Text-Based Network Industries and Endogenous Product Differentiation

Gerard Hoberg and Gordon Phillips*

September 9, 2011

ABSTRACT

We study how firms differ from their competitors using new time-varying measures of product differentiation based on text-based analysis of product descriptions from 50,673 firm 10-K statements filed yearly with the Securities and Exchange Commission. This year-by-year set of product differentiation measures allows us to generate a new set of industries and corresponding new measures of industry competition where firms can have their own distinct set of competitors. Our new sets of industry competitors better explain specific discussion of high competition by management, rivals identified by managers as peer firms and firm characteristics such as profitability and leverage than do existing classifications. We also find evidence that firm R&D and advertising are associated with subsequent differentiation from competitors, consistent with theories of endogenous product differentiation.

*University of Maryland and University of Maryland and National Bureau of Economic Research respectively. Hoberg can be reached at ghoberg@rhsmith.umd.edu and Phillips can be reached at gphillips@rhsmith.umd.edu. We especially thank Dan Kovenock, Steve Martin, John Sutton and seminar participants at Aalto (Helsinki) School of Economics, HEC, IFN (Stockholm), Insead, ISTCE (Lisbon), London Business School, Notre Dame, Northwestern, Stanford, Stockholm School of Economics, University of Amsterdam, University of Southern California, University of Vienna and the Academy of Management meetings for helpful comments. All errors are the authors alone.

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 time-varying industry classifications, we find that firm R&D and advertising are associated with subsequent differentiation from competitors and increased profitability. These results are consistent with Sutton's (1991) theory of endogenous product differentiation.

Our starting point to form new industries is to gather business descriptions from 50,673 firm annual 10-Ks filed with the Securities and Exchange Commission using web crawling algorithms. The vector representations of the text in each firm's product description generate a Hotelling-like product location space for U.S. firms.[1] We process the text in these product descriptions to calculate new industry classifications based on the strong tendency of product market vocabulary to cluster among firms operating in the same markets. Because they are a function of 10-K business descriptions, our classifications are based on the products that firms supply to the market, rather than production processes (as is the case for some existing industry classification schemes).[2]

These tools enable us to examine how industry structure changes over time, and how firms react to such changes within and around their product markets. A key advantage of our analysis 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 measure of product similarity between firms both within and across industries, allowing us to create general network representations of industry competition, with each firm having its own distinct set of competitors. Although numerous studies use industry classifications as control variables, only a few studies examine the clas-

[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. Empirically, the spatial characteristics of our measures can also be viewed as analogous to the patent technology-based space of Jaffe (1986), although Jaffe's space is applicable for patent filing firms and is not generated using product description text.

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

sification schemes themselves and these do not consider the possibility of industry classifications that change materially over time.[3]

We create new industry classification systems based on 10K product similarities using two methods: one historically motivated, and one that allows industry competition to be firm centric and change over time. The first, which we name "fixed industry classifications" (FIC), is analogous to SIC and NAICS industries.[4] Here, firms are grouped together either over fixed periods of time and membership in an industry is required to be transitive. Thus this method requires that if firms B and C are in firm A's 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.

Our second classification system is more general. In this classification, we allow firm competitors to change every year and we relax the membership transitivity requirements of FIC industries and view industries like flexible networks. We name these new generalized network industries "text-based network industry classifications" (TNIC). In this classification system, each firm can have its own set of distinct competitors analogous to a social network, where each individual has a distinct set of friends, with friends of one individual not necessarily being friends of each other. To illustrate why transitivity is restrictive, suppose firms A and B both view firm C as a rival. If A and B have each have products with different distinct features or enhancements that C does not have, then A and B may not compete against each other as they may serve different product segments.

Relative to existing industry classifications, these new text-based classifications offer economically large improvements in their ability to explain managerial discussion of high competition, the specific firms mentioned by managers as being com-

[3]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 fixed 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.

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

petitors, and how advertising and R&D create future product differentiation. Our new industry measures also offer econometric gains in explaining the cross section of firm characteristics. Our empirical tests further 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.

Although it is convenient to use existing industry classifications such as SIC or NAICS for research purposes, these measures have limitations. Neither adjusts significantly over time as product markets evolve, and neither can easily accommodate innovations that create entirely new product markets. 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 that move into different industries, they do not allow for the industries themselves to evolve over time, and they impose transitivity even though two firms that are rivals to a third firm may not compete against each other. Lastly, they do not provide continuous measures of similarity both within and across industries.

Our results are robust to the treatment of firms that report producing in more than one industry (conglomerate firms). When forming fixed classifications, we only use firms that report just one segment 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 in-line with the single industry to which they are most similar.

In our analysis of text-based industry classifications, our ability to update both the product location of a firm and the identity of a firm's competitors over time also allows us to examine whether advertising and research and development are correlated with increasing product differentiation. We find that firms spending more on either advertising or R&D experience significant reductions in measures of ex-post competition and gains in ex-post profitability, consistent with the hypothesis of Sutton (1991) that firms spend on advertising and R&D to create endogenous barriers to

entry. Our results provide evidence across a broad range of industries complementing Ellickson (2007), who analyzes endogenous barriers to entry in the supermarket industry. We note that while our new measures are interesting for research or scientific purposes to examine topics including innovation and the industry life-cycle, they are less useful 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 contributes to existing strands of literature using text analysis to address economic and financial theories, product markets, and mergers and acquisitions. Hoberg and Phillips (2010) show that merging firms with more similar product descriptions in their 10-Ks experience more successful outcomes. Hanley and Hoberg (2010) use document similarity measures to examine prospectus disclosures from the SEC Edgar website to address theories of IPO pricing. In other contexts, papers such as Antweiler and Frank (2004), Tetlock (2007), Tetlock, Saar-Tsechanksy, and Macskassy (2008), Loughran and McDonald (2010), Li (2006) and Boukus and Rosenberg (2006) examine the relation between the types of words in news stories and bulletin boards and stock price movements.

The remainder of the paper is organized as follows. We discuss characteristics and give examples of our new industry classifications in Section I. We describe the data and similarity calculations in Section II. We give methodological details for our new industry classifications in Section III. In Section IV we compare the informativeness of our new industry classifications to existing SIC and NAICS industry groupings. We construct measures of industry competitiveness in Section V, and Section VI examines how industry structure changes over time and examines how these changes relate to theories of product differentiation and endogenous barriers to entry. Section VII concludes.

I Industry Classifications as a Network

In this section we discuss the features of our "unrestricted" text-based network industry classification that are not available using classifications such as SIC and NAICS. We illustrate these new features using examples based on our new industry group-

ings, while postponing the methodological details to Section III. We define our new industry classifications as an unrestricted network as they have features similar to a network where firms are located distinctly in a product space, each surrounded by its own distinct set of competitors, and each having continuous relatedness scores vis-a-vis all other firms.

Unrestricted networks also have a spatial representation, where same-industry firms appear as clusters, akin to cities on a map. Distances from firm to firm within a cluster indicate within-industry product differentiation, and distances across clusters indicate cross-industry similarity. In contrast, existing industry classifications such as SIC or NAICS are restricted in that, while they have a spatial representation, all firms in the same cluster have the same zero distance from each other, all share membership within the cluster, imposing transitivity, and there is no known distance across industry clusters. We now discuss these features in depth and give examples of industries in which the new text-based industries give improvements.

A Ability to Capture Within-Industry Heterogeneity

The concept of product differentiation within industries 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. An ideal classification system should not only identify product markets, but also provide measures of differentiation within industries. Beginning with Berry, Levinsohn, and Pakes (1997), the approach of the product differentiation literature has been to estimate demand and cost parameters in well-defined product markets. For example, Nevo (2000), estimates own- and cross-price elasticities of demand and their effect on post-merger prices in the ready-to-eat cereal market. This approach has been highly informative, especially in understanding the dynamics of industry pricing, competition and substitution in these well-defined industries. However, many theories, especially those related to endogenous barriers to entry and why firms produce across multiple industries, are difficult to test in a single industry setting.

In addition, accurately specifying industry composition is especially difficult in industries where firms offer highly differentiated products or services. This difficulty is readily apparent in the business services industry, SIC code 737. There were over 600 public firms in this industry in 1997 according to Compustat. Using a classification that matches the coarseness of three-digit SIC industries, we find that the markets faced by these firms are quite different. Table I displays sample classifications using our methodology for selected firms in this product area.

[Insert Table I Here]

Table I shows 6 major sub markets within the broad business services industry. They are Entertainment, Medical Services, Information Transmission, Software, Corporate Data Management and Computing Solutions, and Online Retailing and Publishing. Each displayed industry is the TNIC industry surrounding the focal firm listed in each example's header. While SIC codes were not used to make these groupings, we report the codes for illustrative purposes. The SIC codes of rival firms in each market load heavily on 737, but each sub-market also spans firms in other SIC-industries including the three-digit codes 357, 366 and 382. A key theme is that many firms address these markets using the internet and technology, and they often also compete with rivals that have a more traditional brick and mortar presence.

Beyond simply identifying industry clusters, our approach also generates firm-by-firm pairwise relatedness scores. Therefore, our framework can order rivals in terms of their importance to a focal firm, analogous to a network, while also providing simple measures of the overall product differentiation surrounding each firm. Our method can also be used to construct a firm-specific concentration index that can capture the competition that surrounds each firm.

B Ability to Capture Product and Industry Change

The industry classification system should also capture changes to industry groupings over time. Firms often change, introduce and discontinue products over time, and thus enter and exit various industry spaces. This flexibility is directly related to Sutton (1991) and Shaked and Sutton (1987), who suggest that barriers to entry

are endogenous. In particular, advertising and research and development allow firms to differentiate their products and enter into related industries.[5] These theories motivate our examination of advertising and research and development, and their links to future changes in industry membership and competition.

Only industry classifications that frequently recompute product market relatedness can address the changing nature of the product market. Some product areas disappear or change, such as overhead projection systems with vinyl acetates. More common, due to innovation, new product markets like solar power or internet-based products can appear. Our industry classifications are updated annually and can capture rapidly changing product markets. Table II provides examples of two industries that changed dramatically over time.

[Insert Table II Here]

Panel A of Table II displays the TNIC industry surrounding Real Goods Trading Corp, which provides solar technology. In 1997, this market was nascent, and Real Goods had just one rival, Photocomm. By 2008, Real Goods was part of a 9-firm industry group, having a product vocabulary rooted in solar and environmental terminology. Panel B displays the product market surrounding L-1 Identity Solutions in 2008, which provides technological intelligence solutions related to Homeland Security. This entire product market was not in our sample in 1997, and likely emerged after the events of September 11, 2001. The only related firm that was in our sample in 1997, CACI International, migrated from the database management product market to this security-oriented market, as shown in the table.

C Ability to Capture Cross-Industry Relatedness

The industry classification system should also be able to capture cross-industry relatedness. If two product markets are very similar, firms in each product market likely hold a credible threat of entry into the other at low cost. This notion of economies of scope is developed by Hay (1976) and Panzar and Willig (1981). In particular, firms

[5]Lin and Saggi (2002) show that tradeoffs related to product differentiation can affect process innovation and product innovation.

facing this form of cross industry threat might keep prices low to deter entry. Currently, existing research can examine cross-industry relatedness using coarser levels of SIC or NAICS codes or through the Bureau of Economic Analysis's input-output matrix (used to measure vertical relationships). Our methodology uncovers numerous links entirely missed using other classifications. Because our classifications are based on actual product text, we are thus able to detect potential rival firms that offer related products even if they are not direct suppliers or rivals (for example, through economies of scope).

Hoberg and Phillips (2011) is an example of a recent study that explores cross-industry relations using 10-K text-based relatedness scores. The study examines why conglomerates span some industry combinations more frequently than others, and finds that they are most likely to span industry pairs that are closer together in the product space and that surround other highly valued industries. These findings are robust to controls for vertical relatedness and are consistent with conglomerates using industry relatedness to potentially enter nearby high value industries that might otherwise be costly to enter.

D Benefits of Unrestricted Industry Classifications

One of the largest benefits of our approach is that it allows both within-industry and cross-industry relations to be examined. Many empirical studies examining product differentiation focus on single industries.[6] An older literature summarized by Schmalensee (1989) focused on cross-industry relations. Our industry classifications allow for both types of studies. Our classifications are also updated in each year as firms must refile 10-Ks annually, and our industry boundaries can be redrawn using any desired level of coarseness. We are also able to identify a unique set of industry rivals surrounding each firm, which relaxes the restrictive transitivity property of existing classifications. We also define industry competition relative to each firm as in the circular city model of Chamberlin. Analogous to a Facebook circle of friends or a geographic distance, each firm has its own direct competitors identified using a text-based distance from the firm itself on a spatial grid.

[6]For recent examples see Nevo (2000), Mazzeo (2002) and Seim (2006).

Our new classifications can also be used in conjunction with, not in lieu of other data. Although not part of the current study, looking forward, word-by-word mappings can be used to create firm-specific aggregations of BLS price series, BEA input-output data, and patent data. For example, patent filings have a textual description, and this can be used to map how patents are related to each other and across firms - independent of the patent examiner classification. Analogously, if price data is available for verbal product lists, firm-wide price aggregations can also be estimated using various weighting methods based on firm 10-K text.

There are also econometric benefits. For example, many studies examine whether firm actions (such as equity issuance) are related to firm characteristics (such as firm age). Here, the researcher may wish to ensure that any relationship found is due to firm-specific age, and not to a broad industry attribute related to age such as industry life cycle. A solution is to control for industry effects, and it follows that superior industry classifications can improve estimation accuracy. We find that our classifications are able to explain a larger fraction of firm characteristics in cross section than existing classifications, and hence they likely provide better industry controls. Finally, more informative industry classifications can also improve the accuracy of standard errors, as numerous studies use adjusted standard errors to account for clustering at the industry level.

We note that other methods of identifying competitors can also be used in conjunction with our data. In a contemporaneous paper, Rauh and Sufi (2010) use firm self-reported competitors from Capital IQ and show that firm capital structure better reflects that of these competitors than that of firms in the same SIC code. Using our text methods, we obtain similar improvements in predicting capital structure and much larger improvements in predicting operating cash flow. However, Capital IQ peers are currently available only for the most recent year while our classifications are available over many years and offer the flexibility to measure within and across industry similarity using any granularity. Although they are distinct from our measures, self-reported competitors are also useful. For example, we use them as a validation tool to examine whether our industries better overlap with Capital IQ peers relative to other classifications including SIC and NAICS.

II Data and Methodology

Using web crawling and text parsing algorithms, we obtain and construct a database of word business 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 business 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 measure how industry structure changes over time.

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). In our main specification, we restrict attention to words that can be used as a noun (as defined by Webster.com) and proper nouns. We define proper nouns as words that appear with the first letter capitalized at least 90% of the time in our sample of 10-Ks. We also omit common words that are used by more than 25% of all firms, and we omit geographical words including country and state names, as well as the names of the top fifty cities in the US and in the world. As we show later, we choose the word-exclusion method that gives us high explanatory power in some key tests. Our overall results are robust to different word-exclusion / stop-wording screens.

There are many automated processes used in research to evaluate text (see Sebastiani (2002) for a detailed review). However, there is little consensus regarding which method is uniformly best, and hence researchers must often choose a method upon reviewing the unique features of their application. We use the "cosine similarity" method for many reasons. First, its properties are well-understood given its wide usage in studies of information processing, and it is also intuitive given its network and spatial representations. This method is also easy to program and only moderately

computationally burdensome, making it practical for other researchers to replicate. Finally, this method's normalization builds in a natural control for document length. It is called the cosine similarity method because it measures the angle between two word vectors on a unit sphere.

Full details regarding our implementation of the cosine similarity calculation are in Appendix 1. We give a basic description here. Suppose there are N unique words used in the union of the documents used by all firms in our sample. A given firm i's vocabulary can then be represented by an N-vector P_i, each element being populated by the number one if firm i uses the given word, and zero if it does not. The cosine similarity is simply the dot product of normalized vectors for firms i and j as follows.

$$Product\ Cosine\ Similarity_{i,j} = (V_i \cdot V_j), \quad \text{where } V_i = \frac{P_i}{\sqrt{P_i \cdot P_i}} \quad \forall i, j \quad (1)$$

Intuitively, this dot product is higher when firms i and j use more of the same words, as both vectors have positive values in the same elements. This measure is also bounded in [0,1] and has a spatial representation, as each vector V_i has unit length and thus resides on an N-dimensional unit sphere. Because we populate P_i with binary values, our baseline method assigns uniform importance weights to words regardless of their frequency. Following Loughran and McDonald (2010), we also consider an alternative weighting scheme called "total frequency/inverse document frequency" (TF-IDF) in which the P_i vector is instead populated with higher weights for more frequently used words in firm i's own document, and lower weights for words used by a larger fraction of all firms in the economy. Our results later show that uniform weights outperform TF-IDF weights for our application, indicating that a firm's decision to use a given word to describe its products is more important than how frequently the word is used.

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 (domestic firms traded on either NYSE, AMEX, or NASDAQ), we are are able to match (using CIK) 55,326 (97.9% of the CRSP/COMPUSTAT sample).[7] 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 business 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 than 1000 characters. These firms are excluded from our analysis.

[7]We thank the Wharton Research Data Service (WRDS) for providing us with an expanded historical mapping of SEC CIK to COMPUSTAT gvkey. 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.

III Industry Classification Methodology

We first note that industry classifications have a simple network representation. A classification is a complete mapping from any firm-pair (firms i and j) to a real number in the interval $[0, 1]$ describing relatedness. Because the mapping is complete, an industry classification can be succinctly described by an NxN square matrix M (i.e., a network), where N is the number of firms. If the classification is updated yearly, it can further be represented as a time series of such matrices M_t.

We construct classifications using textual pairwise cosine similarity scores as the basis for this mapping, and hence the matrix M_t is populated by applying the aforementioned cosine similarity method to each permutation of firm pairs. The large number of words used in business descriptions, along with the continuous and bounded properties of the cosine similarity method, ensure that the matrix M_t is not sparse, and that its entries are unrestricted real numbers in the interval $[0, 1]$. In contrast, the corresponding network M_t underlying SIC and NAICS industries is heavily "restricted" and must satisfy the following two properties:

Definition: A classification is said to satisfy the ***binary membership transitivity property*** if M_T has binary banded diagonal form ("1" on all banded diagonals and "0" elsewhere). This form satisfies membership transitivity, and hence 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. This form also requires that all firms are homogeneous within industries, and that all industries are entirely unrelated to one another.

Definition: A classification is said to have the ***fixed location property*** if M_t is not updated each year. Intuitively, such industries have a time-fixed product market (they are fixed until the codes are changed or updated).

We use 10-K text to classify firms into industries using two methods. The first method, described in Section A below, is analogous to SIC and NAICS classifications and requires the binary membership transitivity and the fixed location property to hold. We henceforth refer to classifications requiring these two restrictive properties as "Fixed Industry Classifications" (FIC).

Our second method, described in Section B below, relaxes both properties, and we refer to this second class of industries as "Text-Based Network Industry Classifications" (TNIC). A firm's TNIC industry can move across the product space over time as technologies and product tastes evolve. New firms can also appear in the sample, and each firm can have its own distinct set of competitors that may or may not overlap with other firms' competitors. Finally, TNIC industries are sufficiently rich to permit within and across industry similarities to be computed. 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, in our main FIC specification, we form fixed groups of industries by running a clustering algorithm only once using the earliest year of our sample (1997) and 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 also consider a variation where we rerun the clustering algorithm in each year, as this variation imposes the binary membership transitivity property, but relaxes the fixed location property. This allows us to examine the relative economic impact of the two properties separately, and we report later that both properties are about equally important in explaining the difference in explanatory power between FIC industries and TNIC industries.

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. In Appendix 3, we explore this question using Akaike information criterion tests. These tests use likelihood analysis to compare models even when they use varying numbers of parameters (in our case industries). The results suggest that roughly 300 industries best explain firm-level data.

[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 non-geographical nouns and proper nouns that appear in no more than 25% of all 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 500 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.

[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 Network Industry Classifications Based on 10-Ks

We next relax the fixed location and transitivity requirements and construct generalized text-based network industry classifications (TNIC). In addition to offering substantially higher explanatory power (see Section IV), TNIC industries offer many additional advantages. First, the full knowledge of firm pairwise similarities permits calculations of across and within industry similarities. Second, TNIC industries are necessary to test theories predicting dynamic firm and industry movements in the product space over time (see Section VI). Third, industry competitors are defined relative to each firm in the product space - like a geographic radius around each firm - thus each firm will have its own distinct set of closest competitor firms.

We construct TNIC 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.[8] In practice, however, only a small fraction of pairs are actually industry pairs. Although one can use any minimum similarity threshold to construct TNIC-industries, we focus on thresholds generating industries with the same fraction of industry pairs as SIC-3 industries, allowing us to compare SIC and TNIC industries in an unbiased fashion.

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

For three digit SIC codes, 2.05% of all possible firm pairs are industry pairs. A 21.32% minimum similarity threshold generates 10-K based TNIC industries with 2.05% industry pairs (same as SIC-3). We consider one further refinement to further mitigate the impact of document length. For a firm i we compute its median score as the median similarity between firm i and all other firms in the economy in the given year. Intuitively, because no industry is large enough to span the entire economy, this quantity should be calibrated to be near zero. We thus adjust all scores by the median scores of firms comprising the given pair.[9]

Indeed the transitivity property might not hold for these industries. For example, consider firms A and B, which are 25% similar. Because this is higher than 21.32%, A and B are in each other's TNIC industry. Now consider a firm C that is 27% similar to firm A, and 17% 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 22% similar to firm B, then transitivity would hold. Thus, TNIC classifications do not rule out transitivity, but rather transitivity might hold case by case.

We also take into account vertical integration in defining our variable industry classifications. We examine the extent to which firm pairings are vertically related using the methodology described in Fan and Goyal (2006). Based on the four-digit SIC codes of two firms, we use the Use Table of the Benchmark Input-Output Accounts of the US Economy to compute, for each firm pairing, the fraction of inputs that flow between the industries of each pair. If this fraction exceeds 1% of all inputs, we exclude the pairing from TNIC industries regardless of the similarity score. Because just 4% of all pairs are excluded using this screen, and because our results are fully robust to including or excluding this screen, we conclude that firm business descriptions in firm 10-Ks indeed describe firm product offerings, and not firm production inputs.

[9]Our results are robust, though roughly 2% weaker if we omit this step.

IV Comparing Industry Classifications

Our next objective is to examine which industry classifications best explain firm characteristics in cross section, while holding fixed the degree of granularity of the industries we compare. In Section A, we compare the ability of FIC and TNIC industry classifications to explain firm characteristics such as profitability, leverage and stock market Betas. In Section B, we examine which classification systems best explain managerial discussion of high competition, firm self-identified rivals, and which firms are most likely to form product market alliances.

A Econometric Performance of Industry Controls

In this section, we explore industry controls in a panel data setting. As discussed in Section I, more powerful classifications can improve the accuracy of inferences, especially inferences regarding firm characteristics when the researcher needs to control for industry characteristics. From an econometric perspective, improved classifications should explain a larger fraction of total firm heterogeneity (as firms are more similar within industries than they are across industries). We compare explanatory power across many firm characteristics and across our new classification systems as well as existing SIC and NAICS 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 potentially large number of degrees of freedom equal to the number of industries in the classification, leaving fewer for hypothesis testing. Second, industry fixed effects do not account for industry variables that might change over time. To address this second issue, researchers can use industry x year fixed effects. However, this further exacerbates the usage of degrees of freedom given the large number of fixed effects.

Both issues can be addressed using simple industry-averaging methods. Rather than using fixed effects, the researcher can average the given characteristic (the dependent variable) within each industry in each year, and use this average as a single additional control variable. This approach uses only one degree of freedom, and because this average can be computed separately in each year, this approach

also accounts for industry characteristics that might vary over time. This averaging method is also called a kernel method, with equal weights across industry members. This method is general and can be used for both FIC and TNIC classifications.

The averaging method also offers the flexibility to examine the impact of multiple industry firms (conglomerates firms), as weighted averages can positively weight more than one industry when computing a given firm's fixed effect. We consider a conglomerate-adjusted averaging method 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 use the simple single-industry average. 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 average variable to be the average of the N corresponding industry-specific values. Our results discussed below show that conglomerate adjusted averages do not offer material improvements relative to unadjusted averages.

The last method we consider is a similarity-weighted average rather than an equal weighted average.[10] This method can only be used for TNIC industries, as only TNIC industries provide firm-pairwise similarity weights. Table III displays the results.

[Insert Table III Here]

Table III shows that 10-K based industries outperform both SIC and NAICS, especially TNIC industries, which relax both the binary membership transitivity property and the fixed location property. When limiting attention to fixed effects based on FIC industries, the adjusted R-squared for profitability scaled by sales increases by 15.1% from 0.284 to 0.327 when the 10-K based classifications are used rather than the SIC-3 classifications. The improvement is a similar 13.9% when 10K-300 industries are used rather than NAICS-4 industries. The improvement in

[10]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.

explanatory power relative to SIC-3 is even larger at 22.0% for operating income scaled by assets rather than sales.

For other firm characteristics, all except for leverage ratios have stronger results for 10-K based FIC industries. One explanation is that leverage is a managerial policy, and policies might be chosen to target the most readily available industry averages. For example, managers might target SIC or NAICS benchmarks because these targets are easy to obtain.

By comparing the averaging method results in columns 2, 4, and 6 to standard fixed effects in columns 1, 3, and 5, we conclude that the averaging method offers significantly higher explanatory power despite its usage of a single degree of freedom. The main reason is that the averaging method allows the industry controls to vary over time (the average is computed separately in each industry in each year). It is thus more analogous to controlling for industry x year fixed effects than it is to controlling for separate industry and year fixed effects. Its improvement in power can be large, for example its adjusted R-squared is nearly 3x higher for sales growth. This likely reflects the fact that sales growth changes over time more than other characteristics do. In general, the averaging method 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 averaging method performs roughly as well as the unadjusted 10K-300 averaging method. We conclude that these simple conglomerate adjustments do not offer material benefits.

The last four columns display results for TNIC industries: the first two consider raw TNIC industries, and the last two are purged of firm pairs having at least 1% vertical relationships as discussed in Section III.B. Rows one and two show that TNIC 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 43% for the four TNIC specifications is 51.4% higher than the 28.4% adjusted R-squared for standard SIC-3 fixed effects, and 37.8% higher than the SIC-3 averaging method. Perhaps even more striking, the similarity weighted averaging method (third to last column and the last column) performs at this high level even though we exclude the firm itself from the weighted average. This is a

mechanistic disadvantage, as both fixed effects and equal weighted averaging methods include the firm itself in their averages.[11]

As discussed previously, TNIC industries offer two advantages relative to FIC industries: relaxing the fixed location property and relaxing the membership transitivity property. We find that both properties are individually important. Regarding the time fixed location property, comparing the fifth column (time-fixed FIC) to the sixth column (annually-recalculated FIC) shows substantial improvement in explanatory power. For example, the oi/sales R-squared increases from 0.327 to 0.372 when one relaxes just this fixed location property. To assess the impact of the membership transitivity property, the time varying FIC averaging method in the sixth column can be compared to the analogous TNIC averaging method in the eighth column. Here, for example, the oi/sales R-squared increases from 0.372 to 0.458. Because both improvements are similar in magnitude, we conclude that relaxing both properties is important to maximizing explanatory power.

The results also show that controlling for vertical integration has some, but not a large effect on our results, as the last two columns are very similar to the two columns preceding them. We conclude that TNIC industries offer substantial improvements over existing methods used in the literature, and that their focus is mainly on horizontal product scope rather than vertical relationships. For all analysis that follows, we will focus exclusively on the TNIC industry designations that are purged of vertical relatedness (our results are affected little if we instead use raw TNIC industries). Our approach is also conservative because TNIC averaging methods exclude the reference firm.

When comparing industry classifications, it is natural to ask if an optimal level of granularity exists. Because our classifications can be calibrated to an arbitrary level of granularity, we are in a good position to explore this question. To conserve space, we explore this issue in Appendix 3. Using Akaike information criterion tests, we find that roughly 300 industries best describe firm characteristic data in cross

[11] If the reference firm is included using the similarity-weighted average, and it is given a similarity weight of 1, the adjusted R-squared increases to near 70% (not reported). Because this likely overweights the reference firm, we do not recommend using similarity averages that include the reference firm.

section. Hence, our TNIC industries that are calibrated to match SIC-3 industries on granularity are likely to be a good fit for empirical applications. Going further, the fact that SIC-3 and TNIC overlap only partially implies that researchers can absorb even more industry variation using empirical models that control for both TNIC and SIC-3 effects.

B Industry Classifications and Competition

In Section I, we discussed the ideal properties that industry classifications should have. A common theme relates to identifying sources of competition or competitive threat. For example, the concepts of product differentiation, economies of scope, and endogenous barriers to entry all generate implications related to the effects of competition on economic outcomes. We use two data sources to compare industry classifications in terms of their ability to explain competitive pressures. Our approach in this section is to assess competitive pressure directly. This approach may be more accurate than indirect tests such as those based on profitability.

Our first approach follows Ball, Hoberg, and Maksimovic (2011) and we examine the Management's Discussion and Analysis section of each firm's 10-K. A primary source of content in this section is the manager's discussion of his or her firm's performance, and the firm's outlook going forward. For each firm year, we thus define the high competition dummy to be one if the manager cites "high competition", or one of its synonyms, in this section.[12]

[Insert Table IV Here]

Table IV displays the results of logit regressions in which the dependent variable is the high competition dummy. Standard errors are adjusted for clustering at the firm level. We include as independent variables, the sales-based Herfindahl index (sum of squared market shares) based on our TNIC classification - where the competitors vary in each row based on the word exclusion screens as noted - and the sales-based Herfindahl index based on three digit SIC codes. We also standardize all

[12]Synonyms for the word "high" include intense, significant, substantial, significant, vigorous, strong, aggressive, fierce, stiff, extensive, or severe. Synonyms for the word "competition" include compete, competition, or competing.

independent variables to have a mean of zero and a standard deviation of one so that both economic magnitudes and statistical significance levels can be compared across the measures. We conclude that an industry classification more directly measures competitiveness if the HHI implied by the classification is more negatively related to the high competition dummy.

To provide additional information regarding our textual screens, we compare the performance across Herfindahl indices computed using all 10-K words (rows 1 to 4), and those that use non-geographical nouns and proper nouns only (rows 5 to 12). We also explore the role of the common word threshold (i.e., the threshold at which words are discarded if they are used in at least the threshold percentage of all 10-Ks indicated in column 2), and we consider thresholds of 10%, 25% and 100%. Discarding common words and non-nouns changes the sets of words used to compute cosine similarities and thus can change the firms that are identified as competitors. Using each new set of competitors for each firm, we then recalculate the TNIC Herfindahl used in column 3. Lastly, we also consider the total frequency/inverse document frequency (TF-IDF) weighting scheme used in Loughran and McDonald (2010). This method uses a logarithmic ratio to more heavily weight words that are used more frequently in a firm's own-document, and to less heavily weight words that are used by more firms in the overall sample in each year.

Table IV shows that HHIs based on both TNIC classifications and SIC-3 classifications are informative regarding the level of competition perceived by the manager. At a minimum, we conclude that our measures provide new information about measuring competitiveness that is at least as important as information contained in SIC-3 classifications. Going further, the table shows that restricting attention to nouns and proper nouns (also excluding geographical terms) in rows five and later further enhances our results. Finally, we find that the stop word threshold of 25% performs best. The coefficient for this specification (-0.241) is 37.7% larger than the coefficient (-0.175) for the SIC-3 HHI. Hence we conclude that the 10-K based classifications are more informative about competitive pressures than are three digit SIC code classifications.

At the bottom of Table IV, we explore the robustness of this conclusion to various

control variables that might also be related to competitive pressures including firm size, age, profitability, and Tobin's Q. Because it is well known that document size can influence text-based variables, we also control for the size of the firm's Management's Discussion section. In all, we find that both HHI variables weaken somewhat as the new controls are added, however, both variables remain highly significant and the relative importance of the TNIC classification relative to SIC-3 coefficient becomes even larger. The coefficient of -0.170 for the TNIC-based HHI with all controls in row 12 is 97.7% larger than the SIC-3 HHI coefficient of -0.086.

We next consider the approach used by Rauh and Sufi (2010), who gather data from Capital IQ identifying the firms listed by each firm as being rivals. We also note one important limitation in this analysis, as Capital IQ data is not available on a historic basis. Hence, we extract peers using 2011 data, and examine whether industries computed using the last year of our data can better explain the links identified by Capital IQ relative to SIC-3 or NAICS-4 industries.

[Insert Table V Here]

Panel A of Table V displays summary statistics regarding the fraction of Capital IQ competitors that are in the same TNIC industry, as well as the fraction of overlap between our industries and the SIC3 and NAICS classifications. As an additional validation test, we view higher overlap ratios as being superior, as they suggest that the given industry classification better explains the peers that managers themselves identify as being rivals. To ensure a fair comparison, we use TNIC industries that are calibrated to be exactly as coarse as SIC-3 and NAICS-4 industries.

Table V shows that SIC-3 industries have a 47.1% overlap with Capital IQ competitors. TNIC industries reach a maximum overlap with Capital IQ for specifications based on nouns and proper nouns and a 10% stop word threshold, where 62.0% of Capital IQ peers overlap with our TNIC industries. Overall, the table also shows that virtually all TNIC industries, with the sole exception of those using a 100% threshold, outperform SIC-3 and NAICS-4 industries in their ability to explain Capital IQ self-reported peers.

Panel B of Table V repeats this exercise using Capital IQ strategic alliances

rather than Capital IQ competitors. This test is particularly interesting because strategic alliances are likely related to economies of scope, as firms with similar but different technologies can combine their comparative advantages and earn greater profits using alliances. The results show that TNIC industries strongly dominate SIC-3 and NAICS-4 industries along this dimension. The overlap with Capital IQ alliances is just 28.2% for SIC-3, but is in the range of 40.6% to 48.6% for all TNIC industries with the only exception being those with a 100% stop word threshold.

To further inform the calibration of TNIC industries, we also examine the extent to which they overlap with SIC-3 and NAICS-4 industries. The table suggests that this overlap is highest for a 25% common word threshold based on nouns and proper nouns, where overlap reaches a maximum of 52.2% in Panel A. The strong performance of this 25% threshold fits in well with our findings from Table IV. Henceforth, we will focus attention on this TNIC threshold alone to conserve space. However, our key inferences are robust to using other thresholds.

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 consider existing measures based on firm market shares (HHI and C4 indices) and measures based on similarity (summed and average similarity).

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:

[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.

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

$$C4 = \sum_{i=1}^{4largest} \frac{SL_i}{\sum_{i=1}^{N} SL_i} \tag{3}$$

HHI indices and C4 indices can be computed for both FIC and TNIC industries. Our remaining indices are only defined for TNIC industries, as they require the existence of a reference firm. Consider a TNIC industry with N+1 firms, and let one of the firms be the reference firm, and the other N firms are its rivals. 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):

$$TotalSimilarity = \sum_{i=1}^{N} S_i \tag{4}$$

$$AverageSimilarity = \frac{TotalSimilarity}{N} \tag{5}$$

We compute the sales-based Herfindahl (HHI) and C4 indices for each of the three industry classifications we consider: SIC-3, NAICS-4, and our 10-K-based TNIC industries. The average TNIC HHI is 0.191, and the average TNIC C4 is 55.8%. HHI and C4 indices based on SIC-3 and NAICS-4 have means that are similar to each other, and only modestly different from TNIC industries. For example, the average SIC-3 based C4 is 61.3%, which is close to the 61.6% for NAICS-4, but somewhat larger than the 55.8% for TNIC. We also compute total and average similarity for TNIC industries (SIC AND NAICS do not provide analogous measures of product differentiation).

Table VI displays Pearson correlation coefficients for our measures of market structure. The table shows two key findings: (1) 10-K based measures are strongly

[14]Net similarity is the raw pairwise similarity minus the minimum similarity threshold used to form the given TNIC industry. 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.

correlated with each other, and (2) SIC-3 and NAICS-4 measures are strongly correlated with each other, but not with 10K-based variables.

[Insert Table VI Here]

Table VI shows that the 10K-based HHI index is -31.9% correlated with the total similarity variable, suggesting an intuitive link between concentration and product differentiation measures. Furthermore, this correlation is quite far from unity indicating that both measures contain much distinct information. Because analogous similarity measures are not available for SIC or NAICS industries, this fact further illustrates the unique benefits of having network based classifications with known pairwise similarities for all firm pairs.

VI 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. An example from 'An Illustration of Dual Structure' in Sutton's, "Sunk Costs and Market Structure", Section 3.4, illustrates the logic behind our empirical design. In Sutton's example, we observe a firm moving between two industries, as it, and possibly some rivals increase their advertising spending in order to become a small group of leading brands that sell to brand sensitive buyers, thus escaping the large number of non-advertising firms that 'sell on price'. The firm thus uses advertising to move away from non-advertising (or non-R&D) firms.

The main idea is R&D and advertising can create more unique products that appeal to quality-sensitive consumers and make it more expensive for rivals to enter. 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 ex-post changes in our market structure and competitiveness measures on ex-ante advertising and R&D. We recognize that these

tests examine association, as it is difficult to establish causality in this setting. This analysis complements Ellickson (2007) who analyzes the supermarket industry, and further illustrates the challenges that Ellickson notes on providing evidence on endogenous fixed costs.

Importantly, we restrict attention to TNIC industries, as variable membership and variable locations are critical to testing Sutton's theory, which is primarily about trying to prevent entry across industry boundaries. TNIC industry definitions are flexible enough to identify these time-varying effects. SIC-3 and NAICS-4 lack this flexibility because their industry locations are close to fixed, as memberships rarely change.

[Insert Table VII Here]

Table VII 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 predictions across all of our competitiveness measures. For example, rows three and four show that firms spending on advertising and R&D experience substantial improvements in their HHI Index and C4 Index respectively (results significant at the 1% level).

Rows (1) to (6) show that all measures of changes to market structure generate similar results. The C4 index is the most robust variable, and firms spending more on advertising and R&D generate improvements in their ex-post C4 indices. The high relevance of the C4 index is consistent with the larger firms in a given firm's product market playing an important role. Rows (5) and (6) show that advertising and R&D are also positively related to ex-post changes in observed profitability.[15]

Our results are also consistent with Hoberg and Phillips (2010), who show that mergers and acquisitions can also be used to differentiate products from close rivals, and that this is especially relevant when firms face more competition.

[Insert Table VIII Here]

[15]Not reported, the results in Table VII are very similar if we use SIC-3 or NAICS-4 industry controls instead of text-based industry controls.

Table VIII displays the results of tests analogous to those in Table VII, 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 test Sutton's predictions. The table confirms this conjecture, and we find little support using these less powerful measures. Comparing these results to those in Table VII based on dynamic 10K-based TNIC industries, leads us to conclude that time varying network industries are essential in providing the empirical flexibility needed to test the role of endogenous barriers to entry.

VII 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. 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 text-based network industry classifications are based on how firms describe themselves 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 that we term text-based network industry classifications (TNIC) is analogous to social networks, where each individual can have a distinct set of friends, or to geographic networks where the distance between firms

determines whether or not it is a competitor.

Measures of competitiveness based on our new classifications better explain specific discussion of high competition by management, and better explain rivals mentioned by managers as peer firms than do existing classifications. Using our relatedness measures, we create new measures of market structure that capture within-industry competitiveness and better explain firm characteristics.

Our classifications allow us to examine how industry market structure and competitiveness change over time, and whether advertising and research and development serve as endogenous barriers to entry. We find support for Sutton (1991)'s hypothesis that firms spend on advertising and R&D, at least in part, to increase product differentiation and profitability.

References

Antweiler, Werner, and Murray Frank, 2004, Is all that talk just noise? the information content of internet stock message boards, *Journal of Finance* 52, 1259–1294.

Ball, Christopher, Gerard Hoberg, and Vojislav Maksimovic, 2011, Redefining financial constraints: a text-based analysis, Working Paper, University of Maryland.

Berry, Steven, 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).

Ellickson, Paul, 2007, Does sutton apply to supermarkets?, *Rand Journal of Economics* 38, 43–59.

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

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

Hanley, Kathleen, and Gerard Hoberg, 2010, The information content of ipo prospectuses, *Review of Financial Studies* 23, 2821–2864.

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

Hoberg, Gerard, and Gordon Phillips, 2010, Competition and product market synergies in mergers and acquisitions: A text based analysis, forthcoming Review of Financial Studies.

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

Jaffe, Adam, 1986, Technological opportunities and spillovers of r&d: Evidence from firms' patents, profits and market value, *American Economic Review* 76, 984–1001.

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

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

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, 2010, When is a liability not a liability? textual analysis, dictionaries, and 10-ks, forthcoming Journal of Finance.

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.

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

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: 10K-based Classifications of firms in Business Services (SIC3=737)

SubMarket 1 Entertainment (Sample Focal Firm: WANDERLUST INTERACTIVE)

43 Rivals: MAXIS, PIRANHA INTERACTIVE PUBLISHING, BRILLIANT DIGITAL ENTERTAINMENT, MIDWAY GAMES, TAKE TWO INTERACTIVE SOFTWARE, THQ, 3DO, NEW FRONTIER MEDIA INC, ...

SIC CODES OF RIVALS: COMPUTER PROGRAMMING, DATA PROCESSING, AND OTHER COMPUTER RELATED [SIC3=737] (24 RIVALS), MOTION PICTURE PRODUCTION AND ALLIED SERVICES [SIC3=781] (4 RIVALS), MISC OTHER (13 RIVALS)

Core Words: ENTERTAINMENT (42), VIDEO (42), TELEVISION (38), ROYALTIES (35), INTERNET (34), CONTENT (33), CREATIVE (31), PROMOTIONAL (31), COPYRIGHT (31), GAME (30), SOUND (29), PUBLISHING (29), MUSIC (29), PROGRAMMING (29), CABLE (28), FORMAT (28), DEVELOPERS (28), CHANNEL (27), MASS (27), AUDIO (26), FUNCTIONALITY (26), FEATURE (25), FILM (25), TITLE (25), ANIMATION (25), ...

SubMarket 2: Medical Services (Sample Focal Firm: QUADRAMED CORP)

66 Rivals: IDX SYSTEMS, MEDICUS SYSTEMS, HPR, SIMIONE CENTRAL HOLDINGS, NATIONAL WIRELESS HOLDINGS, HCIA, APACHE MEDICAL SYSTEMS, ...

SIC CODES OF RIVALS: COMPUTER PROGRAMMING, DATA PROCESSING, AND OTHER COMPUTER RELATED [SIC3=737] (45 RIVALS), INSURANCE AGENTS, BROKERS, AND SERVICE [SIC3=641] (5 RIVALS), MISCELLANEOUS HEALTH AND ALLIED SERVICES, NOT ELSEWHERE CLASSIFIED [SIC3=809] (4 RIVALS), MANAGEMENT AND PUBLIC RELATIONS SERVICES [SIC3=874] (3 RIVALS), MISC OTHER (9 RIVALS)

Core Words: CLIENT (59), DATABASE (54), SOLUTION (49), PATIENT (47), COPYRIGHT (47), SECRET (47), PHYSICIAN (47), HOSPITAL (46), HEALTHCARE (46), SERVER (45), RESOURCE (44), FUNCTIONALITY (44), BILLING (44), CLIENTS (42), INTERFACE (41), EDUCATION (41), ARCHITECTURE (41), PRODUCTIVITY (41), ENTERPRISE (40), WINDOWS (40), DATABASES (40), REFORM (38), PROFESSIONALS (38), INFRINGEMENT (37), BACKGROUND (36), ...

SubMarket 3: Information Transmission (Sample Focal Firm: FAXSAV)

259 Rivals: OMTOOL LTD, CONCENTRIC NETWORK, PREMIERE TECHNOLOGIES, INTERNATIONAL TELECOMMUNICATION DATA SYSTEMS, IDT CORP, AXENT TECHNOLOGIES, SOLOPOINT, PRECISION SYSTEMS, NETRIX CORP, ...

SIC CODES OF RIVALS: COMPUTER PROGRAMMING, DATA PROCESSING, AND OTHER COMPUTER RELATED [SIC3=737] (112 RIVALS), COMMUNICATIONS EQUIPMENT [SIC3=366] (45 RIVALS), TELEPHONE COMMUNICATIONS [SIC3=481] (38 RIVALS), COMPUTER AND OFFICE EQUIPMENT [SIC3=357] (29 RIVALS), COMMUNICATIONS SERVICES, NOT ELSEWHERE CLASSIFIED [SIC3=489] (7 RIVALS), MISCELLANEOUS BUSINESS SERVICES [SIC3=738] (7 RIVALS), MISC OTHER (15 RIVALS)

Core Words: INTERNET (236), TELECOMMUNICATIONS (211), INTERFACE (194), COMMUNICATION (188), SOLUTION (187), PLATFORM (184), ARCHITECTURE (182), CALL (177), INFRASTRUCTURE (173), VOICE (173), FUNCTIONALITY (173), SERVER (173), COPYRIGHT (166), TRANSMISSION (164), REMOTE (163), WINDOWS (161), CHANNEL (160), CLIENT (160), DATABASE (158), TRAFFIC (156), MICROSOFT (156), INFRINGEMENT (153), CONNECTIVITY (146), EASE (145), USAGE (142), ...

SubMarket 4: Software (Sample Focal Firm: INTUIT)

52 Rivals: NETSCAPE COMMUNICATIONS, MYSOFTWARE, QUARTERDECK, SOFTWARE PUBLISHING CORP, GO2NET, MERIDIAN DATA, MACROMEDIA, MICROSOFT, CE SOFTWARE HOLDINGS, ...

SIC CODES OF RIVALS: COMPUTER PROGRAMMING, DATA PROCESSING, AND OTHER COMPUTER RELATED [SIC3=737] (48 RIVALS), MISC OTHER (4 RIVALS)

Core Words: INTERNET (52), FUNCTIONALITY (48), COPYRIGHT (48), MICROSOFT (48), WINDOWS (46), SOLUTION (45), EASE (44), SECRET (43), DIFFICULTIES (41), VERSION (41), INFRINGEMENT (41), DATABASE (41), CHANNEL (40), COPY (40), PLATFORM (39), SERVER (39), ENVIRONMENTS (38), PROBLEM (37), BACKGROUND (36), INTERFACE (36), DESPITE (36), DEVELOPERS (36), INTRODUCTIONS (36), DESKTOP (36), ENTERPRISE (35), DOCUMENTATION (34), ...

SubMarket 5: Corporate Data Management and Computing Solutions (Sample Focal Firm: HYPERION SOFTWARE)

207 Rivals: ORACLE CORP, FOURTH SHIFT CORP, APPLIX, TIMELINE, PLATINUM TECHNOLOGY, HARBINGER CORP, SANTA CRUZ OPERATION, EDIFY CORP, BANYAN SYSTEMS, ...

SIC CODES OF RIVALS: COMPUTER PROGRAMMING, DATA PROCESSING, AND OTHER COMPUTER RELATED [SIC3=737] (174 RIVALS), COMPUTER AND OFFICE EQUIPMENT [SIC3=357] (22 RIVALS), COMMUNICATIONS EQUIPMENT [SIC3=366] (2 RIVALS), MISC OTHER (15 RIVALS)

Core Words: SERVER (196), CLIENT (194), SOLUTION (193), ENTERPRISE (186), FUNCTIONALITY (185), WINDOWS (183), INTERNET (182), COPYRIGHT (180), MICROSOFT (177), DATABASE (174), ARCHITECTURE (171), INTERFACE (168), ENVIRONMENTS (164), SECRET (159), EASE (152), PLATFORM (151), DATABASES (150), UNIX (143), VENDOR (137), SUITE (134), INFRINGEMENT (131), ORACLE (127), TOOL (127), DESKTOP (127), COMMUNICATION (123), PROGRAMMING (123), ...

SubMarket 6: Retail (Sample Focal Firm: AMAZON.COM INC)

87 Rivals: PREVIEW TRAVEL, YAHOO, DATAMARK HOLDING, NETSCAPE COMMUNICATIONS CORP, WALL DATA, ONSALE, INFOSEEK CORP, IVI PUBLISHING, CASTELLE, CONNECT, NEW ERA OF NETWORKS, V ONE CORP, ...

SIC CODES OF RIVALS: COMPUTER PROGRAMMING, DATA PROCESSING, AND OTHER COMPUTER RELATED [SIC3=737] (66 RIVALS), COMPUTER AND OFFICE EQUIPMENT [SIC3=357] (5 RIVALS), NONSTORE RETAILERS [SIC3=596] (5 RIVALS), COMMUNICATIONS EQUIPMENT [SIC3=366] (4 RIVALS), MISC OTHER (14 RIVALS)

Core Words: INTERNET (84), FUNCTIONALITY (79), COPYRIGHT (78), DATABASE (77), INABILITY (74), SERVER (74), CLIENT (73), INFRINGEMENT (73), SECRET (72), SOLUTION (70), INTRODUCTIONS (70), MICROSOFT (70), ARCHITECTURE (69), DIFFICULTIES (68), DEPENDENCE (68), TELECOMMUNICATIONS (67), DESPITE (67), INFRASTRUCTURE (66), INTERFACE (66), WINDOWS (64), ENTERPRISE (62), COPY (62), EASE (62), CHANNEL (61), PLATFORM (60), VERSION (59), TRAIN (58), ENVIRONMENTS (57), DEVELOPERS (57), VENDOR (56), ALLIANCES (55), ...

Sample TNIC industries centered around firms residing in three digit SIC code 737 in the year 1997.

Table II: Sample Industries that Underwent Changes (TNIC Classifications)

**** Industry Surrounding Real Goods Solar in 1997 ***

Focal Firm: REAL GOODS TRADING CORP (SIC3=596)
1 Rival: PHOTOCOMM INC (SIC=362)

Core Words: ARRAY (2), FUEL (2), BACKUP (2), ELECTRIC (2), NORTHERN (2), REMOTE (2), VOLTAGE (2), UTILITY (2), CONSUMPTION (2), GRID (2), CONVERT (2), WEATHER (2), WIND (2), APPLIANCES (2), SIEMENS (2), AUDIT (2), ELECTRICITY (2), BATTERY (2), CATALOG (2), SPECIALISTS (2), EARTH (2), FOSSIL (2), GREEN (2), SIZING (2), INVERTERS (2), PHOTOCOMM (2)

**** Industry Surrounding Real Goods Solar in 2008 ***

Focal Firm: REAL GOODS SOLAR, INC.(gvkey=179417)(SIC3=362)

9 Rivals: DAYSTAR TECHNOLOGIES INC, AKEENA SOLAR, INC., EVERGREEN SOLAR INC, ASCENT SOLAR TECHNOLOGIES, INC., ENERGY CONVERSION DEVICES INC, SUNPOWER CORP, POWER ONE INC, FIRST SOLAR, INC.

SIC CODES OF RIVALS: ELECTRONIC COMPONENTS [SIC3=367] (6 RIVALS), ELECTRICAL INDUSTRIAL APPARATUS [SIC3=362] (1 RIVAL), RESEARCH AND TESTING SVCS [SIC3=873] (1 RIVAL)

Core Words: ELECTRIC (9), SILICON (9), ELECTRICITY (9), ROOF (9), INTEGRATORS (8), GRID (8), UTILITY (8), FILM (8), OUTPUT (8), SEMICONDUCTOR (8), WATT (8), SUNLIGHT (8), FUEL (7), INSTALLATIONS (7), METAL (7), CELL (7), INCENTIVES (7), FOOT (6), INITIATIVE (6), CONSUMPTION (6), GLASS (6), KYOCERA (6), SURFACE (6), SHARP (6), PEAK (6), TEMPERATURE (6), SUBSIDIES (6), VOLTAGE (6), FOSSIL (6), CADMIUM (6), SUNTECH (6), ...

**** Industry Surrounding L-1 Identity Solutions in 2008 ***

Focal Firm: L-1 IDENTITY SOLUTIONS INC (SIC3=737)
5 Rivals: COGENT, INC., WIDEPOINT CORP, SRA INTERNATIONAL, CACI INTERNATIONAL, ACTIVIDENTITY (All in SIC3=737)
* None of these firms existed as publicly traded firms in 1997 except for CACI International. Although CACI existed in 1997, it was in a different line of business (see below).

Core Words: DEFENSE (6), ARCHITECTURE (6), HOMELAND (6), CAPTURE (6), CLIENT (6), MILITARY (5), ENVIRONMENTS (5), INTEGRATORS (5), MOBILE (5), PROCUREMENT (5), PRIME (5), TRADITIONALLY (5), COPYRIGHT (5), COMBINE (5), DATABASE (5), INTELLIGENCE (5), BUDGET (5), INSTITUTE (5), MISSION (5), IDENTITY (5), INTEGRITY (5), GRUMMAN (5), NORTHROP (5), CONTRACTOR (4), WIRELESS (4), SURVEILLANCE (4), PRIVACY (4), PROCUREMENTS (4), CYBER (4), ...

**** Industry Surrounding CACI International in 1997 ***

SIC CODES OF 60 RIVALS: COMPUTER PROGRAMMING AND DATA PROCESSING [SIC3=737] (48 RIVALS), ENGINEERING AND ARCHITECTURAL [SIC3=871] (2 RIVALS), PERSONNEL SUPPLY SERVICES [SIC3=736] (2 RIVALS), PROFESSIONAL AND COMMERCIAL EQUIPMENT [SIC3=504] (2 RIVALS), MISC OTHER (6 RIVALS)

Core Words: CLIENT (56), SERVER (54), INTERNET (53), SOLUTION (51), ARCHITECTURE (51), DATABASE (51), ENTERPRISE (50), CLIENTS (48), DATABASES (48), PROGRAMMING (47), MICROSOFT (47), ENVIRONMENTS (46), PRODUCTIVITY (43), COPYRIGHT (43), SECRET (43), INTERFACE (42), WINDOWS (42), FUNCTIONALITY (40), TOOL (40), BACKGROUND (39), DOCUMENTATION (39), INTRANET (39), TELECOMMUNICATIONS (38), OBJECT (38), CYCLE (36), LEGACY (36), SUITE (36), VENDOR (36), ...

**** Industry Surrounding CACI International in 2008 ***

SIC CODES OF 18 RIVALS: COMPUTER PROGRAMMING AND DATA PROCESSING [SIC3=737] (8 RIVALS), SEARCH, DETECTION, NAVIGATION, GUIDANCE, AND AERONAUTICAL [SIC3=381] (5 RIVALS), COMMUNICATIONS EQUIPMENT [SIC3=366] (2 RIVALS), MISC OTHER (3 RIVALS)

Core Words: DEFENSE (19), MILITARY (18), MISSION (18), CONTRACTOR (17), HOMELAND (17), PROCUREMENT (17), PRIME (17), QUANTITY (16), INTELLIGENCE (16), ENVIRONMENTS (15), AWARD (15), BUDGET (14), COMMAND (14), ARCHITECTURE (13), SPECTRUM (13), UNDERSTANDING (13), WARFARE (13), SURVEILLANCE (13), TASK (12), LOCKHEED (12), MARTIN (12), SUBCONTRACTOR (12), PROPOSAL (12), PROCUREMENTS (12), RECONNAISSANCE (12), ARMY (11), ...

Sample TNIC industries that changed dramatically between 1997 and 2008.

Table III: Firm Characteristics and Industry Classifications

Row	Variable	Adj R^2 SIC-3 Fixed Effects	Adj R^2 SIC-3 Equal Weighted Average	Adj R^2 NAICS-3 Fixed Effects	Adj R^2 NAICS-4 Equal Weighted Average	Adj R^2 10-K 300 Fixed Effects	Adj R^2 10-K 300 Equal Weighted Average (Annual)	Adj R^2 Conglom. Adjusted 10-K 300 Average	Adj R^2 TNIC Equal Weighted Average	Adj R^2 TNIC Simil. Weighted Average (Ex Self)	Adj R^2 TNIC Equal Weighted Average (Ex Vert)	Adj R^2 TNIC Simil. Weighted Average (Ex Self) (Ex Vert)
(1)	OI/Sales	0.284	0.312	0.287	0.314	0.327	0.372	0.355	0.458	0.414	0.458	0.414
(2)	OI/Assets	0.177	0.208	0.184	0.216	0.216	0.272	0.252	0.375	0.290	0.375	0.290
(3)	Sales Growth	0.023	0.070	0.025	0.082	0.026	0.096	0.088	0.172	0.038	0.172	0.038
(4)	R+D/Sales	0.138	0.169	0.137	0.170	0.191	0.250	0.220	0.203	0.206	0.203	0.206
(5)	Adver./Sales	0.041	0.084	0.061	0.110	0.071	0.169	0.149	0.272	0.159	0.272	0.159
(6)	Book Leverage	0.221	0.245	0.238	0.263	0.209	0.181	0.222	0.327	0.225	0.327	0.225
(7)	Market Leverage	0.277	0.311	0.302	0.337	0.262	0.220	0.280	0.392	0.303	0.392	0.303
(8)	Market Beta	0.096	0.153	0.097	0.160	0.104	0.157	0.160	0.245	0.118	0.245	0.118

Firm characteristics are regressed on various industry industry controls, including fixed-effect-based and industry-averaging method-based controls. All regressions are based on our entire sample from 1997 to 2006, and also include yearly fixed effects. All TNIC industries are based on a 25% stop word threshold.

Table IV: Managerial Indications of High Competition and Industry Competitiveness Measures

Row	Words Used for TNIC Industries	Stop Word Threshold	TNIC HHI	SIC-3 HHI	OI/ Assets	Log Firm Age	Tobin's Q	Log Sales	# Words Bus. Desc.	# Words MD&A	# Obs./ R^2
(1)	All Words	100%	-0.157	-0.202							34,412
			(-6.06)	(-5.10)							0.026
(2)	All Words	25%	-0.218	-0.177							34,412
			(-7.09)	(-4.54)							0.028
(3)	All Words	10%	-0.159	-0.205							34,412
			(-5.33)	(-5.14)							0.026
(4)	All Words	TF-IDF	-0.103	-0.212							34,412
			(-3.76)	(-5.29)							0.024
(5)	Nouns and Proper Nouns	100%	-0.173	-0.199							34,412
			(-6.52)	(-5.04)							0.026
(6)	Nouns and Proper Nouns	25%	-0.241	-0.175							34,411
			(-7.74)	(-4.45)							0.029
(7)	Nouns and Proper Nouns	10%	-0.158	-0.205							34,409
			(-5.18)	(-5.14)							0.026
(8)	Nouns and Proper Nouns	TF-IDF	-0.117	-0.211							34,412
			(-4.35)	(-5.26)							0.025
(9)	Nouns and Proper Nouns	25%	-0.276								34,411
			(-8.65)								0.026
(10)	Nouns and Proper Nouns	25%	-0.241	-0.175							34,411
			(-7.74)	(-4.45)							0.029
(11)	Nouns and Proper Nouns	25%	-0.244	-0.131	-0.109	-0.121	0.036				34,411
			(-7.88)	(-3.43)	(-5.46)	(-4.58)	(1.66)				0.034
(12)	Nouns and Proper Nouns	25%	-0.170	-0.086	0.032	-0.094	0.099	-0.292	-0.335	1.205	34,411
			(-5.18)	(-2.26)	(1.26)	(-2.98)	(4.68)	(-7.23)	(-10.29)	(30.26)	0.156

The table reports the results of logistic regressions where the dependent variable is one if the firm's management mentions high competition (or a synonym thereof) in its Management and Discussion Section of its 10-K in the given year. Independent variables include measures of competitiveness based on TNIC and SIC based classifications (of equal granularity) and additional control variables including sales, age, profitability, Tobin's Q, and document size variables including the number of words in the business description and the MD&A sections of the firm's 10-K. The "Stop Word Threshold" column indicates whether we discard common words defined as those used in at least 10%, 25% or 100% of all documents, or if we instead use TF-IDF to weight common words less heavily as an alternative to discarding them.

Table V: Self Reported Capital IQ Peers and Industry Classifications

		TNIC (set to SIC-3 Granularity)		***TNIC (set to NAICS-4 Granularity)***	
Words Used for TNIC Industry	Stop Word Threshold	TNIC Overlap with Cap IQ	TNIC Overlap with SIC-3	TNIC Overlap with Cap IQ	TNIC Overlap with NAICS-4
		Panel A: Capital IQ Competitors			
All Words	100%	40.9%	46.6%	43.1%	61.8%
All Words	25%	50.6%	50.2%	53.0%	65.8%
All Words	10%	60.1%	49.1%	62.3%	61.5%
All Words	TF-IDF	59.3%	49.0%	61.9%	65.6%
Nouns and Proper Nouns	100%	43.7%	47.3%	46.2%	62.5%
Nouns and Proper Nouns	25%	52.5%	50.2%	55.1%	65.6%
Nouns and Proper Nouns	10%	62.0%	45.8%	63.5%	54.4%
Nouns and Proper Nouns	TF-IDF	58.5%	48.1%	61.0%	64.6%
**Note: The overlap between SIC-3 and Capital IQ Competitors is 47.1%. The overlap between NAICS-4 and Capital IQ Competitors is 44.0%.*					
		Panel B: Capital IQ Alliances			
All Words	100%	35.4%	40.8%	28.4%	41.7%
All Words	25%	40.6%	44.2%	33.6%	47.1%
All Words	10%	43.3%	43.1%	36.5%	47.0%
All Words	TF-IDF	48.3%	42.1%	40.8%	44.5%
Nouns and Proper Nouns	100%	36.6%	40.7%	29.9%	42.7%
Nouns and Proper Nouns	25%	42.2%	43.4%	34.7%	46.9%
Nouns and Proper Nouns	10%	44.3%	42.3%	36.1%	46.3%
Nouns and Proper Nouns	TF-IDF	48.6%	40.2%	40.2%	42.6%
**Note: The overlap between SIC-3 and Capital IQ Alliances is 28.2%. The overlap between NAICS-4 and Capital IQ Alliances is 22.9%.*					

The table reports the fraction of Capital IQ 2011 peers that are also peers as identified by various other industry classifications, including SIC-3, NAICS-4, and TNIC-based classifications constructed to have identical levels of granularity as SIC-3 and NAICS-4. The table also reports the fraction of overlap between SIC-3 and TNIC, and also between NAICS-4 and TNIC. Although Capital IQ data is from 2011 (historical peer data is not available), all SIC, NAICS and TNIC data is from 2008. The "Stop Word Threshold" column indicates whether we discard common words defined as those used in at least 10%, 25% or 100% of all documents, or if we instead use TF-IDF to weight common words less heavily as an alternative to discarding them.

Table VI: Pearson Correlation Coefficients

Row	Variable	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)	Average Similarity (10-K based)	0.812						
(2)	Sales Herfindahl (10-K based)	-0.319	-0.370					
(3)	Sales C4 Index (10-K based)	-0.553	-0.520	0.795				
(4)	Sales Herfindahl (SIC-3 based)	-0.217	-0.202	0.232	0.284			
(5)	Sales C4 Index (SIC-3 based)	-0.300	-0.239	0.243	0.328	0.831		
(6)	Sales Herfindahl (NAICS-4 based)	-0.289	-0.227	0.238	0.310	0.566	0.553	
(7)	Sales C4 Index (NAICS-4 based)	-0.437	-0.343	0.279	0.414	0.524	0.647	0.827

Pearson Correlation Coefficients are reported for our sample of 51,657 observations based on 1997 to 2006. The 10-K based market structure measures are based on 10K-TNIC industries (uses the same number of pairings as three digit SIC codes). All TNIC industries are based on a 25% stop word threshold.

Table VII: Ex-ante investment versus future product differentiation

	Dependent Variable	Positive Adver. Dummy	Positive R&D Dummy	Log Industry Adver. / Sales	Log Industry R&D / Sales	Ind Past Stock Return	Log Assets	Ind. Log B/M Ratio	Adj R^2
(1)	Δ Log Total Summed Similarity	-0.326	0.065	-0.026	0.019	0.181	0.059	-0.131	0.091
		(-3.56)	(0.55)	(-2.24)	(0.81)	(1.51)	(3.71)	(-2.60)	
(2)	Δ Average Similarity	-0.000	0.000	-0.000	-0.000	0.000	0.000	-0.000	0.015
		(-1.12)	(0.39)	(-2.49)	(-0.36)	(0.70)	(1.78)	(-1.59)	
(3)	Δ Sales 10-K Based HHI	0.038	0.016	0.001	0.000	-0.001	-0.002	0.004	0.020
		(6.25)	(3.90)	(1.38)	(0.37)	(-0.74)	(-2.66)	(2.43)	
(4)	Δ Sales 10-K Based C4 Index	0.046	0.014	0.003	0.000	-0.003	0.000	0.002	0.036
		(12.69)	(4.69)	(7.28)	(1.05)	(-2.66)	(0.25)	(1.92)	
(5)	Δ Observed Lerner Index	0.011	0.024	0.002	0.003	-0.013	-0.001	0.003	0.041
		(2.03)	(4.50)	(3.07)	(5.00)	(-5.30)	(-0.76)	(1.91)	
(6)	Δ Observed Firm Profitability	0.010	0.025	0.001	0.003	-0.015	-0.000	0.004	0.019
		(1.75)	(4.44)	(2.57)	(4.92)	(-5.46)	(-0.33)	(2.12)	

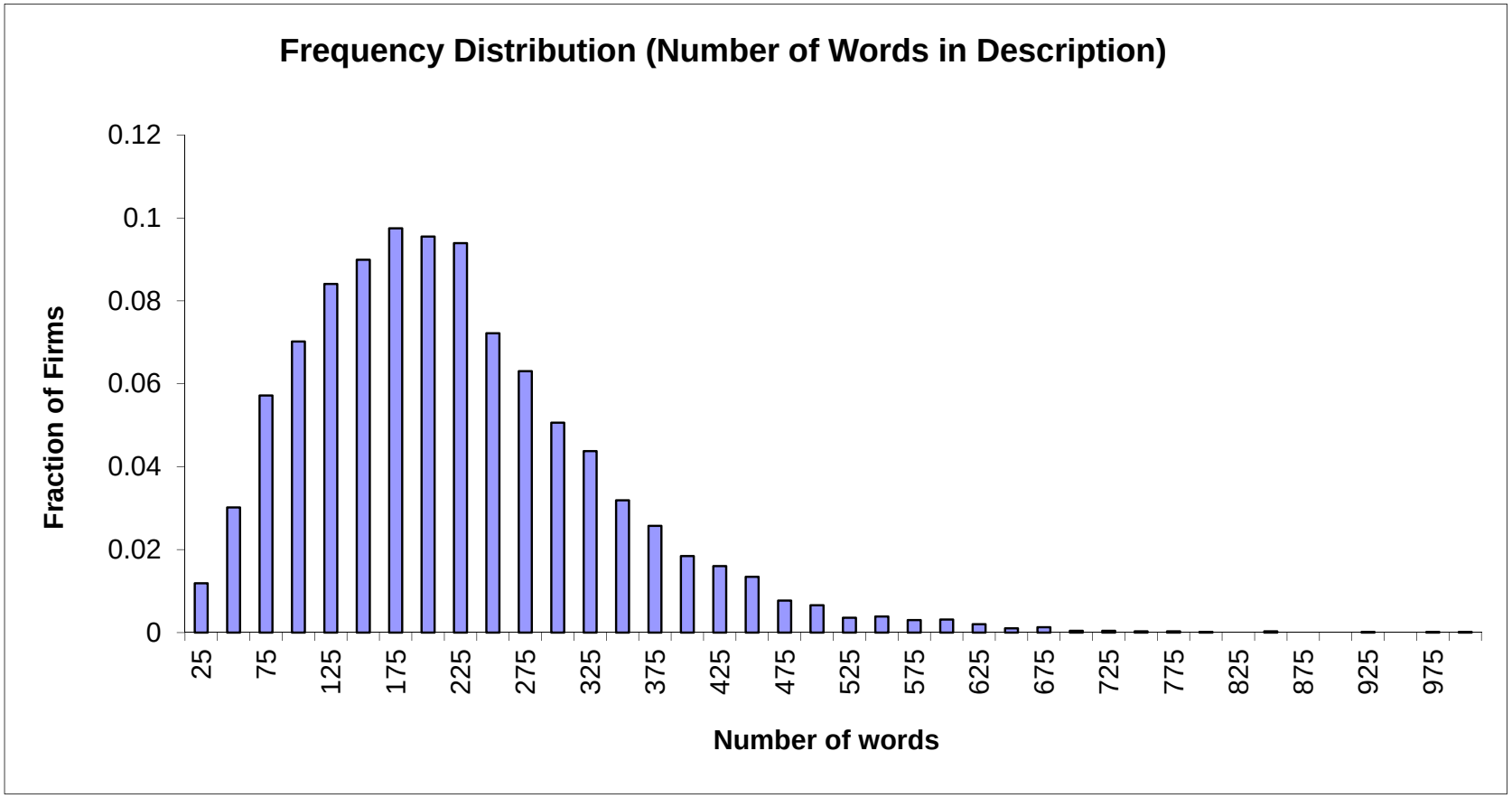
OLS regressions with ex post product changes in market structure (based on 10K-TNIC industries) as the dependent variables. All specifications include industry and yearly fixed effects, and standard errors account for clustering by year and industry (industry controls are based on 10K-300 FIC industries, although results are very similar if we instead use three-digit SIC industries (not reported). The sample has 49,246 observations and is from 1997 to 2006.

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

Dependent Variable	Positive Adver. Dummy	Positive R&D Dummy	Log Industry Adver. / Sales	Log Industry R&D / Sales	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	0.007	-0.020	0.001	-0.000	-0.010	0.008	0.005	0.103
	(0.39)	(-1.35)	(0.59)	(-0.12)	(-2.16)	(1.22)	(0.67)	
(2) Δ Sales SIC-3 C4 Index	0.009	0.005	0.002	0.000	-0.003	-0.001	0.002	0.109
	(1.13)	(0.65)	(1.37)	(0.22)	(-1.64)	(-0.63)	(0.68)	
(3) Δ Observed Firm Profitability	0.006	0.021	0.001	0.003	-0.006	-0.003	0.010	0.094
	(0.52)	(1.64)	(0.49)	(1.40)	(-1.06)	(-0.87)	(1.87)	
Panel B: NAICS-4 Based Market Structure Measures and Industry Controls								
(4) Δ Sales NAICS-4 HHI	-0.006	-0.052	0.000	-0.005	-0.011	0.007	0.023	0.128
	(-0.35)	(-3.32)	(0.02)	(-2.49)	(-2.32)	(1.19)	(2.83)	
(5) Δ Sales NAICS-4 C4 Index	-0.003	-0.012	0.001	-0.002	-0.008	-0.002	0.007	0.116
	(-0.29)	(-1.64)	(0.41)	(-1.41)	(-3.55)	(-0.93)	(2.48)	
(6) Δ Observed Firm Profitability	0.028	0.025	0.003	0.004	-0.006	0.004	0.007	0.175
	(1.19)	(1.43)	(1.17)	(1.60)	(-1.06)	(0.52)	(0.97)	

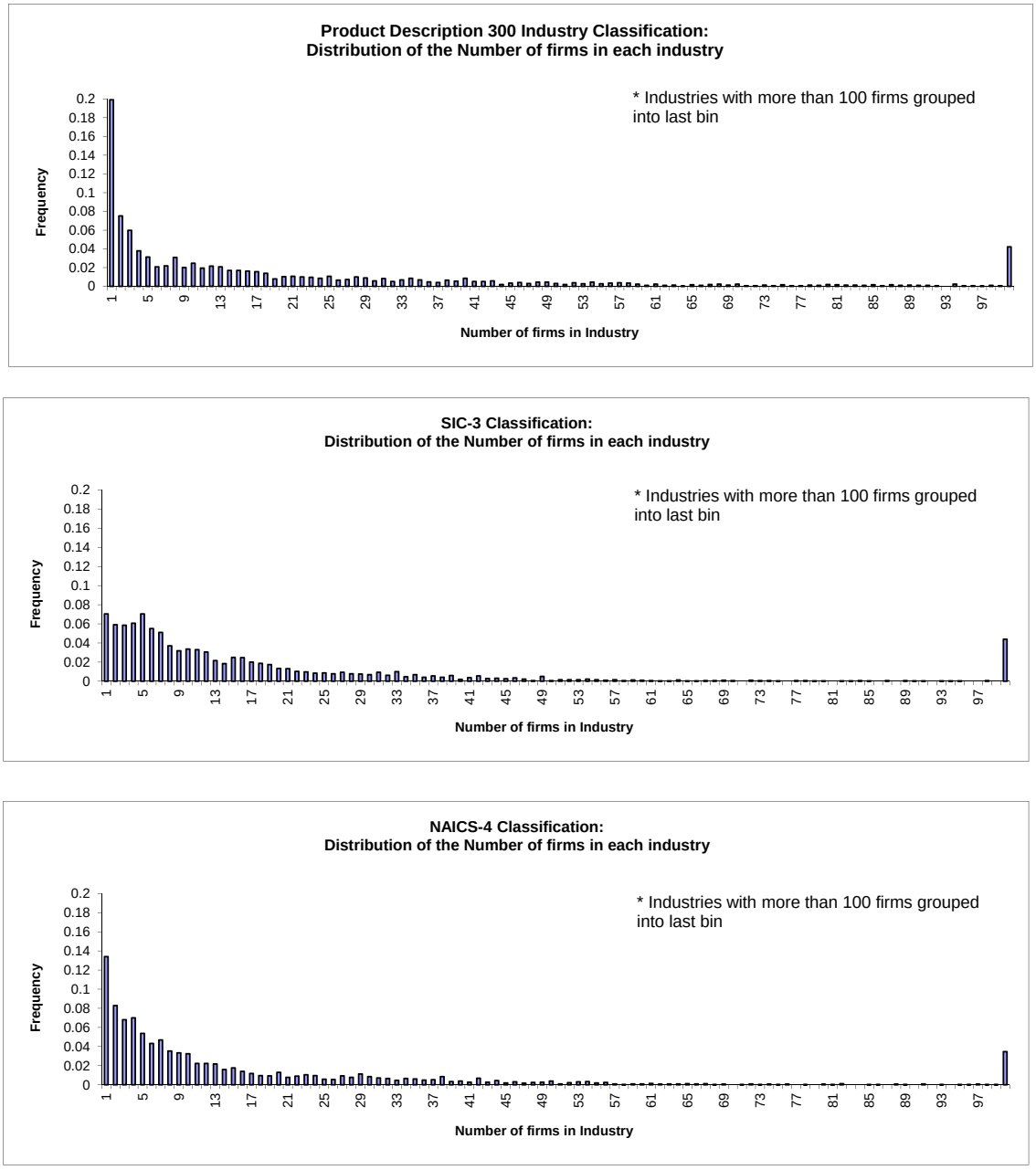
OLS regressions with ex post product changes in market structure (based on three-digit SIC in Panel A, and four-digit NAICS in Panel B) as the dependent variables. All specifications include industry and yearly fixed effects, and standard errors account for clustering by year and industry (industry controls are based on three-digit SIC in Panel A, and four-digit NAICS in Panel B). The sample has 49,246 observations and is from 1997 to 2006.

Figure 1:



Frequency distribution of unique non-common noun and proper noun 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.

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. Specifically, we restrict attention to words that are either nouns or proper nouns, and that also appear in fewer than 25% of all business 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 define the frequency vector V_i to be normalized to unit length.

$$V_i \;=\; \frac{P_i}{\sqrt{P_i \;\cdot\; P_i}} \tag{6}$$

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

$$Product\ Similarity_{i,j} \;=\; (V_i \;\cdot\; V_j) \tag{7}$$

We define product differentiation as one minus similarity.

$$Product\ Differentiation_{i,j} \;=\; 1 \;-\; (V_i \;\cdot\; V_j) \tag{8}$$

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" method, 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 (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 generate a similar number of industries. 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 to have N industries, 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 industry count to $N-1$ industries, we take the maximum pairwise industry similarity as follows

$$\underset{j,k,\ j\neq k}{MAX} \qquad I_{j,k} \tag{9}$$

The two industries with the highest similarity are then combined, reducing the industry count by one. This process is repeated until the number of industries reaches the desired number. 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:

$$I_{l,q} = \sum_{x=1}^{m_l} \sum_{y=1}^{m_q} \frac{S_{x,y}}{m_l \ \ m_q} \tag{10}$$

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 industry count, 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. We do not take further steps to ensure uniqueness due to computational limitations. Also, any departure from the true optimal set of industries would bias our study away from finding significant results, and hence our approach is conservative and might understate the true power of 10-K business 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 25% 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, and our results are virtually unchanged if we further omit 1997 from our sample.

Appendix 3

In this appendix, we further assess the performance of 10K-FIC industries versus SIC and NAICS industries by exploring various levels of granularity. A key advantage of our approach is the ability to set granularity to any arbitrary level. We use the Akaike information criterion to examine which level of granularity is most likely to explain firm characteristic data. Understanding granularity is relevant to understanding the role and breadth of economies of scope.

[Insert Table A3 Here]

Table A3 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 industry counts, we can also broadly compare SIC to NAICS. Panels A and B show that SIC and NAICS are reasonable substitutes for each other. 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 NAICS.

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 2603.1 (10K-300 industries) is broadly lower than the 3091.4 for three digit SIC codes, and the 3097.7 for four digit NAICS codes, even though all three groupings have similar granularity levels.

Although we can conclude that 10K-based industries are more informative than SIC or NAICS industries, Panel C draws only a moderately decisive conclusion that the AIC scores reach a minimum at 300 industries. This minimum is surrounded by only a gradual slope. We conclude that the degree of granularity (roughly 300 industries) used by SIC and NAICS is reasonable, and is also a good benchmark for 10-K based industries.

Table A3: Industry classifications and industry granularity

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	3783.2	0.146	-35.7	-0.000	10	561.0
(2)	SIC-2-digit	3277.7	0.228	-269.3	0.043	72	77.9
(3)	SIC-3-digit	3091.4	0.277	-685.2	0.120	274	20.5
(4)	SIC-4-digit	3039.2	0.301	-808.6	0.167	434	12.9
	Panel B: NAICS based industry definitions						
(5)	NAICS-1-digit	4281.5	0.066	-192.0	0.029	9	623.3
(6)	NAICS-2-digit	3549.2	0.182	-475.9	0.079	23	243.9
(7)	NAICS-3-digit	3219.1	0.238	-750.6	0.133	96	58.4
(8)	NAICS-4-digit	3097.7	0.278	-830.6	0.173	328	17.1
(9)	NAICS-5-digit	3400.1	0.270	-512.5	0.162	672	8.3
(10)	NAICS-6-digit	3602.1	0.271	-299.1	0.161	983	5.7
	Panel C: 10-K product description based industry definitions						
(11)	10K-based-50	2855.8	0.280	-1109.2	0.181	50	112.1
(12)	10K-based-100	2684.5	0.308	-1190.0	0.200	100	56.0
(13)	10K-based-200	2666.6	0.318	-1178.3	0.208	200	28.0
(14)	10K-based-250	2678.7	0.322	-1166.4	0.212	250	22.4
(15)	10K-based-300	2603.1	0.334	-1203.1	0.220	300	18.7
(16)	10K-based-400	2590.9	0.342	-1184.9	0.225	400	14.0
(17)	10K-based-500	2682.0	0.339	-1127.9	0.227	500	11.2
(18)	10K-based-800	2851.5	0.337	-1003.7	0.229	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.