NBER WORKING PAPER SERIES

DYNAMIC TEXT-BASED INDUSTRY CLASSIFICATIONS AND ENDOGENOUS PRODUCT DIFFERENTIATION

Gordon M. Phillips
Gerard Hoberg

Working Paper 15991
http://www.nber.org/papers/w15991

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2010

We especially thank Dan Kovenock, Steve Martin, John Sutton and seminar participants at HEC, IFN (Stockholm), Insead, ISTCE (Lisbon), London Business School, Stockholm School of Economics, University of Amsterdam and University of Vienna for helpful comments. All errors are the authors alone.

Dynamic Text-Based Industry Classifications and Endogenous Product Differentiation
Gordon M. Phillips and Gerard Hoberg
NBER Working Paper No. 15991
May 2010
JEL No. D21,D23,L12,L13,L16,L22,L23

ABSTRACT

We study how firms differ from their competitors using new dynamic measures of product differentiation based on novel text based analysis of 50,673 product descriptions from firm 10-K statements filed yearly with the Securities Exchange Commission. This year-by-year set of firm product differentiation measures allows us to generate a set of dynamic industry classifications and new measures of industry structure and competition. Competitiveness and market structure measures based on these new classifications better correlate with firm profitability than do classifications based on SIC or NAICs. Using these new dynamic industry classifications, we examine endogenous product differentiation. We show that firms use R&D and advertising to differentiate themselves from competitors and increase their profitability.

Gordon M. Phillips
R.H. Smith School of Management
Van Munching Hall
University of Maryland
College Park, MD 20742
and NBER
gphillips@rhsmith.umd.edu

Gerard Hoberg
Robert H. Smith School of Business
University of Maryland
4416 Van Munching Hall
College Park, MD 20742
ghoberg@rhsmith.umd.edu

Defining industry boundaries and industry competitiveness is central to the study of industrial organization. It is also central to broader disciplines in Economics and Finance, where the study of industries, or the need to control for industry, is pervasive. Our paper is based on the premise that product similarity is core to classifying industries, and that empirical work can benefit from the ability to measure industry memberships and product differentiation in every year. Using new dynamic industry classifications, we find that firms use R&D and advertising to differentiate themselves from competitors and increase their profitability. Our results are consistent with Sutton's (1991) theory of endogenous product differentiation.

Our starting point to form new industries is to gather product descriptions from 50,673 firm annual 10-Ks filed with the Securities and Exchange Commission using web crawling algorithms. We then process the text in these product descriptions to calculate new industry classifications based on how firms are related to each other. A key advantage of this framework is that firms must file a 10-K in each year, allowing us to build classifications that change over time. The framework also provides a continuous measures of product similarity between firms both within and across industries. These tools enable us to examine how industry structure changes over time, how firms react to dynamic changes within and around their product markets, and how firms take actions to create product differentiation and barriers to entry.

We measure the similarity of products offered by every unique pair of publicly traded firms every year over a comprehensive ten year sample. The vector representations of the text in each firm's product description generate a Hotelling-like product location space for U.S. firms.[1] The intuition behind our paper is that a firm and its rivals have locations that are very near, as their product descriptions use a common vocabulary. Analogously, firms with unrelated products are far from each other in this space. We apply clustering analysis over these locations to generate industry classifications.

We compare our new industry classifications to existing classifications, and find that our industries generate superior R^2 in explaining the cross section of firm char-

[1]Chamberlin (1933) and Hotelling (1929) famously show that product differentiation is fundamental to profitability and theories of industrial organization, and also that product markets can be viewed as having a spatial representation that accounts for product differentiation.

acteristics. Because they are a function of product descriptions, our classifications are based on the products that firms supply to the market, rather than production processes (as is the case for existing industry classification schemes).[2] Using our product similarity scores between each pair of firms, we then calculate new firm-specific product uniqueness measures and firm-specific measures of industry structure based on the distribution of rivals around each firm. We examine how competition, product differentiation and firm profitability change over time. Consistent with Sutton (1991)'s work on endogenous barriers to entry, we find that product differentiation, product uniqueness, and profitability increase over time as firms advertise and conduct R&D. We also find that measures of market concentration based on firm 10Ks also increase with advertising and R&D.

Although it is convenient to use existing industry classifications such as SIC or NAICS for research purposes, these measures have limitations. Both do not adjust significantly over time as product markets evolve. Innovations can also create new product markets that do not exist in fixed classifications. In the late 1990s, hundreds of new technology and web-based firms were grouped into a large and nondescript SIC-based "business services" industry. More generally, fixed classifications like SIC and NAICS have at least four shortcomings: they only rarely re-classify firms into different industries as firm product offerings change, they do not allow for product markets themselves to evolve over time, they do not allow for the possibility that two firms that are rivals to a third firm, might not directly compete against each another, and lastly, they do not allow for within industry continuous measures of similarity to be computed.

We create new industry classifications based on 10K product similarities using two methods: one historically motivated, and one that allows industry boundaries to dynamically adjust. The first, which we name "fixed industry classifications" (FIC), is analogous to SIC and NAICS industries.[3] Here, industry locations are fixed to remain static over time, and firm membership in an industry is required to be transitive. Thus this method requires that if firms B and C are in firm A's

[2]See http://www.naics.com/info.htm.

[3]We will make these industry classifications and corresponding firm memberships available to researchers via the internet.

industry, then firms B and C are also in the same industry. We assign firms to industries using clustering algorithms that maximize total within-industry similarity where similarity is based on word usage in 10-K product descriptions. These 10-K-based FIC industries provide improvements over SIC and NAICS in explaining a wide array of firm characteristics.

Our second classification is more general, and we relax the fixed location and membership transitivity requirements of FIC industries. First, industry classifications can change each year and second, each firm can have its own distinct industry. Analogous to network measures,[4] each firm can have its own distinct set of competitors and it is possible that firms B and C can be competitors for firm A, while not being direct competitors of each other. We name these generalized industries "variable industry classifications" (VIC). Relative to FIC industries, VIC classifications offer economically large improvements in their ability to explain firm characteristics. They also result in more informative industry competitiveness measures, and given that industry membership can change over time, it allows tests of (1.) how industry boundaries change over time and (2.) whether firm investment in advertising and R&D can create endogenous barriers to entry as in Sutton (1991). Our empirical tests benefit from information about the degree to which specific firms are similar to their competitors, which cannot be derived from zero-one membership classifications such as SIC or NAICS. Generalized competitiveness measures based on VIC industry classifications significantly explain observed firm profitability. We find only weak results in analogous tests based on SIC and NAICS.

Our results are robust to the treatment of firms that report producing in more than one industry. When forming fixed classifications, we only use firms that report one segment (non-conglomerate firms) to identify which industries exist in the economy. Thereafter, we assign conglomerates and non-conglomerates alike to the resulting classifications. Detailed robustness tests show that assigning conglomerates to more than one industry does not generate material improvements in explanatory power, suggesting that multiple industry conglomerate characteristics are strongly

[4]The analogy for social networks is that in networks such as Facebook, each individual can have their own unique set of friends, with friends of one individual not necessarily being friends of each other.

in-line with the single industry to which they are most similar.

In our analysis of VIC classifications, our ability to update both the location and the membership of VIC classifications over time also allows us to examine whether advertising and research and development act as endogenous barriers to entry. We find strong support for the hypothesis of Sutton (1991) that firms will spend on advertising and R&D to reduce competition in their industries. We find that firms spending more on either advertising or R&D indeed do experience significant reductions in ex-post competition, as well as gains in ex-post profitability.

Our new 10K-based VIC industries, given their strong explanatory power and their continuous updating over time, can help future researchers to more precisely examine the predictions of many theories, especially those related to the industry determinants of product innovation, and the industry life cycle. We also note that while our new measures are interesting for research or scientific purposes, they would not be good for policy and antitrust purposes as they could be manipulated by firms fairly easily if firms believed they were being used by policy makers.

Our research also contributes to existing strands of literature using text analysis to address economic and financial theories, product markets, and mergers and acquisitions. Hoberg and Phillips (2009) show that merging firms with more similar product descriptions in their 10-Ks experience more successful outcomes. Rauh and Sufi (2010) use firm self-reported competitors from firm 10-Ks and find that capital structure better reflects that of self reported peers than that of firms in the same SIC code. Hanley and Hoberg (2009) use document similarity measures to examine prospectus disclosures from the SEC Edgar website to address theories of IPO pricing. Loughran and McDonald (2008) show that firms using Plain English have greater small investor participation and more shareholder-friendly corporate governance. In other contexts, papers such as Tetlock (2007), Tetlock, Saar-Tsechanksy, and Macskassy (2008), Li (2006) and Boukus and Rosenberg (2006) find word content to be informative in predicting stock price movements.

The remainder of the paper is organized as follows. We discuss the how our new industry classifications relate to existing literature in Section I. The data and similar-

ity calculation is in section II, and the industry classification methods are in Section III. We then compare the informativeness of a wide array of industry classifications in Section IV. We construct measures of industry competitiveness in Section V, and test their performance in section VI. Section VII examines how competition changes over time and tests theories of product differentiation and endogenous barriers to entry, and section VIII concludes.

I Industries and Product Differentiation

We propose a new method for classifying industries based on product similarities from word product descriptions. The concept of product similarity dates back to Chamberlin (1933), who famously showed that the notion of product differentiation is fundamental to theories of industrial organization, with product differentiation reducing competition between firms. Although Chamberlin (1933) focuses on product substitution by consumers, the impact of product differentiation is taken further by Hay (1976) and Panzar and Willig (1981), who suggest that profit margins can be reduced further even if rival firms do not produce substitutes, if they hold a credible threat of entry at low cost. Firms offering similar products are more likely to hold such a threat due to technological similarities. These studies suggest that industry classifications must be assigned with care, and that a full knowledge of all firm pairwise similarities should be more informative than simple all or nothing classifications.

Although numerous studies use industry classifications as control variables, only a few studies examine the classification schemes themselves and these do not consider the possibility of dynamic industry classifications.[5] In a contemporaneous paper, Rauh and Sufi (2010) use self-reported competitors from firm 10-Ks and show that firm capital structure better reflects that of these self-reported competitors than that

[5]Kahle and Walkling (1996) compare the informativeness of SIC codes obtained from the CRSP and COMPUSTAT databases, and Fama and French (1997) create new industry classifications based on a new way of grouping existing four digit SIC codes. Krishnan and Press (2003) compare SIC codes to NAICS codes, and Bhojraj, Lee, and Oler (2003) also compare various FIC industry classifications. Although these studies are informative, and suggest that existing static classifications can be used in better ways, they do not explore whether the core methodology underlying static classifications can be improved upon.

of firms in the same SIC code. This paper is similar to ours in that it uses information in 10-Ks to identify similar firms, but our paper is unique along three dimensions. First, our paper addresses a different question and the methodology allows us to form industries with varying degrees of similarity. Second, our methodology produces a continuous measure of relatedness based on a full product location space, allowing measurements of the degree of similarity within and across industry groups, and the construction of industries with any arbitrary level of coarseness. Third, our classifications are based on actual product text, and thus we are able to detect rival firms that offer complementary products or additional competitive threats even if they are not direct rivals at present (for example, through economies of scope).

Our approach is further motivated by more recent studies, especially those related to endogenous product differentiation including Mazzeo (2002) and Seim (2006). These studies confirm two key foundations that are important to our approach: (1) there are significant gains to product differentiation, (2) product market locations are dynamic and drift over time as firms continuously invest in product differentiation. Aghion, Bloom, Blundell, Griffith, and Howitt (2005) show that product innovation has an inverse-U-shaped relationship with product market competition, providing additional evidence that product market locations are dynamic, and that the dynamics vary across industries with different characteristics.[6] Berry (1990), through an in-depth analysis of the airline industry, confirms that product differentiation is indeed integral to airline growth strategies and how airlines choose to provide service to different cities. Only dynamic industry definitions are capable of fully addressing the fact that product market locations can move within the product space.

Sutton (1991) and Shaked and Sutton (1987) suggest that barriers to entry are endogenous. In particular, firms can spend resources on advertising and research and development to differentiate their products from potential substitutes. The result is that potential rivals face higher entry costs. These theories motivate our examination of advertising and research and development, and their links to future changes in industry membership and competition.

[6]Lin and Saggi (2002) show that tradeoffs related to product differentiation further affect different types of innovation including process innovation and product innovation.

Many empirical studies examining topics related to product differentiation and competition focus on single industries (see Schmalensee (1978), Kelton and Kelton (1982), and Katz (1978) for example). Beginning with Berry, Levinsohn, and Pakes (1997) the approach of the product differentiation literature has been to estimate demand and cost parameters in differentiated product markets and provide a framework to analyze oligopolistic industries. Empirical studies in this literature involve estimating own- and cross-price elasticities of demand and the effect on post-merger prices in specific markets including the ready-to-eat cereal market (Nevo (2000)). These approaches have been highly informative, especially in understanding the dynamics of industry pricing and competition among firms offering substitute products. However, some theoretical hypotheses, especially those related to inter-industry differences and endogenous barriers to entry, are difficult to test in a single industry setting. Our study helps to fill this void.

II Data and Methodology

Using web crawling and text parsing algorithms, we obtain and construct a database of word product descriptions from 10-K annual filings on the SEC Edgar website from 1997 to 2006. These descriptions are found in a separate section of each 10K filed by each firm. These product descriptions are legally required to be accurate, as Item 101 of Regulation S-K legally requires that firms describe the significant products they offer to the market, and these descriptions must also be updated and representative of the current fiscal year of the 10-K. This recency requirement is important, as our goal is to accurately measure how industry structure changes from year to year.

We merge each firm's text product description to the CRSP/COMPUSTAT data using the central index key (CIK), which is the primary key used by the SEC to identify the issuer. Our resulting database is based on all publicly traded firms (domestic firms traded on either NYSE, AMEX, or NASDAQ) for which we have COMPUSTAT and CRSP data. Our initial sample contains 56,540 firm years from CRSP/COMPUSTAT linked to the sample of filed 10-K annual reports on the SEC's

online Edgar database to their 10K filings.[7]

A Product Similarity

We calculate our firm-by-firm similarity measures by parsing the product descriptions from the firm 10Ks and forming word vectors for each firm to compute continuous measures of product similarity for every pair of firms in our sample in each year (a pairwise similarity matrix). For any two firms i and j, we measure product similarity using each firm's empirical distribution of word usage in its product description - omitting common words that are used by more than 5% of all firms. This method results in a real number in the interval [0,1] describing how similar the words used by firms i and j are, when describing their products and their business. Full details of the pairwise similarity calculation are discussed in Appendix 1. Firms having more common word usage are scored as being more similar (closer to 1). We use the "cosine similarity" method, which is widely used in studies of information processing (see Kwon and Lee (2003)), because it measures the angle between two word vectors on a unit sphere (this unit sphere is based on text vectors and each firm in our sample has a specific spatial address). The cosine method avoids over-scoring larger documents. We make an additional adjustment to exclude very common words from the analysis.

This method generates an empirical product market space on which all firms reside, and it also generates a real number in the interval (0,1) capturing the similarity of words (analogous to geographical distance in this space) used for each pair of firms. The ability to map all firms to specific locations in product market space is the core concept underlying how we compute industry classifications, and improved competitiveness measures in this study. Note that this continuous measure can identify within industries who are the closest competitors to firms.

[7]We thank the Wharton Research Data Service (WRDS) for providing us with an expanded historical mapping of SEC CIK to COMPUSTAT gvkey, as the base CIK variable in COMPUSTAT only contains current links. Although somewhat rare, a small number links have changed over time. Our results are robust to either using the WRDS mapping or the COMPUSTAT mapping.

B The Sample of 10-Ks

We electronically gather 10-Ks by searching the Edgar database for filings that appear as "10-K", "10-K405", "10KSB", "10KSB40". Our primary sample includes filings associated with firm fiscal years ending in calendar years 1997 to 2006. Our sample begins in 1997 as this is when electronic filing with Edgar first became required. Of the 56,540 firm-year observations with fiscal years ending in 1997 to 2006 that are present in both CRSP and COMPUSTAT, we are are able to match (using CIK) 55,326 (97.9% of the CRSP/COMPUSTAT sample).[8] We can also report that our database is well balanced over time, as we capture 97.6% of the eligible data in 1997, and 97.4% in 2006, and this annual percentage varies only slightly in the range of 97.4% in 2006 to 98.3% in 2001. Because we do not observe much time variation in our data coverage, and because database selection can be determined using ex-ante information (ie, the 10-K itself), we do not believe that our data requirements induce any bias. Our final sample size is 50,673 rather than 55,326 because we additionally require that lagged COMPUSTAT data items (assets, sales and operating cash flow) are available before observations can be included in our analysis.

From each linked 10-K, our goal is to extract its product description. This section of the document appears as Item 1 or Item 1A in most 10-Ks. We utilize a combination of PERL web crawling scripts, APL programming, and human intervention (when documents are non-standard) to extract and summarize this section. The web crawling algorithm scans the Edgar website and collects the entire text of each 10-K annual report, and the APL text reading algorithms then process each document and extract each one's product description and its CIK. This latter process is extensively supported by human intervention when non-standard document formats are encountered. This method is highly reliable and we encountered only a very small number of firms (roughly 100) that we were not able to process because they did not contain a valid product description or because the product description had fewer then 1000

[8] We also compute similarities for 1996 (93.5% coverage, electronic filing was optional) and 2007 (98.1% coverage), but only use the 1996 data to compute the starting value of lagged variables, and we only use the 2007 data to compute the values of ex-post outcomes. Also, although we use data for fiscal year endings through 2007, we extract documents filed through December 2008, as many of the filings in 2008 are associated with fiscal years ending in 2007. This is because 10-Ks are generally filed during the 3 month window after the fiscal year ends.

characters. These firms are excluded from our analysis.

III Industry Classification Methodology

We classify firms into industries using two methods based on text clustering. We refer to an "industry classification" as a complete mapping from the set of firms to a set of industries. That is, for each firm, an industry classification identifies which other firms are its rivals in terms of its product market. We consider two key properties that industry classifications might satisfy. These properties are ones which existing SIC and NAICS classifications satisfy and we begin with them to build our first set of classifications. We relax these properties for our more general method of classifying firms.

Definition: A classification is said to have the ***fixed location property*** if, over time, each industry's definition refers to a time-fixed product market.

Definition: A classification is said to have the ***membership transitivity property*** if, for any two firms A and B in the same industry, a firm C that is in A's industry, is also be in B's industry.

The first method we consider is analogous to SIC and NAICS code classifications. These industries satisfy both the fixed location property, and the membership transitivity property. We henceforth refer to industries satisfying both properties as "Fixed Industry Classifications" (FIC). Our second method relaxes both properties, and we refer to this second class of industries as "Variable Industry Classifications" (VIC). VIC industry locations can move across the product space over time as technologies and product tastes evolve. New firms can also appear in the sample, and they can be assigned either to new VIC industries that did not previously exist, or to VIC industries that already do exist. Finally, relaxing transitivity implies that firms competing with one another may not necessarily have the same set of other rivals. We now discuss both methods in detail.

A Fixed Industries Classifications Based on 10-Ks

To maintain consistency with other FIC industry classifications including SIC and NAICS, we form fixed groups of industries only once using the earliest year of our sample (1997). We then hold these industries fixed throughout our sample. We then assign firms to these industries in later years based on their 10-K text similarity relative to the frequency-weighted list of words used in the 1997 10-K product descriptions that were initially assigned to each industry. We provide a detailed description of the text clustering algorithm used to create our FIC classifications in Appendix 2. The main idea is that the clustering algorithm starts by assuming that each of the roughly 5000 firms in 1997 is a separate industry, and then it groups the most similar firms into industries one at a time. The algorithm stops when the desired number of industries remains.

A key virtue of the industry clustering algorithm is that it can generate a classification with any number of industries. We consider industry classifications comprised of 50 to 800 industries in increments of 50. However, we focus most on the 300 industries classification as it is most analogous to popular alternatives including three digit SIC codes and four digit NAICS codes, which have 274 and 328 industries, respectively, in our sample. Although the clustering algorithm's flexibility to pre-specify the number of industries is a virtue, the algorithm is not capable of determining the "optimal" number of industries. We explore this question using Akaike likelihood tests in Section IV.

Table I displays four sample industries created by our algorithm. These four are from the 300 10-K industry grouping, henceforth the "10K-300" industries. The first industry has 26 firms that process payments and cash transfers. Many of these firms are in different SIC codes, some disagreeing even at the one and two digit level. This example illustrates that many technology firms compete with traditional brick and mortar firms, a frequent relationship missed by SIC codes.

[Insert Table I Here]

The second example has thirteen firms that provide geological survey and oil

exploration services. This example shows deeper industry relationships that might often be missed by other classifications. Many of these firms are from SIC codes associated with oil field services, but others such as XOX provide technical 3-D modeling services, and yet others are associated with measurement apparatuses. These components work together to provide the needed toolkits, indicating in a more sophisticated market structure. The example also shows grouping of firms producing similar but not identical products, analogous to some SIC and NAICS groups.[9]

The third example is the second largest 10K-300 industry (the largest is a commercial banking industry with 581 firms). This industry contains 419 firms, and the word list reveals its focus on biotechnology (for example, trials, therapy, drugs, and diseases are key industry terms). The large size of this industry suggests that many biotechnology firms are related despite their spanning different SIC codes. However, because a large number of firms are in the two digit SIC code 28, we also observe significantly agreement with SIC classifications.

The fourth industry includes fourteen firms providing security and detection products. This example is interesting because it depicts many similar firms selling security products to somewhat different markets: business, personal, or computer-based. Although their products are not direct substitutes, each firm could potentially enter another's market at a lower cost than an unrelated firm can. This interpretation is related to Economies of Scope (see Panzar and Willig (1981)). Broadly, coarser classifications should be more related to economies of scope. Because our approach offers the flexibility to define industries as fine or coarsely as desired, it offers a powerful new tool for testing predictions related to Economies of Scope.

[Insert Figure 1 Here]

Our industry classifications are based on the notion that firms in the same industry use many common words to describe their products. Figure 1 displays a histogram showing the number of unique words in firm product descriptions. As noted earlier, we limit attention to words that appear in no more than 5% of all

[9]For example, the SIC-3 industry 737 contains a wide array of technology firms in a single "business services" industry.

product descriptions in order to avoid common words. Typical firms use roughly 200 unique words. The tail is also somewhat skewed, as some firms use as many as 700 to 1000 words, although a few use fewer than 50. Because they are not likely to be informative, we exclude firms having fewer than 20 unique words from our classification algorithm. However, removing this screen or excluding firms with fewer than 50 words generates similar results.

[Insert Figure 2 Here]

Figure 2 displays a histogram showing the distribution of the number of firms in each industry for 10K-300, SIC-3, and NAICS-4 industries. 10K-300 industries (top graph) have firm counts that are similar to those based on SIC-3 (second graph) and to NAICS-4 industries (bottom graph), as most industries have fewer than ten firms. However, they are somewhat different in two ways. First, 10-K groupings have more single-firm industries, and hence some firms have highly unique descriptions. Second, 10-K classifications have more very large industries and are more spread out.

Industry memberships are similar but also quite different. For example (not displayed), the likelihood that two firms in the same SIC-3 industry will also be in the same NAICS-4 industry is 61.3%. The likelihood that they will be in the same 10K-300 industry is a more modest 46.2%. In contrast, when two firms are in the same 10K-300 industry, the likelihood that they will appear in the same SIC-3 and NAICS-4 industry is 44.1% and 54.2%, respectively. We conclude that, 10K-300 industries are quite distinct from both NAICS-4 than SIC-3. However there is also some agreement among all three classifications.

B Variable Industry Classifications Based on 10-Ks

We next relax the fixed location and transitivity requirements and construct generalized variable industry classifications (VIC). In addition to offering substantially higher explanatory power (see Section IV), VIC industries offer additional advantages, and we examine three in this study. First, relaxing industry transitivity is necessary for us to jointly consider how similarity and market shares jointly impact industry competitiveness (see Section V). Second, the full knowledge of firm pair-

wise similarities that accompanies VIC classifications is necessary to efficiently test theories that assume that firms can be partially similar. Third, VIC industries are necessary to test theories predicting dynamic firm and industry movements in the product space over time such as in Sutton (1991) (see Section VII).

We construct VIC classifications using a simple minimum similarity threshold. That is, we simply define each firm i's industry to include all firms j with pairwise similarities relative to i above a pre-specified minimum similarity threshold. A high threshold will result in industries having very few rival firms, and a low threshold results in very large industries.

For two randomly selected firms i and j, we label them as an "industry pair" if, for a given classification, they are in the same industry. Where N denotes the number of firms in the economy, there are $\frac{N^2-N}{2}$ permutations of unique pairs.[10] In practice, however, only a small fraction of pairs are actually industry pairs. Although one can use any minimum similarity threshold to construct VIC-industries, we focus on thresholds generating industries with the same fraction of industry pairs as SIC-3 and SIC-2 industries, allowing us to compare VIC industries to SIC and NAICS in an unbiased fashion.

For three digit SIC codes, 2.05% of all possible firm pairs are industry pairs. For two digit SIC industries, this number is 4.45%. A 7.06% minimum similarity threshold generates 10-K based VIC industries having 2.05% industry pairs (same as SIC-3), and a 5.14% similarity threshold generates industries having 4.45% industry pairs (same as SIC-2). We focus on these two thresholds, and refer to these VIC classifications as 10K-VIC-7.06 and 10K-VIC-5.14.

Indeed the transitivity property might not hold for these industries. For example, consider firms A and B, which are 15% similar. Because this is higher than 7.06%, A and B are in each other's 10K-VIC-7.06 industry. Now consider a firm C that is 9% similar to firm A, and 4% similar to firm B. C is in firm A's industry, but not in firm B's industry, and thus transitivity does not hold. If, alternatively, firm C was 8% similar to firm B, then transitivity would hold. Thus, VIC classifications do not

[10]For a sample of 5000 firms, this is 12.4975 million unique pairs.

rule out transitivity, but rather transitivity might hold case by case.

IV Comparing Industry Classifications

Our next objective is to examine which industry classifications best explain firm characteristics in cross section, while controlling for variation in the degrees of freedom. In Section A, we compare FIC industry classifications, and explore the impact of degrees of freedom usage. In Section B, we compare a broader set of industry classifications.

A Fixed Industry Classifications

We compare three FIC industry classifications (SIC, NAICS, and 10K-based), and consider their ability to explain firm-level profitability in cross section. Higher quality classifications should explain more variation. We ask two questions: (1) which classification method best explains firm profitability holding degrees of freedom fixed? (2) what is the most likely number of industries in the economy?

Industry fixed effects use a moderate to large number of degrees of freedom, and raw R-squared comparisons are biased. Adjusted R-squared addresses this bias. However, adjusted R-squared is silent regarding how many industries best explain the data. That is, does the economy really consist of 200 or 600 industries?

We take a two pronged approach to address these questions. First, we use the Akaike information criterion to examine which models best explain the data when different numbers of degrees of freedom are used. Our 10K-based industries are ideally suited to answer this question because we can generate classifications with any number of industries, allowing us to hone in on the optimum number of industries if such an optimum exists. Our second method is to compare the adjusted R-squared across industry classifications when the number of degrees of freedom used is held fixed. This is useful in determining which classifications are more informative when they are generated using different sources of information (e.g., are SIC, NAICS, or 10-K based classifications more informative?).

[**Insert Table II Here**]

Table II presents the results of the Akaike Information Criterion (AIC) tests. For all four levels of SIC granularity (Panel A), all six levels of NAICS granularity (Panel B), and for product description based industries ranging from 50 to 800 industries (Panel C), we compute the AIC statistic and the adjusted R-squared from regressions in which the dependent variable is profitability scaled by sales or assets, and the independent variable is a set of industry fixed effects based on the given classification. To avoid clustering of firm observations over time, which could bias AIC tests, we run separate cross sectional regressions in each year and we then report the average AIC scores and the average adjusted R-squared calculations based on ten regressions from 1997 to 2006. Classifications with lower AIC scores are more likely to explain the data.

Panel A shows that three and four digit SIC classifications are most informative, and dominate two digit SIC codes. This suggests that the wide usage of three digit SIC codes in existing studies is reasonable. Panel B suggests that four digit NAICS dominate other resolutions, suggesting that NAICS-4 might be a substitute for SIC-3. Because AIC scores are designed to permit comparisons across industries using different information sources and degrees of freedom, we can also broadly compare SIC to NAICS. Panels A and B show that SIC and NAICS are reasonable substitutes. NAICS is marginally better when explaining profitability scaled by assets, and SIC is marginally better when explaining profitability scaled by sales. Our results do not support the conclusion that NAICS dominates SIC, which is perhaps surprising given the more recent establishment of the NAICS system.

Panel C shows that 10K-based industries dominate both SIC and NAICS, as AIC scores in Panel C are broadly lower than those in either Panel A or Panel B. This result is robust to scaling profitability by sales or assets. The AIC score of 2715 (10K-300 industries) is broadly lower than the 3110 for three digit SIC codes, and the 3123 for four digit NAICS codes, even though all three groupings use similar numbers of degrees of freedom.

Although we can conclude that 10K-based industries are more informative than

SIC or NAICS industries, Panel C is only moderately informative regarding how many 10K-based industries best explain the data. There is weak evidence that the AIC scores reach a minimum between 250 and 500 industries. However, this range is broad and the pattern surrounding the minimum is not fully monotonic. We conclude that the degree of granularity (roughly 300 industries) used by SIC and NAICS is reasonable, and that this number is also reasonable for 10-K based industries. Although they might agree on degrees of freedom, however, SIC and NAICS are less informative than 10K-based industries.

B Performance of Industry Controls

In this section, we explore industry controls in a panel data setting. We focus on comparing performance for many firm characteristics, and hold degrees of freedom roughly constant. Superior classifications should explain a larger fraction of the total variation across many characteristics.

We also consider VIC industry classifications. For FIC classifications, industry fixed effects are the most widely used method of industry control. This approach has two limitations. First, it uses a large number of degrees of freedom, leaving fewer for hypothesis testing. Second, in a panel data setting, it does not consider how industry variables change over time. To address this second issue, researchers can use industry x year fixed effects. However, this exacerbates the degrees of freedom problem multiplicatively, and the design matrix might even be too large to invert in practice.

Both issues can be addressed using simple kernel methods. Rather than using fixed effects, one can average the given characteristic (the dependent variable) within each industry, and use this average as a single additional control variable. This approach uses only one degree of freedom in the primary regression, and because this average can be computed separately in each year, this approach also allows industry characteristics to vary over time. We refer to this simple averaging technique as a "flat kernel". This method can be used for both FIC and VIC classifications.

The kernel method also offers the flexibility to examine the impact of multiple

industry firms (conglomerates firms), as kernel weights can be defined to span more than one industry. We construct a conglomerate-adjusted kernel using FIC classifications as follows. First, we use the COMPUSTAT segment tapes to identify how many segments each firm has. For firms with one segment, we do not alter the kernel. For a firm with $N > 1$ segments, we assign the firm to the N 10K-300 industries that it is most similar to, and then follow two steps. First, we compute the average characteristic for each 10K-300 industry. Then, for the conglomerate firm spanning $N > 1$ such industries, we assign its industry characteristic as the average of the N corresponding industry specific values. Our results discussed below show that the impact of conglomerates is small, as conglomerate adjusted kernels do not offer material improvements relative to unadjusted kernels.

The last method we consider is a similarity weighed kernel. Rather than computing an equal weighted average (flat kernel), we use a similarity weighted average.[11] This method can only be used for VIC industries, as VIC industries are defined around a single firm, which provides a natural reference point with which to measure each firm's relative similarity.

Our goal is to compare industry classifications. Results from the prior section suggest that SIC-3 and NAICS-4 should be the appropriate benchmarks for comparing our 10-K based measures. Among 10K-based industries, we then consider 10K-300 FIC industries, as well as 10K-VIC-7.06 industries, as both have attributes calibrated to match SIC-3 based industries (discussed in Section III). Table III displays the results.

[Insert Table III Here]

Table III shows that 10-K based industries outperform both SIC and NAICS, especially VIC industries, which do not require transitivity, and are recomputed in each year. When limiting attention to fixed effects based on FIC industries, adjusted R-squared for profitability scaled by sales increases by 11.6% from 0.285

[11] Technically, we use adjusted similarity weights, where we subtract the similarity threshold used to define the industry from the similarity weights. This way, the weights have the nice property of being bounded below by zero (a firm that just barely gets assigned to the industry will have a weight near zero), allowing similarities to be more informative.

to 0.318 when the 10-K based classifications are used rather than the SIC-3 based classifications. The improvement is a similar 10.4% when 10K-300 industries are used rather than NAICS-4 industries. The improvement in explanatory power is nearly twice as large at 22.5% for operating income scaled by assets rather than sales.

For other firm characteristics, most have stronger results for 10-K based FIC industries (sales growth, R&D/sales, advertising/sales book to market ratios) compared to SIC or NAICS. However, some variables (dividend payer dummy, book leverage, and market leverage) have results that are slightly weaker for 10-K based FIC industries. One explanation is that dividends and leverage are managerial policies, and these policies might be chosen based on easily computed industry averages. For example, managers might target SIC or NAICS benchmarks because these targets are inexpensive to obtain.

By comparing flat kernel results (simple averages over the firm characteristics within an industry group) in columns 2, 4, and 6 to standard fixed effects in columns 1, 3, and 5, we conclude that the flat kernel offers significantly higher explanatory power despite its usage of a single degree of freedom. The main reason is that the kernel method allows the industry controls to vary over time (the kernel average is computed in each industry separately in each year). It is thus more analogous to controlling for industry x year fixed effects (T years x N industry degrees of freedom) than it is to controlling for industry and year fixed effects (T years + N industry degrees of freedom). The kernel method thus has two important advantages: (1) it employs just one degree of freedom, and (2) it has more explanatory power and accounts for industry changes over time. Its improvement in power can be large, for example its adjusted R-squared is nearly 3x higher for sales growth. In general, the kernel dominates fixed effects, and its gains range from a 10% improvement, to much more dramatic gains. Finally, the table also shows that the conglomerate adjusted 10K-300 kernel performs just as well as the unadjusted 10K-300 kernel. We conclude that conglomerates have little influence on our data, and that using the industry of highest overall similarity is an excellent approximation of a conglomerate's characteristics.

The last two columns display results for VIC industries. Rows one and two show

that VIC industries offer substantial improvements in explaining profitability, especially relative to SIC and NAICS codes. For example, the operating income/sales adjusted R-squared of roughly 31% for VIC, is 44% higher than the 28.5% adjusted R-squared for standard SIC-3 fixed effects, and 29% higher than the SIC-3 flat kernel. Perhaps even more striking, the similarity weighted kernel accomplishes this performance even though we exclude the firm itself from the weighted average. This is a mechanistic disadvantage, as both fixed effects and flat kernels include the firm itself in their averages.[12] We conclude that VIC industries offer substantial improvements over existing methods used in the literature. This conclusion is also conservative because the VIC kernel excludes the reference firm.

V Market Structure

In this section, we explain how we construct measures of industry market structure (also sometimes viewed as measures of industry competitiveness) and present summary statistics. We first consider existing measures based on firm market shares alone (HHI and C4 indices) and measures based on similarity alone (summed and average similarity). We also consider a consumer choice approach that results in a unified measure of competitiveness based on both market shares and similarities. We will refer to this new index as the "Product Uniqueness Index" (PUI Index).

A Measuring Market Structure

Consider an industry with N firms, and let SL_i denote firm i's sales. We use the COMPUSTAT database to identify each firm's sales in each year. However, we winsorize firm sales at the 5%/95% level in each year to reduce the impact of outliers, as some firms have substantially higher sales than other firms in our sample.[13] The Herfindahl (HHI) index and the C4 index are defined as follows:

[12] If the reference firm is included using the similarity kernel, and it is given a similarity weight of 1, adjusted R-squared increases to near 70% (not reported). Because this might over-weight the reference firm, we do not recommend using similarity kernels that include the reference firm, although flat kernels are likely ok as they apply equal weight.

[13] Results are similar, but somewhat weaker for HHI and C4 indices if we use non-winsorized sales. Using logged sales rather than winsorized sales also generates similar results.

$$HHI = \sum_{i=1}^{N} \left(\frac{SL_i}{\sum_{i=1}^{N} SL_i} \right)^2 \tag{1}$$

$$C4 = \sum_{i=1}^{4largest} \frac{SL_i}{\sum_{i=1}^{N} SL_i} \tag{2}$$

HHI indices and C4 indices can be computed for both FIC and VIC industries. Our remaining indices are only defined for VIC industries, as they require the existence of a reference firm. Consider a VIC industry with N+1 firms, and let one of the firms be the reference firm, and the other N firms are its rivals. The method for constructing VIC industries generates industries that are firm-specific, which are based on similarities exceeding a minimum threshold relative to a reference firm. Let S_i denote firm i's "net" similarity relative to the reference firm ($i \in 1, ..., N$).[14] Our next two measures are more closely measures of competitiveness rather than market structure, and are functions of similarities alone as follows (Seim (2006) constructs a similar Total Similarity Index):

$$TotalSimilarity = \sum_{i=1}^{N} S_i \tag{3}$$

$$AverageSimilarity = \frac{TotalSimilarity}{N} \tag{4}$$

The four market structure and competitiveness measures discussed above account for either firm market shares or firm similarities, but not both. We now consider a consumer choice model in order to construct a measure that accounts for both. As above, consider an industry with N+1 firms, with one being a reference firm that has N rivals. Let SL_i and S_i be the sales and similarities of the N rivals ($i \in 1, ..., N$), and let SL_{ref} denote the sales of the reference firm. We assume that a single consumer demands a product similar to the one produced by the reference firm. She considers possible substitutes to the reference firm offered by the rivals one at a time before making a decision. Denote the probability that the consumer chooses rival firm i

[14]Net similarity is the raw pairwise similarity minus the minimum similarity threshold used to form the given VIC industry (for example, we focus on the threshold .0706 as discussed earlier). We use net similarities because they have the intuitive property that firms just barely gaining access to the industry would have nearly zero impact on the competitiveness index.

over the reference firm as P_i. She will thus select the reference firm with probability P_{ref}, which can be written:

$$P_{ref} = \prod_{i=1}^{N} 1 - p_i \tag{5}$$

We assume that the consumer's selection rule regarding substitutes proceeds in two simple steps. First, she determines if rival firm i's product is a viable substitute to the reference firm, and we assume this will occur with probability equal to i's similarity relative to the reference firm (S_i). Second, if the given rival is a substitute, then the consumer prefers rival i over the reference firm with probability $\frac{SL_i}{SL_i+SL_{ref}}$. If i is not a substitute, then the consumer prefers the reference firm over firm i. The likelihood of selecting the reference firm is thus:

$$\text{PUI Index} \;=\; P_{ref} \;=\; \prod_{i=1}^{N} \left(1 - \frac{S_i \; SL_i}{SL_i + SL_{ref}}\right) \tag{6}$$

Henceforth, we refer to P_{ref} as the "Product Uniqueness Index" (PUI index). If PUI is close to one then the reference firm will be selected with certainty and the firm has products that are unique and do not face much competition. It captures the main predictions regarding how similarity and market shares jointly influence expected industry profitability, and being a probability, it also has non-extreme distributional properties. For example, existing theories suggest that larger firms face less competition (an implication of the HHI calculation). The PUI index predicts this relationship as a larger SL_{ref} increases the likelihood that the consumer will select the reference firm. As SL_{ref} goes to infinity, the consumer will choose the central firm with probability one. Existing theories also predict that increased within-industry similarities reduce expected profits. The PUI index also captures this prediction, as higher values of S_i increase the likelihood that the consumer will choose a rival over the reference firm. In summary, the PUI index is bounded in the interval [0,1], and firms having a higher PUI index should have higher expected profits.

B Summary Statistics

Panel A of Table IV presents summary statistics for the five 10K-VIC-7.06 industry market structure and competitiveness measures. The Product Uniqueness Index (PUI) defined in equation (6) is indeed bounded in [0,1], and its distribution spans

this entire range with a mean close to the center (0.426). These results confirm that this variable is unlikely to be influenced by outliers. The total summed similarity defined in equation (3) is zero when no rivals exist, and has a mean of 4.791. Average similarity has a mean of 0.029, indicating that the average rival has a similarity of 7.06% + 2.9% = 9.96% relative to the reference firm.[15]

[Insert Table IV Here]

The average Sales Herfindahl (HHI) is 0.108, indicating that industries are somewhat disperse. The C4 index averages 43.2%.

Panel B shows that HHI and C4 indices based on SIC-3 and NAICS-4 have similar means to one another, but they are somewhat different from the 10K-based statistics in Panel A. For example, the average SIC-3 based C4 is 62.2%, which is close to the 62.1% for NAICS-4. These are both larger than the 41.1% in Panel A for 10K-based industries. SIC-3 and NAICS-4 are thus more similar to one another than to 10K-based industries. Panel C shows that the average firm in our sample is profitable and engages in advertising and research and development.

The variables in Panels A and B measure industry market structure and competitiveness. Hence, they should be correlated, and the degree of correlation should indicate how much information is unique to a given measure. Table V displays Pearson correlation coefficients for these measures. The table shows two key findings: (1) 10-K based measures are strongly correlated with each other, and (2) SIC-3 and NAICS-4 variables are strongly correlated with each other, but not with 10K-based variables.

[Insert Table V Here]

Consistent with the first key finding, the PUI Index is highly correlated with the 10K-based HHI (61.1%) and the C4 index (73.7%). It is -59.2% and -59.9% correlated with total and average similarity, respectively. The PUI Index thus contains significant overlap with similarity and concentration, consistent with its goal of summarizing both effects. In contrast, the 10K-based HHI index is only 31% correlated

[15] Although average and total similarity have some moderately extreme values, using logs does not change our results materially.

with the 10K-based similarity variables, suggesting that other measures do not capture both components as well. Finally, the high correlation among the SIC-3 and NAICS-4 variables indicates that NAICS and SIC are similar. However, because the NAICS HHI is more correlated with the 10-K based HHI (10.5% vs 8.7%), NAICS likely offers small improvements over SIC.

VI Market Structure and Profitability

In this section, we examine which measures of market structure best explain observed profitability. In the next section, we examine the dynamic properties of these measures, examining how market structure and competitiveness change over time as firms advertise and conduct R&D

Table VI reports the results of panel data regressions in which firm profitability is the dependent variable. We also include controls for industry and year fixed effects, and all t-statistics are adjusted to control for clustering by year and industry (10K-300 FIC industries). We examine robustness to controlling for SIC-3 and NAICS-4 fixed effects later in this section. We also include controls for the industry's average firm size (the natural logarithm of firm assets) and the industry average value growth orientation (the natural log of the firm's book to market ratio), and the accompanying negative book to market ratio dummy.[16] Book to market ratios are computed as in Davis, Fama, and French (2000).

Table VI shows that three of the five indices reliably explain profitability in the direction predicted by theory. The PUI Index, the sales HHI, and the sales C4 index explain profitability at better than the 1% level in both panels. The C4 index is marginally stronger in Panel A (profitability scaled by sales), and the PUI Index is marginally stronger in Panel B (profitability scaled by assets). Total similarity, and average similarity are the two weakest proxies in both panels, and are more significant for operating income scaled by sales in Panel A. Overall, our results support the conclusion that 10K-VIC-7.06 market structure measures perform well in explaining profitability.

[16]Results are similar if we use firm-level controls instead.

[Insert Table VI Here]

We also reproduced the tests in Table VI using SIC-3 based industry controls rather than 10K-300 controls. The results (which are available from the authors, but are not reported to conserve space) show that all of our key variables are similar or slightly stronger, and all are significant at better than the 1% level. We conclude that our results cannot be explained by traditional SIC-controls.[17]

The market structure measures in Tables VI are based on 10K-VIC-7.06 industries. Table VII tests if measures constructed using SIC or NAICS codes can generate similar results. Firm profitability is the dependent variable, and we include industry and year fixed effects, and adjust standard errors for clustering, as before. The table shows that none of the market structure measures based on SIC-3 or NAICS reliably predicts ex-post profitability. The NAICS-based measures are negatively related to profitability, whereas theory predicts a positive sign. The SIC-3 based HHI is positive and significant in Panel B for industry level regressions, but is only 10% level significant, or not significant, depending on the specification in Panel A. Comparing this table to Table VI suggests that 10-K based measures are considerably stronger than SIC-3 or NAICS-4 based measures.

[Insert Table VII Here]

VII Changes in Industry Market Structure and Competitiveness

In this section, we examine how measures of market structure and competitiveness change over time, and we focus on Sutton (1991), who predicts that advertising and research and development (R&D) can create endogenous barriers to entry. The main idea is that R&D can create more unique products and advertising can make it more expensive for rivals to enter, thwarting entry. A key assumption is that advertising and R&D (which might be geared toward improving product appeal), are actually effective in reducing ex-post competition. We test this assumption by regressing

[17] We also find similar results if we use NAICS-4 industries instead of SIC-3 industries.

ex-post changes in our market structure and competitiveness measures on ex-ante advertising and R&D.

Importantly, we restrict attention to 10K-VIC-7.06 industries, as variable membership and variable locations are critical to testing Sutton's theory, which is primarily about thwarting movement across industry boundaries. For example, a market structure index will change if a large rival enters a given firm's industry, or if a rival is "pushed out" because other firms spent more heavily on advertising and R&D. A push-out could occur either if the given firm simply abandons the market due to its high costs, or if the industry itself moves away from this firm due to R&D-induced product improvements that move the industry away from this non-spending firm. Either way, increases in the product uniqueness measures should improve ex-post profitability, and VIC industry definitions are flexible enough to incorporate this general set of dynamic effects. SIC-3 and NAICS-4 do not offer the flexibility to test this hypothesis because industry locations are fixed, and memberships rarely change.

[Insert Table VIII Here]

Table VIII displays the results. The dependent variable for each row is noted in the first column, and all variables are ex-post changes in the given competitiveness measure. We find overwhelming support for Sutton's assumption across all of our competitiveness measures. For example, row one shows that firms spending more on advertising experience substantial improvements in their ex-post PUI Index (t-statistic of 9.54), and firms spending more on R&D experience similar improvements (t=6.58). The third and fourth column show that firms with zero spending in either category similarly experience a strong decay in the expected profitability.

Rows (2) to (7) show that other measures of changes to market structure generate similar results, but all are somewhat weaker than the results for the PUI Index. The C4 index is the second most robust variable, and firms spending more on advertising and R&D generate improvements in their ex-post C4 indices. Rows (6) and (7) show that advertising and R&D are also positively related to ex-post changes in observed profitability. Finally, Panel B shows that results are marginally stronger when SIC-3 fixed effects are used rather than 10-K based industry fixed effects, confirming that

our results cannot be explained by SIC-3 controls.[18]

Our results are also consistent with Hoberg and Phillips (2009), who show that mergers and acquisitions can also generate product differentiation from close rivals, and that this is especially relevant when firms face more competition.

[Insert Table IX Here]

Table IX displays the results of tests analogous to those in Table VIII, but focuses on measures of market structure constructed from SIC and NAICS codes. As noted earlier, the location and memberships of these industries are fixed over time. This limitation makes it very difficult to examine how market structure changes over time, as firms rarely change their SIC or NAICS classifications. Hence, we expect far less power to measure the impact of potential endogenous barriers to entry. The table confirms this conjecture, and we find little support of Sutton's hypotheses using these less powerful, less dynamic measures. Comparing these results to those in Table VIII based on dynamic 10K-based VIC industries, leads us to conclude that dynamic industries are essential in providing the empirical flexibility and power needed to test the role of endogenous barriers to entry.

VIII Conclusions

We use web crawling and text parsing algorithms to examine product descriptions from annual firm 10-Ks filed with the SEC. The word usage vectors from each firm generate an empirical Hotelling-like product market space on which all firms reside. We use these word usage vectors to calculate how firms are related to each other and to create new industry classifications. Using these new industry classifications, we calculate new measures of market structure and competition and examine their link to firm profitability. These new measures enable us to test theories of product differentiation and whether firms advertise and conduct R&D to create product differentiation, consistent with Sutton (1991)'s work on endogenous barriers to entry.

Our new dynamic industry classifications are based on how firms describe them-

[18]Not reported, our results also cannot be explained by NAICS-4 controls.

selves in each year in the product description section of their 10Ks. Because our classifications are formed in each year, they do not have the staleness and time-fixed location properties associated with SIC and NAICS. In addition, our main classification method is based on relaxing the transitivity requirement of existing SIC and NAICS industries, and thus allows each firm to have its own potentially unique set of competitors. This new method which we term variable industry classifications (VIC) is analogous to social networks where each individual can have a distinct set of friends.

Our new classifications offer substantial improvements in the ability to explain firm characteristics and are able to more precisely test theories of endogenous barriers to entry. Using these new classifications and our relatedness measures, we create new measures of market structure that capture within-industry competitiveness. Using these industry competitiveness measures, we find strong support for the conclusion that increased product differentiation is associated with increased profitability. There is only weak support for a market structure - profitability relation using SIC and NAICS based variables.

Using these new dynamic classification methods also allows us to examine how industry market structure and competitiveness changes over time, and whether advertising and research and development serve as endogenous barriers to entry. We find strong support for Sutton (1991)'s hypothesis that firms spend on advertising and R&D, at least in part, to reduce future competition and entry. Firms spending more on either advertising or R&D experience significant increases in product differentiation and profitability.

Appendix 1

This Appendix explains how we compute the "product similarity" and "product differentiation" between two firms i and j. We first take the text in each firm's product description and construct a binary vector summarizing its usage of English words. The vector has a length equal to the number of unique words used in the set of all product descriptions. For a given firm, a given element of this vector is one if the word associated with the given element is in the given firm's product description. To focus on products, we restrict the words in this vector to less commonly used words. Very common words include articles, conjunctions, personal pronouns, abbreviations, and legal jargon, for example. Hence, we restrict attention to words that appear in fewer than five percent of all product descriptions in the given year. For each firm i, we thus have a binary vector P_i, with each element taking a value of one if the associated word is used in the given firm's product description and zero otherwise.

We next define the normalized frequency vector V_i, which normalizes the vector $P_{x,i}$ to have unit length.

$$V_i \;=\; \frac{P_i}{\sqrt{P_i \;\cdot\; P_i}} \tag{7}$$

To measure how similar the products of firms i and j are, we take the dot product of their normalized vectors, which is then "product similarity".

$$Product\ Similarity_{i,j} \;=\; (V_i \;\cdot\; V_j) \tag{8}$$

We define product differentiation as one minus similarity.

$$Product\ Differentiation_{i,j} \;=\; 1 \;-\; (V_i \;\cdot\; V_j) \tag{9}$$

Because all normalized vectors V_i have a length of one, product similarity and product differentiation both have the nice property of being bounded in the interval (0,1). This normalization ensures that product descriptions with fewer words are not penalized excessively. This method is known as the "cosine similarity", as it measures the cosine of the angle between two vectors on a unit sphere. The underlying unit sphere also represents an"empirical product market space" on which all firms in the sample have a unique location.

Appendix 2

This appendix describes our FIC industry classification methodology based on 10-K text similarities. Our classification goal is to maximize total within-industry product similarity subject to two constraints. First, in order to be comparable to existing methods, a common set of industries must be created and held fixed for all years in our time series, hence we form a fixed set of industries based on our first full year of data, which is 1997. Second, our algorithm should be sufficiently flexible to generate industry classifications for any number of degrees of freedom. This latter requirement is important because, in order to compare the quality of our new classifications relative to alternatives like three or four digit SIC codes, our classifications should utilize a similar number of degrees of freedom. We achieve these goals using a two stage process: (1) an industry formation stage, which is based on the first full year of our sample; and (2) an industry assignment stage, which assigns firms in all years of our sample to the fixed industries determined in stage one.

We begin the first stage by taking the subsample of N single segment firms in 1997 (multiple segment firms are identified using the COMPUSTAT segment database). We then initialize our industry classifications as being N dimensional, with each of the N firms residing within its own one-firm industry. We then compute the pairwise similarity for each unique pair of industries j and k, which we denote as $I_{j,k}$.

To reduce the dimensionality to $N-1$ industries, we take the maximum pairwise industry similarity as follows

$$\underset{j,k,\ j\neq k}{MAX} \qquad I_{j,k} \tag{10}$$

The two industries with the highest similarity are then combined, resulting in a one dimension reduction in the number of industries. This process is repeated until the number of industries reaches the desired number of degrees of freedom. Importantly, when two industries with m_j and m_k firms are combined, all industry similarities relative to the new industry must be recomputed. For a newly created industry l, for example, its similarity with respect to all other industries q is computed as the average firm pairwise similarity for all firm pairs in which one firm is in industry l

and one in industry q as follows:

$$I_{l,q} = \sum_{x=1}^{m_l} \sum_{y=1}^{m_q} \frac{S_{x,y}}{m_l \quad m_q} \tag{11}$$

Here, $S_{x,y}$ is the firm-level pairwise similarity between firm x in industry l and firm y in industry q.

Although this method guarantees maximization of within-industry similarity after one iteration, it does not guarantee this property after more than one iteration. For example, a firm that initially fits best with industry j after one iteration might fit better with another industry k after several iterations because industry k was not an option at the time the initial classification to industry j was made. Thus, we recompute similarities ex-post to determine whether within industry similarity can be improved by moving firms to alternative industries. If similarity can be improved, we reclassify suboptimally matched firms to their industry of best fit.

Once this process is complete, the set of industries generated by the algorithm will have the desired number of degrees of freedom, and will have the property that within industry similarity cannot be maximized further by moving any one firm to another industry. It is important to note, however, that industry classifications fitting this description are not necessarily unique. It is plausible that multiple simultaneous firm reassignments can further improve within-industry similarity. Although we do not take further steps to ensure uniqueness due to computational limitations, we believe the quality of our classifications is rather good, especially given their empirical performance. Moreover, any departure from the truly optimal set of industries would only bias our study away from finding significant results, and hence our approach is conservative and might understate the true power of product descriptions.

The industry assignment stage takes the industries formed in the first stage as given, and assigns any given firm in any year to the industry it is most similar to. We begin by computing an aggregate word usage vector for each industry. Each vector is based on the universe of words appearing in fewer than 5% of all firms in 1997 as before. The vector is populated by the count of firms in the given industry using the given word, and this vector is then normalized to have unit length (similar to how we compute firm pairwise similarities in Appendix 1). This normalization ensures

that industries using more words are not rewarded on the basis of size, but rather are only rewarded on the basis of similarity. For a given firm that we wish to classify, we simply compute its similarity to all of the candidate industries, and assign the firm to the industry it is most similar to. A firm's similarity to an industry is simply the dot product of the firm's normalized word vector to the industry's normalized word vector.

Although we use the first full year of our sample, 1997, to form industries, we do not believe that this procedure generates any look ahead bias. The industry formation itself is purely a function of the text in product descriptions and the definition of a multiple segment firm obtained from COMPUSTAT. We use multiple segment identifiers from 1996, which precedes our sample. We examine profitability from 1997 to 2006, and we can further report that our results are virtually unchanged if we omit 1997 from this analysis. It is also relevant to note that the NAICS industry classification was created in the middle of our sample, and we find that our industry classifications outperform NAICS despite the propensity for NAICS classifications to have a recency advantage.

References

Aghion, Philippe, Nicholas Bloom, Richard Blundell, Rachel Griffith, and Peter Howitt, 2005, Competition and innovation: an inverted u relationship, *Quarterly Journal of Economics* 120, 701–28.

Berry, Steven, 1990, Airport presence as product differentiation, *American Economic Review* 80, 394–399.

——— , James Levinsohn, and Ariel Pakes, 1997, Automobile prices in market equilibrium, *Econometrica* 63, 841–890.

Bhojraj, Sanjeev, Charles Lee, and Derek Oler, 2003, What's my line? a comparison of industry classifications for capital market research, *Journal of Accounting Research* 41, 745–774.

Boukus, Ellyn, and Joshua Rosenberg, 2006, The information content of fomc minutes, Yale University working paper.

Chamberlin, EH, 1933, *The Theory of Monopolistic Competition* (Harvard University Press: Cambridge).

Davis, James, Eugene Fama, and Kenneth French, 2000, Characteristics, covariances, and average returns: 1929-1997, *Journal of Finance* 55, 389–406.

Fama, Eugene, and Kenneth French, 1997, Industry costs of equity, *Journal of Financial Economics* 43, 153–193.

Hanley, Kathleen, and Gerard Hoberg, 2009, The information content of ipo prospectuses, Forthcoming Review of Financial Studies.

Hay, D.A., 1976, Sequential entry and entry-deterring strategies in spatial competition, *Oxford Economic Papers* 28, 240–257.

Hoberg, Gerard, and Gordon Phillips, 2009, Competition and product market synergies in mergers and acquisitions: A text based analysis, Working Paper, University of Maryland.

Hotelling, H., 1929, Stability in competition, *Economic Journal* pp. 41–57.

Kahle, Kathleen, and Ralph Walkling, 1996, The impact of industry classifications on financial research, *Journal of Financial and Quantitative Analysis* 31, 309–335.

Katz, Barbara, 1978, Territorial exclusivity in the soft drink industry, *Journal of Industrial Economics* 27, 85–96.

Kelton, Christine, and David Kelton, 1982, Advertising and intraindustry brand shift in the u.s. brewing industry, *Journal of Industrial Economics* 30, 293–303.

Krishnan, Jayanthi, and Eric Press, 2003, The north american industry classification system and its implications for accounting research, *Contemporary Accounting Research* 20, 685–717.

Kwon, Oh-Woog, and Jong-Hyeok Lee, 2003, Text categorization based on k-nearest neighbor approach for web site classification, *Information Processing & Management* 39, 25–44.

Li, Feng, 2006, Do stock market investors understand the risk sentiment of corporate annual reports?, University of Michigan Working Paper.

Lin, Ping, and Kamal Saggi, 2002, Product differentiation, process r&d, and the nature of market competition, *European Economic Review* 46, 201–211.

Loughran, Tim, and Bill McDonald, 2008, Plain english, Notre Dame University working paper.

Mazzeo, Michael, 2002, An empirical model of firm entry with endogenous product choices, *Rand Journal of Economics* 33, 221–42.

Nevo, Aviv, 2000, Mergers with differentiated products: the case of the ready to eat cereal industry, *Rand Journal of Economics* 31, 395–421.

Panzar, J., and R. Willig, 1981, Economies of scope, *American Economic Review* 71, 268–272.

Rauh, Joshua, and Amir Sufi, 2010, Explaining corporate capital structure: Product markets, leases, and asset similarity, Northwestern University Working Paper.

Schmalensee, Richard, 1978, Entry deterrence in the ready-to-eat breakfast cereal industry, *Bell Journal of Economics* 9, 305–327.

Seim, Katja, 2006, An empirical model of firm entry with endogenous product choices, *Rand Journal of Economics* 37, 619–40.

Shaked, Avner, and John Sutton, 1987, Product differentiation and industrial structure, *Journal of Industrial Economics* 26, 131–146.

Sutton, John, 1991, *Sunk Costs and Market Structure* (MIT Press: Cambridge, Mass).

Tetlock, Paul, Maytal Saar-Tsechanksy, and Sofus Macskassy, 2008, More than words: Quantifying language to measure firms' fundamentals, *Journal of Finance* 63, 1437–1467.

Tetlock, Paul C., 2007, Giving content to investor sentiment: The role of media in the stock market, *Journal of Finance* 62, 1139–1168.

Table I: Sample Industries Classified Using Product Similarity

INDUSTRY WITH 26 FIRMS, LISTED AS (SIC GROUP) LIST OF FIRM NAMES:
(286 INDUSTRIAL ORGANIC CHEMICALS): NOVA CORP, (602 COMMERCIAL BANKS): US BANCORP, (609 FUNCTIONS RELATED TO DEPOSITORY BANKING): CONCORD EFS, ELECTRONIC CLEARING HOUSE, NATIONAL PROCESSING, MONEYGRAM PAYMENT SYSTEMS, EURONET SERVICES, (641 INSURANCE AGENTS, BROKERS, AND SERVICE): NATIONAL DATA CORP, ENVOY CORP , (732 CONSUMER CREDIT REPORTING AGENCIES, MERCANTILE REPORTING AGENCIES,): FIRST USA PAYMENTECH INC, (737 COMPUTER PROGRAMMING, DATA PROCESSING, AND OTHER COMPUTER RELATED): ELECTRONIC DATA SYSTEMS CORP , FIRST DATA CORP, IFS INTERNATIONAL, AFFILIATED COMPUTER SERVICES, SIMS COMMUNICATIONS, TRANSACTION SYSTEMS ARCHITECTS, CYBERCASH, SABRE GROUP HOLDINGS, GENISYS RESERVATION SYSTEMS, GALILEO INTERNATIONAL, PEGASUS SYSTEMS, (738 MISCELLANEOUS BUSINESS SERVICES): SPS TRANSACTION SERVICES, PMT SERVICES, BA MERCHANT SERVICES, (872 ACCOUNTING, AUDITING, AND BOOKKEEPING SERVICES): CERIDIAN CORP, (999 NONCLASSIFIABLE ESTABLISHMENTS): CRW FINANCIAL

industry words listed as (number of firms using word) word list:
(20) debit, (19) terminals, (18) visa, (17) processors, merchant, (16) electronically, (15) terminal, outsourcing, capture, checks, (14) issuer, (13) processor, issuing, cardholder, (12) interfaces, transmitted, interchange, merchants, travel, express, host, clearing, (11) verification, fraud, teller, (10) transmit, transmits, recurring, mainframe, issuers, (9) link, bill, explore, technologically, portfolios, smart, (8) deployment, connect, switch, returned, inquiries, linked, online, authorizations, dial, accepting, remittance, desk, discover, fraudulent, atms, cardholders, (7) protocols, stop, club, purchaser, runs, niche, delivers, staffing, lost, branded, telemarketing, automate, compare, hotels, eliminates, databases, batch, fulfillment, (6) airline, airlines, adapt, distance, installing, telecommunication, wire, match, popularity, coordination, assert, bills, nation, accessible, membership, derives, facilitates, gaming, transferring, load, messages, retrieval, authorizing, completing, inquiry, message, checking, handled, outsource, supermarkets, travelers, authorizes, split, vertically, payroll, precautions, legacy, owed, hospitality
.* words with frequency 5 or less omitted to conserve space.

INDUSTRY WITH 13 FIRMS, LISTED AS (SIC GROUP) LIST OF FIRM NAMES:
(138 OIL AND GAS FIELD SERVICES): VERITAS DGC, TGC INDUSTRIES, UNIVERSAL SEISMIC ASSOCIATES, WESTERN ATLAS, 3-D GEOPHYSICAL, EAGLE GEOPHYSICAL, OMNI ENERGY SERVICES CORP, (353 CONSTRUCTION, MINING, AND MATERIALS HANDLING MACHINERY AND EQUIPME): BOLT TECHNOLOGY CORP, (381 SEARCH, DETECTION, NAVIGATION, GUIDANCE, AERONAUTICAL, AND NAUTICA): LABARGE, (382 LABORATORY APPARATUS AND ANALYTICAL, OPTICAL, MEASURING, AND CONTR): INPUT OUTPUT, GEOSCIENCE CORP, (735 MISCELLANEOUS EQUIPMENT RENTAL AND LEASING): MITCHAM INDUSTRIES, (737 COMPUTER PROGRAMMING, DATA PROCESSING, AND OTHER COMPUTER RELATED): XOX CORP

industry words listed as (number of firms using word) word list:
(13) seismic, exploration, geophysical, (11) surveys, (10) drilling, recording, (9) subsurface, survey, geophones, (8) cables, signals, crews, earth, reservoir, dimensional, (7) geographical, transmitted, hole, marine, holes, atlas, terrain, (6) deployed, rock, magnetic, explosive, drill, shallow, geoscience, (5) bidding, analog, transmit, penalty, output, trucks, succeeding, vessel, zone, technique, vessels, petroleum, positioning, sensors, geological, interpretation, explosives, formations, geologic, dependability, strata, streamers, dynamite, (4) imaging, finding, instrumentation, depreciation, revolving, copyrights, failures, varied, denominated, deploy, navigation, fleet, repairs, boxes, precision, preserve, characterization, attempted, libraries, waves, geoscientists, oilfield, towed, zones, reservoirs, shipboard, intervals, telemetry, meters, geophysicists, hydrophones, streamer
* words with frequency 3 or less omitted to conserve space.

Sample industries were created using the 10-K based product similarity clustering algorithm described in Appendix 2. They are based on an industry classification that created 300 industries in total, and are based on the classification year 1997.

Table I: Sample Industries Classified Using Product Similarity (Cont)

INDUSTRY WITH 419 FIRMS, LISTED AS (SIC GROUP) LIST OF FIRM NAMES:
Firms span 4 key two digit SIC codes, listed in declining frequency (28 DRUGS), (38 SURGICAL/MEDICAL SUPPLIES), (87 RESEARCH AND TESTING SERVICES), (80 MEDICAL AND DENTAL LABS):

(369) trials, (345) efficacy, (342) therapeutic, (338) drugs, (333) commercialization, (331) blood, (328) diseases, (307) therapy, (306) trial, (303) animal, (293) preclinical, (292) cells, (289) biotechnology, (288) investigational, (284) indications, (282) cell, (274) commercialize, (273) cancer, (262) tissue, (258) inventions, (256) compounds, (254) therapies, (253) pharmaceuticals, (249) biological, (246) cosmetic, (245) discovery, collaborative, (243) treatments, (233) collaboration, (232) academic, (230) humans, novel, (227) clearance, (222) evidence, (220) protein, (219) death, (217) indication, (216) treating, proteins, (211) compound, (209) dose, (207) subjects, (206) valid, chemistry, (204) scientists, medicine, (203) chronic, (201) disorders, (199) rigorous, molecular, (197) milestone, (196) healthy, surgery, physician, (195) universities, (194) validity, collaborators, (192) protocol, formulation, (189) withdrawal, therapeutics, (188) dosage, (186) questions, (185) vitro, infection, (184) succeed, (183) lengthy, molecules, (182) tissues, (181) immune, (180) oral, (179) infectious, (178) surveillance, optimal, (177) animals, collaborations, (176) approaches, (175) confidential, discovered, payors, (174) screening, (173) prosecution, (172) causes, surgical, acute, (171) prevention, (169) clinically, (167) milestones, heart, (166) infringing, discoveries, (165) protocols, (164) toxicity, (162) cardiovascular, (161) diagnosis, researchers, molecule, (160) causing, breach, metabolism, (158) commercialized, (157) advisors, candidate, (156) undergo, factual, (155) skin, (154) doses, (153) remedies, (152) synthetic, biology, (151) infections, preventing, unknown, (150) aids, tumor, genetic, threatening, (149) gene, (148) interference, achievement, circumvented, complications, (146) binding, radioactive, invention, mechanism, (145) notification, antibodies, symptoms, (144) lung, (143) innovations, incidence, bone, (142) biopharmaceutical, diagnosed, recall, (141) substance, (140) inhibit, (139) exposed, insurers, (138) clearances, (137) literature, acid, accumulated, cleared, (136) injection, pain, (135) cancers, (134) institutes, (133) encounter, sufficiently, liver, (132) tumors, (131) defend, volunteers, (130) formulations, recombinant, (129) inflammatory, suffer, (128) completely, experimental, sciences, (127) mechanisms, invalidated, invasive, breast, (126) refusal, withdrawn, expend, launch, tolerance, efficacious, (125) undergoing, brain, viable, investigator, (123) breached, attempting, deficit, pharmacology
.* words with frequency 122 or less omitted to conserve space.

INDUSTRY WITH 14 FIRMS, LISTED AS (SIC GROUP) LIST OF FIRM NAMES:
(357 COMPUTER AND OFFICE EQUIPMENT): MICROTOUCH SYSTEMS, INTERLINK ELECTRONICS, (366 COMMUNICATIONS EQUIPMENT): CHECKPOINT SYSTEMS, DETECTION SYSTEMS, NAPCO SECURITY SYSTEMS, SENSORMATIC ELECTRONICS CORP, INTERNATIONAL ELECTRONICS, ITI TECHNOLOGIES, SENTRY TECHNOLOGY CORP, ENSEC INTERNATIONAL, STRATESEC, (382 LABORATORY APPARATUS AND ANALYTICAL, OPTICAL, MEASURING, AND CONTR): APOLLO INTERNATIONAL OF DELAWARE, (737 COMPUTER PROGRAMMING, DATA PROCESSING, AND OTHER COMPUTER RELATED): LORONIX INFORMATION SYSTEMS, (738 MISCELLANEOUS BUSINESS SERVICES): PROTECTION ONE

industry words listed as (number of firms using word) word list:
(12) detection, (11) alarm, (9) station, door, alarms, sensors, detect, (8) attached, zone, readers, activated, sensor, (7) configurations, hard, panels, signals, signature, magnetic, installations, mounted, inside, intrusion, surveillance, (6) glass, interfaces, wireless, cameras, motion, configured, microprocessor, programmed, optional, manual, detectors, false, infrared, (5) audio, doors, installing, panel, passive, transmit, wire, wired, images, assembles, afford, compatibility, electronically, receiver, controllers, verification, communicate, heat, recording, break, command, (4) imaging, premises, switch, terminals, engineered, film, games, soft, vertical, motors, window, coordinates, proximity, recurring, kinds, labels, intelligence, micro, peripheral, staffing, frequencies, causing, deactivation, installs, integrating, microwave, expended, vibration, smoke, workforce, lock, keyboard, deter, activation, touch, speakers, technicians, integrators, sophistication, inexpensive, graphical, buttons, sensing, lights, perimeter, theft
* words with frequency 3 or less omitted to conserve space.

Sample industries were created using the 10-K based product similarity clustering algorithm described in Appendix 2. They are based on an industry classification that created 300 industries in total, and are based on the classification year 1997.

Table II: Industry classifications and profitability

Row	Industry Definition	oi/sales		oi/assets		# of Industries	Avg # Firms per Industry
		Akaike Information Criterion	Adj R^2	Akaike Information Criterion	Adj R^2		
	Panel A: SIC-code based industry definitions						
(1)	SIC-1-digit	3806.8	0.147	-38.8	-0.000	10	562.1
(2)	SIC-2-digit	3299.2	0.229	-269.7	0.042	72	78.1
(3)	SIC-3-digit	3110.1	0.278	-687.6	0.120	274	20.5
(4)	SIC-4-digit	3051.4	0.303	-815.6	0.167	434	13.0
	Panel B: NAICS based industry definitions						
(5)	NAICS-1-digit	4311.4	0.066	-192.2	0.029	9	624.5
(6)	NAICS-2-digit	3571.0	0.183	-479.5	0.079	23	244.4
(7)	NAICS-3-digit	3247.0	0.237	-755.5	0.133	96	58.6
(8)	NAICS-4-digit	3123.3	0.278	-835.3	0.173	328	17.1
(9)	NAICS-5-digit	3418.6	0.271	-520.0	0.162	672	8.4
(10)	NAICS-6-digit	3622.1	0.272	-308.6	0.162	984	5.7
	Panel C: 10-K product description based industry definitions						
(11)	10K-based-50	3026.2	0.260	-1015.1	0.164	50	112.4
(12)	10K-based-100	2998.2	0.267	-1064.7	0.175	100	56.2
(13)	10K-based-200	2762.1	0.306	-1231.8	0.209	200	28.1
(14)	10K-based-250	2761.7	0.310	-1204.1	0.209	250	22.5
(15)	10K-based-300	2715.2	0.320	-1231.6	0.218	300	18.7
(16)	10K-based-400	2854.9	0.309	-1111.4	0.208	400	14.1
(17)	10K-based-500	2744.0	0.330	-1178.8	0.226	500	11.2
(18)	10K-based-800	2682.8	0.355	-1231.4	0.253	800	7.0

The table reports average Akaike Information Criterion (AIC) for cross sectional regressions in which profitability is regressed on a specified set of industry fixed effects. To avoid clustering over time (which would bias AIC tests), we run separate regressions in each year from 1997 to 2006 and report average AIC scores.

Table III: Firm Characteristics and Industry Classifications

Row	Variable	Adj R^2 SIC-3 Fixed Effects	Adj R^2 SIC-3 Kernel	Adj R^2 NAICS-3 Fixed Effects	Adj R^2 NAICS-4 Kernel	Adj R^2 10-K 300 Fixed Effects	Adj R^2 10-K 300 Kernel	Adj R^2 Conglom. Adjusted 10-K 300 Kernel	Adj R^2 VIC Flat Kernel	Adj R^2 VIC Sim. Kernel (Excl Self)
(1)	OI/Sales	0.285	0.318	0.288	0.321	0.318	0.349	0.354	0.405	0.421
(2)	OI/Assets	0.178	0.213	0.184	0.222	0.218	0.253	0.255	0.317	0.301
(3)	Sales Growth	0.023	0.075	0.025	0.086	0.027	0.068	0.067	0.097	0.045
(4)	R&D/Sales	0.139	0.170	0.138	0.171	0.151	0.181	0.187	0.191	0.217
(5)	Adver./Sales	0.040	0.086	0.061	0.114	0.076	0.133	0.134	0.200	0.162
(6)	CAPX/Sales	0.045	0.077	0.053	0.091	0.041	0.061	0.065	0.072	0.068
(7)	Book Leveage	0.220	0.250	0.238	0.269	0.214	0.234	0.225	0.257	0.244
(8)	Market Leveage	0.275	0.316	0.301	0.344	0.273	0.304	0.294	0.327	0.325
(9)	COGS/sales	0.098	0.121	0.101	0.130	0.109	0.129	0.134	0.134	0.139
(10)	SG+A/sales	0.068	0.104	0.070	0.111	0.083	0.111	0.113	0.177	0.136
(11)	Market Beta	0.096	0.159	0.097	0.166	0.097	0.155	0.152	0.190	0.142
(12)	Log(B/M) Ratio	0.188	0.250	0.194	0.262	0.198	0.255	0.250	0.283	0.238
(13)	Dividend payer Dummy	0.362	0.384	0.352	0.374	0.348	0.364	0.351	0.404	0.374

Firm characteristics are regressed on various industry industry controls, including fixed-effect-based and kernel-based controls. All regressions are based on our entire sample from 1997 to 2006, and also include yearly fixed effects.

Table IV: Summary Statistics

Variable	Mean	Std. Dev.	Minimum	Median	Maximum
Panel A: 10K based Market Structure Measures					
Product Uniqueness Index (PUI)	0.426	0.329	0.000	0.411	1.000
Total Summed Similarity	4.791	7.732	0.000	1.932	64.790
Average Similarity	0.029	0.019	0.000	0.024	0.352
Sales 10K-Based Herfindahl	0.097	0.122	0.005	0.054	1.000
Sales 10K-Based C4 Index	0.411	0.247	0.030	0.359	1.000
Panel B: SIC-3 and NAICS 4 Market Structure Measures					
SIC-3 HHI	0.178	0.156	0.036	0.120	1.000
SIC-3 C4	0.622	0.196	0.278	0.568	1.000
NAICS-4 HHI	0.183	0.167	0.022	0.120	1.000
NAICS-4 C4	0.621	0.209	0.175	0.580	1.000
Panel C: Industry and Firm Characteristics					
Industry Advertising/Sales	0.013	0.047	0.000	0.005	2.419
Industry R&D/Sales	0.389	1.296	0.000	0.009	37.258
Industry Log B/M Ratio	-0.720	0.606	-8.463	-0.660	2.565
Industry oi/sales	0.049	0.262	-1.000	0.094	0.982
Firm oi/sales	0.056	0.353	-1.000	0.113	1.000
Firm Log Assets	5.486	2.113	-3.079	5.407	14.210

Summary statistics are reported for our sample of 50,673 observations based on 1997 to 2006. The market structure measures in Panel A are based on VIC-7.06 industries (uses the same number of pairings as three digit SIC codes). Those in Panel B are based on existing three digit SIC and four digit NAICS industries.

Table V: Pearson Correlation Coefficients

Row	Variable	Product Uniqueness Index (10-K based)	Total Summed Similarity (10-K based)	Average Similarity (10-K based)	Sales Herfindahl Index (10-K based)	Sales C4 Index (10-K based)	Sales Herfindahl Index (SIC-3 based)	Sales C4 Index (SIC-3 based)	Sales Herfindahl Index (NAICS-4 based)
		Correlation Coefficients							
(1)	Total Summed Similarity (10-K based)	-0.592							
(2)	Average Similarity (10-K based)	-0.599	0.765						
(3)	Sales Herfindahl (10-K based)	0.611	-0.308	-0.306					
(4)	Sales C4 Index (10-K based)	0.737	-0.434	-0.414	0.824				
(5)	Sales Herfindahl (SIC-3 based)	0.253	-0.219	-0.155	0.087	0.107			
(6)	Sales C4 Index (SIC-3 based)	0.287	-0.284	-0.164	0.075	0.091	0.829		
(7)	Sales Herfindahl (NAICS-4 based)	0.247	-0.267	-0.189	0.105	0.132	0.569	0.559	
(8)	Sales C4 Index (NAICS-4 based)	0.326	-0.385	-0.263	0.129	0.170	0.526	0.656	0.828

Pearson Correlation Coefficients are reported for our sample of 50,673 observations based on 1997 to 2006. The 10-K based market structure measures are based on VIC-7.06 industries (uses the same number of pairings as three digit SIC codes).

Table VI: Market Structure Measures and Profitability

Row	Dependent Variable	Product Uniqueness Index	Total Summed Similarity	Average Similarity	Sales 10K-Based Herfindahl Index	Sales 10K-Based C4 Index	Log Assets	Log B/M Ratio	Negative B/M Dummy	Year+ Industry Fixed Effects
					Panel A: Profitability scaled by sales					
(1)	oi/sales	0.113					0.048	0.008	-0.115	Yes
		(5.45)					(26.28)	(2.02)	(-10.30)	
(2)	oi/sales		-0.004				0.051	0.008	-0.115	Yes
			(-2.99)				(21.99)	(2.13)	(-10.16)	
(3)	oi/sales			-0.754			0.051	0.008	-0.115	Yes
				(-2.05)			(22.28)	(1.98)	(-10.08)	
(4)	oi/sales				0.086		0.052	0.007	-0.114	Yes
					(3.69)		(22.26)	(1.78)	(-10.14)	
(5)	oi/sales					0.085	0.054	0.007	-0.112	Yes
						(6.08)	(21.49)	(1.76)	(-10.08)	
					Panel B: Profitability scaled by assets					
(6)	oi/assets	0.086					0.030	-0.004	-0.084	Yes
		(7.82)					(16.46)	(-1.43)	(-8.02)	
(7)	oi/assets		-0.001				0.032	-0.004	-0.083	Yes
			(-1.87)				(15.53)	(-1.43)	(-7.91)	
(8)	oi/assets			-0.197			0.032	-0.005	-0.083	Yes
				(-1.11)			(15.52)	(-1.49)	(-7.86)	
(9)	oi/assets				0.055		0.033	-0.005	-0.083	Yes
					(3.52)		(15.27)	(-1.63)	(-7.92)	
(10)	oi/assets					0.054	0.034	-0.005	-0.082	Yes
						(5.62)	(14.97)	(-1.65)	(-7.87)	

OLS regressions with profitability defined as operating income divided by sales (Panel A) or assets (Panel B) as the dependent variable. All specifications include year and industry fixed effects, and standard errors account for clustering across year and industries. Industry fixed effects are based on the set of 10-K based 300 FIC industries. The sample has 50,673 observations and is from 1997 to 2006. The market structure variables based on VIC-7.06 industries.

Table VII: SIC-3 and NAICS-4 Measures of Market Structure and Profitability

Row	Dependent Variable	Sales SIC-3 HHI	Sales SIC-3 C4 Index	Sales NAICS-4 HHI	Sales NAICS-4 C4 Index	Log Assets	Log B/M Ratio	Negative B/M Dummy	Year+ Industry Fixed Effects	# Obs.
					Panel A: Firm Level Regressions					
(1)	oi/sales	0.041				0.055	0.015	-0.119	Yes	50,673
		(1.65)				(25.10)	(2.95)	(-9.13)		
(2)	oi/sales		-0.014			0.055	0.015	-0.119	Yes	50,673
			(-0.28)			(25.11)	(2.94)	(-9.13)		
(3)	oi/sales			0.037		0.055	0.013	-0.112	Yes	50,673
				(1.00)		(8.74)	(1.71)	(-4.95)		
(4)	oi/sales				0.031	0.055	0.013	-0.112	Yes	50,673
					(0.59)	(8.74)	(1.71)	(-4.96)		
(5)	oi/assets	0.029				0.034	-0.002	-0.086	Yes	50,673
		(1.75)				(16.47)	(-0.35)	(-7.38)		
(6)	oi/assets		-0.018			0.034	-0.002	-0.086	Yes	50,673
			(-0.56)			(16.44)	(-0.36)	(-7.38)		
(7)	oi/assets			0.013		0.034	-0.002	-0.085	Yes	50,673
				(0.52)		(6.31)	(-0.23)	(-3.32)		
(8)	oi/assets				-0.023	0.034	-0.002	-0.084	Yes	50,673
					(-0.67)	(6.30)	(-0.22)	(-3.33)		
					Panel B: Industry Level Regressions					
(9)	oi/sales	0.058				0.065	-0.012	-0.132	Yes	2,638
		(2.96)				(8.73)	(-1.86)	(-2.06)		
(10)	oi/sales		-0.028			0.064	-0.013	-0.129	Yes	2,638
			(-1.14)			(8.57)	(-1.88)	(-1.99)		
(11)	oi/sales			0.016		0.033	0.010	0.072	Yes	2,811
				(0.56)		(4.87)	(1.28)	(1.87)		
(12)	oi/sales				0.005	0.033	0.010	0.070	Yes	2,811
					(0.12)	(4.91)	(1.27)	(1.83)		
(13)	oi/assets	0.048				0.043	-0.030	-0.000	Yes	2,638
		(3.16)				(7.71)	(-5.28)	(-0.00)		
(14)	oi/assets		-0.003			0.042	-0.030	0.003	Yes	2,638
			(-0.15)			(7.60)	(-5.34)	(0.11)		
(15)	oi/assets			0.023		0.026	0.004	0.023	Yes	2,811
				(1.02)		(4.48)	(0.31)	(0.49)		
(16)	oi/assets				0.031	0.025	0.004	0.022	Yes	2,811
					(0.89)	(4.18)	(0.30)	(0.47)		

OLS regressions with profitability as the dependent variable. Panel A regressions are firm level, and Panel B regressions are industry level. All specifications include year and industry fixed effects, and standard errors account for clustering across year and industries. Industry fixed effects are based on three digit SIC or four digit NAICS, based on which market structure measure is included in the regression. The sample has 50,673 observations and is from 1997 to 2006. The market structure variables based on three digit SIC code or four digit NAICS industries.

Table VIII: Ex-ante investment versus future product differentiation

Dependent Variable	Log Industry Ad / Sales	Log Industry R&D / Sales	Zero Adver. Dummy	Zero R&D Dummy	Ind Past Stock Return	Log Assets	Ind. Log B/M Ratio	Adj R^2
Panel A: 10-K 300 Based Industry Controls								
(1) Δ Product Uniqueness Index	0.005	0.003	-0.070	-0.041	-0.004	-0.000	0.001	0.027
	(9.54)	(6.58)	(-12.79)	(-8.73)	(-2.13)	(-0.32)	(0.60)	
(2) Δ Log Total Summed Similarity	-0.055	-0.004	0.692	0.667	0.065	0.049	-0.144	0.108
	(-6.70)	(-0.40)	(8.16)	(7.50)	(1.44)	(2.47)	(-2.36)	
(3) Δ Average Similarity	-0.000	-0.000	0.001	0.000	0.000	0.000	-0.000	0.020
	(-3.04)	(-0.04)	(1.85)	(1.15)	(1.46)	(2.65)	(-0.78)	
(4) Δ Sales 10-K Based HHI	0.001	0.001	-0.033	-0.016	-0.000	-0.001	0.003	0.018
	(2.40)	(2.61)	(-4.83)	(-3.84)	(-0.01)	(-1.17)	(2.26)	
(5) Δ Sales 10-K Based C4 Index	0.004	0.002	-0.062	-0.036	-0.001	0.000	0.005	0.026
	(7.28)	(5.13)	(-10.14)	(-7.97)	(-1.04)	(0.61)	(3.24)	
(6) Δ Observed Lerner Index	0.003	0.002	-0.019	-0.019	-0.014	-0.000	0.005	0.050
	(3.24)	(3.74)	(-1.99)	(-3.25)	(-3.80)	(-0.29)	(2.22)	
(7) Δ Observed Firm Profitability	0.003	0.002	-0.019	-0.019	-0.016	-0.000	0.007	0.019
	(3.07)	(3.54)	(-1.77)	(-2.68)	(-4.47)	(-0.15)	(2.44)	
Panel A: SIC-3 Based Industry Controls								
(8) Δ Product Uniqueness Index	0.005	0.003	-0.065	-0.042	-0.003	0.001	0.002	0.025
	(8.69)	(7.16)	(-11.70)	(-9.40)	(-1.81)	(1.64)	(1.08)	
(9) Δ Log Total Summed Similarity	-0.043	-0.020	0.602	0.516	0.076	0.018	-0.071	0.116
	(-6.66)	(-2.20)	(7.69)	(5.48)	(2.20)	(1.83)	(-2.30)	
(10) Δ Average Similarity	-0.000	0.000	0.001	-0.000	0.000	0.000	-0.000	0.017
	(-2.91)	(0.33)	(1.84)	(-0.40)	(1.88)	(1.23)	(-0.44)	
(11) Δ Sales 10-K Based HHI	0.001	0.001	-0.031	-0.014	0.000	-0.000	0.004	0.009
	(2.24)	(1.87)	(-5.22)	(-3.62)	(0.19)	(-0.51)	(2.45)	
(12) Δ Sales 10-K Based C4 Index	0.003	0.002	-0.056	-0.035	-0.001	0.002	0.006	0.023
	(7.11)	(5.14)	(-10.47)	(-8.14)	(-0.72)	(2.82)	(4.34)	
(13) Δ Observed Lerner Index	0.003	0.003	-0.020	-0.029	-0.014	-0.001	0.004	0.044
	(2.74)	(6.06)	(-2.17)	(-5.18)	(-4.08)	(-1.43)	(1.74)	
(14) Δ Observed Firm Profitability	0.003	0.003	-0.022	-0.028	-0.017	-0.001	0.006	0.017
	(2.79)	(5.00)	(-2.10)	(-4.00)	(-4.64)	(-1.13)	(2.08)	

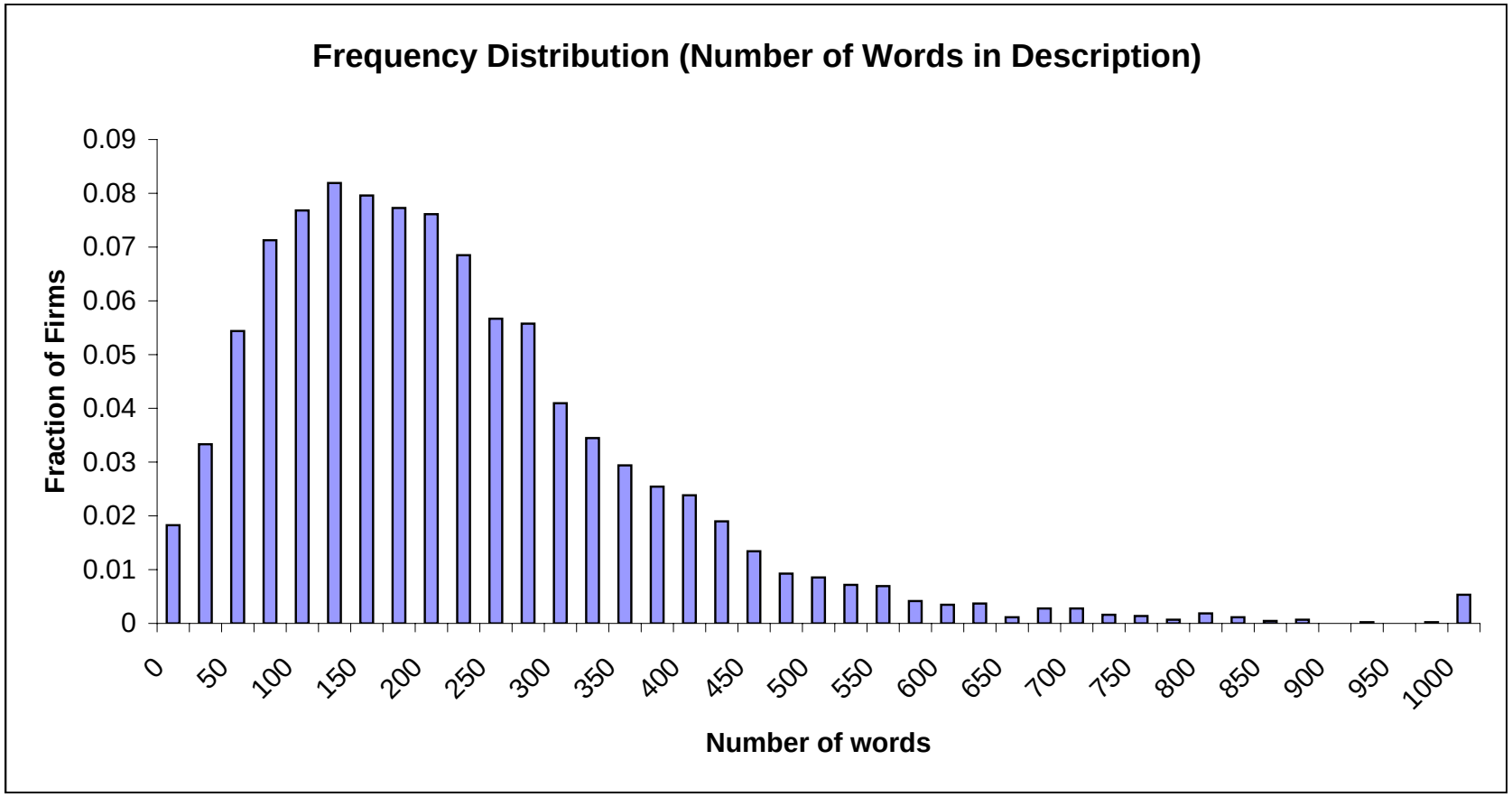
OLS regressions with ex post product changes in market structure (based on VIC-7.06 industries) as the dependent variables. All specifications include industry and yearly fixed effects, and standard errors account for clustering across year and industry (industry controls are based on 10K-300 FIC industries in Panel A, and three-digit SIC industries in Panel B). The sample has 48,572 observations and is from 1997 to 2006.

Table IX: Ex-ante investment versus future product differentiation (SIC-3 and NAICS-4 Industry Definitions)

Dependent Variable	Log Industry Ad / Sales	Log Industry R&D / Sales	Zero Adver. Dummy	Zero R&D Dummy	Ind Past Stock Return	Log Assets	Ind. Log B/M Ratio	Adj R^2
Panel A: SIC-3 Based Market Structure Measures and Industry Controls								
(1) Δ Sales SIC-3 HHI	0.002	-0.000	-0.009	0.018	-0.006	0.001	0.009	0.091
	(0.67)	(-0.13)	(-0.49)	(1.11)	(-1.38)	(0.17)	(1.35)	
(2) Δ Sales SIC-3 C4 Index	0.002	0.001	-0.005	-0.004	-0.003	-0.001	0.002	0.109
	(1.32)	(0.49)	(-0.68)	(-0.55)	(-2.12)	(-0.64)	(0.78)	
(3) Δ Observed Firm Profitability	0.002	0.001	-0.016	-0.009	-0.006	-0.007	0.009	0.091
	(1.30)	(0.39)	(-1.53)	(-0.77)	(-1.43)	(-1.80)	(2.05)	
Panel B: NAICS-4 Based Market Structure Measures and Industry Controls								
(4) Δ Sales NAICS-4 HHI	0.000	-0.005	0.005	0.058	-0.009	0.002	0.026	0.134
	(0.10)	(-2.07)	(0.25)	(2.80)	(-2.24)	(0.38)	(3.77)	
(5) Δ Sales NAICS-4 C4 Index	0.001	-0.001	0.003	0.010	-0.007	-0.002	0.006	0.114
	(0.41)	(-1.12)	(0.39)	(1.21)	(-3.53)	(-0.95)	(2.48)	
(6) Δ Observed Firm Profitability	0.005	0.001	-0.042	-0.005	-0.005	-0.003	0.003	0.175
	(1.70)	(0.39)	(-1.70)	(-0.24)	(-0.97)	(-0.50)	(0.36)	

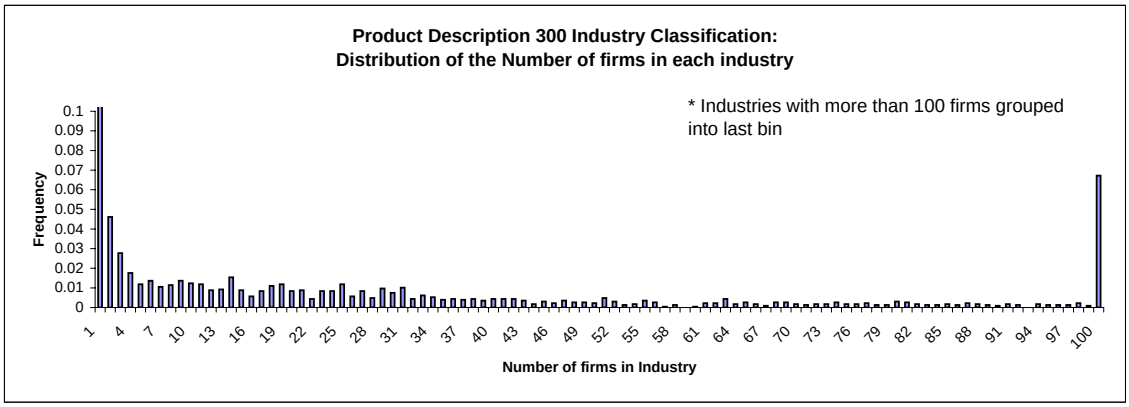
OLS regressions with ex post product changes in market structure (based on three-digit SIC in Panel A, and four-digit NAICS in Panel B) as the dependent variables. All specifications include industry and yearly fixed effects, and standard errors account for clustering across year and industry (industry controls are based on three-digit SIC in Panel A, and four-digit NAICS in Panel B). The sample has 48,572 observations and is from 1997 to 2006.

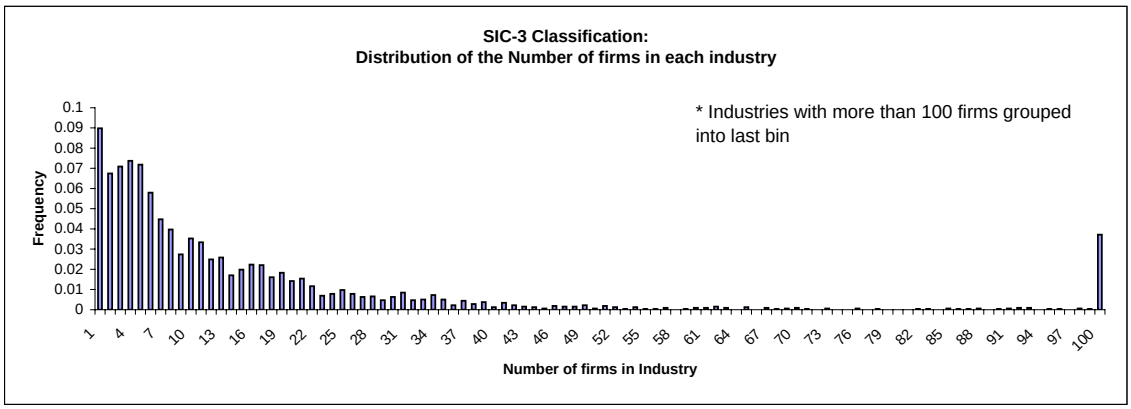
Figure 1:

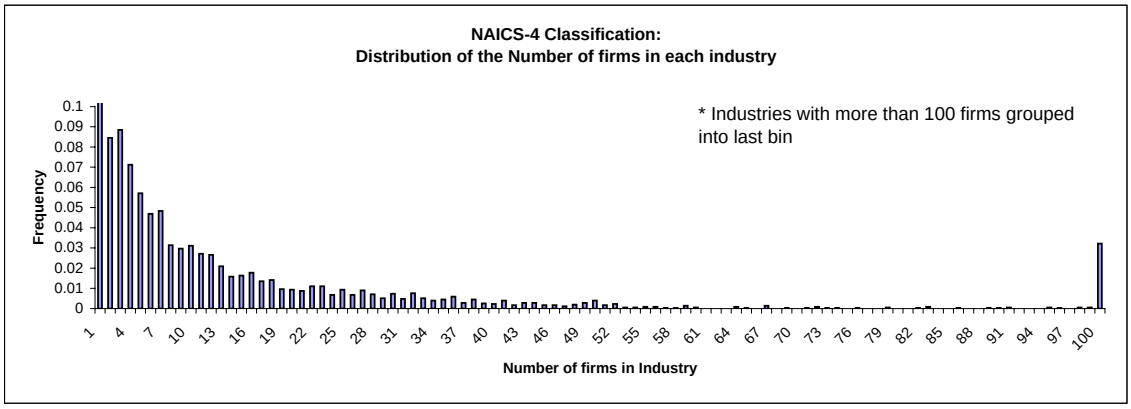


Frequency distribution of unique non-common words in 10-K product descriptions.

Figure 2:







Frequency distribution of the number of firms in each industry based on three FIC industry classification methods: 10K-300 industries, three digit SIC industries, and four digit NAICS industries. All three classifications have close to 300 industries in our sample.