

Conglomerate Industry Spanning

Gerard Hoberg and Gordon Phillips*

August 21, 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 to 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 and that span high value industries tend to trade at modest premia. The conglomerates that are most easy to replicate trade at discounts relative to matched pure-play firms. These findings are consistent with conglomerate firms generating product synergies when producing in related profitable industries not spanned by single-segment firms.

*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 Denis Sosyura and seminar participants at Rotterdam School of Management, Erasmus University, Stanford University, Tilberg University, University of Illinois and the University of Mannheim for helpful comments. All errors are the authors alone. Copyright ©2011 by Gerard Hoberg and Gordon Phillips. All rights reserved.

Why do firms operate in multiple industries? This question has been the focus of much research that takes the industries that firms operate in 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 use fundamental product market characteristics to analyze in which industries conglomerate firms produce. We examine whether conglomerate firms may create value through operating in related industries that allow for synergies, new products to be created and potential entry into new markets. 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 to operate and to explore conglomerate valuations in cross-section. 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 allow measurement of how close product markets are to each other. Our framework also allows an assessment of which in-

dustries in the 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 the industry pair is to each other. 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 between them, and in industries with lower within-industry product similarity. Conglomerate firms are also less likely to produce in industries that span other 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 industries that are between them.

We use text-based analysis to assess conglomerate valuation. Our primary contribution is to redefine the universe of competitor single-segment firms used to construct conglomerate valuation benchmarks, and then to use best-fit vocabulary matching of 10-K business descriptions of single-segment firms to conglomerate firms assigning differential weights to each single-segment firm. We also consider enhanced benchmarks that additionally seek to match the conglomerate to single-segment firms along five key accounting dimensions including profitability and growth. We present strong evidence that our weighted benchmarks provide economically large improvements in the accuracy of conglomerate reconstructions.

Once we have these best matches, we then seek to explain differences between the valuation of the matched firms and the actual conglomerate valuation. Our primary focus is then on understanding cross-sectional variation in conglomerate excess valuations - not in assessing the average discount or premium of conglomerate firms.¹ We find that, on average, conglomerates do not trade at a discount relative

¹For articles on average or median discount of conglomerate firms see 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). 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.

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 discounts. We also find that conglomerate valuation is higher when conglomerates span high-value industries. These findings are consistent with conglomerate firms producing in related industries that have product synergies when these segments are combined, and in industries that are more difficult to enter for competing single-segment firms.

The main contributions of our paper are two-fold. Our first 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 and to capture potential product differentiation. We also calculate pairwise 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 between their operating segments.

Our second contribution is to further explain cross-sectional conglomerate valuation. We use text-based analysis to form a more accurate matched set of pure-play competitors such that each firm and each segment has its own allocated set of competitors. This new “network” method of competition draws on a firm- and segment-centric notion of competition analogous to a Facebook circle of friends. In order to find competitor firms, we use the text 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 similar firms receive a higher weight based on text-based distances. Both the set of weights and overall goodness of fit of the pure-play competitors in matching the conglomerate firm 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 (which is derived based on the R^2 from the vocabulary decomposition) is a more accurate measure of the conglomerate's overall competitive position. Firms that cannot be easily replicated are highly differentiated, and likely face less competition for their overall product offerings. We find that conglomerate firms that are more difficult to replicate and that span high-value industries are more valuable than conglomerates that are easier to replicate and span low-value industries.

We add to the existing literature in two central ways. First, we examine which industries conglomerates are more likely to operate within and we are the first to explore the link between conglomerate structure and the overall industry structure surrounding the each conglomerate. Previous research has examined the investment decisions of conglomerate firms including Shin and Stulz (1998), Rajan, Servaes, and Zingales (2000), Scharfstein and Stein (2000) and Maksimovic and Phillips (2002). Maksimovic and Phillips examine how long-run industry characteristics effect acquisition decisions by conglomerate and single-segment firms. Hann, Ogneva, and Ozbas (2011) show that producing in different industries lowers a firm's cost of capital consistent with a coinsurance effect. We examine which industry factors explain the distribution of conglomerate industry segments.

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 main contribution is to show that the discount varies with how the firm is positioned with respect to its competitors and industry characteristics, especially the ability to easily replicate a conglomerate using pure plays.

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 is whether the potential for product market synergies and industry competition influence in which industries conglomerate firms produce. Our foundation is similar to that of Rhodes-Kropf and Robinson (2008) and Hoberg and Phillips (2010) who examine whether asset complementarities and synergies are important for mergers. Going beyond the previous analysis, we also consider the role of industry opportunities and competition in surrounding and between industries.

We consider whether the industries a given industry pair "spans" influences whether conglomerate firms produce within that industry pair. Industry spanning is the idea that there may be a product space or industry that is more similar to each industry in a given industry pair than the two industries in the pair are similar to each other. In a spatial sense, such an industry lies between the two industries in the pair. We hypothesize that a conglomerate is more likely to produce in a particular industry pair if that pair spans other highly valued, less competitive industries. Producing in such an industry pair may allow the multiple industry firm to more easily enter and produce products in these highly-valued concentrated product markets.

We generate industry pair characteristics using text-based analysis of business descriptions from 10-Ks filed with the SEC. We then examine these industry characteristics to understand in which industries conglomerates are most likely to 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 use to capture how industries are related to each other. We consider these new measures in addition to existing industry-relatedness measures including 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. To construct industry relatedness measures, we begin by using the relatedness of each pairwise set of firms that operate either within an industry or across any given industry pair. These product relatedness measures are constructed for each pair of firms using the words from each firm’s business description from their 10K filed with the SEC. We discuss the specific word relatedness measures we use - the cosine similarity measure - later in the subsequent data and methodology section.

One crucial part of our analysis is to identify industry membership. Given that we will be comparing our text-based measures with existing methods, we begin by using the existing Compustat segment-level industry designations to identify which firms are conglomerates and to assess conglomerate industry configurations. To construct industry attributes, we focus on single-segment firms and omit the words from the product descriptions from the conglomerate firms themselves, to avoid producing a mechanistic relation. We then use the conglomerate firm’s Compustat segment definitions to examine how the observed conglomerate industry configurations and the words conglomerate firms use in their product description relate to these text-based industry attributes from single-segment firms. Because conglomerate segments are reported using SIC codes, our initial analysis relates to industry configurations and their incidence is based on three-digit SIC code industry definitions. In later analysis, we relax this initial reliance on SIC-3 industry definitions and also include firms identified as competitors using text-based methods when exploring conglomerate valuations and benchmark reconstructions.²

The first industry relatedness variable we calculate is *Across Industry Similarity (AIS)*. This measures how close industry i and industry j are in the product space.

²Due to the high dimensionality of the calculations underlying the initial industry spanning tests, we do not consider alternative industry definitions for this part of our analysis as doing so would render our calculations infeasible due to the exponential growth in the number of industry pairwise permutations.

Suppose industry i has N_i pure play firms, and industry j has N_j pure play firms. Across industry similarity is the average textual cosine similarity of all pairwise permutations of the N_i and N_j firms, where textual similarity is based on word vectors from firm business descriptions (see Section II.C for a discussion of the cosine similarity method). Simply put, it captures the proportion of product words the two firms in a pair have in common. Each pair in a permutation includes one firm from industry i and one from industry j and we average the cosine similarity measures across all such pairs from the two industries. The AIS measure is meant to capture the similarities between products that two industries produce. Industries that are closer together share similar sets of words in their constituent firm product descriptions and are potentially likely to share asset complementarities.

Second, we measure *Within Industry Similarity (WIS)*. Suppose industry i has N_i pure play firms. Within industry similarity is the average cosine similarity of the business descriptions for all pairwise permutations of these N_i firms. Firms in industries with higher within industry similarity are likely to have less unique products, and likely face more significant competition from their rivals due to the absence of product differentiation. When assessing this measure for a given industry pair, we compute WIS as the average within industry similarity of the two industries i and j .

Third, we also measure the fraction of other industries between a pair of industries i and j , which we label *Between Industries*. Because our text based measures give all firms, and hence all industries, a unique location on a high dimensional unit sphere, we can use geometry to assess whether other industries lie in the space between a given industry pair. This novel measure can be used to examine whether conglomerates benefit from business opportunities lying between their segments, perhaps through asset complementarities. The AIS measure discussed above 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.

$$AIS_{k,i} \leq AIS_{i,j} \quad \text{AND} \quad AIS_{k,j} \leq AIS_{i,j} \quad (1)$$

The fraction of industries between a given pair of industries i and j is therefore the number of industries k (excluding i and j) satisfying this condition divided by the total number of industries in the database in the given year (excluding i and j). We also consider a dummy variable identifying industry pairs that have zero industries lying between them.

We use these measures to 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. These firms are likely easier to manage and have more potential synergies.

We test this hypothesis by examining the number of conglomerate firms that operate in each pairwise set of industries and examine whether this number of conglomerate firms is increasing in the pair's across industry similarity (AIS).

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

We test this hypothesis by examining whether the number of conglomerate firms operating in a pair is decreasing when the industry pair has a high within industry similarity (WIS).

H3: Between-Industry Spanning: Conglomerate firms are more likely to operate in an industry pair when the pair of industries spans more high-value, less competitive industries.

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

H4: Conglomerate Valuations and Spanning: Conglomerate valuation will increase the harder the conglomerate firm is to replicate with pure-play single-segment firms. Conglomerate valuation will increase relative to benchmark valuation if it spans industries that are highly valued.

II Data and Methodology

In this section we describe our conglomerate database, the construction of key text-based variables used to examine where conglomerates produce in the product space, and our identification of pure-play, single-segment conglomerate competitors.

A The COMPUSTAT Industry Sample

We construct our COMPUSTAT sample using the industrial annual files to identify the universe of publicly traded firms, and the COMPUSTAT segment files to identify which firms are 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 (2010). 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.

C Word Vectors and Cosine Similarity

We employ methods used in Hoberg and Phillips (2010a) and Hoberg and Phillips (2010) to construct word vectors and measure similarity. The first step is to form word vectors for each firm based on the text in product descriptions of each firm.

To construct each firm’s word vector, we first omit common words that are used by more than 25% of all firms. Following Hoberg and Phillips (2010a), we further restrict our universe in each year to words that are either nouns or proper nouns.⁴ Let M_t denote the number of such words. For a firm i in year t , we define its word

³We 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.

⁴We identify nouns using Webster.com as words that can be used in speech as a noun. We identify proper nouns as words that appear with the first letter capitalized at least 90% of the time in the corpus of all 10-K product descriptions. Previous results available from the authors did not impose this restriction to nouns. These results were qualitatively similar.

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 Conglomerate Competitors

We use our text-based analysis of firms to redefine and augment traditional identification of conglomerate competitors, which allows us to construct new measures of excess conglomerate valuation. As a baseline calculation, we begin by following the existing literature (Lang and Stulz (1994) and Berger and Ofek (1995)) and we consider existing pure-play single-segment firms as identified the Compustat segment tapes as the set of competitors for each segment of a conglomerate. We refer to this set of competitors as the Compustat set of rivals, and this set includes single-segment firms with the same 3-digit SIC code as the reported Compustat segment. We then consider several enhancements. First, rather than computing valuation ratios as the simple median of all candidate single-segment rivals for a given conglomerate segment, we weight each single segment firm using information from a vocabulary decomposition of the conglomerate relative to the vocabulary used by all pure play

candidate rivals in their business descriptions. Firms with more words in common with the conglomerate will thus have higher weights.

Our second augmentation is to expand the set of “pure-play” single-segment rivals by adding the single-segment firms that are in the conglomerate’s “Text-Based Network Industry Classification” (TNIC) industry as defined in Hoberg and Phillips (2010a). These firms have a high product word cosine similarity score relative to the conglomerate firm and are likely rivals. Importantly, the TNIC industry classification is equally as coarse as are SIC-3 industries, so our results are not due to any changes in industry coarseness. The number of potential competitor benchmarks for each conglomerate segment $N_{it,bench}$ is thus either as large (if no pure play TNIC peers exist) or larger (if pure play TNIC peers do exist) than the Compustat set of rivals. We refer to this method as the “SIC+TNIC Universe” universe.

After identifying the potential competitors for each segment, we again weight the potential firms based on the vocabulary decomposition to construct conglomerate valuation benchmarks and excess valuation measures. We also consider further enhancing the weighting scheme of the set of potential competitors so that pure play benchmarks can match the conglomerate on accounting characteristics including profitability and sales growth. We discuss these valuation metrics and how we use them to construct excess valuation measures for each conglomerate firm fully in Section IV when we examine the cross-sectional and time-series dimensions of conglomerate valuation.

E Conglomerate Restructuring

We examine whether our spatial industry variables can explain how conglomerates restructure over time, 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 con-

glomerate’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 Likely 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 $t + 1$. We define an analogous set of variables to assess conglomerate segment pairs that newly appear in year t that were not part of the conglomerate in year $t - 1$.

F 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.

G Summary Statistics

Table I displays summary statistics for our conglomerate and pure play firms, and industry pair databases. Panel A shows that conglomerate firms are generally larger than the 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 TNIC HHI.

Panel B of the table compares randomly drawn pairs of SIC-3 industries to the SIC-3 industries comprising a conglomerate configuration. The panel shows that a randomly drawn pair of three digit SIC industries has 0.147 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 of random pairs is 0.017, which closely matches the average firm similarity reported in Hoberg and Phillips (2010a). This quantity is nearly double for actual conglomerates at 0.032, indicating that conglomerates are far less diversified than previously thought. This conclusion is reinforced by comparing the fraction of all other industries lying between the given pair, which is 32.5% for random pairs, and just 9.7% for actual conglomerates. Conglomerate industry pairs are in regions of the product space that are substantially closer together than randomly chosen industries. The average within industry similarity, intuitively, is much higher at 0.086. This quantity is somewhat lower at .073 for actual conglomerates.

[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 table 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 produce in a wider range of the product market space, as they have lower across industry similarity. They also tend to produce in industries with more industries between them, and industries that have higher within industry similarity. In Panel C of Table III, we observe that most conglomerates (30,525) are stable from one year to the next, although 3,259 of them reduce in size by one segment, and

600 conglomerates reduce in size by two or more segments. Analogously, 4,741 firms increase in size by one segment, and 1,644 firms increase in size by two segments.

In Panel D, we observe that vertically related 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. Be-

cause 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 are added to conglomerates in a given year. We consider raw segment additions for growing or stable conglomerates, and we also consider the SDC mergers and acquisitions database. This allows us to separately consider segments likely added through growth, or those potentially acquired in a transaction. One observation is one pair of segments in an existing conglomerate in year t . We require the conglomerate firm itself to exist in year t and year $t + 1$.

The dependent variable varies by Panel. The dependent variable in Panel A is the **Number of Newly Added Conglomerates Spanning Pairs**, which is the number of conglomerates having new segments in both industries associated with a given pair in a given year (where the conglomerate did not have this segment in the previous year). In Panel B, we restrict attention to new segments in conglomerates that previously had fewer segments in the previous year. Intuitively, these new segments were likely added through acquisition or organic investment. In Panel C, we restrict attention to new segments in conglomerates that were the acquirer in an acquisition in the SDC database for a transaction amounting to at least ten percent of the firm's assets. The independent variables include various measures of the product market features of the industry pair, and within the industries comprising the pair.

[Insert Table V Here]

The results in Panel A of Table V show that segment pairs are likely to be added 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. We also see that conglomerate firms are more likely to add segments when the fraction of industries between the conglomerate pair is high and the average within-industry similarity is low. These findings are present especially in concentrated and highly-valued industry pairs.

The results in Panels B and C further show that conglomerate segments are more likely to be added through growth or acquisition when concentrated and highly valued industries lie between the segment pairs. In particular, conglomerate firms add such segments when the resulting industry pairs have high across industry similarity, a high fraction of industries between the industry pair, and low average within industry similarity.

The results in this table are consistent with conglomerates choosing to expand into industries where (a) they face less competition, (b) have more high valued industries lying between the industry pair, (c) have high cross industry similarity, and (d) low within industry similarity. The results are consistent with conglomerates choosing to expand into industries with the potential for new products and related-industry synergy gains.

Table VI examines how industry characteristics influence which industry pairs disappear from conglomerates. Using raw segment changes and the SDC mergers and acquisitions database, we further examine segment pairs that likely were sold or closed. 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 again 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 Likely 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 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 VI Here]

The results in Panels A and B of Table VI 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.

Row (7) in Panel B further shows that segments are less likely to be sold or closed when there are high value concentrated industries lying between the given segment and the other segments of the conglomerate. This supports the conclusion that conglomerates see added value (such as the possibility of low cost entry) from the industries that lie between their segments.

Given these strong results on which industries conglomerate firms choose to produce in and their link to potential strategic gains relating to concentration and industry valuation, we now turn to the question of how industry composition affects the cross-sectional variation in conglomerate valuation.

IV Conglomerate Valuation

In this section, we explore whether information in firm product descriptions can be used to construct more informative pure-play or single-segment benchmarks. We consider both product market identification and the weighting method of single-segment

firms when reconstructing conglomerate benchmarks based on pure play firms. Following the existing literature, we then compare actual conglomerate valuations to the valuation of our pure play benchmarks.

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).⁵ 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 ratio 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

⁵Many studies including Campa and Kedia (2002) and Villalonga (2004b) use a BO-based method.

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 denotes 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 ($M_t \times 1$) 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 ($M_t \times N_{it,bench}$) 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.

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

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 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 TNIC industry as defined in Hoberg and Phillips (2010a). These firms have products that are similar to the conglomerate’s product description, and the TNIC 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 TNIC peers exist) or larger (if pure play TNIC peers do exist). We refer to this method as the “SIC+TNIC 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 $N_{char} \times 1$ vector containing the conglomerate’s actual characteristics for these five variables. Let Z_{it} denote a $N_{it,bench} \times N_{char}$ 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:

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

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:

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

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

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+TNIC 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 is identical to the “SIC+TNIC 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 $N_{it,seg} \times 1$ 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 $N_{it,bench} \times N_{it,seg}$ 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 TNIC 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 $N_{it,bench} \times N_{it,seg}$ 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+TNIC 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 VII 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 premia (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 in Panels B and C, we discard an excess value calculation if it is outside the range $\{-1.386, +1.386\}$ (in actual levels instead of natural logs this range is $\{\frac{1}{4}, 4\}$), to reduce the effect 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. In Panel A of VII we include all observations. In Panel B, following the convention in the literature, we restrict attention to excess values less than 400% and greater than 25% (screen applied separately for each method). In Panel C, we omit all firm-years for a conglomerate in which its estimated excess value is outside this range using any calculation method we consider (as this allows a comparison that holds the sample size fixed).

[Insert Table VII Here]

Panel A of Table VII shows that more refined text-based valuation methods generate smaller conglomerate discounts. For excess valuations based on sales, the 8.2% discount for the Berger and Ofek benchmark in row one declines to just 1.2% using the text-based method that addresses all three limitations. The most basic text-based benchmark, which holds fixed the same SIC-universe of pure play candidates, results in a decline in the excess value discount to 5.8%. Therefore, just changing the weighting of single segment firms alone is partially but not fully responsible for our ability to explain the discount. Row 3 of Panel A expands the universe to include TNIC pure play rivals of the conglomerate. This expansion reduces the discount to 4.6%. Finally, using the five key accounting characteristics in Row 4 reduces the discount to 1.2%. In row 5 of Panel A, we see that further constraining the weights to match segment-specific sales ratios increases the discount to just 1.8%.

When excess valuation is based on assets in the fourth column, we see that the discount of -2.7% using the Berger and Ofek benchmark declines analogously to nearly zero (0.1%) using the constrained text-based benchmark in row four. We conclude that our ability to explain the benchmark is due to three factors: (1) Using weights based on textual decompositions, (2) improving the benchmark candidates

to include both SIC and TNIC peers, and (3) constraining the benchmark to have similar accounting characteristics relative to the conglomerate being reconstructed.

Columns two and four, which report mean squared error statistics, strongly support the conclusion that the constrained model based on the enlarged SIC+TNIC universe offers the most accurate conglomerate pricing. When based on sales, the mean squared error in row 4 of .320 is 32.4% smaller than the mean squared error of .474 associated with the Berger and Ofek benchmark. When based on assets, this improvement is 27.7%.

In Panels B and C, we omit excess valuations outside the interval $\{-1.386, +1.386\}$. Panel B omits just the method-specific conglomerate-year observations in which excess valuations are outside this range and Panel C omits the firm if any of the five valuation method places the value outside this range. The results are similar to Panel A. We see the discount in excess value disappearing using our text-based methods. In Panel C in particular, the excess value discount entirely disappears for both the sales based and the asset based methods. We also see large decreases in mean squared error using our text-based methods. The results in Panel C are especially clean because the sample size is held fixed across methods.

We conclude that improving conglomerate benchmarks alone can explain the previously reported conglomerate discount, and dramatically improve valuation accuracy. The intuition behind this result squares well with the original intent: a portfolio of pure plays that matches the conglomerate in operations and assets should be a valid benchmark to the conglomerate itself. 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 VIII, 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 investment intensities.

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) benchmark 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 VIII Here]

Table VIII reports correlations between the actual conglomerate characteristic and the implied characteristic for each characteristic noted in the first column using each valuation method noted in the remaining columns. Comparing correlations between the single-segment constructed benchmarks and the actual conglomerate firms using the Berger and Ofek benchmark to the text-based benchmarks 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 constructed conglomerate benchmark. For example, the 28.9% correlation between the OI/Assets of the actual conglomerate and the Berger and Ofek benchmark increases dramatically to (35.7% to 42.1%) even using unconstrained text-based weights. As indicated in the methodology section, the text-based 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 surprising that these characteristic correlations are higher. We conclude that text based measures offer substantial improvements over existing methods.

It is also natural to ask which type of pure play firms are weighted more than others when reconstructing conglomerates and giving differential weights to component single-segment firms. Panel A of Table IX explores this question and displays average characteristics for firms assigned weights in the highest and lowest quartile using the text-based conglomerate benchmarks. Panel B further examines how the weights on the single-segment benchmark firms vary when we examine how well the vocabularies of the pure plays fit the vocabulary of the conglomerate (we define the “Difficulty of Pure Plays to Replicate the Conglomerate” as one minus the R^2 from the vocabulary decomposition).

The first three columns of Panel A are based on the “SIC+TNIC universe (unconstrained)” method. This method is text based and uses an enhanced set of eligible pure plays (SIC and TNIC peers) to reconstruct the conglomerates. In the second three columns in Table IX, we repeat the same exercise using the “SIC+TNIC universe (constrained)” method, which also holds fixed key accounting variables as discussed earlier.

[Insert Table IX Here]

Panel A shows that pure play firms receiving higher weights using text decompositions tend to be older, are more mature firms, and have lower sales growth. These firms also have less research and development, and are more profitable than those pure plays assigned lower weights. Because mature firms have lower valuation ratios, this might explain why conglomerates appear undervalued using earlier methods.

The results in the latter three columns are similar to those in the first three columns, but notably sharper. For example, the average difference in age is nearly 7.5 years using the constrained text method, compared to just 4.4 years using the unconstrained text method. We conclude that equal weighting all pure plays, as was done using the Berger and Ofek benchmark, will overweight high growth firms and thus generate the inappropriate conclusion that conglomerates are undervalued. Our results in the next section formally confirm this conjecture.

Panel B Table IX provides a similar comparison of characteristics but splits the

sample based on which firms have above or below median “Difficulty of Pure Plays to Replicate” measures ($1 - R^2$ from the textual decomposition regression in equation (3)). A conglomerate that is difficult to replicate has an R^2 of zero and a “Difficulty of Pure Plays to Replicate” value of 1. This constrained regression is run once per conglomerate-year, as this provides us with the weights used to construct the excess values year by year as discussed in the previous section. This same calculation thus provides one difficulty to replicate statistic for each conglomerate in each year..

Panel B shows that conglomerates with concentrated industries and high-value industries between their segments have the sharpest correlations between difficulty to replicate and the fractions of industries lying between the pair. This finding adds to our earlier evidence that conglomerates spanning these high value industries generating more product market synergies, as these same synergies can explain why the pure plays cannot replicate the conglomerate as easily.

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 are harder to replicate and face less competition have higher valuations relative to our pure-play based benchmarks. To explore this question, we regress conglomerate excess valuation on the text-based variables that capture these factors.

We regress conglomerate-year excess valuations on the “Difficulty of Pure Plays to Replicate” variable, along with the fraction of industries between the conglomerate industry segments, our concentration measures 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.

Our variable for the “Difficulty of Pure Plays to Replicate” captures how easily the conglomerate can be reconstructed using the set of pure play firms that exist in its markets. The intuition underlying this calculation is that a conglomerate that

is more difficult to replicate is more protected and differentiated, 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.

The fraction of industries between the conglomerate industry segment captures the potential gains a conglomerate can reap by expanding its product offerings into these between markets. Average across industry similarity captures the potential for synergies between industries. Within industry similarity and the average concentration of a conglomerate's pure play markets capture the competition that a conglomerate firm faces within its pure-play markets.

[Insert Table X Here]

Table X 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. 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 also find that the reported R^2 s 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. If reconstructed benchmarks are perfect, fewer variables should be capable of explaining valuations in the cross section. The table also shows that the level of significance of our key variable, difficulty of pure plays to replicate, is quite similar across panels.

Other findings in the table are 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 (6), although they are both significant when the Difficulty to Replicate variable is excluded in rows (2) and (4). These results are 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 the effect of this concentration variable on average.

Table XI 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 X with the exception of the difficulty to replicate variable. The table shows that raw excess valuations are modestly higher for the highest quintile (+5.4% using the text-based model) relative to the lowest quintile (-2.3%). This effect is magnified for average residual excess valuations (+9.1% versus -4.8%). 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 XI Here]

Our last table examines conglomerate excess valuation for subsamples based on which industries conglomerate firms operate within. Panel A of Table XII considers conglomerates with high-value industries (above median) between their segment pairs, and Panel B examines conglomerate excess value for conglomerates with low-value industries (below median) between their segment pairs.

[Insert Table XII Here]

The results in Panel A, as compared to Panel B, show that the “Difficulty of Pure Plays to Replicate” variable is significant in both panels but significantly larger for high-value industries. We can also see that the fraction of industries between is positive and significant for high-value industries but insignificant for conglomerates with low-value industries between their industry pairs. As in Table X, the difficulty to replicate variable subsumes other product market variables, as this is a comprehensive measure of how differentiated the conglomerate is from potential rivals in all of its markets.

Overall these results, combined with those in Table X, are consistent with conglomerates having higher valuations when their products are difficult to replicate with the best possible single-segment peers. Our results are also consistent with conglomerates trading at a discount when their products are very similar to those produced by single-segment peers. The results are consistent with conglomerates trading at premia and adding value when they are able to generate synergies and produce products that are not produced by single-segment peers.

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 to 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 related industries being able to enter profitable industries that lie between them.

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 using product-word vocabulary matching. We also use text based analysis to weight these pure-play benchmarks to match the conglomerate firm on multiple accounting characteristics in addition to product word matching.

We find that on average conglomerates do not trade at a discount relative to text-matched single segment firms. More importantly, 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 discounts. These findings are consistent with higher valued conglomerate firms producing in related highly-valued industries that have product synergies.

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, 2010, Product market synergies in mergers and acquisitions: A text based analysis, *Review of Financial Studies* 23, 3773–3811.
- , 2010a, Text-based network industry classifications and endogenous product differentiation, University of Maryland Working Paper.
- 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).
- Matusaka, 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, *ACMCS* 34, 1–47.
- Servaes, Henri, 1996, The value of diversification during the conglomerate merger wave, *Journal of Finance* 51, 1201–1225.
- Shin, Hyun, and Rene Stulz, 1998, Are internal capital markets efficient?, *Quarterly Journal of Economics* 113, 531–552.
- Stein, Jeremy, 1997, Internal capital markets and the competition for corporate resources, *Journal of Finance* 52, 111–133.
- 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 conglomerate and pure play firms (Panel A), Industry Pairs and conglomerates (Panel B), and Conglomerate Segment Pairs (Panel C) for our sample from 1996 to 2008. The variables in Panel A include the TNIC 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. The **Fraction of Industries Between Pair** is the fraction of all other SIC-3 industries residing in the product market space “between” or closer in product space than 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 and j, and average the two measures. The **Average HHI** is computed analogously by averaging the “Text-Based Network Industry Classification” (TNIC) 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 if 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.

| Variable | Mean | Std. Dev. | Minimum | Median | Maximum |
|---|-------|-----------|---------|--------|---------|
| <i>Panel A: Conglomerates (15,373 obs) and Pure-Plays (56,491 obs)</i> | | | | | |
| Firm Value (Conglomerates) | 12430 | 48462 | 0.483 | 1228 | 1036340 |
| Firm Value (Pure-Plays) | 2450 | 18863 | 0.003 | 215. | 1038648 |
| TNIC HHI (Conglomerates) | 0.140 | 0.219 | 0.006 | 0.059 | 1.000 |
| TNIC HHI (Pure-Plays) | 0.111 | 0.153 | 0.006 | 0.058 | 1.000 |
| <i>Panel B: Industry Pairs (312,240 obs) and Conglomerates (15,373 obs)</i> | | | | | |
| Number of Conglomerates Spanning Pair (Ind. Pairs) | 0.147 | 0.855 | 0.0 | 0.0 | 57.0 |
| Across Industry Similarity (Ind. Pairs) | 0.017 | 0.010 | 0.000 | 0.014 | 0.169 |
| Across Industry Similarity (Conglomerates) | 0.032 | 0.019 | 0.000 | 0.025 | 0.138 |
| Fraction of Industries Between Pair (Ind. Pairs) | 0.325 | 0.257 | 0.000 | 0.267 | 0.992 |
| Fraction of Industries Between Pair (Conglomerates) | 0.097 | 0.133 | 0.000 | 0.042 | 0.992 |
| Within Industry Similarity (Ind. Pairs) | 0.086 | 0.038 | 0.000 | 0.081 | 0.433 |
| Within Industry Similarity (Conglomerates) | 0.073 | 0.030 | 0.010 | 0.066 | 0.188 |
| Same 2-digit SIC Dummy (Ind. Pairs) | 0.018 | 0.133 | 0.000 | 0.000 | 1.000 |
| Same 2-digit SIC Dummy (Conglomerates) | 0.228 | 0.371 | 0.000 | 0.000 | 1.000 |
| Vertical Relatedness (Ind. Pairs) | 0.003 | 0.014 | 0.000 | 0.000 | 0.536 |
| Vertical Relatedness (Conglomerates) | 0.027 | 0.066 | 0.000 | 0.006 | 0.536 |

| Variable | Obs | Percentage | Std. Dev. |
|--|-------|------------|-----------|
| <i>Panel C: Change in Conglomerate Segment Pair Variables (32,181 obs)</i> | | | |
| 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 | Aver- age HHI | Pair Likelihood if Random | Same 2-digit SIC Dummy |
|---------------------------------|-------------------------------------|--|----------------------------------|--|--------------------------------------|----------------------------------|---------------------|------------------------------------|---------------------------------|
| <i>Correlation Coefficients</i> | | | | | | | | | |
| (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.

| Sub Sample | Across Industry Similarity | Within Industry Similarity | Average HHI | Fraction of Industries Between | # Obs. |
|--|----------------------------|----------------------------|-------------|--------------------------------|---------|
| <i>Panel A: Overall</i> | | | | | |
| 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 |
| <i>Panel B: By Conglomerate Size</i> | | | | | |
| 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 |
| <i>Panel C: Shrinking, Stable, and Growing Conglomerates</i> | | | | | |
| 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 |
| <i>Panel D: Vertical and Same SIC-2 Conglomerates</i> | | | | | |
| 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 based on the competitiveness or the valuations of industries lying between the industry pair. Panel C displays results based on subsamples of valuations and competitiveness. The independent variables include various measures of the product market features of 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 |
|---------------------------------------|-----------------------------------|----------------------------|-------------------------------------|-------------------------|---------------------------------|-----------------|---------------------------|-----------------------|----------------------|------------------|
| <i>Panel A: Full Sample</i> | | | | | | | | | | |
| (1) | All Industry Pairs | 14.060 (19.98) | 0.060 (4.85) | 0.410 (6.20) | -1.347 (-13.90) | | 0.084 (9.32) | 0.943 (18.94) | 8.669 (7.00) | 312,240 0.128 |
| (2) | All Industry Pairs | 12.809 (18.94) | 0.045 (3.45) | 0.423 (6.40) | | 0.181 (4.05) | 0.085 (9.42) | 0.973 (19.14) | 8.869 (7.13) | 312,240 0.125 |
| <i>Panel B: Univariate Subsamples</i> | | | | | | | | | | |
| (3) | Concentrated Industry Pairs | 27.374 (11.15) | 0.249 (6.32) | | -1.034 (-18.74) | | 0.086 (6.65) | 0.638 (8.57) | 3.715 (7.56) | 154,324 0.110 |
| (4) | Competitive Industry Pairs | 12.730 (16.76) | -0.050 (-1.89) | | -1.625 (-11.18) | | 0.076 (6.07) | 1.044 (20.31) | 8.033 (7.67) | 154,321 0.103 |
| (5) | High Firm Value Industry Pairs | 21.110 (12.96) | 0.190 (5.51) | | -1.260 (-11.80) | | 0.063 (6.19) | 1.199 (19.43) | 5.695 (4.14) | 154,326 0.100 |
| (6) | Low Firm Value Industry Pairs | 11.380 (15.52) | -0.010 (-1.43) | | -1.453 (-10.86) | | 0.120 (5.91) | 0.743 (12.26) | 8.491 (12.13) | 154,319 0.124 |
| <i>Panel C: Bivariate Subsamples</i> | | | | | | | | | | |
| (7) | Concentrated and High Value Pairs | 38.414 (6.04) | 0.425 (4.29) | | -0.865 (-12.71) | | 0.066 (3.99) | 0.779 (5.53) | 3.207 (4.11) | 65,904 0.113 |
| (8) | Competitive and High Value Pairs | 19.416 (12.14) | 0.160 (3.11) | | -1.534 (-10.37) | | 0.062 (5.93) | 1.294 (16.84) | 6.165 (3.93) | 88,422 0.097 |
| (9) | Concentrated and Low Value Pairs | 22.061 (8.88) | 0.146 (4.20) | | -1.153 (-13.12) | | 0.113 (4.06) | 0.595 (7.94) | 3.937 (7.39) | 88,420 0.114 |
| (10) | Competitive and Low Value Pairs | 8.544 (9.38) | -0.258 (-14.40) | | -1.813 (-9.73) | | 0.124 (4.70) | 0.817 (13.45) | 10.600 (8.77) | 65,899 0.127 |

Table V: New Conglomerate Segments

OLS regressions which examine the number of new conglomerate segments in each three-digit SIC codes pairs in the given year in the COMPUSTAT segment tapes. Panel A counts the number of new conglomerates producing in both industries of a given three-digit pair. In Panel B, new segments are from conglomerates that previously had fewer segments in the previous year. Panel C restricts new segments to conglomerates that were the acquirer in an acquisition in the SDC database for a transaction amounting to at least ten percent of the firm's assets. The independent variables include product market features of the industry 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 Industries Between Pair | Avg. Within Industry Simil. | Pair Likelihood if Random | Same 2-digit SIC Code | Vertical Relatedness | Obs. /RSQ |
|---|----------------------|----------------------------|----------------------------------|-----------------------------|---------------------------|-----------------------|----------------------|------------------|
| <i>Panel A: Dep. Var = New Segment Pairs</i> | | | | | | | | |
| (1) | All Industry Pairs | 2.409 (5.56) | 0.010 (3.23) | -0.268 (-4.41) | 0.017 (3.86) | 0.124 (4.69) | 0.756 (2.61) | 312,240 0.052 |
| (2) | Concen. + High Value | 6.016 (4.13) | 0.065 (3.11) | -0.160 (-4.78) | 0.016 (2.50) | 0.130 (2.61) | 0.519 (1.96) | 65,904 0.051 |
| (3) | Concen. + Low Value | 2.348 (7.02) | -0.009 (-0.84) | -0.337 (-4.49) | 0.014 (3.42) | 0.155 (5.85) | 0.541 (2.08) | 88,422 0.046 |
| (4) | Compet. + High Value | 3.051 (4.15) | 0.021 (2.63) | -0.176 (-4.39) | 0.019 (3.56) | 0.076 (2.58) | 0.758 (2.91) | 88,420 0.038 |
| (5) | Compet. + Low Value | 1.416 (3.16) | -0.036 (-3.69) | -0.341 (-3.75) | 0.024 (2.77) | 0.111 (3.60) | 0.829 (2.45) | 65,899 0.049 |
| <i>Panel B: Dep. Var = New Segment Pairs Likely Obtained through Growth</i> | | | | | | | | |
| (6) | All Industry Pairs | 1.994 (4.92) | 0.009 (3.84) | -0.213 (-3.99) | 0.014 (3.61) | 0.101 (4.29) | 0.614 (2.37) | 312,240 0.048 |
| (7) | Concen. + High Value | 4.249 (4.36) | 0.042 (3.37) | -0.125 (-4.06) | 0.014 (2.30) | 0.130 (2.61) | 0.415 (1.60) | 65,904 0.046 |
| (8) | Concen. + Low Value | 1.826 (5.26) | -0.010 (-1.01) | -0.272 (-4.14) | 0.012 (3.25) | 0.125 (5.74) | 0.479 (2.18) | 88,422 0.043 |
| (9) | Compet. + High Value | 2.384 (3.63) | 0.016 (2.29) | -0.134 (-4.20) | 0.015 (3.05) | 0.062 (2.50) | 0.602 (2.76) | 88,420 0.033 |
| (10) | Compet. + Low Value | 1.260 (3.17) | -0.028 (-3.11) | -0.272 (-3.31) | 0.022 (2.67) | 0.087 (3.06) | 0.640 (2.07) | 65,899 0.046 |
| <i>Panel C: Dep. Var = New Segment Pairs Linked to SDC Acquisitions</i> | | | | | | | | |
| (11) | All Industry Pairs | 0.239 (3.99) | 0.002 (2.09) | -0.019 (-3.13) | 0.001 (2.43) | 0.004 (1.79) | 0.073 (2.31) | 312,240 0.007 |
| (12) | Concen. + High Value | 0.605 (2.44) | 0.007 (2.22) | -0.009 (-3.20) | 0.001 (1.81) | -0.001 (-0.35) | 0.036 (1.24) | 65,904 0.005 |
| (13) | Concen. + Low Value | 0.260 (4.19) | 0.001 (0.46) | -0.021 (-2.88) | 0.001 (2.36) | 0.006 (1.78) | 0.046 (1.53) | 88,422 0.007 |
| (14) | Compet. + High Value | 0.278 (2.41) | 0.002 (1.32) | -0.012 (-3.27) | 0.001 (3.49) | 0.003 (1.07) | 0.040 (2.27) | 88,420 0.004 |
| (15) | Compet. + Low Value | 0.176 (2.18) | -0.002 (-0.72) | -0.033 (-2.91) | 0.001 (1.52) | 0.005 (2.03) | 0.091 (1.35) | 65,899 0.007 |

Table VI: Which Segments Exit

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 . The conglomerate firm itself has 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 Likely 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 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).

| Row | Sample | Across Industry Similarity | Fraction Industries Between Pair | Avg. Within Industry Simil. | Pair Likelihood if Random | Same 2-digit SIC Code | Vertical Relatedness | Obs. /RSQ |
|---|----------------------|----------------------------|----------------------------------|-----------------------------|---------------------------|-----------------------|----------------------|-----------------|
| <i>Panel A: Dep. Var = Segment Pair Disappears</i> | | | | | | | | |
| (1) | All Pairs | -6.557 (-2.94) | 0.282 (1.75) | 0.362 (0.32) | 0.004 (0.67) | -0.043 (-1.23) | -1.724 (-4.13) | 32,181 0.015 |
| (2) | Concen. + High Value | -17.662 (-2.83) | -0.047 (-0.15) | -2.029 (-1.20) | 0.013 (1.53) | -0.146 (-1.16) | -0.213 (-0.17) | 7,387 0.015 |
| (3) | Compet. + High Value | -10.653 (-3.45) | -0.120 (-0.13) | 0.498 (0.31) | -0.004 (-0.43) | -0.135 (-2.15) | -2.493 (-2.27) | 6,976 0.024 |
| (4) | Concen. + Low Value | -15.653 (-2.40) | 0.024 (0.09) | 0.259 (0.13) | 0.006 (0.51) | -0.131 (-1.50) | 1.076 (0.60) | 8,706 0.011 |
| (5) | Compet. + Low Value | -3.574 (-0.78) | 2.919 (1.33) | 0.896 (0.68) | -0.014 (-0.69) | 0.005 (0.09) | -1.192 (-2.17) | 5,636 0.013 |
| <i>Panel B: Dep. Var = Segment Pair Likely Sold or Closed</i> | | | | | | | | |
| (6) | All Pairs | -8.521 (-3.14) | -0.004 (-0.02) | 1.166 (0.98) | 0.008 (1.17) | -0.137 (-2.37) | -1.692 (-3.48) | 32,181 0.009 |
| (7) | Concen. + High Value | -22.566 (-3.18) | -0.507 (-2.03) | -1.110 (-0.50) | 0.015 (1.67) | -0.424 (-2.43) | 0.124 (0.19) | 7,387 0.011 |
| (8) | Compet. + High Value | -14.508 (-2.46) | -1.009 (-0.99) | 0.065 (0.04) | -0.003 (-0.24) | -0.221 (-1.91) | -2.499 (-1.99) | 6,976 0.016 |
| (9) | Concen. + Low Value | -14.934 (-2.09) | -0.262 (-0.94) | 1.662 (1.01) | 0.009 (0.83) | -0.198 (-2.92) | 1.336 (0.91) | 8,706 0.007 |
| (10) | Compet. + Low Value | -6.072 (-0.88) | 1.303 (0.67) | 2.356 (1.22) | -0.004 (-0.20) | -0.082 (-0.91) | -1.312 (-1.55) | 5,636 0.007 |
| <i>Panel C: Dep. Var = Segment Pair Sold Off</i> | | | | | | | | |
| (11) | All Pairs | -2.326 (-0.25) | 0.186 (0.54) | -0.229 (-0.07) | 0.004 (0.29) | 0.085 (0.31) | 0.120 (0.12) | 32,181 0.004 |
| (12) | Concen. + High Value | -27.734 (-1.61) | -0.305 (-0.73) | -8.540 (-1.18) | 0.021 (2.51) | 0.655 (1.26) | 2.888 (0.71) | 7,387 0.009 |
| (13) | Compet. + High Value | 11.600 (0.84) | 0.115 (0.04) | 1.604 (0.37) | 0.010 (0.33) | -0.385 (-0.97) | -2.983 (-0.80) | 6,976 0.006 |
| (14) | Concen. + Low Value | 2.023 (0.15) | -0.258 (-0.24) | 3.912 (0.77) | -0.009 (-0.26) | 0.046 (0.15) | 0.726 (0.21) | 8,706 0.007 |
| (15) | Compet. + Low Value | -0.911 (-0.04) | 10.004 (2.43) | -0.353 (-0.07) | -0.552 (-1.15) | 0.147 (0.39) | 0.945 (0.39) | 5,636 0.004 |

Table VII: 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. Panel B discards an excess value calculation if it is outside the range $\{-1.386, +1.386\}$ to reduce the affect of outliers (screen applied separately for each valuation method). Panel C discards the observation if any one of the five valuation methods generates a value outside this range (this holds the sample size fixed). 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 compute the average and MSE of the excess values. 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. 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+TNIC 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 TNIC industry as the conglomerate. The **HP: SIC+TNIC 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+TNIC 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.

| 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 |
|---|---|-------------------------------|----------------------------------|-------------------------|--------------------------------|-----------------------------------|--------------------------|-------------------|
| <i>Panel A: Raw Data</i> | | | | | | | | |
| 1 | Berger+Ofek Baseline (ss) | -0.082 | 0.474 | 12714 | -0.027 | 0.288 | 10916 | |
| 2 | HP: SIC Universe (wf): Unconstrained | -0.058 | 0.463 | 12714 | -0.038 | 0.268 | 12714 | 0.041 |
| 3 | HP: SIC+TNIC Universe (wf): Unconstrained | -0.046 | 0.402 | 12733 | -0.008 | 0.242 | 12733 | 0.031 |
| 4 | HP: SIC+TNIC Universe (wf): Constrained | -0.012 | 0.320 | 12773 | -0.001 | 0.208 | 12773 | 0.047 |
| 5 | HP: SIC+TNIC Universe (ss): Constrained, Segment-by-Segment | -0.018 | 0.377 | 12675 | 0.020 | 0.282 | 10902 | 0.058 |
| <i>Panel B: Restrict to Excess Valuations to interval [-1.386,+1.386] (Berger and Ofek)</i> | | | | | | | | |
| 6 | Berger+Ofek Baseline (ss) | -0.069 | 0.334 | 11892 | -0.066 | 0.212 | 8761 | |
| 7 | HP: SIC Universe (wf): Unconstrained | -0.047 | 0.342 | 11912 | -0.033 | 0.216 | 8805 | 0.041 |
| 8 | HP: SIC+TNIC Universe (wf): Unconstrained | -0.038 | 0.314 | 12079 | -0.014 | 0.194 | 8823 | 0.031 |
| 9 | HP: SIC+TNIC Universe (wf): Constrained | -0.012 | 0.252 | 12213 | -0.009 | 0.166 | 8844 | 0.047 |
| 10 | HP: SIC+TNIC Universe (ss): Constrained, Segment-by-Segment | -0.012 | 0.281 | 12053 | -0.017 | 0.191 | 8744 | 0.058 |
| <i>Panel C: Uniformly Restrict to interval [-1.386,+1.386]</i> | | | | | | | | |
| 11 | Berger+Ofek Baseline (ss) | -0.065 | 0.306 | 11152 | -0.049 | 0.183 | 7716 | |
| 12 | HP: SIC Universe (wf): Unconstrained | -0.040 | 0.308 | 11152 | -0.018 | 0.190 | 7748 | 0.041 |
| 13 | HP: SIC+TNIC Universe (wf): Unconstrained | -0.028 | 0.274 | 11152 | -0.001 | 0.171 | 7766 | 0.030 |
| 14 | HP: SIC+TNIC Universe (wf): Constrained | 0.004 | 0.210 | 11152 | 0.002 | 0.143 | 7778 | 0.045 |
| 15 | HP: SIC+TNIC Universe (ss): Constrained, Segment-by-Segment | 0.000 | 0.244 | 11152 | -0.003 | 0.169 | 7720 | 0.056 |

Table VIII: 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. The last two columns are based on constrained text regressions where the reconstructed conglomerate matches as closely as possible the actual conglomerate on five key characteristics: Sales Growth, Log Age, OI/Sales, OI/Assets, and R&D/Sales. These correlations are not 100% because the conglomerate reconstruction is based on a weighted median calculation.

| Row | Variable | Berger + Ofek (Baseline) | Text-based SIC only No Constr. | Text-based SIC+TNIC No Constr. | Text-based SIC+TNIC Constrained | Text-based SIC+TNIC Constrained (Seg by Seg) |
|---------------------------------|--------------|--------------------------------|--------------------------------------|--------------------------------------|---------------------------------------|---|
| <i>Correlation Coefficients</i> | | | | | | |
| 1 | Assets | 0.110 | 0.194 | 0.291 | 0.409 | 0.399 |
| 2 | Sales | 0.156 | 0.229 | 0.385 | 0.387 | 0.315 |
| 3 | OI/Sales | 0.375 | 0.425 | 0.479 | 0.850 | 0.675 |
| 4 | OI/Assets | 0.289 | 0.357 | 0.421 | 0.832 | 0.690 |
| 5 | R&D/Sales | 0.473 | 0.673 | 0.705 | 0.908 | 0.821 |
| 6 | Tobin's Q | 0.366 | 0.442 | 0.469 | 0.551 | 0.502 |
| 7 | Sales Growth | 0.241 | 0.270 | 0.309 | 0.825 | 0.683 |
| 8 | TNIC HHI | 0.325 | 0.430 | 0.535 | 0.516 | 0.387 |
| 9 | Log Age | 0.268 | 0.298 | 0.436 | 0.924 | 0.731 |

Table IX: Which Pure Plays Match with Conglomerates?

The table displays summary statistics for groups of pure play firms used with high versus low weights to reconstruct conglomerates (Panel A), and conglomerates with high and low difficulty to replicate using pure plays (Panel B). In Panel A, summary statistics are displayed for pure plays in the lowest quartile of fitted weights, and for those in the highest quartile. The table uses two text-based reconstructions: **HP: SIC+TNIC Universe: Whole Firm, Unconstrained**, and **HP: SIC+TNIC Universe: Whole Firm, Constrained**. Table VII presents details on these text-based weight methods. Summary statistics are computed by first sorting all pure play benchmarks for each conglomerate in each year into quartiles based on each pure play's computed weight from the textual decomposition and then averaging the characteristic for all firms in the lowest and highest quartiles, respectively. In Panel B, summary statistics are displayed for conglomerates with above and below median difficulty to replicate using pure plays and within the four subsamples as noted in the section headers. For both panels, *t*-statistics adjusted for clustering at the firm level test whether the means are different.

| <i>Panel A: Benchmark Portfolio Weights vs Characteristics</i> | | | | | | | |
|--|--------------------------|---|---------------------------------|-----------------------------------|--|---------------------------------|-----------------------------------|
| | | <i>SIC+TNIC Universe: Whole Firm, Un-constrained</i> | | | <i>SIC+TNIC Universe: Whole Firm, Constrained</i> | | |
| Row | Variable | Lowest Weights Quartile | Highest Weights Quartile | <i>t</i> -statistic of Difference | Lowest Weights Quartile | Highest Weights Quartile | <i>t</i> -statistic of Difference |
| 1 | Assets | 3466.56 | 4723.34 | 6.52 | 3564.72 | 4934.27 | 6.85 |
| 2 | Sales | 1563.12 | 2147.15 | 9.49 | 1580.38 | 2213.75 | 10.53 |
| 3 | oi/sales | 0.07 | 0.08 | 6.59 | 0.07 | 0.08 | 8.00 |
| 4 | oi/assets | 0.07 | 0.07 | 0.55 | 0.07 | 0.07 | 1.20 |
| 5 | R+D/sales | 0.11 | 0.09 | -14.15 | 0.11 | 0.09 | -14.20 |
| 6 | Tobin's Q | 2.05 | 1.92 | -4.84 | 2.04 | 1.88 | -6.50 |
| 7 | Sales Growth | 0.17 | 0.16 | -9.10 | 0.18 | 0.16 | -17.92 |
| 8 | TNIC HHI | 0.08 | 0.07 | -12.58 | 0.08 | 0.07 | -9.48 |
| 9 | Firm Age | 25.32 | 29.19 | 18.63 | 24.06 | 30.32 | 29.14 |
| <i>Panel B: Difficulty of Pure Plays to Replicate vs Characteristics</i> | | | | | | | |
| Row | Variable | Lowest Difficulty to Replicate | Highest Difficulty to Replicate | <i>t</i> -statistic of Difference | Lowest Difficulty to Replicate | Highest Difficulty to Replicate | <i>t</i> -statistic of Difference |
| | | <i>Conglomerates with Concentrated Industries Between</i> | | | <i>Conglomerates with Competitive Industries Between</i> | | |
| 10 | Fraction of Ind. Between | 0.142 | 0.224 | 5.951 | 0.046 | 0.054 | 1.940 |
| 11 | Across Ind. Similarity | 0.023 | 0.019 | -7.193 | 0.041 | 0.032 | -7.913 |
| 12 | Within Ind. Similarity | 0.063 | 0.047 | -8.153 | 0.091 | 0.063 | -12.133 |
| 13 | TNIC HHI | 0.053 | 0.082 | 10.054 | 0.029 | 0.059 | 17.454 |
| | | <i>Conglomerates with High Value Industries Between</i> | | | <i>Conglomerates with Low Value Industries Between</i> | | |
| 14 | Fraction of Ind. Between | 0.079 | 0.203 | 10.213 | 0.071 | 0.131 | 6.561 |
| 15 | Across Ind. Similarity | 0.035 | 0.023 | -11.465 | 0.034 | 0.024 | -9.119 |
| 16 | Within Ind. Similarity | 0.084 | 0.058 | -11.289 | 0.073 | 0.050 | -10.204 |
| 17 | TNIC HHI | 0.032 | 0.068 | 15.182 | 0.047 | 0.075 | 11.201 |

Table X: Conglomerate Excess Valuations

This table presents OLS regression estimates of conglomerate excess value on text-based variables. One observation is one conglomerate from 1997 to 2008. 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+TNIC Universe: Constrained" model as illustrated in Table VII. 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. 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. **Across Segment Similarity** is the average textual similarity of pure play firms operating in the same three digit SIC segments as the conglomerate. **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 TNIC 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. Year fixed effects are included in each regression. All standard errors are adjusted for clustering by firm.

| Row | Difficulty of Pure Plays to Replicate | Fraction of Indust. Between | 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 |
|---|---------------------------------------|-----------------------------|---------------------------|---------------------------|--------------------------------|---------------------|----------------------|------------------|-----------------|------------------|------------------|----------------|
| <i>Panel A: Excess Value (Text-based Constrained Valuation Model)</i> | | | | | | | | | | | | |
| (1) | 0.406 (7.59) | . | . | . | . | -0.026 (-1.35) | -0.338 (-1.44) | 1.287 (6.42) | 0.467 (7.08) | 0.603 (8.58) | 0.044 (9.07) | 9,201 0.100 |
| (2) | . | 0.120 (1.92) | . | . | . | -0.075 (-3.97) | -0.282 (-1.19) | 1.212 (5.97) | 0.453 (6.97) | 0.600 (8.50) | 0.039 (7.95) | 9,201 0.088 |
| (3) | . | . | -0.888 (-1.98) | . | . | -0.073 (-3.88) | -0.259 (-1.08) | 1.147 (5.53) | 0.453 (6.98) | 0.594 (8.41) | 0.040 (8.07) | 9,201 0.088 |
| (4) | . | . | . | -0.616 (-2.48) | . | -0.068 (-3.58) | -0.309 (-1.31) | 1.102 (5.29) | 0.459 (6.99) | 0.600 (8.48) | 0.040 (8.20) | 9,201 0.089 |
| (5) | . | . | . | . | 0.508 (2.06) | -0.067 (-3.48) | -0.355 (-1.51) | 1.144 (5.58) | 0.463 (7.03) | 0.597 (8.47) | 0.040 (8.17) | 9,201 0.088 |
| (6) | 0.389 (7.04) | 0.060 (0.87) | 0.156 (0.28) | -0.229 (-0.80) | 0.009 (0.04) | -0.023 (-1.14) | -0.286 (-1.18) | 1.263 (6.11) | 0.473 (7.09) | 0.604 (8.59) | 0.045 (9.14) | 9,201 0.101 |
| <i>Panel B: Excess Value (Berger + Ofek Valuation Model)</i> | | | | | | | | | | | | |
| (7) | 0.549 (8.23) | . | . | . | . | 0.056 (2.41) | 0.003 (0.01) | 2.454 (10.55) | 0.688 (7.58) | 1.113 (12.36) | 0.063 (10.84) | 8,951 0.198 |
| (8) | . | 0.099 (1.31) | . | . | . | -0.012 (-0.54) | 0.022 (0.08) | 2.314 (9.77) | 0.659 (7.32) | 1.105 (12.18) | 0.055 (9.49) | 8,951 0.179 |
| (9) | . | . | -1.273 (-2.37) | . | . | -0.007 (-0.29) | 0.116 (0.39) | 2.222 (9.26) | 0.664 (7.38) | 1.096 (12.11) | 0.056 (9.68) | 8,951 0.180 |
| (10) | . | . | . | -1.130 (-3.94) | . | 0.005 (0.21) | 0.077 (0.26) | 2.122 (8.85) | 0.681 (7.55) | 1.106 (12.25) | 0.058 (9.88) | 8,951 0.182 |
| (11) | . | . | . | . | 0.348 (1.10) | -0.008 (-0.33) | -0.041 (-0.14) | 2.264 (9.53) | 0.665 (7.36) | 1.102 (12.18) | 0.056 (9.52) | 8,951 0.179 |
| (12) | 0.542 (7.96) | -0.007 (-0.08) | 0.045 (0.07) | -0.720 (-2.19) | -0.473 (-1.52) | 0.057 (2.33) | 0.054 (0.18) | 2.392 (9.96) | 0.688 (7.57) | 1.118 (12.43) | 0.063 (10.78) | 8,951 0.199 |

Table XI: 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. 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+TNIC Universe: Constrained” model as illustrated in Table VII. The residual excess valuations are residuals from a regression of excess valuation on all of the variables included in Table X 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. |
|---------------------------------------|-------------------------|-----------------------------------|------------------------------------|--|---|-------|
| <i>Summary Statistics by Quintile</i> | | | | | | |
| Lowest Difficulty | 0.630 | -0.045 | -0.035 | -0.059 | -0.058 | 2,331 |
| Quintile 2 | 0.729 | -0.003 | -0.043 | 0.002 | -0.009 | 2,339 |
| Quintile 3 | 0.795 | -0.007 | -0.093 | -0.000 | -0.027 | 2,337 |
| Quintile 4 | 0.858 | -0.004 | -0.094 | 0.015 | -0.008 | 2,339 |
| Highest Difficulty | 1.028 | 0.047 | -0.029 | 0.085 | 0.107 | 2,333 |

Table XII: Conglomerate Excess Valuations (Various Subsamples)

The table presents OLS regression estimates of conglomerate excess value on our text based variables. One observation is one conglomerate from 1997 to 2008. We consider various subsamples as noted in the panel headers. The dependent variable is the conglomerate’s excess valuation using the “HP: SIC+TNIC Universe: Constrained” model as illustrated in Table VII. 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. 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. **Across Segment Similarity** is the average textual similarity of pure play firms operating in the same three digit SIC segments as the conglomerate. **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 TNIC 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 regressions contain time fixed effects. All standard errors are adjusted for clustering by firm.

| Row | Difficulty of Pure Plays to Replicate | Fraction of Indust. Between | 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 |
|---|---------------------------------------|-----------------------------|---------------------------|---------------------------|--------------------------------|---------------------|----------------------|-----------------|-----------------|-----------------|-----------------|----------------|
| <i>Panel A: Excess Value (Conglomerates with High Value Industries Between)</i> | | | | | | | | | | | | |
| (1) | | 0.183 (2.34) | . | . | | -0.075 (-2.97) | 0.200 (0.60) | 1.391 (4.70) | 0.486 (6.49) | 0.561 (6.39) | 0.041 (6.20) | 4,178 0.097 |
| (2) | 0.473 (5.08) | 0.098 (1.13) | -0.084 (-0.09) | -0.396 (-0.96) | -0.120 (-0.32) | -0.015 (-0.55) | 0.238 (0.68) | 1.426 (4.70) | 0.507 (6.61) | 0.567 (6.44) | 0.047 (7.12) | 4,178 0.111 |
| <i>Panel B: Excess Value (Conglomerates with Low Value Industries Between)</i> | | | | | | | | | | | | |
| (3) | | 0.062 (0.74) | . | . | | -0.073 (-2.96) | -0.520 (-1.67) | 0.806 (2.70) | 0.455 (3.86) | 0.664 (5.36) | 0.036 (5.80) | 4,172 0.079 |
| (4) | 0.323 (5.11) | -0.107 (-1.10) | -1.726 (-1.83) | -0.089 (-0.23) | -0.214 (-0.74) | -0.036 (-1.40) | -0.544 (-1.73) | 0.784 (2.61) | 0.455 (3.79) | 0.640 (5.20) | 0.043 (6.79) | 4,172 0.092 |