the average firm similarity reported in Hoberg and Phillips (2010a). The average within industry similarity, intuitively, is much higher at 0.086. The table also shows that a randomly drawn pair of industries is sufficiently far apart such that 32.5% of all other industries lie between them.

[**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 spanning conglomerate pairs. The first column of this tables shows that this variable is positively related to across industry similarity, and negatively related to within industry similarity and the fraction of industries between a given pair. Although these univariate results hold for across industry similarity 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 conglomerates are more likely to span 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.

The table also shows that the average HHI variable and the within industry similarity variable are modestly correlated at -48.7%. This result is consistent with findings in Hoberg and Phillips (2010a), and confirms that concentrated product markets generally have more product differentiation. Aside from these modest to high correlations, 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 conglomerates populate industries with across industry similarity of .0304, which is 79% higher than the 0.017 of randomly chosen industry pairs. Conglomerates 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 conglomerates (two or three segments) compared to those of larger conglomerates. The table suggests that larger conglomerates tend to cast a wider footprint across the product market space, as they have lower across industry similarity. They also tend to reside in industries with more industries residing between them, and industries that have higher within industry similarity. In Panel C of Table III, we observe that most conglomerates (30,525) are stable from one year to the next, although 3,259 reduce in size by one segment, and 600 reduce in size by two or more segments. Analogously, 4,741 increase in size by one segment, and 1,644 increase in size by two segments.

In Panel D, we observe that vertically related conglomerates have average across industry similarities that are close to the average for all conglomerate pairs. This finding mirrors findings in Hoberg and Phillips (2010a), who show that industry classifications based on business descriptions do not correlate with vertical relationships (rather they focus on horizontal distances or economies of scope). In contrast, across industry similarities are somewhat higher for industries having the same two digit SIC code, as SIC codes are measures of horizontal relatedness. Both vertical industries and those in the same two-digit SIC code also have fewer than average industries between them.

III Results: Conglomerate Spanning

In this section we examine whether we can predict whether conglomerates produce in particular industry pairs. We test whether across industry similarity and within industry similarity matter for the number of conglomerate firms producing in a particular industry pair.

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 Conglomerates Spanning Pair**, which is the number of conglomerates 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 cross industry similarity increases the number of conglomerate firms producing in a particular industry, while average within industry similarity decreases the conglomerate firms producing in a particular industry. Because within industry similarity and the average HHI are moderately correlated, we examine their effects separately. The table shows that conglomerates broadly tend to span more concentrated markets, ie, those with higher product differentiation and higher concentration. However, within industry similarity matters more and we include only this variable henceforth. Panel A also shows that the fraction of industries between a given pair also matters, and its sign depends on the characteristics of the industries between.

Panels B and C show that when high value and concentrated industries are between, conglomerates span 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 asset complementarities to span technologies that might permit entry into previously concentrated product markets.

Table V examines how industry characteristics influence which industry pairs disappear from conglomerates. Using the SDC mergers and acquisitions database, we examine when the segment pair is likely sold or closed as well as potentially reclassified. One observation is one pair of segments in an existing conglomerate in year t. We require the conglomerate firm itself to exist in year t and year $t + 1$.

The dependent variable varies by Panel. In Panel A, the dependent variable is **Segment Pair Disappears**, which is a dummy equal to one if the given pair does not exist in the conglomerate's structure in the following year. In Panel B, the dependent variable is **Segment Pair Like Sold or Closed**, which is a dummy equal to one if the given pair does not exist in the conglomerate's structure in the following year, and the conglomerate has fewer segments in year $t + 1$ relative to year t. In Panel C, the dependent variable is **Segment Pair Likely Reclassified**, which is a dummy equal to one if the given pair does not exist in the conglomerate's structure in the following year, and the conglomerate has at least as many segments in year $t+1$ relative to year t. In Panel D, the dependent variable is **Segment Pair Like Sold Off**, which is a dummy equal to one if the given pair does not exist in the conglomerate's structure in the following year, and the conglomerate was the target of an acquisition of at least ten percent of its assets in year $t + 1$.

[**Insert Table V Here**]

The results in Panel A of V show that segment pairs are less likely to be sold or closed if the across industry similarity is high. This result also has the largest coefficient if 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 conglomerate firms using two related industries to maximize asset complementarities and to produce products in highly concentrated industries.

The results in Panels A and B of V show that segment pairs are less likely to be sold or closed if the across industry similarity is high. This result has the largest coefficient if the industries between two industry pairs are highly concentrated. This result is consistent with conglomerate firms using two related industries to maximize asset complementarities and to produce products in highly concentrated industries.

The results in Panel C show that segment pairs are more likely to be reclassified when there are other industries between them. This result is consistent with these pairs reclassifying in order to potentially enter the markets between the given pair. For example, because the given conglomerate has technologies that produce goods on either side of the between industry, it is likely that the given conglomerate has access to potential low cost entry into the between industry.

Given these strong results on which industries conglomerate firms choose to produce in and the characteristics of these industries being high concentration, high value industries, we now turn to the question of how industry composition affects the cross-sectional variation in conglomerate valuation.

IV Conglomerate Valuation

The study of conglomerate valuations, especially compared to non-conglomerate firm valuations, has a rich history in Finance. Conventional wisdom suggests that conglomerates are better diversified and can better survive downturns in any one of its product markets. Recently, Villalonga (2010) shows that conglomerates faired better in the financial crisis. Therefore, when Lang and Stulz (1994), Berger and Ofek (1995), and Servaes (1996) famously established that conglomerate firms have stock market valuations that appear to be low relative to single segment firms, it was viewed as a major puzzle. Maksimovic and Phillips (2002) show that conglomerate discounts can arise when firms can reallocate assets over the business cycle, and when productivity levels vary. Other studies including Campa and Kedia (2002) and Villalonga

(2004b) use econometrics, including self selection and propensity score methods, and find that conglomerates are fairly valued. Regardless of the conclusion, most existing studies rely heavily on SIC-based industry classifications to identify a conglomerate's peers, and the counter factual it would experience under a non-conglomerate structure. Hoberg and Phillips (2010a) find that static industry classifications including SIC codes are flawed, and that text-based alternatives perform significantly better and offer more research flexibility. Because conglomerates are modeled using a sum of parts approach, the potential gains from improved classification can be especially large in this setting.

In this section, we explore whether information in firm product descriptions can be used to construct more informative benchmarks, both in terms of product market identification and in terms of identifying the universe of pure play firms that are best suited to serve as a counter-factual to operating under a conglomerate structure.

A Existing Methods

Although we depart significantly from the literature in some of our conglomerate valuation methods, we begin by considering a modified algorithm based on Lang and Stulz (1994) (LS) and Berger and Ofek (1995) (BO).[4] LS and BO begin by defining a universe of candidate pure plays for each conglomerate segment. In BO, this universe is initially defined as all pure plays operating in the firm's four digit SIC industry. However, if the number of firms in this universe is less than five, then the pure plays in the given segment's three-digit industry are used. Finally, coarseness is increased to the two digit or even the one digit level until a universe of at least five pure plays is identified. Because changing the level of coarseness can alter the economic information contained in the benchmark (due to economies of scope or irrelevant peers), we exclusively use three-digit SIC industries as our starting point following the broader literature on industry analysis in Finance. However, we can report that using variable levels of coarseness as used in BO produce materially similar results.

The second step following BO's framework is to compute the firm value to sales ra-

[4] Many studies including Campa and Kedia (2002) and Villalonga (2004b) use a BO-based method.

tio for each pure play firm in each segment's universe, and then compute the median. The given segment's imputed value is then the segment's actual sales multiplied by this median ratio. Medians are used to reduce the impact of outliers, as firm value to sales ratios can become extreme, especially when firms have low sales or high growth options. Finally, the imputed value of the conglomerate firm is the sum of the imputed values of the given conglomerate's segments. Excess value is the natural logarithm of the conglomerate's imputed firm value divided by the conglomerate's actual firm value. This calculation can also be done using assets as an alternative to sales. A negative excess value, intuitively, suggests that the conglomerate is valued less than it might otherwise be valued if it were to operate under separate pure-play structures. We refer to this method as the "Berger+Ofek Baseline" method.

B Unconstrained Text-Based Methods

We note three key limitations of the LS and BO methods. A first is the equal treatment of all firms in a given segment's pure play universe in the median calculation. This assumption can reduce accuracy, as additional information exists regarding the nature of the products each pure play produces, and their comparability to a given conglomerate. Methods that weight more relevant pure plays more heavily should perform better. A second limitation is the use of SIC codes to identify the universe of relevant pure play benchmarks. Methods that enhance the set of pure plays beyond traditional SIC boundaries, if the additional pure plays are relevant, should perform better. A third limitation of the LS and BO method is the focus on a single accounting characteristic such as sales or assets. Candidate pure play firms likely vary along many other dimensions that can also explain valuation differences. For example, some pure plays might have very high sales growth, and might not be relevant as a benchmark for a given mature conglomerate. Henceforth, we refer to these three limitations as the "equal weighting limitation", the "limited universe limitation", and the "single characteristic limitation", respectively. Text-based methods offer a solution to all three limitations. In this section, we first examine vocabulary decompositions that directly address the first two limitations. We address the third limitation in the next section.

Although we consider many text-based methods, we adopt the approach of changing one degree of research freedom at a time. Our most basic text-based conglomerate reconstruction method therefore holds fixed the set of pure-play benchmarks used in BO (those in the same three-digit SIC code). However, we use a textual decomposition to determine which pure plays use product vocabulary that best matches that of the conglomerate. This decomposition provides us with a set of weights, which we use to replace the BO equal-weighted median calculation with a weighted median calculation. To determine the weights, we use least squares to decompose the business description of the conglomerate into parts observed in the pure play firms. Using the same notation from Section II, M_t denote the number of unique words in the corpus, i denotes a given conglomerate being reconstructed, t denotes the year of the given conglomerate observation, and $N_{i,t}$ is the conglomerate's (M_t x 1) normalized word vector. Further suppose that the given conglomerate-year observation has $N_{it,bench}$ candidate benchmark pure play firms to use in its reconstruction. Each benchmark has its own normalized word vector. Let $BENCH_{it}$ denote a (M_t x $N_{it,bench}$) matrix in which the normalized word vectors of the benchmark pure plays are appended as columns. We thus identify the set of pure play weights (w_{it}) that best explains the conglomerate's observed product market vocabulary as the solution to the following least squares problem.

$$\underset{w_{it}}{MIN}(N_{it} - BENCH_{it} \cdot w_{it})^2 \tag{2}$$

The solution to this problem (w_{it}) is simply the regression slopes associated with a no-intercept regression of the conglomerate's observed word usage N_{it} on the word usage vectors of the $N_{it,bench}$ pure plays. Importantly, unlike the BO method where pure plays are treated equally, this method assigns greater weight to pure plays whose product vocabulary best matches that of the conglomerate. Imputed value is therefore computed by first computing the weighted median value to sales ratio for over all $N_{it,bench}$ pure plays using the weights w_{it}. We then multiply the resulting value to sales ratio by the conglomerate's total sales to get the conglomerate's imputed value, and excess value is then equal to the natural logarithm of the imputed value to actual firm value ratio. We refer to this most basic text reconstruction, which addresses the "equal weighting limitation", as the "SIC Universe: Unconstrained" method.

We next consider an analogous method with a single enhancement that also addresses the "limited universe limitation". In this case, we add to the pure play universe by adding pure play firms that are in the conglomerate's VIC-7.06 industry as defined in Hoberg and Phillips (2010a). These firms have products that are similar to the conglomerate's product description, and the VIC-7.06 industry classification is equally as coarse as are SIC-3 industries. The calculation follows as described above, except in this case the number of benchmarks $N_{it,bench}$ is as large (if no pure play VIC-7.06 peers exist) or larger (if pure play VIC-7.06 peers do exist). We refer to this method as the "SIC+VIC Universe: Unconstrained" method.

C Constrained Text-Based Methods

We next consider the third limitation, the "single characteristic limitation". The LS and BO method has an underlying assumption that a single firm characteristic, for example sales or assets, is a sufficient statistic to explain a pure play's firm value. Because asset valuations are forward looking and depend on fundamentals (such as profitability), this limitation is quite severe. We consider a constrained least squares approach to construct a pure-play based imputed value that holds any number of accounting characteristics fixed to those of the conglomerate itself.

Using the same notation, suppose a conglomerate has $N_{it,bench}$ candidate pure play firms. Suppose the researcher identifies N_{char} accounting characteristics they wish to hold fixed when computing imputed valuations. In our case, we consider $N_{char} = 5$, and account for the following five accounting characteristics: Sales Growth, Log Age, OI/Sales, OI/Assets, and R&D/Sales. Let C_{it} denote a N_{char} x 1 vector containing the conglomerate's actual characteristics for these five variables. Let Z_{it} denote a $N_{it,bench}$ x N_{char} matrix in which one row contains the value of these five characteristics for one of the pure play benchmark candidates. We then consider the set of weights w_{it} that solve the following constrained optimization:

$$\underset{w_{it}}{MIN}(N_{it} - BENCH_{it} \cdot w_{it})^2 \text{ such that } \; Z'_{it}w_{it} = C_{it} \qquad (3)$$

The solution to this problem (w_{it}) is simply the slopes associated with a no-intercept constrained regression of the conglomerate's observed word usage N_{it} on the word usage vectors of the $N_{it,bench}$ pure plays. The closed form solution for the weights is:

$$w_{it} = (BENCH'_{it}BENCH_{it})^{-1}(BENCH'_{it}N_{it} - Z_{it}\lambda), \text{ where} \tag{4}$$

$$\lambda = [Z'_{it}(BENCH'_{it}BENCH_{it})^{-1}Z_{it}]^{-1}[Z'_{it}(BENCH'_{it}BENCH_{it})^{-1}BENCH'_{it}N_{it} - C_{it}]$$

Intuitively, this set of weights identifies the set of pure plays that use vocabulary that can best reconstruct the conglomerate's own vocabulary, and that also exactly match the conglomerate on the N_{char} characteristics. We refer to this method as the "SIC+VIC Universe: Constrained" method.

D Accounting for Segment Sales

The LS and BO method computes imputed values segment-by-segment, and therefore utilizes information contained in reported segment-by-segment sales. To the extent that sales explains valuations better than other characteristics, this information might be useful. The basic text-based methods described above do not use segment-by-segment sales, and instead rely on the weights obtained from the textual reconstruction to derive imputed value. We believe that it is an empirical question as to whether textual weights or sales weights best explain valuations. However, it is important to explore this question. We therefore consider a method that identical to the "SIC+VIC Universe: Constrained" method described above, except that we add an additional set of constraints based on the segment sales to ensure that the imputed value is weighted by sales across segments as is the case for the BO method.

Consider a conglomerate having $N_{it,seg}$ segments, and let S_{it} denote the $N_{it,seg}$ x 1 vector of sales weights (one element being a given segment's sales divided by the total sales of the conglomerate). To compute imputed values that impose segment sales-based weights, we make two modifications to the constrained optimization. First, we append the vector S_{it} to the vector C_{it}. Second, we create a $N_{it,bench}$ x $N_{it,seg}$ matrix of ones and zeros. A given element is one if the pure play associated with the given row is in the industry space corresponding to the given segment of the conglomerate associated with the given column. This matrix is populated based on how the pure-play benchmarks are selected. If the benchmark is selected due to its residing in a three digit SIC industry of a given segment, then the given pure play firm is allocated to that segment. If the benchmark was selected due to its residing in the VIC-7.06 industry of the conglomerate itself, then it is allocated to the segment whose SIC-benchmarks it is most similar (as measured using the cosine similarity method). We then append this $N_{it,bench}$ x $N_{it,seg}$ matrix of ones and zeros to the matrix Z_{it}. The solution to the resulting constrained optimization is a set of new weights w_{it} that has the property that the sum of weights allocated to each segment equals the given segment's sales divided by the total conglomerate sales ratio. Therefore, imputed values can be computed segment by segment. We refer to this method as the "SIC+VIC Universe: Constrained, Segment-by-Segment" method.

V Results: Conglomerate Valuation

In this section, we first assess the quality of conglomerate reconstruction using the several different reconstruction methods discussed earlier. We focus on the accuracy of valuation relative to the observed conglomerate valuations, and we also readdress the question regarding whether or not conglomerates trade at a discount relative to what they might trade at under a non-conglomerate structure. We conclude this section by examining hypotheses regarding which types of conglomerates have high or low valuations, and explore conglomerate valuations in cross section.

A Methodological Validation

Following the methodology discussion in Section IV, we examine excess valuations using five different conglomerate reconstruction methods. In particular, we consider the Berger and Ofek (1995) benchmark, and four text based methods aimed at addressing key limitations in the BO method. Table VI displays average excess valuations, and mean squared error statistics based on these five methods. Mean excess value calculations are useful to explore if conglomerates trade at discounts (negative excess valuations) or uremia (positive excess valuations), and mean squared error statistics are useful to compare the relative valuation accuracy of valuation methods. A method with a lower MSE generates excess valuations that are closer to the mean excess valuation, and are therefore more accurate. Following convention in the literature, we discard an excess value calculation if it is outside the range $\{-1.386, +1.386\}$ to reduce the affect of outliers. Therefore, the Observation counts available for each valuation method vary slightly. In particular, more accurate valuation methods generate excess valuations outside this range less often, and thus have higher observation counts. The table reports mean excess value, MSE statistics, and observation counts for excess value calculations based on sales (first three columns) and assets (last three columns).

Following conventions in the literature, we apply many screens to the conglomerate sample included in this part of our study. In particular, we require lagged COMPUSTAT data for our control variables, we drop firms with sales less than $20 million, firms with zero assets, and firms for which summed segment sales disagrees with the overall firm's sales by more than 1%. We also require that 10-K text data is available, and also that a sufficient number of pure play firms exist in segment industries to compute excess valuations. We are left with 6,225 observations of Berger and Ofek excess valuations, and 4,942 firms for which we can run our cross sectional excess value regressions with a full set of control variables.

[**Insert Table VI Here**]

Table VI shows that more refined text-based valuation methods generate smaller conglomerate discounts. For excess valuations based on sales, the 8.1% discount for the Berger and Ofek benchmark in row one declines to just 1.6% using the text-based method that addresses all three limitations. In particular, the most basic text-based benchmark, which holds fixed the same SIC-universe of pure play candidates, generates a modest reduction in the discount to 7.9%. Expanding the universe to include VIC-7.06 pure play rivals of the conglomerate reduces the discount to 4.9%, and holding fixed the five key accounting characteristics using the constrained model reduces the discount to 1.6%. In the final row, we see that further constraining the weights to match segment-specific sales ratios further reduces the discount to just 0.2%. However, unlike other enhancements, this last enhancement results in a loss of accuracy. When excess valuation is based on assets in the fourth column, we see that the discount of -2.5% using the Berger and Ofek benchmark declines analogously to +0.3% using the constrained text-based benchmark in row four.

Columns two and four, which report mean squared error statistics, strongly support the conclusion that the constrained model based on the enlarged SIC+VIC universe offers the most accurate conglomerate pricing. When based on sales, the mean squared error of .257 is 24.2% smaller than the mean squared error of .339 associated with the Berger and Ofek benchmark. When based on assets, this improvement is 22.8%. As observation counts using the constrained model are also highest, we further conclude that this model generates excess valuations outside the interval $\{-1.386, +1.386\}$ less often, further confirming its ability to value conglomerates more accurately. We conclude that improving conglomerate valuation accuracy, and matching benchmarks on the basis of both vocabulary usage and accounting variables known to explain valuations, both contribute to explaining the previously reported conglomerate discount. Our results therefore do not support the conclusion that conglomerate firms trade at discounts. These findings are in line with other recent studies that draw the same conclusion using other methods (see Campa and Kedia (2002), Villalonga (2004b), and Graham, Lemmon, and Wolf (2002)).

In Table VII, we assess whether conglomerates reconstructed using the various methods discussed above have similar characteristics as the conglomerates themselves. As the objective of these methods is to rebuild an identical replica of what the conglomerate would look like under a non-conglomerate structure, better benchmarks should match the conglomerate along more dimensions. For example, they should have similar sales growth, should be equally as mature, should be as profitable, and they should have similar expense structures.

To address this question, we first compute implied characteristic values using the same methods used to compute imputed valuations in the excess value valuations. For example, the implied Sales Growth of a Berger and Ofek (baseline) valuation is computed as the sales weighted average of the segment-by-segment computed median sales growth of the pure plays in each segment's three digit SIC industry. For a text-based benchmark, the weighted median sales growth is the implied sales growth of the conglomerate.

[Insert Table VII Here]

Table VII reports these correlations for each characteristic noted in the first column using each valuation method noted in the remaining columns. Comparing correlations using the Berger and Ofek benchmark to the other models reveals that the text-based benchmarks strongly outperform the Berger and Ofek baseline in terms of matching characteristics. The simplest text based methods that do not constrain accounting characteristics (columns two and three) have higher correlations than the Berger and Ofek benchmark. For example, the 26.9% correlation for the Berger and Ofek benchmark and oi/sales rises dramatically to 42.2% using unconstrained text-based weights. As indicated in the methodology section, these weights are purely a function of the vocabulary used by the pure plays and the conglomerate, and are not mechanistically related to the accounting numbers that these methods are better able to match. In the last two columns, not surprisingly, we observed that Pearson correlations rise dramatically when we use the text-based constrained optimization. As these weights use five key accounting characteristics to better fit each conglomerate's mapping, it is not as surprising that these weights are higher. We conclude that the text based measures offer many significant gains over existing methods.

B Determinants of Conglomerate Valuations

In this section, we examine whether conglomerate valuations vary in cross section. As discussed in our hypotheses section (Section I), we focus on examining whether conglomerates that face less competition have higher valuations relative to our pure-play based benchmarks. We explore this question in two ways. First, we use tools available in the existing literature and measure competition based on the average concentration of a conglomerate's pure play markets. In particular, we use the same method discussed in the previous section and compute the weighted median VIC-7.06 HHI (using firm-HHI data from Hoberg and Phillips (2010a)) of the pure plays used to reconstruct each conglomerate using the constrained text-based reconstruction.[5] Second, we consider the conglomerate as a whole, and ask how easily the

[5] Our results for this HHI variable are not sensitive to which reconstruction method is used to compute the conglomerate's HHI.

conglomerate can be reconstructed using the set of pure play firms that exist in its markets. The intuition here is that a conglomerate that is more difficult to replicate is more protected, and hence faces less of a competitive threat. For example, any asset complementarities or product market synergies created through its conglomerate structure cannot be easily raided by any new conglomerates that might form based on existing pure plays.

Our measure of how difficult a conglomerate is to replicate obtains directly as the R^2 from the constrained regression equation in equation (3). In particular, this constrained regression is run once per conglomerate-year observation, as this provided us with the weights used to construct the excess values discussed in the previous section. This same calculation thus provides our measure of how difficult each conglomerate is to reconstruct in each year. We explain a panel of conglomerate-year excess valuations using this "Difficulty of Pure Plays to Replicate" variable, along with our baseline concentration measure as discussed above, and our across and within industry similarity measures. We also include controls for document length, vertical relatedness, and a number of accounting measures used in the existing literature.

[**Insert Table VIII Here**]

Table VIII displays the results of OLS panel data regressions in which one observation is one conglomerate in one year, and the dependent variable is the excess valuation using the constrained text based valuation method (Panel A) and the Berger and Ofek (1995) valuation method (Panel B). t-statistics are shown in parentheses, and standard errors are adjusted for clustering by firm.

Our first key finding is that the difficulty of pure plays to replicate variable is positive and highly statistically significant in both panels. This is our main result. Conglomerates that are harder to replicate have high valuations relative to pure play benchmarks. As this variable captures the uniqueness of the conglomerate's products relative to the pure play benchmarks, one would not expect its affect on valuation to be negated out in the difference as was the case for the average HHI variable. This finding, which is robust at the 1% level of significance in all rows, is consistent with these firms earning higher rents due to the inability of other firms to enter their product markets. Hence the product market synergies or asset complementarities that the given firm enjoys under the conglomerate structure are not vulnerable. Our control variables indicate that conglomerates are also valued more when they have more investment (R+D and Capital Expenditures), when they are more profitable, and when they are larger. Conglomerates are also less valuable when their segments are vertically related.

We find the reported R2s are higher in Panel B than in Panel A. This result arises because our text-based valuation model is a better fit as shown previously than the Berger and Ofek method. There exists less unexplained cross-sectional variation, our dependent variable, in Panel A than in Panel B. In general the significance levels of the key variable, difficult of pure plays to replicate the conglomerate, are however quite similar across panels.

Other findings in the table is that traditional segment-by-segment average within segment similarity and the average concentration ratio (Conglomerate Average Concentration)are not significantly related to excess valuations in the full model in Row (1), although they are both significant when the first Difficulty to Replicate variable is excluded in rows (3) and (5). The result is not surprising because the pure play firms used to construct the excess valuation benchmark enjoy the same level of concentration on average. Therefore, the excess valuation, which is a difference, would come close to negating any effect of this concentration variable on average.

Table IX displays the economic magnitudes of our findings regarding the difficulty of pure plays to replicate variable. In each year, we sort firms into quintiles based on this variable, and we compute the average excess valuation for each group. We also compute the average residual excess valuation, where residuals are from a regression of excess valuation on all of the variables in Table VIII with the exception of the difficulty to replicate variable. The table shows that raw excess valuations are modestly higher for the highest quintile (+4.3% using the text-based model) relative to the lowest quintile (-1.2%). This effect is magnified for average residual excess valuations (+9.6% versus -4.7%). We conclude that the impact of a conglomerate's difficulty to reconstruct is meaningful, and that conglomerates that are more difficult to replicate trade at modest premia relative to their pure play benchmarks.

[Insert Table IX Here]

VI Conclusions

We use text-based analysis of conglomerate and pure play business descriptions from 10-Ks filed with the SEC to examine in which industries conglomerates are most likely operate and to understand cross-sectional conglomerate valuation. We find that conglomerate firms are more likely to operate in industry pairs that are closer together in the product space, in industry pairs that have profitable opportunities "between" them, and in industries with lower within industry product similarity. These findings are consistent with product synergies, from related industry production and also from conglomerates producing in two related profitable industries being able to enter profitable related industries.

We also find that conglomerate firms are less likely to produce in industries with high within industry similarity and in industries that span competitive industries. These findings are consistent with conglomerate firms choosing to produce in the more concentrated industries with higher profitability.

We examine the cross-sectional valuation effects of conglomerate industry production. Using text-based analysis we redefine benchmark single-industry segment "pure-play" firms for each industry segment of conglomerate firms. Our methodology does not just identify pure-play benchmarks. It uses text based analysis to weight these pure-play benchmarks to match the conglomerate firm on multiple accounting characteristics in addition to product word similarity. Our analysis allows us to better understand the cross-sectional variation in conglomerate valuation premia and discounts.

We find that on average conglomerates do not trade at a discount relative to text-matched single segment firms. However, this average effect masks important cross-sectional variation. We find that conglomerates that are more difficult to reconstruct using pure-play firms tend to trade at modest premia and those conglomerates that are easier to replicate trade at small discounts. These findings are consistent with higher valued conglomerate firms producing in related industries that have product synergies and in industries that are also harder to enter for pure-play firms.

References

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

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

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

Comment, Robert, and Gregg Jarrell, 1995, Corporate focus and stock returns, *Journal of Financial Economics* 37, 61–87.

Custodio, Claudia, 2010, Mergers and acquisitions accounting can explain the diversification discount, Arizona State University Working Paper.

Denis, David, Diane Denis, and Atulya Sarin, 1997, Agency problems, equity ownership and corporate diversification, *Journal of Finance* 52, 135–160.

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

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

Hann, Rebecca, Maria Ogneva, and Oguzhan Ozbas, 2011, Corporate diversification and the cost of capital, University of Maryland and University of Southern California Working Paper.

Hoberg, Gerard, and Gordon Phillips, 2010a, New dynamic product based industry classifications and endogenous product differentiation, University of Maryland Working Paper.

——— , 2010b, Competition and product market synergies in mergers and acquisitions: A text based analysis, forthcoming Review of Financial Studies.

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.

Laeven, Luc, and Ross Levine, 2007, Is there a diversification discount in financial conglomerates?, *Journal of Financial Economics* 85, 331–367.

Lamont, Owen, and Christopher Polk, 2002, Does diversification destroy value? evidence from the industry shocks, *Journal of Financial Economics* 63, 51–77.

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

Lins, Karl, and Henri Servaes, 1999, International evidence on the value of corporate diversification, *Journal of Finance* 54, 2215–2240.

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).

——— , 2008, The industry life-cycle, acquisitions and investment: Does firm organization matter?, *Journal of Finance* 63, 673–709.

Matsusaka, John, 2001, Corporate diversification, value maximization, and organizational capabilities, *Journal of Business* 74, 409–431.

Rajan, Raghuram G., Henri Servaes, and Luigi Zingales, 2000, The cost of diversity: the diversification discount and inefficient investment, *Journal of Finance* 55, 35–80.

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.

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

Servaes, Henri, 1996, The value of diversification during the conglomerate merger wave, *Journal of Finance* 51, 1201–1225.

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

Villalonga, Belen, 2004a, Diversification discount or premium? new evidence from business information tracking series, *Journal of Finance* 59, 479–506.

——— , 2004b, Does diversification cause the diversification discount, *Financial Management* 33, 5–27.

Wernerfelt, Birger, and Cynthia Montgomery, 1988, Diversification, ricardian rents, and tobin's q, *Rand Journal of Economics* 19, 623–632.

Table I: Summary Statistics

Summary statistics are reported for our sample of firms (Panel A), Industry Pairs (Panel B), and Conglomerate Segment Pairs (Panel C) for our sample from 1996 to 2008. The variables in Panel A include the VIC-7.06 HHI and the total firm value (book debt plus market value of equity). The variables in Panel B include product market measures describing an industry pair. The **Number of Conglomerates Spanning Pair** is the number of conglomerates having segments in both industries associated with the given pair. **Across Industry Similarity** is the average pairwise similarity between firms in one of the industries in the pair, and firms in the other industry. To compute "between" variables, let i and j denote the two industries comprising the given industry pair observation. Let k denote a third industry under consideration. Industry k is "between" i and j if (1) the average pairwise distance between firms in industry i and k is less than the average pairwise distance between firms in industry i and j, and (2) the average pairwise distance between firms in industry j and k is less than the average pairwise distance between firms in industry i and j. The **Fraction of Industries Between Pair** is the fraction of all other SIC-3 industries residing in the product market space "between" the two industries comprising the pair. **Zero Industries Between** is a dummy equal to one if no industries are between i and j. To compute **Average Within Industry Similarity**, we first compute the average pairwise similarity of firms in industry i. We recompute this quantity for industry j. Average Within Industry Similarity is the average of the two. The **Average HHI** is computed analogously by averaging the VIC-7.06 HHI of firms in each industry, and taking the average of the two. The **Pair Likelihood if Random** is a control variable equal to the fraction of all pure play firms in industry i, multiplied by the fraction of all pure play firms in industry j (multiplied by 10,000 for convenience). The **Same 2-digit SIC Dummy** is a dummy equal to one of industries i and j share the same two digit SIC code. **Vertical Relatedness** is the average fraction of input the two industries in an industry pair obtain from one another (from the input-output tables). The variables in Panel C identify changes in conglomerate structures using the Compustat segment definitions and the SDC acquisition database. One observation is a pair of segments in an existing conglomerate in year t, and we require that the conglomerate exist in year t and $t+1$.

		Std.			
Variable	Mean	Dev.	Minimum	Median	Maximum
Panel A: Conglomerates (15,373 obs) and Pure-Plays (56,491 obs)					
Firm Value (Conglomerates)	12430	48462	0.483	1228	1036340
VIC HHI (Conglomerates)	0.140	0.219	0.006	0.059	1.000
Firm Value (Pure-Plays)	2450	18863	0.003	215.	1038648
VIC HHI (Pure-Plays)	0.111	0.153	0.006	0.058	1.000
Panel B: Industry Pair Variables (312,240 obs)					
Number of Conglomerates Spanning Pair	0.147	0.855	0.0	0.0	57.0
Across Industry Similarity	0.017	0.010	0.000	0.014	0.169
Fraction of Industries Between Pair	0.325	0.257	0.000	0.267	0.992
Zero Industries Between Dummy	0.012	0.107	0.000	0.000	1.000
Within Industry Similarity	0.086	0.038	0.000	0.081	0.433
Average HHI	0.118	0.060	0.015	0.107	0.611
Pair Likelihood if Random	0.198	1.405	0.001	0.031	119.3
Same 2-digit SIC Dummy	0.018	0.133	0.000	0.000	1.000
Vertical Relatedness	0.003	0.014	0.000	0.000	0.536

Variable	Obs	Percentage	Std. Dev.
Panel C: Change in Conglomerate Segment Pair Variables (32,181 obs)			
Segment Pair Disappears	4,566	14.2%	34.9%
Segment Pair Likely Sold or Closed	3,415	10.6%	30.8%
Segment Pair Likely Reclassified	1,096	3.4%	18.1%
Segment Pair Likely Sold Off	330	1.0%	10.1%

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. The variables include various measures of the product market topography between the industry pair, and within the industries comprising the pair. Please see Table I for a description of the variables.

Row	Variable	Number of Spanning Conglom. Pairs	Across Industry Similarity	Zero Industries Between Dummy	Fraction of Industries Between	Within Industry Similarity	Average HHI	Pair Likelihood if Random	Same 2-digit SIC Dummy
				Correlation Coefficients					
(1)	Across Industry Similarity	0.229							
(2)	Zero Industries Between Dummy	0.160	0.446						
(3)	Fraction of Industries Between Pair	-0.132	-0.691	-0.137					
(4)	Within Industry Similarity	-0.044	0.184	0.058	-0.092				
(5)	Average HHI	-0.011	-0.176	-0.042	0.088	-0.487			
(6)	Pair Likelihood if Random	0.144	-0.009	0.020	-0.002	-0.020	-0.031		
(7)	Same 2-digit SIC Dummy	0.231	0.315	0.200	-0.135	-0.030	0.020	0.012	
(8)	Vertical Relatedness	0.200	0.165	0.078	-0.124	-0.049	0.055	0.028	0.155

Table III: Conglomerate Summary

Summary statistics showing various mean characteristics across various subsamples of industry pairs from 1996 to 2008. Industries are based on three-digit SIC industries. Results are based on our sample of 312,240 industry pair x year permutations, and 40,769 observed conglomerate industry pair x year observations. In Panel A, we display summary statistics for all observed conglomerate pairs, and we compare them to the statistics of randomly drawn industry pairs. In Panel B, we display summary statistics for conglomerates of varying size. In Panel C, we show results for conglomerates that are growing, stable, or shrinking, as noted in the first column. In Panel D, we show results for vertically integrated segments, and segments in the same two-digit SIC code. Please see Table I for a description of the variables displayed.

	Across	Within		Fraction of	
	Industry	Industry	Average	Industries	
Sub Sample	Similarity	Similarity	HHI	Between	# Obs.
Panel A: Overall					
All Conglomerates	0.0296	0.0768	0.1150	0.1293	40,769
Randomly Drawn SIC-3 Industries	0.0167	0.0862	0.1183	0.3255	312,240
Panel B: By Conglomerate Size					
2 Segments	0.0341	0.0738	0.1192	0.0867	6,365
3 Segments	0.0311	0.0750	0.1164	0.1132	11,672
4-5 Segments	0.0289	0.0786	0.1130	0.1366	15,794
6+ Segments	0.0247	0.0785	0.1133	0.1790	6,938
Panel C: Shrinking, Stable, and Growing Conglomerates					
Shrink by 2+ Segments	0.0268	0.0788	0.1097	0.1490	600
Shrink by 1 Segment	0.0295	0.0779	0.1119	0.1296	3,259
Stable Conglomerate	0.0301	0.0769	0.1160	0.1260	30,525
Add 1 Segment	0.0282	0.0760	0.1117	0.1414	4,741
Add 2+ Segments	0.0262	0.0739	0.1135	0.1485	1,644
Panel D: Vertical and Same SIC-2 Conglomerates					
Vertically Related Segments	0.0319	0.0717	0.1212	0.0739	15,007
Same SIC-2 Segments	0.0471	0.0829	0.1085	0.0291	8,015

Table IV: Where Conglomerates Exist

OLS regressions with 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 pairings of observed SIC-3 industries in the given year in the COMPUSTAT segment tapes. The dependent variable is the **Number of Conglomerates Spanning Pair**, which is the number of conglomerates 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. Panel C displays results based on subsamples divided on the basis of both valuations and competitiveness. The independent variables include various measures of the product market topography between the industry pair, and within the industries comprising the pair. Please see Table I for a description of the independent variables. Panel regressions are estimated with year fixed effects and standard errors are clustered by year (t-statistics are in parentheses).

Row	Sample	Across Industry Similarity	Fraction of Industries Between Pair	Zero Industries Between	Avg. Within Industry Similarity	Average HHI	Pair Likelihood if Random	Same 2-digit SIC Code	Vertical Relatedness	# Obs. / RSQ
				Panel A: Full Sample						
(1)	All Industry Pairs	14.060	0.060	0.410	-1.347		0.084	0.943	8.669	312,240
		(19.98)	(4.85)	(6.20)	(-13.90)		(9.32)	(18.94)	(7.00)	0.128
(2)	All Industry Pairs	12.809	0.045	0.423		0.181	0.085	0.973	8.869	312,240
		(18.94)	(3.45)	(6.40)		(4.05)	(9.42)	(19.14)	(7.13)	0.125
				Panel B: Univariate Subsamples						
(3)	Concentrated Industry Pair	27.374	0.249		-1.034		0.086	0.638	3.715	154,324
		(11.15)	(6.32)		(-18.74)		(6.65)	(8.57)	(7.56)	0.110
(4)	Competitive Industry Pair	12.730	-0.050		-1.625		0.076	1.044	8.033	154,321
		(16.76)	(-1.89)		(-11.18)		(6.07)	(20.31)	(7.67)	0.103
(5)	High Firm Value Industry	21.110	0.190		-1.260		0.063	1.199	5.695	154,326
		(12.96)	(5.51)		(-11.80)		(6.19)	(19.43)	(4.14)	0.100
(6)	Low Firm Value Industry P	11.380	-0.010		-1.453		0.120	0.743	8.491	154,319
		(15.52)	(-1.43)		(-10.86)		(5.91)	(12.26)	(12.13)	0.124
				Panel C: Bivariate Subsamples						
(7)	Concentrated and High Val	38.414	0.425		-0.865		0.066	0.779	3.207	65,904
		(6.04)	(4.29)		(-12.71)		(3.99)	(5.53)	(4.11)	0.113
(8)	Competitive and High Value	19.416	0.160		-1.534		0.062	1.294	6.165	88,422
		(12.14)	(3.11)		(-10.37)		(5.93)	(16.84)	(3.93)	0.097
(9)	Concentrated and Low Value	22.061	0.146		-1.153		0.113	0.595	3.937	88,420
		(8.88)	(4.20)		(-13.12)		(4.06)	(7.94)	(7.39)	0.114
(10)	Competitive and Low Value	8.544	-0.258		-1.813		0.124	0.817	10.600	65,899
		(9.38)	(-14.40)		(-9.73)		(4.70)	(13.45)	(8.77)	0.127

Table V: Which Conglomerates Split

Logit regressions with standard errors clustered by year for our sample of 32,181 industry pairs from 1997 to 2008. One observation is one pair of segments in an existing conglomerate in year t. We require the conglomerate firm itself to exist in year t and year $t + 1$. The dependent variable varies by Panel. In Panel A, the dependent variable is **Segment Pair Disappears**, which is a dummy equal to one if the given pair does not exist in the conglomerate's structure in the following year. In Panel B, the dependent variable is **Segment Pair Like Sold or Closed**, which is a dummy equal to one if the given pair does not exist in the conglomerate's structure in the following year, and the conglomerate has fewer segments in year $t + 1$ relative to year t. In Panel C, the dependent variable is **Segment Pair Likely Reclassified**, which is a dummy equal to one if the given pair does not exist in the conglomerate's structure in the following year, and the conglomerate has at least as many segments in year $t + 1$ relative to year t. In Panel D, the dependent variable is **Segment Pair Like Sold Off**, which is a dummy equal to one if the given pair does not exist in the conglomerate's structure in the following year, and the conglomerate was the target of an acquisition of at least ten percent of its assets in year $t + 1$. Please see Table I for a description of the independent variables. All regressions are estimated with year fixed effects and standard errors are clustered by year (t-statistics are in parentheses).

			Fraction	Avg.	Pair	Same		
		Across	Industries	Within	Likeli-	2-digit	Vertical	
		Industry	Between	Industry	hood if	SIC	Relat-	Obs.
Row	Sample	Similarity	Pair	Simil.	Random	Code	edness	/RSQ
			Panel A: Dep. Var = Segment Pair Disappears					
(1)	All Pairs	-6.557	0.282	0.362	0.004	-0.043	-1.724	32,181
		(-2.94)	(1.75)	(0.32)	(0.67)	(-1.23)	(-4.13)	0.015
(2)	Concen. + High Value	-17.662	-0.047	-2.029	0.013	-0.146	-0.213	7,387
		(-2.83)	(-0.15)	(-1.20)	(1.53)	(-1.16)	(-0.17)	0.015
(3)	Compet. + High Value	-10.653	-0.120	0.498	-0.004	-0.135	-2.493	6,976
		(-3.45)	(-0.13)	(0.31)	(-0.43)	(-2.15)	(-2.27)	0.024
(4)	Concen. + Low Value	-15.653	0.024	0.259	0.006	-0.131	1.076	8,706
		(-2.40)	(0.09)	(0.13)	(0.51)	(-1.50)	(0.60)	0.011
(5)	Compet. + Low Value	-3.574	2.919	0.896	-0.014	0.005	-1.192	5,636
		(-0.78)	(1.33)	(0.68)	(-0.69)	(0.09)	(-2.17)	0.013
			Panel B: Dep. Var = Segment Pair Likely Sold or Closed					
(6)	All Pairs	-8.521	-0.004	1.166	0.008	-0.137	-1.692	32,181
		(-3.14)	(-0.02)	(0.98)	(1.17)	(-2.37)	(-3.48)	0.009
(7)	Concen. + High Value	-22.566	-0.507	-1.110	0.015	-0.424	0.124	7,387
		(-3.18)	(-2.03)	(-0.50)	(1.67)	(-2.43)	(0.19)	0.011
(8)	Compet. + High Value	-14.508	-1.009	0.065	-0.003	-0.221	-2.499	6,976
		(-2.46)	(-0.99)	(0.04)	(-0.24)	(-1.91)	(-1.99)	0.016
(9)	Concen. + Low Value	-14.934	-0.262	1.662	0.009	-0.198	1.336	8,706
		(-2.09)	(-0.94)	(1.01)	(0.83)	(-2.92)	(0.91)	0.007
(10)	Compet. + Low Value	-6.072	1.303	2.356	-0.004	-0.082	-1.312	5,636
		(-0.88)	(0.67)	(1.22)	(-0.20)	(-0.91)	(-1.55)	0.007
			Panel C: Dep. Var = Segment Pair Likely Reclassified					
(11)	All Pairs	0.755	0.945	-2.311	-0.015	0.179	-1.317	32,181
		(0.20)	(2.78)	(-1.27)	(-1.44)	(3.40)	(-1.47)	0.011
(12)	Concen. + High Value	-1.966	1.155	-3.873	-0.002	0.415	0.383	7,387
		(-0.20)	(1.98)	(-1.81)	(-0.16)	(1.73)	(0.11)	0.017
(13)	Compet. + High Value	-0.059	2.191	1.589	-0.008	0.141	-1.939	6,976
		(-0.01)	(1.58)	(0.52)	(-0.43)	(1.04)	(-1.63)	0.012
(14)	Concen. + Low Value	-15.880	0.860	-4.737	-0.009	-0.011	0.853	8,706
		(-1.18)	(1.54)	(-1.44)	(-0.30)	(-0.05)	(0.26)	0.011
(15)	Compet. + Low Value	1.570	5.747	-4.486	-0.072	0.154	-0.346	5,636
		(0.20)	(1.95)	(-1.50)	(-1.99)	(0.85)	(-0.66)	0.013
			Panel D: Dep. Var = Segment Pair Likely Sold Off					
(16)	All Pairs	-2.326	0.186	-0.229	0.004	0.085	0.120	32,181
		(-0.25)	(0.54)	(-0.07)	(0.29)	(0.31)	(0.12)	0.004
(17)	Concen. + High Value	-27.734	-0.305	-8.540	0.021	0.655	2.888	7,387
		(-1.61)	(-0.73)	(-1.18)	(2.51)	(1.26)	(0.71)	0.009
(18)	Compet. + High Value	11.600	0.115	1.604	0.010	-0.385	-2.983	6,976
		(0.84)	(0.04)	(0.37)	(0.33)	(-0.97)	(-0.80)	0.006
(19)	Concen. + Low Value	2.023	-0.258	3.912	-0.009	0.046	0.726	8,706
		(0.15)	(-0.24)	(0.77)	(-0.26)	(0.15)	(0.21)	0.007
(20)	Compet. + Low Value	-0.911	10.004	-0.353	-0.552	0.147	0.945	5,636
		(-0.04)	(2.43)	(-0.07)	(-1.15)	(0.39)	(0.39)	0.004

Table VI: Quality of Excess Valuation Calculations Across Methods

This table displays comparative summary statistics regarding conglomerate valuations and valuation accuracy across several different methods for computing conglomerate valuations. All of the conglomerate valuation methods we consider are based on reconstructions of a conglomerate firm using the valuation ratios of existing pure play firms operating in the same industries as each segment. A conglomerate's excess value is the natural logarithm of its firm value divided by the implied firm value using the pure play reconstruction. Following convention in this literature, we discard an excess value calculation if it is outside the range $\{-1.386, +1.386\}$ to reduce the affect of outliers. The **Excess value** column reports the average sample-wide excess valuation using the valuation method reported in the first column. The **MSE Excess Value** is the mean squared error of excess valuations using the given valuation method (lower values indicate more accurate valuations). The **Observation** counts column reports the number of conglomerates used to compete the average and MSE excess value. Observation counts vary slightly because more accurate valuation methods produce valuations going outside the permissible range $\{-1.386, +1.386\}$ less often. We report these three columns using excess valuation metrics computed using sales to value ratios (first three columns) and asset to value ratios (second three columns). The final column, **Standard Deviation of Weights** is computed for the text-based valuation methods, where the text is used to compute differential weights for the pure play firms used to compute excess values. For a detailed description of the valuation methods, please see Section IV. We provide a basic description here. The first method, **Berger+Ofek Baseline** is a replication of the calculation used in Berger and Ofek (1995), where each segment is valued by computing the median firm value to sales ratio of pure play firms operating in the three digit SIC code of each segment, and then multiplying this median by the segment's reported sales. Adding these implied segment valuations gives the overall conglomerate's implied value and is the key benchmark compared to the actual conglomerate firm value used to compute excess valuation. The **HP: SIC Universe: Whole Firm, Unconstrained** uses text-based weights to reconstruct the conglomerate. The median firm value to sales ratio is computed using a weighted median calculation, where the weights are given by the text decomposition regression (conglomerate vocabulary is decomposed into the text of the available pure plays to construct a more precise product market replica). The **HP: SIC+VIC Universe: Whole Firm, Unconstrained** method extends this method by expanding the set of available pure plays for the text decomposition regression to include pure plays residing in the same VIC-7.06 industry as the conglomerate. The **HP: SIC+VIC Universe (wf): Whole Firm, Constrained** extends the method further using constrained regression, where the best-fit text-based reconstruction uses constrained regression methods to require that the reconstructed conglomerate matches the actual conglomerate on five key characteristics: Sales Growth, Log Age, OI/Sales, OI/Assets, and R&D/Sales. The **HP: SIC+VIC Universe: Constrained, Segment-by-Segment** method is analogous, but also requires that the pure plays allocated to each segment contribute to total sales of the reconstructed firm according to the actual sales ratios of the conglomerate. Note that the Berger+Ofek method is by definition Unconstrained, as the benchmark it creates does not not attempt to match the conglomerate on any characteristics beyond sales.

Row	Benchmark	Excess Value (Sales Based)	MSE Excess Val. (Sales based)	# Obs. (Sales based)	Excess Value (Assets Based)	MSE Excess Val. (Assets based)	# Obs. (Assets based)	Std. Dev. Weights
1	Berger+Ofek Baseline:	-0.081	0.339	6225	-0.025	0.224	5611	
2	HP: SIC Universe: Unconstrained	-0.079	0.343	6234	-0.037	0.222	5663	0.036
3	HP: SIC+VIC Universe: Unconstrained	-0.049	0.312	6321	-0.010	0.205	5676	0.025
4	HP: SIC+VIC Universe: Constrained	-0.016	0.257	6426	0.003	0.173	5688	0.047
5	HP: SIC+VIC Universe: Constrained, Segment-by-Segment	-0.002	0.280	6326	0.043	0.211	5619	0.059

Table VII: Characteristic Correlations (Conglomerate vs. Benchmark)

The table displays Pearson Correlation coefficients between actual conglomerate characteristics and implied characteristics using several different conglomerate valuation methods. The characteristic being analyzed is identified in the first column, and the remaining columns present correlations using the valuation methods noted in the column headers. Implied characteristics using each method are computed using the same weighting scheme used to compute the excess valuations. For example, the implied Sales Growth of a Berger and Ofek (baseline) valuation is computed as the sales weighted average of the segment-by-segment computed median sales growth of the pure plays in each segment's three digit SIC industry. For a text-based benchmark, the weighted median (using text reconstruction weights) sales growth is the implied sales growth of the conglomerate. Higher correlations imply that the reconstructed conglomerate matches the true conglomerate on characteristics. This information allows us to compare reconstruction methods using information beyond valuation alone. Importantly, the last three columns are based on constrained text regressions where the aim is to have the reconstructed conglomerate match the actual conglomerate on five key characteristics: Sales Growth, Log Age, OI/Sales, OI/Assets, and R&D/Sales. Intuitively, correlations for these variables jumps in these latter three columns. These correlations are not 100% because the conglomerate reconstruction is based on a weighted median calculation and not a weighted average. We use weighted medians as is the convention in the literature as this mitigates the impact of highly skew value to sales ratios in the valuation reconstruction.

Row	Variable	Berger + Ofek (Baseline)	Text-based SIC only No Constr.	Text-based SIC+VIC No Constr.	Text-based SIC+VIC Constrained	Text-based SIC+VIC Constrained (Seg by Seg)
				Correlation Coefficients		
1	Tobin's Q	0.353	0.449	0.468	0.566	0.558
2	Sales Growth	0.261	0.298	0.330	0.845	0.807
3	Log Age	0.289	0.282	0.397	0.924	0.906
4	OI/Sales	0.268	0.379	0.422	0.880	0.862
5	OI/Assets	0.293	0.360	0.441	0.883	0.862
6	SG&A/Sales	0.503	0.566	0.605	0.768	0.742
7	COGS/Sales	0.463	0.502	0.539	0.747	0.734
8	CAPX/Sales	0.476	0.510	0.527	0.681	0.637
9	R&D/Sales	0.369	0.640	0.673	0.770	0.758
10	Advertising/Sales	0.211	0.330	0.378	0.327	0.303
11	Market Leverage	0.417	0.422	0.478	0.507	0.470
12	Book Leverage	0.372	0.389	0.429	0.447	0.414
13	Sales	0.109	0.205	0.307	0.370	0.360
14	Assets	0.134	0.212	0.277	0.381	0.333

Table VIII: Conglomerate Excess Valuations

The table displays OLS regressions with time fixed effects. One observation is one conglomerate form 1997 to 2008. Observations having operations in financial industries (SIC-3 600 to 699), and observations lacking adequate data to compute the independent variables, are excluded. The dependent variable is the conglomerate's excess valuation using the best text-based reconstruction (Panel A) or using the Berger and Ofek reconstruction (Panel B) as the dependent variable. The best text-based reconstruction is the "HP: SIC+VIC Universe: Constrained" model as illustrated in Table VI. The independent variables include five text based variables, and four control variables used in the literature (R&D/Sales, CAPX/Sales, OI/Sales, and Log Assets). All variables are winsorized at the 1%/99% level . The **Difficulty of Pure Plays to Replicate** variable is one minus the R^2 from the text decomposition regressions used to rebuild each conglomerate. Conglomerates for which this is high are difficult to reconstruct using pure plays and are likely using asset complementarities to generate more unique products. **Across Segment Similarity** is the average textual similarity of pure play firms operating in the same three digit SIC segments as the conglomerate. When this is high, the conglomerate's segments reside in product market locations that are closer together. **Within Segment Similarity** is the average similarity of pure play firms operating within the industries occupied by the given conglomerate. **Conglomerate Average Concentration** is the weighted average VIC-7.06 HHI of all firms used to reconstruct the given conglomerate. **Log Document Length** is a control variable equal to the natural logarithm of the number of words in the given conglomerate's product description. All standard errors are adjusted for clustering by firm.

Row	Difficulty of Pure Plays to Replicate	Across Segment Similarity	Within Segment Similarity	Conglom. Average Concentration	Log Document Length	Vertical Relatedness	R&D/ Sales	CAPX/ Sales	OI/ Sales	Log Assets	# Obs. / RSQ
					Panel A: Excess Value (Text-based Constrained Valuation Model)						
(1)	0.303	-0.000	-0.583	0.083	-0.040	-0.634	1.214	0.709	0.627	0.044	4,972
	(4.14)	(-0.00)	(-1.49)	(0.24)	(-1.43)	(-2.71)	(5.20)	(7.74)	(6.54)	(6.78)	0.107
(2)	0.325				-0.049	-0.637	1.275	0.680	0.619	0.043	4,972
	(4.61)				(-1.81)	(-2.81)	(5.53)	(7.52)	(6.46)	(6.68)	0.105
(3)		-1.158			-0.075	-0.551	1.203	0.678	0.599	0.041	4,972
		(-1.59)			(-2.78)	(-2.33)	(5.14)	(7.48)	(6.17)	(6.29)	0.095
(4)			-0.883		-0.066	-0.609	1.141	0.692	0.607	0.041	4,972
			(-2.48)		(-2.45)	(-2.64)	(4.86)	(7.60)	(6.24)	(6.33)	0.097
(5)				0.626	-0.067	-0.615	1.191	0.676	0.596	0.042	4,972
				(1.94)	(-2.37)	(-2.69)	(5.12)	(7.43)	(6.14)	(6.42)	0.096
					Panel B: Excess Value (Berger + Ofek Valuation Model)						
(6)	0.433	2.025	-1.537	-0.426	0.067	-0.967	2.321	0.788	1.018	0.065	4,814
	(5.49)	(1.81)	(-3.17)	(-1.05)	(2.05)	(-3.42)	(8.59)	(6.19)	(7.72)	(7.73)	0.183
(7)	0.430				0.060	-0.869	2.403	0.782	1.003	0.065	4,814
	(5.45)				(1.91)	(-3.10)	(9.00)	(6.36)	(7.61)	(7.87)	0.178
(8)		-0.111			0.017	-0.830	2.300	0.737	0.966	0.060	4,814
		(-0.11)			(0.52)	(-2.84)	(8.52)	(5.82)	(7.32)	(7.21)	0.162
(9)			-1.331		0.039	-0.829	2.169	0.804	0.986	0.062	4,814
			(-3.07)		(1.20)	(-2.91)	(7.96)	(6.58)	(7.51)	(7.48)	0.167
(10)				0.263	0.022	-0.836	2.289	0.745	0.967	0.061	4,814
				(0.66)	(0.66)	(-2.91)	(8.51)	(6.04)	(7.33)	(7.29)	0.162

Table IX: Economic Magnitudes and Excess Valuation

This table displays average excess valuation statistics for quintiles based on the difficulty of pure plays to replicate variable. The **Difficulty of Pure Plays to Replicate** is one minus the R^2 from the text decomposition regressions used to rebuild each conglomerate. Conglomerates for which this is high are difficult to reconstruct using pure plays and are likely using asset complementarities to generate more unique products. For each quintile, we report the average difficulty variable, and average raw excess valuations based on both the text-based and Berger and Ofek methods in the first three columns. The best text-based reconstruction is the "HP: SIC+VIC Universe: Constrained" model as illustrated in Table VI. The residual excess valuations are residuals from a regression of excess valuation on all of the variables included in Table VIII excluding the Difficulty to Replicate variable. These residual excess valuations thus reflect the conditional impact of the difficulty to replicate on the excess valuation.

Difficulty to Replicate Quintile	Difficulty to Replicate	Raw Excess Valuation (text-based)	Raw Excess Valuation (Berger+Ofek)	Residual Excess Valuation (text-based)	Residual Excess Valuation (Berger+Ofek)	Obs.
			Summary Statistics by Quintile			
Lowest Difficulty	0.791	-0.012	0.004	-0.047	-0.016	1,221
Quintile 2	0.866	-0.010	-0.102	-0.015	-0.047	1,229
Quintile 3	0.906	-0.016	-0.122	0.012	-0.013	1,228
Quintile 4	0.947	-0.030	-0.165	-0.003	-0.063	1,229
Highest Difficulty	1.162	0.043	-0.020	0.096	0.138	1,224