Forthcoming Journal of Political Economy

Text-Based Network Industries and Endogenous Product Differentiation

Gerard Hoberg and Gordon Phillips*

*University of Southern California and University of Southern California and National Bureau of Economic Research respectively. Hoberg can be reached at hoberg@marshall.usc.edu and Phillips can be reached at gordon.phillips@marshall.usc.edu. We especially thank our Editor, Sam Kortum, and three anonymous referees. We also thank Dan Kovenock, Steve Martin, Nagpurnanand Prabhala, John Sutton and seminar participants at Aalto (Helsinki) School of Economics, HEC-Paris, IFN (Stockholm), Insead, ISTCE (Lisbon), London Business School, Notre Dame, Northwestern, Stanford, Stockholm School of Economics, University of Amsterdam, University of Chicago, University of Southern California, University of Vienna and the Academy of Management meetings for helpful comments. All errors are the authors alone. Copyright ©2009 by Gerard Hoberg and Gordon Phillips. All rights reserved.

ABSTRACT

We study how firms differ from their competitors using new time-varying measures of product similarity based on text-based analysis of firm 10-K product descriptions. This year-by-year set of product similarity measures allows us to generate a new set of industries where firms can have their own distinct set of competitors. Our new sets of competitors explain specific discussion of high competition, rivals identified by managers as peer firms and changes to industry competitors following exogenous industry shocks. We also find evidence that firm R&D and advertising are associated with subsequent differentiation from competitors, consistent with theories of endogenous product differentiation.

I. Introduction

Defining industry boundaries and industry competitiveness is central to the study of industrial organization. It is also central to broader disciplines in Economics, Finance and Management Strategy, where the study of industries, or the need to control for industry, is pervasive. We develop new time-varying industry classifications using text-based analysis of firm product descriptions filed with the Securities and Exchange Commission. Our paper is based on the premise that product similarity is core to classifying industries, and that empirical work can benefit from more flexible measures of industry membership, product differentiation, and changes in both as they occur in each year. We use these new industries to show how industries and their competitors change both in competitive intensity and in product offerings following industry shocks. We also show that firm research and development (R&D) and advertising are associated with subsequent differentiation from competitors and increased profitability.

Our starting point is to gather business descriptions from 50,673 firm annual 10-Ks filed with the Securities and Exchange Commission (SEC) using web crawling algorithms. We process the text in these business descriptions to form new industry classifications based on the strong tendency of product market vocabulary to cluster among firms operating in the same market. 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 existing industry classification schemes).[1] Using the traditional industry groups, firms are placed within predefined industry groups, instead of using the information that firms provide to determine who they compete against. To identify related firms, our methods use the business description section of the 10-K where firms give detail on the products they offer. In particular, the business description section of the 10-K is mandated by SEC regulations requiring firms to describe the significant products they offer to their customers. Our new classifications are thus based on the products sold by firms that arise from underlying consumer preferences and demand.

There are two central ideas in our paper. The first is that 10-K product words describe the features and bundles of products each firm offers. Thus, we use the text in each firm's 10-K business description to assign each firm a spatial location based on product words, generating a Hotelling-like product location space for publicly traded U.S. firms.[2] Each firm has a unique spatial location and its own potential set of nearby competitors in this space based on product word overlaps. Groups of likely competitors are thus located in clusters in space analogous to cities on a map. Larger distances within a cluster indicate product differentiation, and distances across clusters indicate across-industry similarity.

[1] See http://www.naics.com/info.htm. The Census Department states "NAICS was developed to classify units according to their production function. NAICS results in industries that group units undertaking similar activities using similar resources but does not necessarily group all similar products or outputs."

[2] 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.

The second central idea relates to networks, as we calculate how similar each firm is to every other firm by calculating firm-by-firm pairwise word similarity scores using the 10-K product words. We thus reduce high-dimensional word vectors to a simple matrix of firm pairwise similarity scores. Using these pairwise similarity scores, we then group firms into industries. Our general industry classification can then be represented as an unrestricted network of firms. Because firms update their 10-Ks, the network is time-varying. In this network, a firm's competitors are analogous to a Facebook circle of friends, where each firm can have its own distinct set of competitors. Because each firm-pair has a continuous degree of relatedness, the analogy is that some pairs are close friends and some pairs are more distant acquaintances.

What makes our analysis possible is that publicly traded firms must file a 10-K in each year, allowing us to build classifications that change over time. Using this time-varying feature, we examine how firms react to changes within and around their product markets over time.[3] For example, we assess the extent to which firms adjust product offerings following large industry shocks. Although numerous studies use industry classifications as control variables, only a few studies examine the classification schemes themselves and these do not consider the possibility of industry classifications that change materially over time.[4]

[3] Note that only publicly traded firms file 10-Ks. However the methods can be applied to a broader set of product descriptions using firm Internet web pages.

[4] 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.

We create 10-K based industry classifications 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"*, is analogous to SIC and NAICS industries. Firms are grouped together using fixed product market definitions and industry membership is constrained to be transitive. Therefore, 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. To implement this method, we assign firms to industries using a clustering algorithm that maximizes total within-industry similarity based on word usage in 10-K product descriptions.

Our second more general network classification is unconstrained. We now allow product market definitions to change every year and we relax the membership transitivity requirement of fixed industry classifications. We thus view industries as time-varying intransitive networks. We name these new industries *"text-based network industry classifications"* or TNIC on our external web site where we maintain the data.[5] In this classification system, each firm has its own set of distinct competitors. To illustrate why transitivity is restrictive, suppose firms A and B both view firm C as a rival. If A and B 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, our text-based classifications offer economically large improvements in their ability to explain differences in key charac-

[5] Our new industry network groupings and underlying data are available for download at http://cwis.usc.edu/projects/industrydata/.

teristics such as profitability, sales growth, and market risk across industries. They also better explain the extent to which managers mention high competition in the Management's Discussion and Analysis section of the 10-K, the specific firms mentioned by managers as being competitors, and how advertising and R&D investments relate to future product differentiation. Our empirical tests benefit from information regarding the degree to which specific firms are similar to their competitors and how this changes over time, neither of which can be derived from static zero-one membership classifications such as SIC or NAICS.[6]

Using our ability to identify both the time-varying product market location of a firm and the time-varying identity of its rivals, we examine how these items change following large exogenous industry shocks. We focus on the September 11, 2001 shock to the military goods and services industry, and also on the post-2000 collapse of the software industry. Following these exogenous shocks, our new industry classifications capture significant changes in the membership of each firm's rivals, the degree of product similarity, and also in the nature and type of products offered. Our results suggest that a positive shock to the military industry led to increases in competition and increases in product market similarity as rivals relocated in the product market space to areas of common high demand. In contrast, a negative industry shock in the software industry led to reductions in similarity and movement toward more differentiated products.

[6] 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.

We also examine whether advertising and research and development are correlated with decreasing ex-post product similarity. We find that firms spending more on either advertising or R&D experience significant reductions in measures of ex-post similarity and significant gains in ex-post profitability. These findings are consistent with Sutton (1991)'s hypothesis that firms spend on advertising and R&D to differentiate themselves and 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.[7]

One of the benefits of our approach is that it allows both within-industry and across-industry relations to be examined. Many empirical studies examining product differentiation focus on single industries,[8] while an older literature summarized by Schmalensee (1989) focuses on cross-industry relations. We are able to identify a unique set of industry rivals surrounding each firm over time as in the circular city model of Chamberlin, which relaxes the restrictive transitivity property of existing classifications. We also identify groups of other firms that share some similar vocabulary as the initial group of firms, thus capturing across-industry relatedness.

Although it is convenient to use existing industry classifications such as SIC or NAICS for research purposes, these measures have at least four limitations. First, neither reclassifies firms significantly over time as the product market evolves. Second,

[7] We note that while our new measures are interesting for research or scientific purposes, for example to examine innovation, shocks, or industry life-cycles, they are less useful for policy and antitrust purposes as they could be manipulated by firms if they believed they were being used for this purpose.

[8] For recent examples see Nevo (2000), Mazzeo (2002), Seim (2006) and Gowrisankaran and Rysman (2012).

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. Third, SIC and NAICS impose transitivity even though two firms that are rivals to a third firm might not be rivals. Lastly, they do not provide continuous measures of similarity both within and across industries. There are also econometric benefits to using our new industries as they are more informative in tests of external validity that we conduct in this paper.

Our new classifications can also be used in conjunction with, not in lieu of other data. Although not part of the current study, 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-specific price levels can also be estimated using various weighting methods based on firm 10-K text.

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, Gentzkow and Shapiro (2010) measure the Republican vs Democratic slant of U.S. daily newspapers and document a preference among readers for like-minded news. Papers such as Antweiler and Frank (2004), Tetlock (2007), Tetlock, Saar-Tsechanksy, and Macskassy (2008), and Loughran and McDonald (2011) examine the tone of various documents and link them for example to stock price movements.

The remainder of the paper is organized as follows. We describe the verbal data and similarity calculations in Section II. We give methodological details for our new industry classifications in Section III. We discuss central properties of new industry classifications and give new industry examples in Section IV. In Section V we examine the external validation of our new industry classifications. Section VI examines how industry similarity and competitors change over time following large exogenous industry shocks. Section VII tests theories of endogenous barriers to entry and examines how R&D and advertising are associated with subsequent changes in similarity and profitability. Section VIII concludes.

II. Objective and Methodology: From Words to Industry Classifications

Our industry classifications are based on the notion that firms in the same industry use many of the same words to identify and describe their products. In this section, we describe our objective in building new measures of industry relatedness. We also describe the underlying data structures that define our new industries.

A. Objective

Our overall objective is to capture the relatedness of firms based on their product offerings to customers using a flexible network approach. This approach provides a measure of distance between firm pairs in product space and does not impose transitivity between members of the network. Our approach allows competitors of each firm to be identified based on the similarity of their product offerings. Competitors can be different for each firm - even when firms may share some overlapping competitors. Thus an industry can be thought of as a cluster in a network with porous boundaries. This more general classification of industries can be used to test a larger set of economic hypotheses (see Section IV).

Our objective differs from that of traditional SIC or NAICS industry classifications, which is to place firms in predefined industry categories based on production processes, not the products they offer to customers.[9] SIC and NAICS industries also impose transitivity among group members and provide no measure of similarity between firms within an industry, or between firms in neighboring industries.

Our methods come closer to capturing the fact that cross-price elasticities of demand between firms may be different for different pairs of firms within the same industry. In many industry studies in industrial organization, researchers obtain detailed price and quantity data in order to measure these intra-industry cross-price elasticities. Our classifications provide measures of distance for all firm pairs simultaneously without having to obtain detailed price and quantity data - which for many

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

competitors with differentiated products can be difficult to obtain.

A second major part of our objective is to allow for frequent annual updating. Our industries are updated every year as firm product offerings change. In contrast, firm SIC and NAICS codes update infrequently despite the fact that firm products change materially over time, as can be seen in their 10Ks.

The last part of our objective is to uniquely capture horizontal relatedness between firms, not vertical relatedness. As we describe later in the paper, we thus remove pairs from our related pairs that are in two-different traditional industries classified as shipping to each other using the Input - Output tables of the BEA. The pairs we remove turn out to be relatively scarce, representing just 4% of the pairs in our data. We thus conclude that our methods naturally capture horizontally related product offerings and not vertical links or vertical production processes.

B. Capturing Relatedness Between Firms

Our primary building block is the set of unique words that firms use to describe their products in their business description sections from 10-K annual filings on the SEC Edgar website from 1996 to 2008. These descriptions are found in a separate section of each 10-K filed by each firm. 10-K business descriptions are legally required to be accurate, as Item 101 of Regulation S-K requires that firms describe the significant products they offer to the market. 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 industries change over time.

We use the 10-K business descriptions to compute pairwise word similarity scores for each pair of firms in a given year. In our main specification, we limit attention to nouns (defined by Webster.com)[10] and proper nouns that appear in no more than 25% of all product descriptions in order to avoid common words. 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 discuss later, our results are robust to altering these stop word thresholds.

Figure 1 displays a histogram showing the number of unique words in firm product descriptions. 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 some 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 1 Here**]

We map firms into industries using word vectors and firm pairwise cosine similarity scores based on the words used by each firm. Full details regarding our implementation of the cosine similarity calculation are in Appendix 1. We give a basic description here. Suppose there are W unique words used in the union of the documents used by all firms in our sample. In our sample, W is 61,146 unique nouns and proper nouns

[10] When a word can be used as more than one part of speech, we include the word in our universe if it has at least one use as a noun.

in 1996 and 55,605 in 2008. A given firm i's vocabulary can then be represented by a W-vector P_i, with each element being populated by the number one if firm i uses the given word, and zero if it does not. We then normalize each vector to have unit length as follows:

$$V_i \;=\; \frac{P_i}{\sqrt{P_i \;\cdot\; P_i}} \quad \forall i, j \tag{1}$$

Given that each vector has length W, and because we normalize these vectors to have unit length, all firms in a given year thus reside in a space shaped as the surface of a W-dimensional unit sphere. We define Q_t as the matrix containing the set of normalized vectors V_i for all firms i in year t. Q_t is thus an N_t x W matrix, where N_t is the number of firms in year t. Each row i of Q_t contains the normalized vector V_i defined above for firm i in year t. Q_t is thus a complete description of the firm-to-word spatial representation of firms in product space over time.

In order to derive the firm-to-firm network representation of our industries, we use the vectors V_i and V_j for a pair of firms i and j to calculate the product cosine similarity or the firm pairwise similarity score as follows:

$$Product\ Cosine\ Similarity_{i,j} \;=\; (V_i \;\cdot\; V_j) \tag{2}$$

The network representation of firms is fully described by an $N_t x N_t$ square matrix M_t (i.e., a network), where an entry of this matrix for row i and column j is the $Product\ Cosine\ Similarity_{i,j}$ for firms i and j defined above. The large number of words used in business descriptions ensures that the matrix M_t is not sparse, and that its entries are unrestricted real numbers in the interval $[0, 1]$. Because firms update their 10-Ks annually, M_t is time-varying.

Intuitively, the cosine similarity is higher when firms i and j use more of the same words, as both vectors will then have positive values in the same elements. Because we populate P_i with binary values, our baseline method weights words equally regardless of their frequency.[11]

We use the "cosine similarity" method for many reasons (see Sebastiani (2002) for a detailed review of related methods). 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 only moderately computationally burdensome, making it practical to replicate or extend. Finally, this method's normalization builds in a natural control for document length. It is called cosine similarity because it measures the angle between two word vectors on a unit sphere.

III. Industry Classification Methods and Firm 10-Ks

We construct network industry classifications using the matrix of firm pairwise cosine similarity scores (M_t) as the basic building block. We consider two methods. First, we consider a fixed industry classification method where we impose transitivity on firm membership such that if firm A and firm C are in the same industry as firm B, then firms A and C are in the same industry. Second, we relax transitivity and allow

[11] Following Loughran and McDonald (2011), 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 suggest that uniform weights outperform TF-IDF weights for our application, indicating that a firm's use of a given word to describe its products is more important than how frequently the word is used.

firms to have different sets of competitors.

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 real numbers in the interval $[0, 1]$. In contrast, the analogous matrix 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 implies that for any two firms A and B in the same industry, a firm C that is in A's industry is also 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).

Our first method, described in Section B below, is analogous to SIC and NAICS classifications and requires both the binary membership transitivity and the fixed location properties to hold. We refer to classifications requiring these two restrictive properties as "Fixed Industry Classifications".

Our second method, described in Section C below, relaxes both properties. We refer to this class of industries as "Text-Based Network Industry Classifications". Both firms and entire industries can move in the product space over time as technologies and product tastes evolve. New firms can appear in the sample, and each firm can have its own distinct set of competitors. Finally, these industries are sufficiently rich to permit within and across industry similarities to be computed.

A. The Sample of 10-Ks and the Business Descriptions

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 2008. Our sample begins in 1997 as this is when electronic filing with Edgar first became required. We link 10-K data from Edgar to Compustat using the unique SEC firm identifier, the central index key, and the Compustat gvkey. Among firm-year observations with fiscal years ending in 1997 to 2008 that are present in both CRSP and COMPUSTAT (domestic firms traded on either NYSE, AMEX, or NASDAQ), we have 10-K coverage for 97.9% of the CRSP/COMPUSTAT sample.[12] We can also report that our database is well balanced over time, as we capture between 97.4% and 98.3% in all years of our primary sample from 1997 to 2008. Because database selection can be determined using ex-ante information (ie, the 10-K itself), we do not believe that our data requirements indicate any bias.

[12] We thank the Wharton Research Data Service (WRDS) for providing us with an expanded historical mapping from the firm-level central index key (used by the SEC) to the COMPUSTAT gvkey. We also compute similarities for 1996 (93.5% coverage, electronic filing was optional), but only use the 1996 data to compute the starting value of lagged variables. Also, although we use data for fiscal year endings through 2008, we extract documents filed through December 2009, as many of the filings in 2009 are associated with fiscal years ending in 2008. This is because 10-Ks are generally filed during the 3 month window after the fiscal year ends.

Our full sample of 10-Ks from 1997 to 2008 is comprised of 68,302 observations, which declines to 63,875 when we exclude firms without valid Compustat data or firms with non-positive sales, or firms with assets of less than $1 million. This declines further to 50,673 if we additionally require one year of lagged Compustat data and exclude financial firms (SIC codes in the range 6000-6999).

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 central index key. This latter process is extensively supported by human intervention when non-standard document formats are encountered. This method is reliable and we encountered only a 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.

As described earlier, we then parse the words in the business description and exclude common words and words that are not nouns or proper nouns. Using the resulting word vectors for each firm, we then form the matrix of pairwise similarity scores for the firms in our sample in each year (M_t) as described in the last section.

B. Fixed Industry Classifications Based on 10-Ks

To maintain consistency with other fixed classifications such as SIC and NAICS, we form 10-K based fixed 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 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 our fixed industry classification and our more general textual network industries.

We provide a detailed description of the text clustering algorithm used to create our fixed industry 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 positive feature of the 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 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 331 industries, respectively, in our sample. Although the algorithm's flexibility to pre-specify the number of industries is a positive feature, it 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 to 400 industries best explain firm-level data.

[**Insert Figure 2 Here**]

Figure 2 displays a histogram showing the distribution of the number of firms in each industry for 10K-based 300 industries, SIC-3, and NAICS-4 industries. 10K-based 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, given some firms have highly unique descriptions. Second, 10-K classifications have more very large industries and are more spread out.

Industry memberships have roughly one half to two-thirds overlap. 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-based industry is a more modest 46.2%. In contrast, when two firms are in the same 10K-based 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-based industries are quite distinct from both NAICS-4 than SIC-3.

C. 10-K Based Textual Network Industry Classifications

We next relax the fixed location and transitivity requirements and construct text-based network industry classifications. We use a simple minimum similarity threshold, and define each firm i's industry to include all firms j with pairwise cosine similarities relative to i above a pre-specified minimum 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 a "membership 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.[13] In practice, however, only a small fraction of pairs are actually membership pairs. Although one can use any minimum similarity threshold to construct a classification, we focus on thresholds generating industries with the same fraction of membership pairs as SIC-3 industries, allowing us to compare our industries to SIC-3 in an unbiased fashion.

For three-digit SIC codes, 2.05% of all possible firm pairs are membership pairs. A 21.32% minimum similarity threshold (where we define firms i and j as being in the same industry when $100 \cdot V_i \cdot V_j > 21.32$) generates 10-K based industries with 2.05% membership pairs, which is the same as SIC-3. We consider one further refinement to

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

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 achieve this by subtracting these median scores from the raw scores to obtain our final scores used for each firm.[14]

Note that the transitivity property does not hold for these firm-centric 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 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.

We also take into account vertical relatedness 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, 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 our network industries regardless of the similarity score. Because just 4% of all pairs are excluded using this screen, and because our results are robust to including or excluding this screen, we conclude that firm business descriptions in firm 10-Ks indeed focus heavily on horizontal firm product offerings, and not on vertical

[14] Our results are robust, although our external validation tests are slightly weaker if we omit this step.

firm production inputs.

IV. Qualitative Assessment of our New Industry Classifications

In this section we discuss central properties of our "unrestricted" text-based network industry classification and illustrate these properties using sample industries based on the new classifications. We focus on properties of networks that are not captured by traditional fixed industry classifications. The properties are (1) capturing within industry heterogeneity, (2) capturing product and industry change, and (3) capturing cross-industry relatedness.

A. Capturing Within-Industry Heterogeneity

The concept of product differentiation within industries dates back to Chamberlin (1933), who famously showed that differentiation is fundamental to theories of industrial organization and reduces competition between firms. An informative classification should not only identify product markets, but also measure differentiation within industries. Beginning with Berry, Levinsohn, and Pakes (1997), who consider the automobile industry, the approach in 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. Gowrisankaran and Rysman (2012) consider a dynamic setting, and examine the effect of transitory price shocks on one-month and one-year price elasticities. Holmes and Stevens (2004) show that industries are not as homogeneous as standard classifications might suggest, as smaller and larger firms might specialize or exhibit different degrees of differentiation.

These latter studies motivate the need for more refined industry classifications offering the flexibility to assess industries in a dynamic way, or to assess the distribution of rivals using multiple degrees of granularity. Text-based network industries address both issues. The standard single-industry approach used in the literature has been highly informative, especially in understanding the dynamics of industry pricing, competition and substitution in well-defined industries.[15] However, many theories, especially those related to endogenous barriers to entry and multiple industry production, are difficult to test in a single industry setting.

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 publicly traded 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 1 displays sample classifications using our methodology for selected firms in this product area.

[Insert Table 1 Here]

Table 1 shows 6 major sub markets within this broad business services industry: Entertainment, Medical Services, Information Transmission, Software, Corporate

[15] For additional recent examples, see Mazzeo (2002) and Seim (2006).

Data Management and Computing Solutions, and Online Retailing and Publishing. Each sub-industry is the text-based network industry surrounding the focal firm listed in each example's header. Although SIC codes were not used to make these groupings, we report them for illustrative purposes. The SIC codes 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 theme is that many firms address these markets using the internet but they often also compete with rivals having 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 aggregated measures of overall product differentiation surrounding each firm. We examine measures of competition in Section V.

B. Ability to Capture Product and Industry Change

An informative industry classification should also capture changes to industry groupings over time. Firms 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.[16] These theories motivate our exami-

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

nation of advertising and research and development, and their links to future changes in industry membership and competition (see Section VII).

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 the use of mainframes and work stations in the software industry (see Section VI). 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 2 provides examples of two industries that changed dramatically over time.

[**Insert Table 2 Here**]

Panel A of Table 2 displays the text-based network 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.

Our ability to assess product market change is also important in our examination of exogenous shocks to the military and software industries. In Section VI, we examine not only changes in the degree of competition and in industry membership, but also changes in the type and features of products offered following the shock.

C. Ability to Capture Cross-Industry Relatedness

An informative industry classification should also be able to capture across-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 facing this form of cross industry threat might keep prices low to deter entry. Currently, existing research examines across-industry relatedness using coarser levels of SIC or NAICS codes or through the Bureau of Economic Analysis's input-output matrix, which is used to measure vertical relations.

Our methodology uncovers numerous links entirely missed using other classifications. For example, text-based network similarities reveal that firms in the newspaper publishing and printing industry (SIC-3 is 271) are highly similar to firms in the radio and broadcasting stations industry (SIC-3 is 483). This example illustrates the fact that both industries likely cater to the same customers, who demand advertising, and thus exhibit at least some substitution. A different firm-specific example is the merger in 2006 of Disney and Pixar. Disney was classified in the business services industry (SIC-3 is 737), while Pixar was classified in the motion pictures industry (SIC-3 is 781). Movies are produced by both but using different production methods, explaining in part why they are in different SIC codes. Our methodology indicates that they are similar and are in the same text-based network industry.

These examples are interesting given that these firms have SIC codes that disagree even at the one or two-digit level, suggesting that traditional classifications treat these industries as entirely unrelated. Because our classifications are based on actual product text, we are able to detect potential rival firms that offer related products even if they are not currently direct rivals.

Hoberg and Phillips (2013) is an example of another recent study that explores cross-industry relations using language overlaps based on 10-K relatedness scores. The study examines why conglomerates operate in some industry combinations more frequently than others, and finds that they are most likely to operate in industry pairs that are close together in the product space, and that spatially surround other highly valued industries. These findings are consistent with conglomerate firms using industry relatedness to potentially enter nearby high value industries that might otherwise be costly to enter.

V. External Validation

Our next objective is to compare our industry classifications to existing SIC and NAICS industry classifications. Throughout, we hold fixed the degree of granularity of the industries we examine, and we compare classifications on their ability to generate higher across-industry variation in key firm characteristics (Subsection A). We then examine which classification best explains managerial discussions of high competition and firm disclosures of self-identified rivals (Subsection B).

A. Across-Industry Variation

We compare the informativeness of industry classifications based on the extent to which they generate higher levels of across-industry variation in profitability, sales growth and stock market risk (market betas). Because we hold fixed the degree of granularity in the classifications we compare, we conclude that a classification generating a higher degree of across industry variation is more informative.

To ensure consistency in our tests, we compute the degree of across-industry variation using both a firm-weighted approach and an industry weighted approach. For the firm weighted approach, we first compute the given firm's industry value of a given characteristic as the mean of the given characteristic among its industry peers. We then compute across industry variation as the standard deviation of these industry characteristics across all firm-year observations in our sample. The firm-weighted calculation is particularly relevant for the intransitive text-based network classification, as each firm has a unique set of peers. The firm-weighted approach thus allows us to compare text-based network industries to other classifications including 10-K fixed industry classifications, three-digit SIC, and four-digit NAICS industries.

Industry-weighted estimates of across-industry variation are computed, for each industry-year observation, by first computing the average characteristic among all firms in the given industry in the given year. Industry-weighted variation of a given characteristic is then the standard deviation of these industry characteristics across all industry-year observations in our sample. We note that this calculation is only feasible for industry classifications that satisfy the industry transitivity property, as intransitive industries are firm-specific and observations exist at the firm-year level.

[**Insert Table 3 Here**]

Panel A of Table 3 displays the results of the firm-weighted calculations. The table shows that 10-K-based fixed industries have more across industry variation than SIC-3 or NAICS-4 industries for all five characteristics we consider. Regarding profitability defined as operating income/sales (oi/sales), across industry variation is 0.204 and 0.205 for SIC-3 and NAICS-4, respectively. This increases 12.7% to 0.231 for 10-K-based fixed industries. We observe similar gains regarding oi/sales, sales growth, market beta, and the unlevered asset beta. We conclude that 10-K-based fixed industries are more informative than both SIC-3 and NAICS-4. It is also important to note that the 10-K-based fixed classification is calibrated to have the same granularity as SIC-3, and both classifications are also constrained to have fixed locations and to be transitive. These gains reveal the informativeness of the information source (text in the 10-K) and not the network methods.

Panel A also compares the transitive classifications (SIC-3, NAICS-4, and 10K-based 300 industries) to the intransitive text-based network classification. Because transitivity limits the flexibility of a network, we expect the intransitive network classification to have higher across-industry variation. The table confirms this prediction. Regarding oi/sales, the across industry variation of 0.231 for 10-K-based fixed industries increases by another 7.4% to 0.248 for the equal weighted text-based network classification, and by 15.6% to 0.267 for the similarity weighted network classification. We observe similar, and in some cases larger, gains for the other characteristics. For example, we see a nearly 30% increase for asset betas when comparing 10K-based fixed industries and similarity weighted network industries. We conclude overall that 10-K based classifications are more informative than SIC and NAICS classifications, and also that the intransitive text-based network classification is even more informative when compared to all three fixed industry classifications.

Panel B of Table 3 repeats the tests in Panel A using industry weighted calculations rather than firm weighted calculations. Althought this calculation is only feasible for transitive classifications, the conclusions are the same as those in Panel A. The 10-K-based fixed classification is more informative than the SIC-3 and NAICS-4 classifications. The improvements are also economically large. For oi/sales, the gain is nearly 30% relative to SIC-3, and nearly 20% relative to NAICS-4. The results in Panel B thus ensure that the results in Panel A are not highly influenced by firm versus industry weighting.

B. Competition and Reported Peers

In this section, we examine whether firms with more text-based network rivals, and more similar such rivals, are more likely to disclose competitive pressures in the Management's Discussion and Analysis section of their 10-K.

Our approach follows Hoberg and Maksimovic (2014) 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.[17]

Table 4 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. Our primary independent variables are measures of the total similarity surrounding each firm. Total similarity is a global measure and is the sum of the pairwise similarities between the given firm and all other firms in our sample in the given year. We predict that a manager of a firm with higher total similarity will be more likely to disclose discussions noting higher levels of competition in his or her firm's Management's Discussion and Analysis section of its 10-K.

We also decompose total similarity into respective components based on how far given rivals are from the focal firm in the product market space. Our first measure, Total Similarity (Top 2% peers), only considers total similarity summed over the set of firm pairs with pairwise similarities above the 98th percentile in the given year. This measures the degree to which the firm has "close rivals": those with a degree of similarity analogous to the SIC-3 classification (where roughly 2% of all firm pairs are in the same SIC-3 classification). Overall, we consider summed similarity in the following bands: 0.0-2.0%, 2.0% to 5.0%, 5.0% to 10.0%, and 10.0% to 25.0%. This

[17] 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.

test is made possible by the fact that the text-based network provides a pairwise similarity score for every firm pair in our sample, and the objective is to identify the extent to which competitive pressures are felt by managers even when firms are more distant in the product space. Evidence of competitive pressures from more distant peers would support the conclusion that issues relating to potential entry threats, and issues relating to incentives to pursue limit pricing, likely apply in the universe of US publicly traded firms.

[**Insert Table 4 Here**]

Table 4 shows the results of this test, where the dependent variable is an indicator variable indicating that the manager discusses competitive pressures. Row (1) shows that firms with higher global total similarity are far more likely to discuss competitive pressures in their management's discussion. This result suggests that information in the text-based network classification is informative regarding the presence of firms that managers themselves perceive to actually be rivals. In particular, these rivals pose competitive threats that managers feel obliged to mention when interpreting their firm's performance and future prospects.

Rows (2) to (6) further show that these competitive pressures extend beyond the range typically studied by researchers. In particular, many studies control for three-digit SIC industries, which has a granularity of roughly 2%. Rows (3) and (6) show that managers facing more distant rivals in the 2.0% to 5.0% band, on the margin, also are more likely to disclose issues relating to higher competition in their 10-Ks. Because all independent variables are standardized, we can also compare magnitudes. The results suggest that competition in the nearest band is roughly 50% more important than competition in the second band, however both bands are highly significant. These findings are consistent with the conclusion that managers react not only to current rivals in their immediate markets, but also to rivals in neighboring markets that might pose potential entry threats.

In rows (7) to (12) of Table 4, 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 and Analysis section. In all, we find that our primary total similarity variables weaken somewhat as the new controls are added. However, all variables remain highly significant and our primary conclusions obtain with or without the controls.

We also consider the approach used by Rauh and Sufi (2012), who gather data from Capital IQ identifying the firms listed by each firm in its 10-K as being a rival. We note one important limitation in this analysis is that Capital IQ data is not available on a historic basis. We thus extract Capital IQ peers using 2011 data, and we examine whether text-based network industries computed using the last year of our data (2008) can better explain the Capital IQ peers relative to SIC-3 or NAICS-4 industries. We display the results in Table 5. The table reveals that our network industries outperform SIC-3 and NAICS-4 industries in their ability to explain Capital IQ self-reported peers. For example, or our baseline approach that uses nouns and proper nouns and a 25% stop word threshold generates 52.5% overlap with Capital

IQ peers. This compares to 47.1% for SIC-3. Our peers perform even better relative to NAICS-4: 55.1% overlap compared to 44.0%. Given that our comparisons hold industry granularity constant, we conclude that these results are driven by economic informativeness and are not due to technical differences in granularity.

VI. Capturing Industry Change

In this section we assess changes in firm similarity and product market location following two major exogenous industry shocks. We examine the impact of these shocks on the total similarity of rivals around a given firm, the number of industry rivals, and the extent of product market dislocation (changes in the products offered). We first consider the military goods and services industry following the shock to military spending after the events of September 11, 2001. Our second analysis is based on the software industry following the technology stock market collapse (dot com shock) beginning in March 2000.

We choose these two shocks due to their importance in shaping the trajectory of large numbers of firms in these industries, and also because these examples best illustrate unique features of text-based network industries that make them well-suited to address research questions involving time-varying changes that are difficult to analyze using other industry classifications. Because one shock is a positive demand shock (defense) and the other is a negative demand shock (software), these choices also allow us to explore if product markets react differently to positive and negative shocks.

A. Military Intelligence and Battlefield Products

Figure 3 displays the scale and timing of the positive military spending demand shock. Prior to 2001, U.S. military spending gradually declined from nearly 6% of GDP to just 3% of GDP. After 2001, this trend sharply reversed and spending increased to just below 5% of GDP by 2009. Although much of the observed increase from 3% to 5% is gradual, we expect that industry participants knew at the time of the September 11th attacks that future spending would grow and the growth would be large and long-lasting. We focus on how the product market changed before and after 2001. We also note that this shock is large magnitude (1-2% of annual U.S. GDP) and it is also exogenous, as market participants did not expect the 2001 attacks. This shock presents a nearly ideal laboratory for examining the impact of a large demand shock on product market structure and dislocation.

Product market research assessing the impact of September 11th attacks on the military industry is difficult using existing classifications for three reasons. First, military products span many different SIC and NAICS-based industries including heavy equipment, transportation, electronics, software, security, and medical equipment. Second, only a fraction of the firms in each market actually sells to the military. A manual review of all firms would be necessary to identify the relevant set engaged in selling to the military.

Third, existing industry classifications are not updated materially over time. A researcher would have to manually assess how firms supplying to the military changed over time. For example, some firms might begin or terminate sales to the military. Also, updated product offering data is not available using other classifications, making the issue of product market dislocation very difficult to study. Because text-based network industries are based on the actual vocabulary used in annual firm 10-Ks, we are able to solve these problems directly.

We use the series of matrices Q_t containing the product words (defined in Section II) to identify the set of firms likely providing goods and services to the military. For a given year t, Q_t is the set of normalized product word vectors for all firms in the sample in that year. Each firm's product word vector has length W (the number of unique 10-K words), and each element identifies whether or not the given firm uses the specific word corresponding to that element.

We consider two queries on Q_t. We first identify the set of "Military Intelligence Firms" as the set of all firms that use all three of the following terms in their business descriptions in year t: "Military", "Defense", and "Intelligence". In our second test, we identify the set of "Military Battlefield Firms" as the set of firms using all of the following four terms in year t: "Military", "Defense", "Battlefield" and "Equipment". We consider both to explore potential differential treatment effects in these related markets. Because Q_t is updated and comprehensive, these queries immediately yield the set of firms addressing each market in each year.

Our hypotheses are that (A) the 9/11 attacks generate a positive demand shock in these markets and we expect an increase in entry and total similarity following the shock and (B) we expect a shift in product location to reflect changing needs of the military after this shock. To test the first hypothesis, we compute the following averages over firms in each product market in time series: the average total text-based network similarity surrounding each firm and the average number of network rivals faced by each firm. We assess each statistic in time-series, and we test the first hypothesis by examining if each variable experiences a structural break after 2001.

To test the second hypothesis, we use the information in Q_t in a different way. For each word w_k corresponding to the W columns in Q_t, we compute the extent to which w_k has grown in usage in the given market after the shock. Words experiencing the highest positive change indicate product words that increased most in use after the shock. Words receiving the most negative change indicate product words that experienced the most severe decline. Assessment of these word change vocabularies illustrates the extent to which product markets experienced a location shift after the September 11th attacks.

[**Insert Table 6 and Table 7 Here**]

Tables 6 and Table 7 display results for “Military Intelligence Firms” and “Military Battlefield Firms”, respectively. Panel A in both tables reports the time series statistics describing the number of firms in each product market, the degree of total similarity, and the average number of text-based network rivals faced by firms in each market. The footer of each table reports the results of tests examining whether each industry experienced a structural break in total similarity and the number of rivals after 2001.

In Table 6, a structural break test on the level of total similarity is significant at the 1% level. A similar test regarding the average number of rivals is just below standard levels of statistical significance. In Table 7, both structural break tests are significant at the 5% level. We conclude that the 2001 shock resulted in an increase in total similarity, indicating increased competition. The positive demand shock likely triggered entry (or movement by existing firms) into newly demanded military intelligence and battlefield products. We conclude that industry participants likely benefited from the positive demand shock. However, these gains were likely offset, at least in part, by increased competition.

Panels B and C display the results of the product market location change tests. Panel B reports the words that most increased in usage after the shock (2002 versus 2000). For Military Intelligence firms in Table 6, we find an increased focus on surveillance (47.2% of firms use the term in 2002, compared to just 22.5% in 2000). This is consistent with information gathering being a central theme for the military after the shock. We also observe an increased focus on disasters, email, integrators, and the specific firm Northrop Grumman. These results also support an increased focus on information gathering, especially the gathering of non-battlefield intelligence. The results for Military Battlefield Firms in Table 7 are similar, and we also see an increased focus on surveillance, optics, simulation, artillery, infrastructure, and cleanup. These results support the conclusion that military battlefield firms became more focused on possible ground conflicts, and that simulations, infrastructure and artillery likely played a central role.

Panel C reports words that became less utilized after the shock. In Table 6, key terms include broadcast, microwave, microelectronics, transmitters, and the specific firm Litton. These results indicate a declining focus on information transmission and electronic devices. Put together with Panel B, we find a likely dislocation away from information transmission and towards non-battlefield information gathering technologies. For Military Battlefield Firms in Table 7, the following terms declined in usage: subassemblies, broadcast, radar, algorithms, airframes, and the specific firms Motorola and Lockheed. Together with Panel B, these results suggest a shift away from traditional air power and communication, and toward a focus on potential ground conflicts.

Overall, we conclude that the September 11th shock resulted in two important changes for both Military industries: an increase in entry into new high demand markets, and product changes towards non-battlefield information gathering and products intended for potential ground conflicts. Because the shock was exogenous, it is likely that the shock caused these changes.

B. Software Industry

We next consider the post-2000 software market collapse. This shock is also large in size, and is also difficult to examine using traditional industry classifications (software firms are typically lumped with many other firms into a "business services" SIC code, and this market also experienced rapid change over time). It is also important to note that this is a negative demand shock, whereas the military shock was a positive shock. Our empirical approach is identical to that used for the Military industry above. We define the software market broadly to include firms with all three of the following terms in their business descriptions in year t: “Software”, “Computer”, and “Program”.

[Insert Table 8 Here]

Panel A of Table 8 shows a decline in the total similarity surrounding each firm and in the average number of text-based network rivals in the software industry. The decline is large in magnitude and long-lasting. This suggests that the negative shock led to firm exit, or movement by firms into other product markets. We conclude that this negative demand shock created a symmetric-opposite result relative to the positive shock in the military industry. However, we also note that the process of consolidation following the negative shock is more gradual and slow moving relative to the re-alignment following the positive shocks in the military industry.

Panels B and C of Table 8 show that the software industry also experienced product market dislocation after the shock. Panel B shows a movement toward a focus on Email, intelligence, Linux, portals, and lifecycles. Panel C shows a dislocation away from Unix, mainframes, intranet, telemarketing and workstations.

Our study suggests that major exogenous demand shocks impact industries in ways not well-documented in the existing literature. We observe large changes in market structure and product similarity, and also product market dislocation, indicating that industry boundaries are undergoing large changes over time.

VII. Endogenous Barriers to Entry

In this section, we examine how measures of industry similarity and profitability change over time following 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 firms that do not advertise but rather 'sell on price'.

The main idea is that R&D and advertising can create unique products that appeal to quality-sensitive consumers, making 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 similarity. We test this assumption by regressing changes in ex-post similarity and profitability on ex-ante advertising and R&D levels. 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.

We focus on text-based network industries for this test, as observing changes in industry memberships and industry locations is critical to testing Sutton's theory, which is primarily about trying to solidify industry boundaries and prevent entry across industry boundaries. Fixed industry classifications lack this flexibility because their industry locations are fixed over time, and memberships rarely change.

Table 9 displays the results. We consider text-based network industry results in Panel A. One observation is one firm in one year. The dependent variable for each row is noted in the first column. We consider three dependent variables: the change in ex-post total similarity, the change in the number of text-based network rivals, and the change in average profitability of firms in the focal firm's network industry. The independent variables include dummies for industries having zero advertising and R&D expenditures. We also include the natural logarithm of the average R&D/sales and advertising/sales for industries that have non-zero spending. This allows us to examine the effects of having R&D and advertising programs, and also the effect of intensity in each category. Finally, we include controls for size, stock returns, and the book to market ratio. In Panel B, we repeat these same tests using firm-level advertising and R&D.

[Insert Table 9 Here]

Table 9 shows strong support for Sutton's predictions using both text-based network industry level (Panel A) and firm-level (Panel B) R&D and advertising. For example, row (1) shows that firms in network industries with non-zero R&D and advertising, and also those with more intense spending, experience ex-post reductions in their total similarity. We observe similar results in row (2) regarding the number of rivals, where results are particularly strong for advertising but are not significant for

R&D. Examining ex-post profitability in row (3), we find support for the conclusion that both advertising and R&D are associated with increasing ex-post profitability. The results are similar in Panel B, suggesting that these effects hold at both the firm and the industry level.

To understand the intuition for these results, we conduct a detailed analysis of one industry: the online education industry. We define this industry as the set of text-based network peer firms surrounding the Apollo Group, which runs the University of Phoenix, a large provider of online education programs.

Panel C of Table 9 displays time series statistics for the network industry surrounding Apollo Group during our sample period. We chose this example because the improvements in (and the declining costs of) technology likely reduced the natural barriers to entry in this industry. We thus predict that firms providing online education would have strong incentives to invest in advertising to reinforce the eroding natural barriers to entry. We expect a movement toward niche markets established and defended by higher advertising.

Panel C supports this intuition. The last two columns, which display the average advertising to sales ratio for firms in Apollo's network industry, and the total advertising to total sales ratio for these firms in aggregate, confirm that advertising among Apollo's industry rivals increased dramatically during our sample period. Consistent with a link to barriers to entry, we also observe declines both in the average number of network rivals and the average total similarity among firms in this market as advertising ramped up. We conclude that Apollo and its rivals likely had high incentives to spend on advertising as natural barriers to entry eroded, and our evidence suggests that this strategy was successful.

VIII. Conclusions

We use web crawling and text parsing algorithms to build new measures of firm pairwise product similarity based on business descriptions from 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 vectors to calculate how firms are related to each other and to create new industry classifications. These new classifications enable us to assess product market changes, the impact of major shocks, and 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 the product description section of their 10-Ks. Because our classifications are updated in each year, they do not have the time-fixed location restrictions associated with SIC and NAICS. In addition, our main classification method is based on relaxing the transitivity restriction of existing SIC and NAICS industries, and thus allows each firm to have its own potentially unique set of competitors. This new method, which we term text-based network industry classifications (TNIC), is analogous to a social network where each individual can have a distinct set of friends, or to geographic networks where the distance from a firm determines whether or not it is a competitor.

Measures of similarity in industry groups based on our new classifications better explain specific discussions of high competition by management and better explain rivals mentioned by managers as peer firms than do existing classifications. Our new industry classifications also allow us to examine how industry membership and rival similarity change over time in response to exogenous industry shocks, and whether advertising and research and development serve as endogenous barriers to entry. We find that major demand shocks in the military and software product markets are followed by economically large changes in the number of similar firms, the degree of product differentiation, and the type of products firms offer to the market. We also find support for Sutton (1991)'s hypothesis that firms spend on advertising and R&D, at least in part, to increase ex-post 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." *J. Finance* 52: 1259–1294.

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." *J. 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 J. Econ.* 38: 43–59.

Fama, Eugene, and Kenneth French. 1997. "Industry Costs of Equity." *J. Financial Econ.* 43: 153–193.

Fan, Joseph, and Vidhan Goyal. 2006. "On the Patterns and Wealth Effects of Vertical Mergers." *J. Business* 79: 877–902.

Gentzkow, Matthew, and Jesse Shapiro. 2010. "What Drives Media Slant? Evidence from U.S. Daily Newspapers." *Econometrica* 78, 35–71.

Gowrisankaran, Gautam, and Marc Rysman. 2012. "Dynamics of Consumer Demand for New Durable Goods." *J.P.E.* 120: 1173–1219.

Hanley, Kathleen, and Gerard Hoberg. 2010. "The Information Content of IPO Prospectuses." *Rev. Financial Studies* 23: 2821–2864.

Hay, D.A. 1976. "Sequential Entry and Entry-Deterring Strategies in Spatial Competition." *Oxford Econ. Papers* 28: 240–257.

Hoberg, Gerard, and Vojislav Maksimovic. 2014. "Redefining Financial Constraints: A Text-Based Analysis." *Rev. Financial Studies* Forthcoming.

Hoberg, Gerard, and Gordon Phillips. 2010. "Competition and Product Market Synergies in Mergers and Acquisitions: A Text-Based Analysis." *Rev. Financial Studies* 23: 3773–3811.

——— . 2013. "Industry Choice and Product Language." Working Paper, University of Southern California.

Holmes, Thomas, and John Stevens. 2004. "Spatial Distribution of Economic Activities in North America." *Handbook of Urban and Regional Economics.*

Hotelling, H. 1929. "Stability in Competition." *Econ. J.* 41–57.

Jaffe, Adam. 1986. "Technological Opportunities and Spillovers of R&D: Evidence from Firms' Patents, Profits and Market Value." *A.E.R.* 76: 984–1001.

Kahle, Kathleen, and Ralph Walkling. 1996. "The impact of Industry Classifications on Financial Research." *J. 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.

Lin, Ping, and Kamal Saggi. 2002. "Product Differentiation, Process R&D, and the Nature of Market Competition." *European Econ. Rev.* 46: 201–211.

Loughran, Tim, and Bill McDonald. 2011. "When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *J. Finance* 66: 35–65.

Mazzeo, Michael. 2002. "An Empirical Model of Firm Entry with Endogenous Product Choices." *Rand J. Econ.* 33: 221–42.

Nevo, Aviv. 2000. "Mergers with Differentiated Products: The Case of the Ready to Eat Cereal Industry." *Rand J. Econ.* 31: 395–421.

Panzar, J., and R. Willig. 1981. "Economies of Scope." *A.E.R.* 71: 268–272.

Rauh, Joshua, and Amir Sufi. 2012. "Explaining Corporate Capital Structure: Product Markets, Leases, and Asset Similarity." *Rev. Finance* 16: 115–155.

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 J. Econ.* 37: 619–40.

Shaked, Avner, and John Sutton. 1987. "Product Differentiation and Industrial Structure." *J. Industrial Econ.* 26: 131–146.

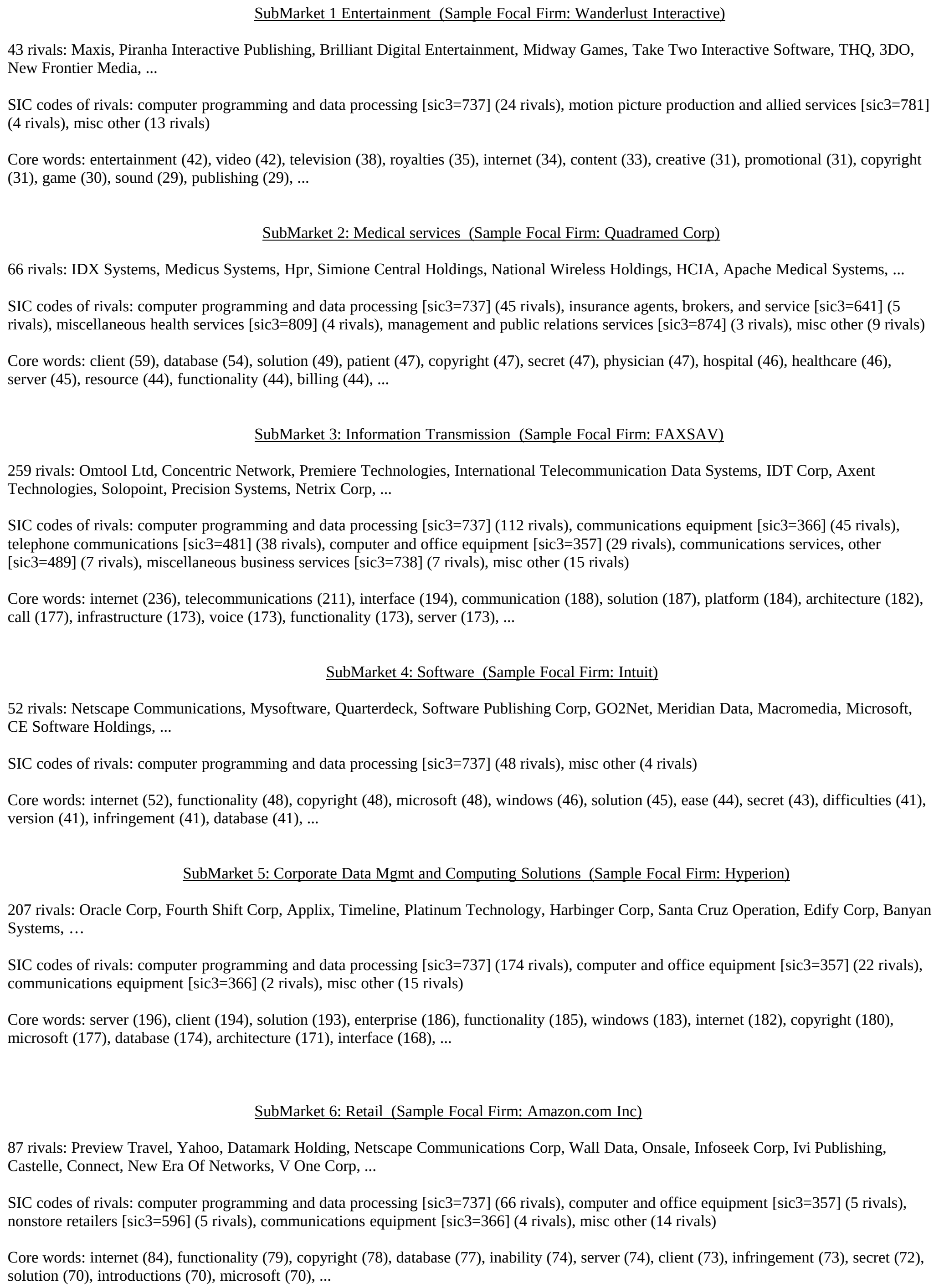
Sutton, John. 1991. *"Sunk Costs and Market Structure"* (MIT Press: Cambridge, Mass).

Tetlock, Paul, Maytal Saar-Tsechanksy, and Sofus Macskassy. 2008. "More Than Words: Quantifying Language to Measure Firms' Fundamentals." *J. Finance* 63: 1437–1467.

Tetlock, Paul C. 2007. "Giving Content to Investor Sentiment: The Role of Media in the Stock Market." *J. Finance* 62: 1139–1168.

TABLE 1

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, ...

SIC codes of rivals: computer programming and data processing [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), ...

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 and data processing [sic3=737] (45 rivals), insurance agents, brokers, and service [sic3=641] (5 rivals), miscellaneous health services [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), ...

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 and data processing [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, other [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), ...

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 and data processing [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), ...

SubMarket 5: Corporate Data Mgmt and Computing Solutions (Sample Focal Firm: Hyperion)

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 and data processing [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), ...

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, ...

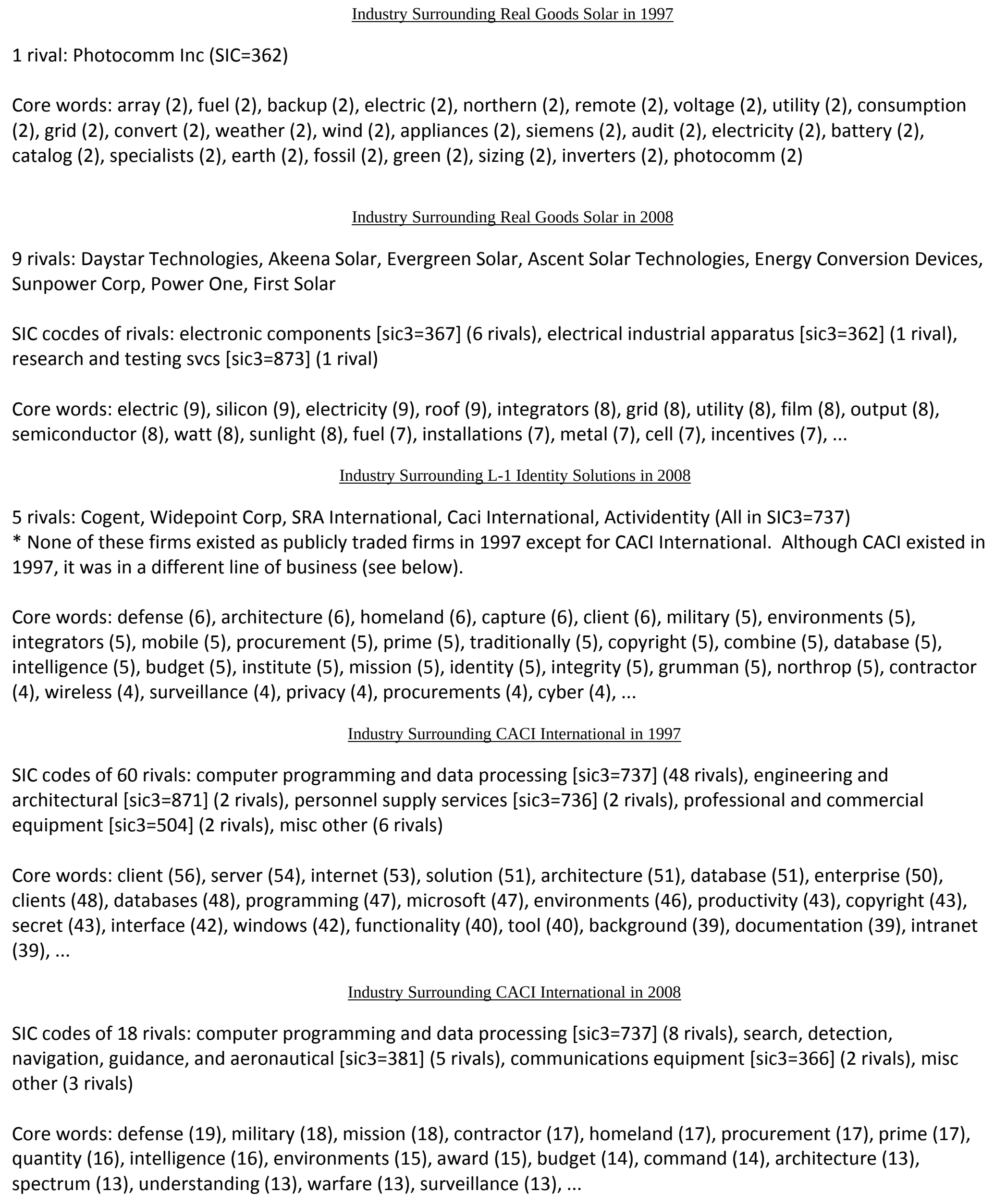
SIC codes of rivals: computer programming and data processing [sic3=737] (66 rivals), computer and office equipment [sic3=357] (5 rivals), nonstore retailers [sic3=596] (5 rivals), communications equipment [sic3=366] (4 rivals), misc other (14 rivals)

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), ...

Sample text-based network industries centered around firms residing in three-digit SIC code 737 in the year 1997.

TABLE 2

Sample Industries that Underwent Changes (text-based network Classifications)



Industry Surrounding Real Goods Solar in 1997

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

9 rivals: Daystar Technologies, Akeena Solar, Evergreen Solar, Ascent Solar Technologies, Energy Conversion Devices, Sunpower Corp, Power One, First Solar

SIC cocdes 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), ...

Industry Surrounding L-1 Identity Solutions in 2008

5 rivals: Cogent, 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), ...

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), ...

Sample text-based network industries that changed dramatically between 1997 and 2008.

TABLE 3
Firm Characteristics and Industry Classifications

Row	Industry Controls	OI/Sales	OI/Assets	Sales Growth	Market Beta	Asset Beta
Panel A: Across Industry Standard Deviations (Firm-Weighted Results) (All Industry Classifications)						
(1)	SIC-3 Fixed Effects	0.204	0.111	0.126	0.283	0.271
(2)	NAICS-4 Fixed Effects	0.205	0.112	0.136	0.289	0.276
(3)	10K-based 300 Fixed Effects	0.231	0.128	0.157	0.298	0.285
(4)	TNIC Equal Weighted Average	0.248	0.142	0.163	0.332	0.324
(5)	TNIC Similarity Weighted Avg (Ex Self)	0.267	0.153	0.199	0.384	0.369
Panel B: Across Industry Standard Deviations (Industry-Weighted Results) (Transitive Industry Classifications Only)						
(1)	SIC-3 Fixed Effects	0.156	0.111	0.179	0.347	0.308
(2)	NAICS-4 Fixed Effects	0.169	0.126	0.210	0.414	0.362
(3)	10K-based 300 Fixed Effects	0.202	0.139	0.224	0.469	0.432

For a given variable indicated in the first column, across industry standard deviations are computed as the standard deviation of the industry average of the given variable across all firms in our sample (Panel A) and across all industries (Panel B). TNIC refers to text-based network industries.

TABLE 4
Managerial Indications of High Competition and Industry Similarity Measures

Row	Total Similarity (TSimm)	TSimm (Top 2% Peers)	TSimm (Top 3% to 5% Peers)	TSimm (Top 6% to 10% Peers)	TSimm (Top 11% to 25% Peers)	OI/ Assets	Log Firm Age	Tobin's Q	Log Sales	# Words Bus. Desc.	# Words MD&A	# Obs./ R^2
(1)	0.527											41,823
	(16.37)											0.047
(2)		0.387										41,823
		(12.41)										0.039
(3)			0.244									41,823
			(8.82)									0.030
(4)				0.022								41,823
				(0.81)								0.024
(5)					-0.344							41,823
					(-11.04)							0.032
(6)		0.347	0.234	0.001	-0.237							41,823
		(10.90)	(8.09)	(0.04)	(-6.64)							0.049
(7)	0.400					0.038	-0.061	0.061	-0.211	-0.360	1.111	41,823
	(10.63)					(1.52)	(-2.06)	(3.03)	(-5.51)	(-11.01)	(30.92)	0.148
(8)		0.286				0.026	-0.060	0.079	-0.255	-0.343	1.136	41,823
		(8.54)				(1.08)	(-2.04)	(3.95)	(-6.75)	(-10.76)	(31.40)	0.145
(9)			0.142			0.015	-0.068	0.095	-0.233	-0.268	1.148	41,823
			(4.76)			(0.61)	(-2.35)	(4.78)	(-6.14)	(-8.80)	(31.65)	0.141
(10)				0.024		0.010	-0.066	0.107	-0.256	-0.268	1.164	41,823
				(0.81)		(0.40)	(-2.28)	(5.41)	(-6.74)	(-8.85)	(31.96)	0.140
(11)					-0.252	0.016	-0.064	0.088	-0.237	-0.281	1.151	41,823
					(-7.68)	(0.66)	(-2.19)	(4.47)	(-6.24)	(-9.06)	(31.71)	0.143
(12)		0.253	0.139	-0.036	-0.212	0.036	-0.060	0.054	-0.218	-0.345	1.111	41,823
		(7.52)	(4.51)	(-1.05)	(-5.42)	(1.46)	(-2.02)	(2.69)	(-5.68)	(-10.63)	(30.83)	0.149

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's Discussion and Analysis section of its 10-K in the given year. Independent variables include measures of similarity using the text-based network pairwise network. In addition to the total similarity of all peers (TSimm), we also consider the total similarity based on stepwise groups of the most similar peers (Top 2%, 5%, 10%, and 25%). Each localized peer group is orthogonalized in a stepwise fashion (from 2% to 25%). Additional control variables include sales, age, profitability, Tobin's Q, Year fixed effects, and document size variables including the number of words in the business description and in the Management's Discussion and Analysis (MD&A) section of the firm's 10-K. All independent variables are standardized for ease of comparison. (*t*-statistics in parentheses are based on standard errors adjusted for clustering by firm).

TABLE 5
Self Reported Capital IQ Peers and Industry Classifications

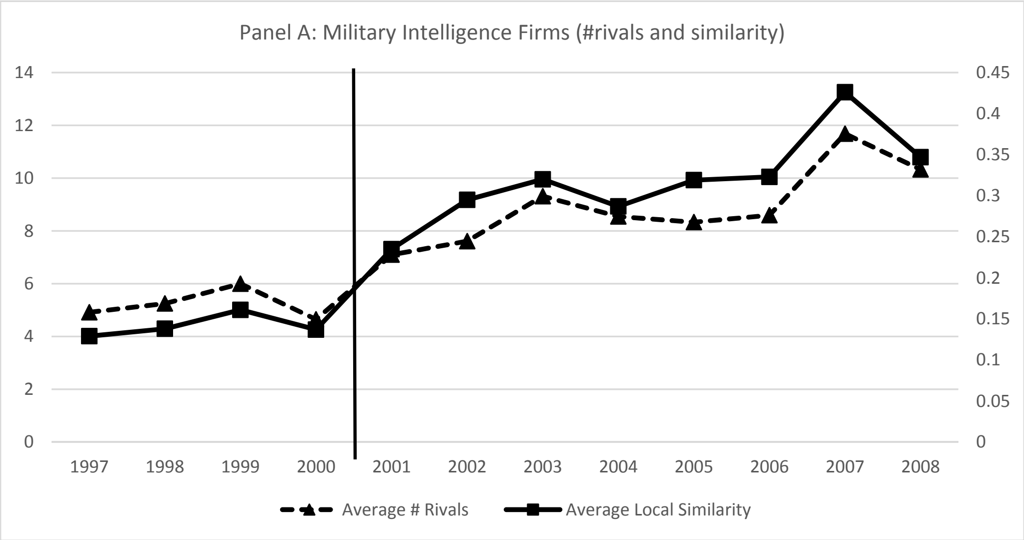
		Text-based industries (SIC-3 Granularity)		***Text-based industries (NAICS-4 Granularity)***	
Words Used for TNIC Industry	Stop Word Threshold	TNIC Overlap with Cap IQ	TNIC Overlap with SIC-3	TNIC Overlap with Cap IQ	TNIC Overlap with NAICS-4
		Panel A: Capital IQ Competitors			
All Words	100%	40.9%	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%.*

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 Text-based network classifications (TNIC) 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 6

Military Intelligence Firms: Competitor Changes


Panel A: Military Intelligence Firms (#rivals and similarity)

Panel B: Words that became more prominent after the shock (2000 vs 2002)

surveillance (22.5%-47.2%), legacy(7.5%-31.9%), priorities(7.5%-29.2%), grumman(15.0%-36.1%), northrop(15.0%-34.7%), transformation(2.5%-19.4%), boeing(20.0%-36.1%), integrators(15.0%-30.6%),disasters(2.5%-18.1%), running(0.0%-15.3%), cancellations(7.5%-22.2%), disaster(2.5%-16.7%), allegations(0.0%-13.9%), cancel(10.0%-23.6%), conflict(5.0%-18.1%), email(2.5%-15.3%), streamline(2.5%-15.3%), task(15.0%-27.8%), thales(0.0%-12.5%), desktop(12.5%-25.0%), visibility(10.0%-22.2%), engagement(7.5%-19.4%), path(7.5%-19.4%), assured(5.0%-16.7%), deficit(2.5%-13.9%)

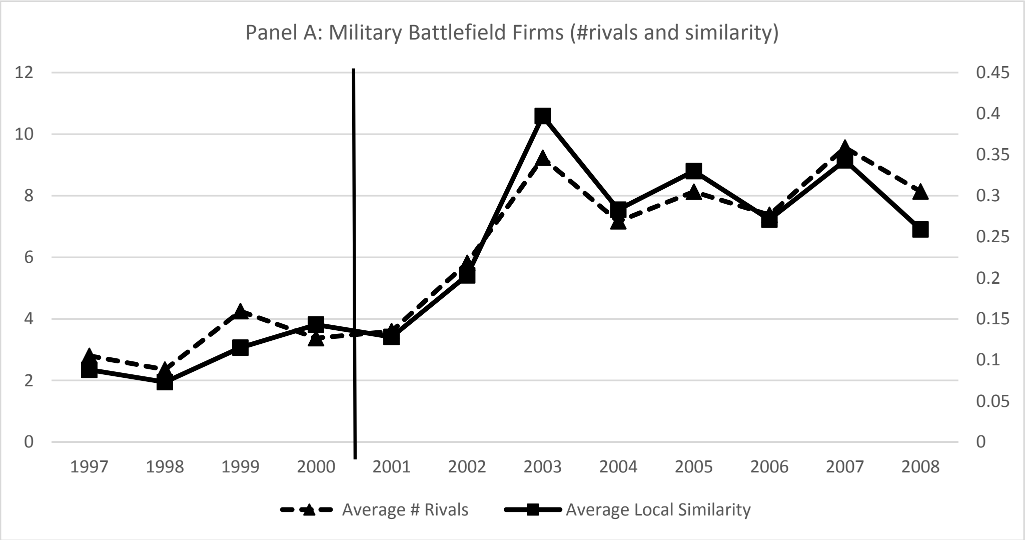
Panel C: Words that became less prominent after the shock (2000 vs 2002)

litton(20.0%-1.4%), band(32.5%-16.7%), microwave(37.5%-23.6%), broadcast(27.5%-13.9%), microelectronics(17.5%-4.2%), transmitter(20.0%-6.9%), semiconductor(30.0%-18.1%), pollution(25.0%-13.9%),hughes(15.0%-4.2%), transmitters(20.0%-9.7%), promise(10.0%-0.0%), physics(12.5%-2.8%), receiver(25.0%-15.3%), antenna(27.5%-18.1%), semiconductors(15.0%-5.6%), significance(17.5%-8.3%), navigation(30.0%-20.8%), cockpits(10.0%-1.4%), cubic(10.0%-1.4%), daimler(10.0%-1.4%), recourse(10.0%-1.4%), scanner(10.0%-1.4%), shipboard(25.0%-16.7%), marconi(12.5%-4.2%), substrates(12.5%-4.2%)

Time series changes in industry attributes for firms stating the following three words in their business descriptions: Military, Defense, and Intelligence. Note that structural break tests regarding whether local similarity of rivals changed before and after 2001 is significant at the 1% level. A similar test based on the number of rivals is not significant. We also report the words that grew (shrank) in prominence before and after the 2001 military demand shock (we display the fraction of firms using the given word before and after the shock).

TABLE 7

Military Battlefield Firms: Competitor Changes



Panel B: Words that became more prominent after the shock (2000 vs 2002)

surveillance(50.0%-81.8%), transformation(0.0%-31.8%), optic(25.0%-54.5%), simulation(43.8%-72.7%), learning(6.3%-31.8%), corps(25.0%-50.0%), grumman(25.0%-50.0%), northrop(25.0%-50.0%), ceiling(12.5%-36.4%), imagery(12.5%-36.4%), artillery(0.0%-22.7%), beam(0.0%-22.7%), cleanup(0.0%-22.7%), infrastructures(0.0%-22.7%), omissions(0.0%-22.7%), disaster(6.3%-27.3%), incumbent(6.3%-27.3%), recruit(6.3%-27.3%), throughput(6.3%-27.3%), congressional(25.0%-45.5%), threat(25.0%-45.5%), smart(12.5%-31.8%), allegations(0.0%-18.2%), blank(0.0%-18.2%), defense(0.0%-18.2%)

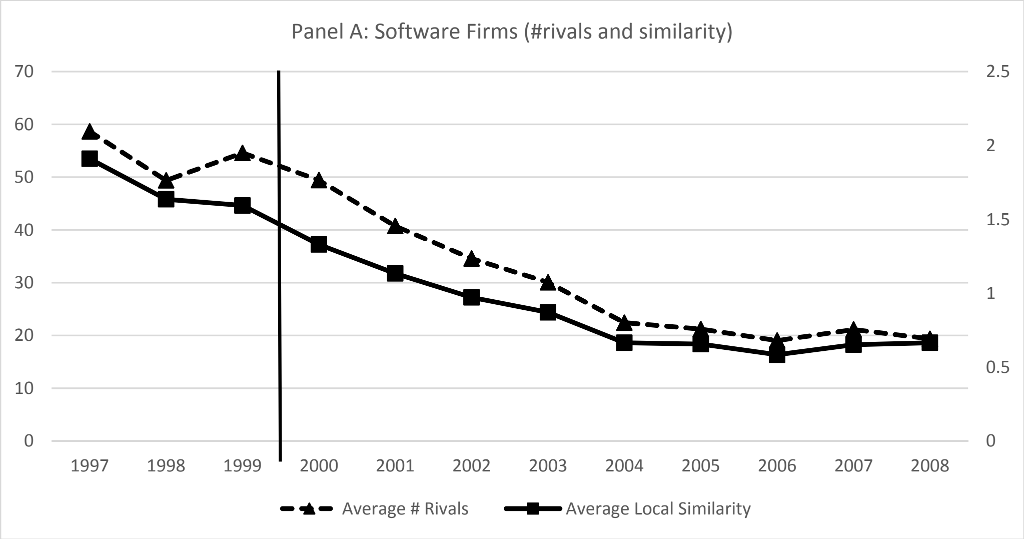
Panel C: Words that became less prominent after the shock (2000 vs 2002)

subassemblies(37.5%-4.5%), literature(31.3%-4.5%), motorola(31.3%-4.5%), rack(31.3%-4.5%), lockheed(75.0%-50.0%), compact(31.3%-9.1%), memory(31.3%-9.1%), broadcast(43.8%-22.7%), radar(68.8%-50.0%), algorithm(18.8%-0.0%), blood(18.8%-0.0%), microprocessors(18.8%-0.0%), panasonic(18.8%-0.0%), procurements(50.0%-31.8%), airframe(31.3%-13.6%), aperture(31.3%-13.6%), israel(31.3%-13.6%), hughes(25.0%-9.1%), radiation(25.0%-9.1%), warner(25.0%-9.1%), workstation(25.0%-9.1%), absorption(18.8%-4.5%), asics(18.8%-4.5%), assign(18.8%-4.5%), backup(18.8%-4.5%)

Time series changes in industry attributes for firms stating the following three words in their business descriptions: Military, Defense, and Intelligence. Note that structural break tests regarding whether local similarity of rivals changed before and after 2001 is significant at the 1% level. A similar test based on the number of rivals is not significant. We also report the words that grew (shrank) in prominence before and after the 2001 military demand shock (we display the fraction of firms using the given word before and after the shock).

TABLE 8

Software Firms: Competitor Changes



Panel B: Words that became more prominent after the shock (2000 vs 2002)

email(2.0%-19.1%), portal(0.3%-15.8%), intelligence(7.0%-21.6%), lifecycle(1.2%-13.7%), metrics(1.3%-13.0%), linux(0.2%-11.7%), insight(3.8%-14.8%), accountability(5.0%-15.4%), subscription(7.4%-17.2%), portals(0.0%-8.9%), disaster(5.8%-14.2%), players(6.9%-15.3%), verizon(0.0%-8.3%), validation(7.0%-15.3%), homeland(0.0%-8.2%), dell(3.5%-11.6%), portability(6.6%-14.6%), sciences(7.8%-15.7%), bandwidth(9.4%-17.2%), google(0.0%-7.7%), hybrid(5.9%-13.5%), download(6.4%-13.8%), cisco(5.6%-13.0%), streaming(1.2%-8.6%), interoperability(7.2%-14.5%)

Panel C: Words that became less prominent after the shock (2000 vs 2002)

unix(19.4%-7.7%), mainframe(14.3%-3.4%), intranet(14.5%-4.0%), telemarketing(18.4%-9.7%), compaq(9.0%-1.5%), workstation(14.5%-7.1%), announcements(14.8%-7.8%), workstations(16.2%-9.3%), novell(9.4%-2.9%), lotus(6.4%-0.5%), diversion(12.6%-6.8%), disk(16.3%-10.7%), object(15.0%-9.4%), intranets(8.0%-2.5%), peripheral(14.8%-9.4%), netscape(5.6%-0.3%), dial(11.1%-6.0%), lucent(9.7%-4.8%), vars(10.1%-5.3%), strain(7.7%-3.0%), pentium(5.8%-1.2%), microprocessor(10.4%-5.9%), peripherals(13.4%-8.9%), lans(7.7%-3.4%), packard(20.3%-16.3%)

Time series changes in industry attributes for firms stating the following three words in their business descriptions: Military, Defense, and Intelligence. Note that structural break tests regarding whether local similarity of rivals changed before and after 2001 is significant at the 1% level. A similar test based on the number of rivals is not significant. We also report the words that grew (shrank) in prominence before and after the 2001 military demand shock (we display the fraction of firms using the given word before and after the shock).

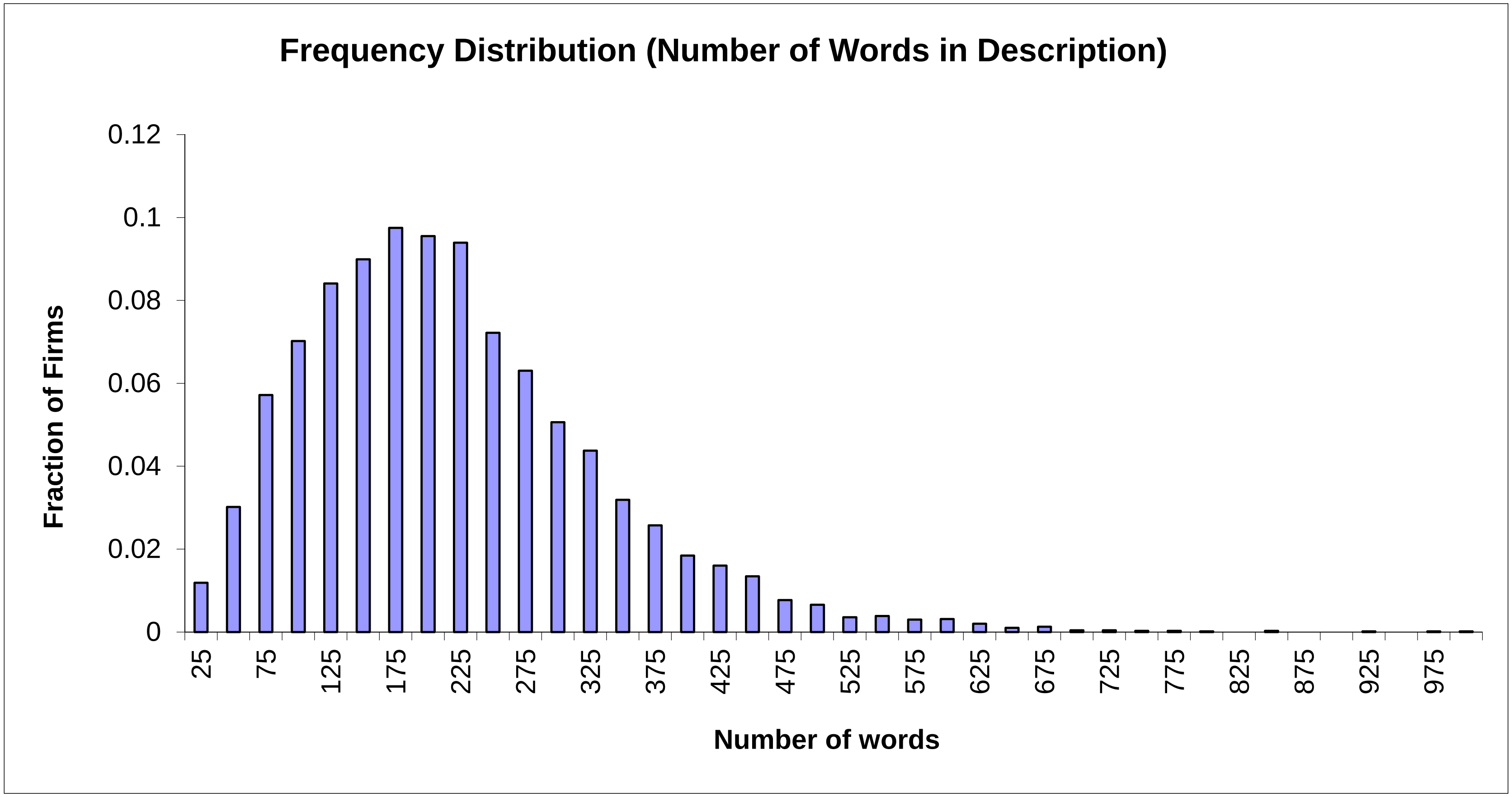
TABLE 9
Ex-ante advertising and R&D versus future similarity

	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: Text-based network industry Regressions								
(1)	Δ Total Similarity	-0.414	-0.152	-0.034	-0.005	0.055	0.018	-0.004	0.127
		(-13.72)	(-5.80)	(-8.55)	(-1.37)	(1.63)	(2.71)	(-0.25)	
(2)	Δ Number of Rivals	-12.301	-1.997	-1.195	0.156	2.184	1.636	-1.616	0.102
		(-6.94)	(-1.70)	(-4.83)	(0.87)	(1.41)	(4.38)	(-1.23)	
(3)	Δ Profitability	0.038	0.039	0.004	0.005	-0.022	-0.010	0.014	0.078
		(4.61)	(6.61)	(5.15)	(7.38)	(-4.92)	(-8.43)	(3.74)	
	Panel B: Industry-Adjusted Firm-Level Regressions								
(4)	Δ Total Similarity	-0.037	-0.116	-0.005	-0.020	0.059	0.016	0.020	0.121
		(-1.33)	(-5.15)	(-0.83)	(-4.25)	(1.76)	(2.40)	(1.29)	
(5)	Δ Number of Rivals	-0.775	-3.768	-0.103	-0.738	2.208	1.464	-1.291	0.100
		(-0.62)	(-4.03)	(-0.37)	(-3.71)	(1.42)	(3.99)	(-0.96)	
(6)	Δ Profitability	0.031	0.053	0.007	0.013	0.009	0.015	0.001	0.015
		(4.83)	(9.22)	(5.21)	(10.87)	(2.97)	(11.15)	(0.27)	

Year	# Firms in Group	In Group Avg Total Similarity	In Group Avg # TNIC Rivals	In Group Avg Adver/Sales	In Group Tot Adver/ Tot Sales
Panel C: Advertising and total similarity surrounding Apollo					
1997	20	102.4	8.9	0.013	0.003
1998	19	98.2	11.6	0.011	0.008
1999	23	95.5	11.6	0.031	0.029
2000	24	86.3	11.0	0.081	0.028
2001	17	78.9	9.765	0.031	0.023
2002	15	72.9	9.3	0.037	0.033
2003	17	66.6	7.8	0.041	0.035
2004	15	62.8	7.9	0.047	0.040
2005	14	61.1	7.9	0.057	0.067
2006	11	54.8	6.0	0.060	0.081
2007	12	54.5	6.0	0.063	0.064
2008	14	54.4	7.1	0.058	0.060

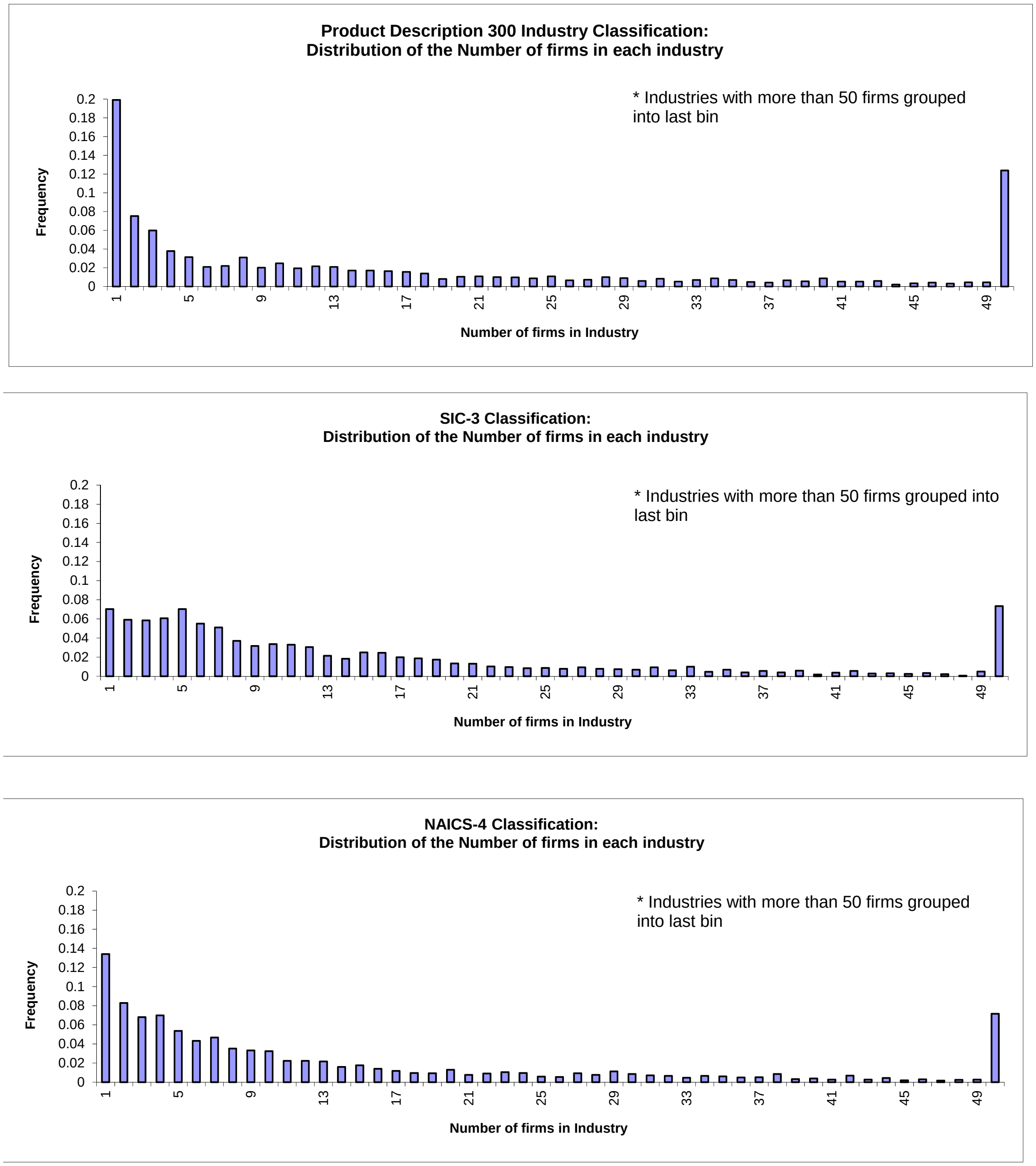
OLS regressions with ex post product changes in total similarity, and the number of rivals and profitability as the dependent variables. Panel A is based on advertising and R&D computed at the Text-based network industry level. Panel B is based on firm-level network-industry-adjusted advertising and R&D. All specifications include 10-K based fixed industry classification and yearly fixed effects. (t-statistics in parentheses are based on standard errors adjusted for clustering by year and industry). The sample has 49,246 observations. Panel C displays time series statistics including similarity measures and advertising activity for firms in Apollo's network industry during our sample period.

Figure 1



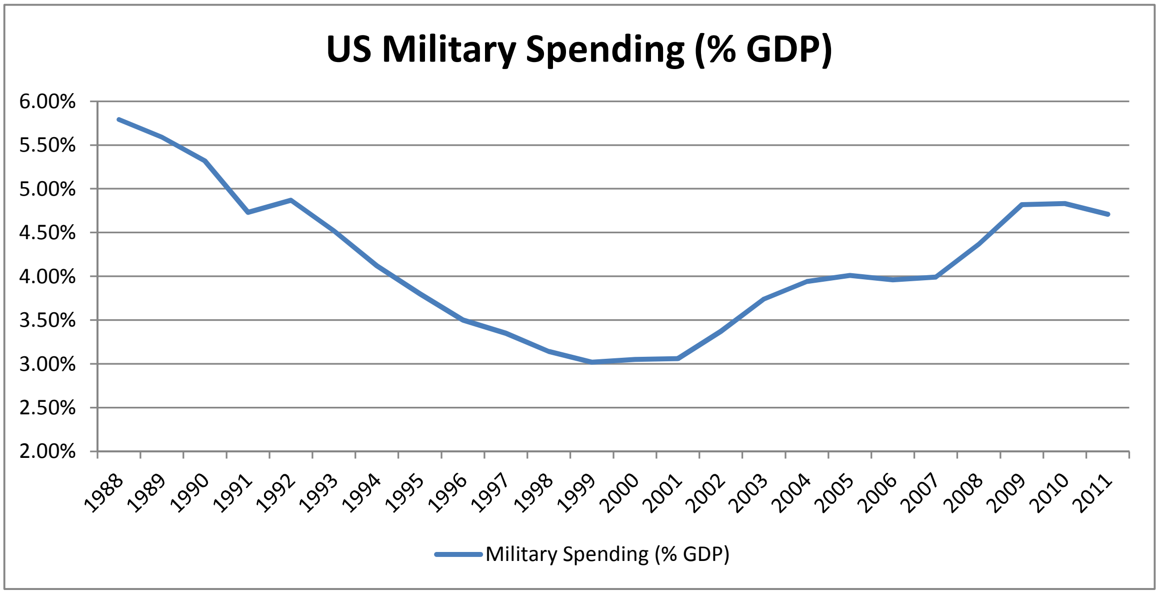
Frequency distribution of unique non-common noun and proper noun words in 10-K product descriptions.

Figure 2



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

Figure 3



Annual military spending by the United States as a fraction of GDP (source: World Bank).

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.

$$V_i \ = \ \frac{P_i}{\sqrt{P_i \ \cdot \ P_i}} \tag{3}$$

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".

$$Product\ Similarity_{i,j} \ = \ (V_i \ \cdot \ V_j) \tag{4}$$

We define product differentiation as one minus similarity.

$$Product\ Differentiation_{i,j} \ = \ 1 \ - \ (V_i \ \cdot \ V_j) \tag{5}$$

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 fixed 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. Thus, 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

$$\underset{j,k,\ j\neq k}{MAX} \qquad I_{j,k} \tag{6}$$

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:

$$I_{l,q} = \sum_{x=1}^{m_l} \sum_{y=1}^{m_q} \frac{S_{x,y}}{m_l \quad m_q} \tag{7}$$

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 thus 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-based fixed 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 (AIC) 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 A1 Here**]

Table A1 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-based 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 A1
Industry classifications and industry granularity

		oi/sales		oi/assets			
		Akaike		Akaike			Avg #
		Information	Adj	Information	Adj	# of	Firms per
Row	Industry Definition	Criterion	R^2	Criterion	R^2	Industries	Industry
		Panel A: SIC-code based industry definitions					
(1)	SIC-1-digit	3459.7	0.143	-59.1	-0.000	10	538.1
(2)	SIC-2-digit	2990.9	0.224	-303.8	0.048	72	74.7
(3)	SIC-3-digit	2827.1	0.273	-730.4	0.132	274	19.6
(4)	SIC-4-digit	2782.9	0.298	-864.7	0.183	434	12.4
		Panel B: NAICS based industry definitions					
(5)	NAICS-1-digit	3897.9	0.070	-215.9	0.031	9	597.9
(6)	NAICS-2-digit	3249.1	0.177	-504.7	0.086	23	234.0
(7)	NAICS-3-digit	2960.4	0.230	-778.8	0.142	97	55.5
(8)	NAICS-4-digit	2843.7	0.272	-890.5	0.190	331	16.3
(9)	NAICS-5-digit	3143.4	0.262	-575.5	0.176	680	7.9
(10)	NAICS-6-digit	3334.4	0.263	-381.1	0.178	1002	5.4
		Panel C: 10-K product description based industry definitions					
(11)	10K-based-50	2581.7	0.278	-1172.3	0.198	50	107.6
(12)	10K-based-100	2413.8	0.308	-1272.9	0.222	100	53.8
(13)	10K-based-200	2438.5	0.312	-1221.9	0.223	200	26.9
(14)	10K-based-250	2417.1	0.321	-1245.1	0.232	250	21.5
(15)	10K-based-300	2370.4	0.328	-1258.7	0.237	300	17.9
(16)	10K-based-400	2348.0	0.338	-1253.0	0.243	400	13.5
(17)	10K-based-500	2440.7	0.333	-1196.1	0.244	500	10.8
(18)	10K-based-800	2603.0	0.329	-1076.9	0.245	800	6.7

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 2008 and report average AIC scores.