What You Say Your Strategy Is and Why It Matters: Natural Language Processing of Unstructured Text

Anoop Menon
The Wharton School, University of Pennsylvania

Jaeho Choi
The Wharton School, University of Pennsylvania

Haris Tabakovic
Brattle Group and Harvard Business School

ABSTRACT

This study explores how new text analysis tools can be used in strategic management research that examines unstructured textual data. We build on two established natural language processing (NLP) techniques, vector space models and topic modeling, to create text-based measures of several core constructs in strategy – namely strategic change, positioning, and focus. These techniques are applied to the entire sample of 52,392 business descriptions in 10-K annual reports from 1996 to 2016. Results show that these new methods produce innovative yet meaningful measures of firm strategy which open up previously unexplored avenues of research to strategy scholars. The study advances emerging strategy research utilizing text analysis methods, demonstrates that NLP techniques can overcome some of the limitations of traditional text analysis methods such as keyword counts and mapping analysis, and provides a template for how other machine learning techniques could be introduced into strategy.

Keywords:
Natural language processing; text analysis; strategic change; positioning; focus; annual reports

1. INTRODUCTION

Research in strategic management has long investigated the drivers of firm performance (Rumelt, Schendel, & Teece, 1992). Researchers have been primarily concerned with how various aspects of firm strategy relate to competitive advantage. Among the theories and constructs that have been put forward, a large body of literature has investigated the role of strategic change (e.g., Haveman, 1992; Singh, House, & Tucker, 1986), positioning against rivals (e.g., Porter, 1980; 1985), and focus of products and business portfolios (e.g., Wernerfelt & Montgomery, 1988) in driving competitive advantage.

Empirical testing of these theories requires measures of varied aspects of strategy. However, measurement of the theoretical constructs has repeatedly proven to be a difficult task (Boyd, Dess, & Rasheed, 1993). Large-scale empirical tests have typically employed proxy measures of strategy using accounting and financial data sources (e.g., Finkelstein & Hambrick, 1990), but often yielded inconsistent findings (Rajagopalan & Spreitzer, 1997). Lack of longitudinal data which measures multiple dimensions of the strategic characteristics of firms often hinders researchers in their attempts to run large-scale empirical analysis that could be generalizable across different industries and firms. This paper suggests that researchers may make use of text analysis to address this methodological gap and shows that text-based measures of strategy can be a powerful complement to existing archival measures of strategy.

Over the past three decades, a class of methods called text analysis has become a widely used methodological approach in various streams of strategic management research. With text analysis, researchers have been able to extract meaningful information from unstructured textual data and perform qualitative as well as quantitative analyses with it. In recent years, developments in computational tools have significantly improved the reliability, replicability, and scalability of this method (Kabanoff, 1997; Grimmer & Stewart, 2013). With increasing availability of textual data, the value of textual analysis has increased, providing researchers the opportunities to address old questions in new ways. However, the existing tools for analyzing

text have not been well suited for *measuring strategy* due to their limited scope in generating relevant data.

Broadly speaking, two types of text analysis techniques are usually used in the literature: "*keyword counts*" (e.g., Cho & Hambrick, 2006; Kaplan, 2008) and "*mapping analysis*" (e.g., Huff, Narapareddy, & Fletcher, 1990; Carley, 1993). The first method, keyword counts, typically uses the computer to automatically measure the frequency of keywords which have been ex-ante identified by researchers as being associated with a construct of interest, and has been the most widely used technique to date (Duriau, Reger, & Pfarrer, 2007). The appeal of this technique derives from the fact that frequency of keywords provides an objective and intuitive indicator of a construct's magnitude or relative importance (Knoke & Kuklinski, 1982; Weber, 1990). The second approach, called "mapping analysis" (Axelrod, 1976; Carley & Palmquist, 1992), extracts representations of relational linkages between concepts in texts (Huff, 1990), and is mostly a complementary technique to the former. Typically, human coders are trained to determine relevant concepts and relations in texts and read through a set of documents until they reach a reliable representation of the cognitive relationships underlying the texts (Huff, Narapareddy, & Fletcher, 1990).

While both these techniques have become accepted textual analysis methods within the literature, the two conventional techniques present an apparent tradeoff. By counting keywords, researchers can extract narrow measures of the content in a text (e.g., managerial attention to an emerging technology), but are unable to exploit the rich and valuable information that resides in the broader structure of the words in texts (Carley & Palmquist, 1992). In contrast, mapping analysis provides richer information underlying a text (e.g., managerial beliefs of action-outcome linkages), but the technique requires a significant amount of human interpretation to extract an agreeable level of data, restricting the use the method to analyze a sizable amount of textual data. In this regard, the current tools for analyzing textual data have been limited in generating data for large-scale empirical analysis in strategy research.

Recent advances in natural language processing (NLP) methods from computer science provide new opportunities to utilize textual data (Grimmer & Stewart, 2013). Most importantly, NLP enables researchers to extract deeper meaning structures from a voluminous amount of text through computational algorithms. In this study, we explore some NLP tools to shed new light on some core questions in strategy research. To this end, we introduce two complementary techniques of increasing sophistication to utilize textual data and develop three new measures of core constructs in strategy research.

First, building on a widely used NLP framework called "vector space models," we create two measures of firm strategy: "change in strategy" and "differentiation from industry competitors." We show how researchers can employ a summary statistic of text called "*term frequency-inverse document frequency* (*tf-idf*)" to measure the relative distance between texts and further use this measure to create positional variables of strategy. Second, developing on an NLP technique called "*topic modeling,*" which extracts the latent topics in a set of documents and measures the composition of topics within texts, we create a measure of "strategic focus" (Siggelkow, 2003). Using the compositional weights of topics inferred from texts, we demonstrate that researchers can use this technique to measure the concentration or dispersion of topics within texts and further test how this measure correlates to important firm characteristics.

Applying these methods to corporate annual reports, we find evidence that our new text-based measures of change, differentiation, and focus in strategy are associated with the actual strategies of the firms. We demonstrate a few examples where our constructs identified periods of sharp strategic change, increased differentiation, or increased focus in firms that were clearly linked to those events in the actual timelines of these firms, as seen in the historical accounts. We also demonstrate how these constructs are significantly associated with future firm performance, in line with what has been suggested by the existing literature. For instance, we find that larger strategic change is, on average, correlated with lower future performance, and that a moderate degree of differentiation from rivals is correlated with increased future performance than high or low differentiation, thus providing empirical support for some fundamental claims in strategy.

These findings also point to the potential of using textual analysis of corporate texts to measure and run large-scale empirical analyses of important constructs in strategy research.

2. LITERATURE

In this section, we review the existing literature on strategic change, positioning, and focus. While summarizing how the prior literature has theorized and empirically tested these constructs in relation to firm competitive advantage, we briefly discuss how a text-based measure of the respective construct may complement the existing empirical studies.

2.1 Strategic change

Change in strategy is one of the most recurrently studied topics in the strategic management literature. Accordingly, the discussion has been widely scattered across different streams of research (see Ginsberg, 1988; Rajagopalan & Spreitzer, 1997 for a review). Based on the theoretical lenses adopted in the studies, one may categorize the literature on strategic change in terms of the ecological (e.g., Kelly & Amburgey, 1991; Amburgey, Kelly, & Barnett, 1993), institutional (e.g., Kraatz & Zajac, 1996), adaptation (e.g., March, 1981; Lant & Mezias, 1992; Greve, 1998), and interpretive (e.g., Dutton & Duncan, 1987; Gioia & Chittipeddi, 1991) perspectives.

In relating strategic change to organizational performance outcomes, two major theoretical views offer diametrically opposing predictions (Zajac & Kraatz, 1993). The adaptation perspective suggests that organizations attempt to adapt to changing environments in order to secure performance and survival (March & Simon, 1958). Although the performance outcome of the adaptive efforts is not guaranteed (March, 1981), it is expected that strategic change leads to performance improvement by increasing the fit between the organization and the environment (e.g., Virany, Tushman, & Romanelli, 1992). The ecological perspective, in contrast, tends to view strategic change as not only rarely occurring, but also dysfunctional and disruptive to organizations (Hannan & Freeman, 1977; 1984; Singh et al., 1986). It is argued that

change in organizational features introduces disruption to the operational reliability and thus impairs financial performance and aggravates the likelihood of organizational failure.

In accordance with these contradictory predictions, prior empirical studies find varying results with regard to the effect of strategic change on organizational performance (Rajagopalan & Spreitzer, 1997). For example, some studies find a positive relationship between change and financial performance (e.g., Zajac & Kraatz, 1993; Haveman, 1992), supporting the prediction that change resolves the incongruity between the organization and the environment (Tushman & Romanelli, 1985). However, a different set of studies find a negative relationship between strategic change and organizational performance, in line with the prediction offered by the ecological perspective (e.g., Singh et al. 1986). Still other studies find a non-significant relationship between change and performance (e.g., Kelly & Amburgey, 1991), casting doubt that change has a definitively negative or positive effect on performance.

The inconclusiveness in the empirical findings could be attributed to the differences in the operationalization of strategic change and organizational performance and/or to the differences in the empirical settings (Rajagopalan & Spreitzer, 1997). Studies typically focus on firms in a specific industry and use industry-specific measures of change in some particular characteristics of the firm. For instance, examining the U.S. airline industry, Kelly and Amburgey (1991) measured the changes in the product mix and the level of diversification as the business- and corporate-level changes and the probability of business failure as an indicator of firm performance. Haveman (1992) studied the savings and loan industry and used change in the firm's investment portfolio on eight different submarkets as the change variable and two financial measures (net worth and net income) and the rate of failure as the performance variables.

While industry-specific studies provide rich detail about each industry, it is not easy to generalize the research findings to a broader set of industries. Thus, for a large sample study to test the theory of interest, one needs a measure of strategic change that can be measured across a

diverse set of firms in different industries. To this end, we suggest that a text-based measure of strategic change can overcome the methodological limitation.

2.2 Strategic positioning

The positioning school of strategy (Porter, 1980; 1985; 1991; Ghemawat, 1991) argues that a firm's competitive advantage derives from its position within the market vis-à-vis its rivals. It is argued that the firms that create and capture additional value, through lower cost and/or higher quality positions, than the rivals generate competitive advantage (Porter, 1985; Brandenburger & Stuart; 1996). A firm's position is achieved through the activities that the firm performs along its value chain (Porter, 1985; 1996). The earlier work argued that firms can perform three stylized types of strategies (cost leadership, differentiation, or focus) to achieve a superior position (Porter, 1985). However, more recent arguments have moved towards viewing a firm as a system of interconnected activities and suggest that the mutual reinforcement of activities is crucial in creating and sustaining competitive advantage (Porter, 1996; Siggelkow, 2001; 2002).

According to this perspective, when firms are crowded around a position within the market, increased rivalry likely erodes competitive advantage (Porter, 1985), unless industry participants collude to reduce rivalry by forming an informal group (Caves & Porter, 1977; Porter, 1979). Increased number of undifferentiated rivals erodes profitability because firms would need to share the fixed amount of value created in the market (Brandenburger & Stuart, 1996; Hatten & Hatten, 1987). Thus, sustained competitive advantage is achieved when firms can continuously create strategic distance from their rivals through repositioning their business strategy (Menon & Yao, 2017).

While the competitive advantage from superior positioning is now an old argument in the strategic management literature, empirical research on this argument has been largely limited because of the difficulty of measuring competitive positions of market players (Smith, Grimm, Gannon, Wally & Young, 1997). Only a few studies were performed on select industries, such as the airline industry (e.g., Peteraf, 1993; Baum & Korn, 1996; Chen, 1996; Smith et al., 1997),

where the set of competitors and the boundaries of competition (single business) are clearly defined and the information on the competitive positions of the players in the market are publicly available. Because firms often operate in multiple businesses and information regarding their market profiles is not readily available, it is difficult to apply methods utilized in this kind of studies to firms in other industries.

Departing from models that assess market positions in terms of price and costs (Brandenburger & Stuart, 1996) or industry-specific factors (e.g., Chen, 1996), we adopt a landscape perspective of competition (e.g., Hotelling, 1929; Lee, Lee, & Roh, 2002) and, using a text-based measure of strategic position, examine how strategic distance between industry participants relate to the competitive advantage of firms in the marketplace.

2.3 Strategic focus

The degree to which a firm should be focused in its business and product portfolio has also been one of the most central topics examined in strategic management research (see Ramanujam & Varadarajan, 1989; Palich, Cardinal, & Miller, 2000). Evidenced by the work in industrial organization (Edwards, 1955) and agency theory (Jensen, 1986), the economic literature has heavily studied this topic as well (Lang & Stulz, 1994; Montgomery, 1994). However, similar to the discourse on strategic change, there has been no convergence in the theoretical frameworks and empirical findings that explain the relationship between firm focus and performance (Palich et al., 2000; Zahavi & Lavie, 2013).

In the spirit of agency theoretic explanations, a set of studies argues that the level of diversification would have a negative effect on firm performance due to the agency costs incurred by firm managers (Amihud & Lev, 1981; Jensen, 1986). In support of this argument, empirical studies find evidence of a negative relationship between the degree of diversification and firm performance, particularly in terms of Tobin's *q* (Wernerfelt & Montgomery, 1988; Lang & Stulz, 1994) and return on invested capital (Montgomery, 1985). In contrast, the resource-based perspective argues that firms may diversify their business to utilize the excess capacity in productive factors (Penrose, 1959; Wernerfelt, 1984), implying that diversification is

not necessarily bad for firm performance. Departing from the argument about the performance consequences of the level of diversification, a standard argument in this stream of research is that firms may obtain economies of scope through related diversification (Markides & Williamson, 1994; Montgomery, 1994). In line with this reasoning, a set of empirical findings support that a firm's performance is maximized when the firm holds a portfolio of related businesses (Palich, Cardinal, & Miller, 2000).

While empirical studies on strategic focus abound in the literature, measurement of the level and type of diversification has been a persistent problem for empirical testing of the theories (Chatterjee & Blocher, 1992; Hoskisson, Hitt, Johnson, & Moesel, 1993). In many studies, the degree of focus is typically measured as a Herfindahl-Hirschman Index (HHI)-like measure of sales based on the Standard Industry Classification (SIC) or the North American Industry Classification System (NAICS) codes (e.g., Montgomery, 1985; Ravenscraft & Scherer, 1987). A simple count of the industry codes is frequently used as well (Palich et al., 2000). However, the heavy reliance on industry classification codes for measuring diversification can be a problem for empirical testing because neither the SIC nor the NAICS codes properly reflect the distance between industrial boundaries over time since the classifications are static. This introduces major difficulties in effectively gauging the level or type of diversification.

Moving away from measures that rely upon traditional industry classifications, in this study, we suggest using a text-based measure of strategic focus. As a complement to the industry-focused measures, we adopt a topic focused view which is both more temporally dynamic, as well as generally applicable across industry contexts.

3. DATA

For the analysis to follow, we use the business descriptions in corporate annual reports to measure the various constructs of strategy discussed in the literature section. Annual reports are an important and relevant source of data that periodically disclose the current state of the strategy of public firms. Since Bowman (1978; 1984) pioneered the use of annual reports to identify

corporate strategies, researchers have used portions of the annual reports to study various types of firm characteristics (e.g., Bettman & Weitz, 1983; Fiol, 1989; Guo, Yu, & Gimeno, 2017). For instance, Gavetti and Rivkin (2007) used the business descriptions in the annual reports of Lycos and Yahoo to examine the difference and similarity in their representations of the firm's strategy in the late 90s.

Annual reports, technically referred to as the "Form 10-K" filings, are required by the federal securities laws for all public companies and are archived by the U.S. Securities and Exchange Commission (SEC). Item 1 of the 10-K filings concerns the description of the business. It is the part where firms are required to comprehensively overview the current state of their business, mostly discussing the major business areas, including their product and geographic scope. While it is argued that the reports are primarily prepared by the public relations department in firms (e.g., Barr, Stimpert, & Huff, 1992), prior studies find that there is a reasonable correspondence between the written information and the objective reality (e.g., Bowman, 1984; Gavetti & Rivkin, 2007; Guo, Yu, & Gimeno, 2017). Thus, in this analysis, we use the Item 1 section of the annual reports to create our text-based strategy measures.

We gathered the entire collection of 10-K filings from the EDGAR database of SEC[1]. The EDGAR database contains 10-K filings starting with 1994, but the data is sparse for the earlier years because it was only on May 6, 1996 that firms were fully required to make their filings electronically available through the database. Thus, we restricted the sample of annual reports to those that were filed between 1996 and 2016. After merging the 10-K data with firm financial data from COMPUSTAT, our final sample size is 52,392 firm-year observations.

After retrieving the 10-K data from the database, we performed a few steps of text processing prior to applying our text analysis methods. First, we extracted the portion of business descriptions (Item 1) from the raw text files. While 10-K filings are divided into four parts with

[1] We retrieved various types of 10-K filings from the EDGAR database, including "10-K," "10-K405," "10KSB," and "10KSB40." "10-K405" was a form filed by firms which did not disclose their internal trading activities within a required time frame and "10KSB" and "10KSB40" were forms filed by small businesses. The use of these special forms has been discontinued after 2002 for "10-K405" and after 2009 for "10KSB" and "10KSB40".

15 items, the electronic versions do not provide the individual items separately. We created a custom algorithm to identify the sections in the 10-K and used it to extract the Item 1 for each filing.

Next, we employed a few text preprocessing techniques that are standard in text analysis studies (see Jurafsky & Martin, 2014, and Manning, Raghavan & Schütze, 2008, for an overview of NLP). The preprocessing steps are used for decreasing the complexity of the textual data, while preserving the substantive content for the analysis (Denny & Spirling, 2017). As a first step, we parsed each document into words and removed non-alphabetical expressions, such as punctuations, special characters (e.g., #, %, &, $, and ₩), and numbers. Then, we lowercased (de-capitalized) the words and removed "*stop words*," words that primarily function as grammatical fillers in text. Examples of stop words include 'there', 'here', 'about', 'which', 'just', 'or', 'nor', and 'such'.[2] While stop words can be used for the analysis of, for example, the style of writing, its removal does not lead to significant difference in our analyses since it does not contain semantic meaning.

Most text analysis methods take the individual words as the unit of analysis. However, since there are multi-word expressions, such as "business portfolio" or "global expansion strategy," that deliver a particular meaning different from the individual words, we take into account these expressions with a process called "*noun-phrase chunking*." The chunking process first identifies how each word functions in a sentence (using a function called part-of-speech tagging) and then extracts the combinations of words that function as noun-phrases in each sentence.

Words are often written in various forms while they essentially convey the same meaning (e.g., "focus" and "foci"). Stemming and lemmatization are typically used as alternative techniques to reduce the variant forms of vocabularies in the corpus (Manning et al., 2008). Stemming refers to a heuristics-based process of reducing words to their basic form (i.e., word stem). Lemmatization refers to the process of reducing words to their dictionary form (i.e.,

[2] The full list of stop words can be accessed at *http://www.nltk.org/book/ch02.html*.

word's lemma). In our analysis, after comparing the results of stemming and lemmatization, we elected to use the latter since it more reliably converted the words to interpretable forms.

As a final step for preprocessing, we filtered out a portion of the vocabularies by its frequency of appearance in different documents. Prior work in NLP has found that both extremely frequent and extremely infrequent words do not contribute much when analyzing the patterns in a large set of documents and, practically, that the filtering significantly reduces the size of words analyzed (Denny & Spirling, 2017). We used a cutoff rule whereby words that appeared in more than 75% of documents and those that appeared in less than 20 documents were discarded.[3]

4. TECHNIQUES

In this section, we explain how we create our text-based measures of strategic change, positioning, and focus, based on two types of established techniques in the field of NLP, called "vector space models" and "topic modeling."

4.1 Vector Space Models

Every NLP application requires decisions about how to convert the textual data to a numerical form (Manning et al., 2008). One basic but robust approach is a class of models called "vector space models," which transforms a text into a vector form with each word appearing in the text as a feature in a vector space. Given a collection of documents, called the corpus, each document can be represented as a vector consisting of D dimensions, where D is the size of the dictionary or the total number of unique words (and multi-word expressions) that appear in the corpus. In this model, the order of words is disregarded since it imposes minimal cost on inference (see Grimmer & Stewart, 2013 for discussion).

Each feature in the vector representation of a text includes a numerical value of a word which may be calculated in various ways. One way is to assign a binary value, 0 or 1, for each

[3] In a robustness test, we find that the unfiltered version produces qualitatively similar results to the filtered version. We adopted the filtered version for our analyses in order to extract easily interpretable LDA topic keywords.

feature in the vector, indicating the presence of a word in each document (e.g., Hoberg & Phillips, 2010). While this approach can significantly reduce the time required to process the textual data, it is a very coarse method that drops a key piece of information in the data, namely, the frequency of words.

As evidenced in prior work (e.g., Kaplan, 2008), the frequency of words in a text reflects the degree to which a concept is invoked. Moreover, information about how frequently a word appears in different texts in the corpus also reveals the word's degree of generality; words that appear in a small portion of the corpus often have significant discriminating power in the text. To account for these factors, one can use a measure called the *term frequency - inverse document frequency* (*tf-idf*), which is a numerical weight of the word proportional to the frequency of word in a text and inversely proportional to the log of the frequency of word within the corpus. Formally, the *tf-idf* value of a term in a document is calculated as:

$$tf\text{-}idf_{t,d} = tf_{t,d} * idf_t,$$

where $tf_{t,d}$ denotes the number of occurrences of term t in document d and idf_t denotes the log of the inverse fraction of the documents that contain the term t in the corpus ($idf_t = \log(N / df_t)$, where N is the total number of documents and df_t is the number of documents containing term t).

Using the vector representations, one can calculate the distance or similarity between documents through various vector calculation methods. One of the most frequently used methods is to calculate the cosine similarity between two document vectors (Salton & Buckley, 1988; Manning & Schütze, 1999). The cosine similarity is computed as the cosine of the angle between two *tf-idf* vectors of texts (V_1, V_2):

$$CosineSimilarity\ (V_1, V_2) = (\sum_i V_{1,i} \cdot V_{2,i}) / (\sum_i V_{1,i}^2 \cdot \sum_i V_{1,i}^2)^{1/2}.$$

Once the cosine similarity is calculated, one can easily compute the cosine distance between the two vectors V_1 and V_2 as:

$$CosineDistance\ (V_1, V_2) = 1 - CosineSimilarity\ (V_1, V_2).$$

As an alternative to the cosine distance, one may use the Euclidean distance between two vectors, which is calculated as:

$$EuclideanDistance\ (V_1, V_2) = (\sum_i (V_{1,i} - V_{2,i})^2)^{1/2}.$$

4.2 Topic Modeling

Topic modeling is another powerful application of NLP which is used to extract latent or hidden themes that pervade a set of documents. The term refers to a class of algorithms that performs the extraction of topics through various computational processes. In our analysis, we use a topic model called the Latent Dirichlet Allocation (LDA) (Blei, Ng & Jordan, 2003) which is one of the simplest but most robust methods in topic modeling. Topic modeling, and particularly the LDA, has recently gained some acceptance in the management literature (e.g., Kaplan & Vakili, 2015), but its application has mostly remained to extract keywords for latent topics in a large body of texts. Below, we briefly explain the basic intuitions of topic modeling and its potential for analyzing large textual data sets (see Blei (2012) for a comprehensive overview of the LDA method).

The LDA as well as other topic models are based on the assumption that a document contains a set of topics, and the topic itself consists of a set of keywords. From a statistical perspective, topic models assume that a given document is generated by selecting topics from the distribution of topics and then selecting a set of words from the probability distribution of words that are associated with those particular topics. This assumed structure is referred to as the document generative process. Thus, a document is a collection of words that are associated with the topics present in that document, with the words associated with the "heavier" topics in the document being present more, and vice versa. Given this framework, a topic model typically infers the probability distributions of topics and keywords by analyzing the pattern of words appearing in the corpus. The LDA algorithm estimates the hidden parameters (i.e., the set of topics in the corpus, the probability of words associated with a topic, the topic of each word in a document, and the topic distribution of each document) in the statistical framework through a Bayesian inference method by inverting the document generative process (Blei, 2012).

The LDA algorithm is an unsupervised method, meaning that it generates topics without any intervention by the researcher. Thus, it eliminates bias introduced by human coders and

enables researchers to discover unknown thematic structures latent in a large textual data set. The only input parameter for a standard LDA is the optimal number of topics to be generated from the corpus. Given a set of preprocessed documents, the algorithm automatically analyzes the co-occurrence of words across the data set and produces information regarding 1) the probability distribution of words in each topic and 2) the probability distribution of topics in each document. Using this output information, a researcher can infer a set of topics to classify documents (e.g., Hasan, Ferguson, & Koning, 2015) or identify a document that spawns a new topic in the corpus data (e.g., Kaplan & Vakili, 2015).

5. CONSTRUCTS

5.1 Measure of strategic change

Using the vector space model, we measure a firm's *strategic change* as the degree of change in its business description from year to year. In particular, we use the cosine distance between the firm's 10-K Item 1 *tf-idf* vector in year *t* and that in year *t-1*. As discussed earlier, prior studies have demonstrated that there is a close correspondence between the written description and the reality of firm strategies (e.g., Bowman, 1984). For instance, Gavetti and Rivkin (2007), examining Lycos's 1996 and 1999 10-K filings, found that the firm's business description dramatically changed over the three-year period, reflecting its shift in strategy: from a technology-oriented search engine company to a media-oriented company. They also found that Yahoo, in contrast to Lycos, persisted in using similar business descriptions which reflected the firm's strategy as a "media company" since early 1996 (p. 430). Given such evidence, the degree of change in a firm's strategy can be measured by the dissimilarity of business descriptions over different periods.

5.2 Measure of strategic positioning

Analogous to our measure of strategic change, we measure a firm's *strategic positioning* within its industry as the firm's average distance from its rivals in the same primary industry. This approach is similar to Hoberg and Philips (2016), who used a text-based distance measure

of firms to create a new industry classification scheme. Rather than identifying an arbitrary cluster of firms in the entire text-based space, we use the relative positioning of the firms within an industry boundary to measure its level of differentiation. For instance, if most of the participants in an industry are more or less similar in their strategy and a particular firm positions itself to be differentiated from its competitors, the average distance against one's rivals would be high for the differentiator and low for the rest. Technically, to reflect this idea, we first calculate the cosine distance of the text vectors for all the pairs between the firm against all its rivals within the same four-digit SIC industry in year *t* and next take the average of the pairwise distance values.

5.3 Measure of strategic focus

Using the LDA topic model, we measure a firm's strategic focus by calculating the sum of squared topic weights from the LDA output. This is similar to a HHI measure of diversification, which computes the focus of a firm by taking the sum of the squared business segment shares (Palich et al., 2000). After running the LDA algorithm on our 10-K data, setting 100 topics to be extracted, we found that keywords from a topic in the output data roughly corresponded to an industry topic. Therefore, we reasoned that the distribution of topic weights in each 10-K filing reflects the degree of concentration/diversification of the firm's business portfolio.

5.4 The soft drink industry

As a way to check how the proposed text-based measures of strategy capture different aspects of firm strategy, we show below the measures computed for the major firms in the beverage industry (SIC = 2080). We selected three major soft drink manufacturers, namely Coca-Cola Company (hereafter Coca-Cola), PepsiCo, and Dr. Pepper Snapple Group, because their histories are relatively well known to the public and have exhibited several shifts in their strategy over the past two decades.

First, in Figure 1, we show the degree of strategic change for the three major soft drink manufacturers. The upper panel shows, for each firm, the cosine distance between the current

and previous years' text vectors in dots and the historical trend in a median spline. The bottom panel overlays the median splines of the three firms for direct comparison. Higher values in the tf-idf cosine distance suggests that a company has used significantly different language to depict their business in a given year compared to the previous year.

In the figure, we find that PepsiCo has gone through some significant changes over the past two decades in comparison to its two rivals. In particular, the median spline shows that PepsiCo had two increased periods of change during the early and the late 2000s. These two periods correspond remarkably well to major changes in PepsiCo's recent history. First, in 2001, PepsiCo acquired and merged with Quaker Oats Company, which held various food brands in the breakfast cereal and snack categories. The acquisition characterized a major move in PepsiCo's business portfolio since the firm then had only a limited presence in the food industry, through its snack business Frito-Lay, and the sports drink category. In the late 2000s (2009-2010), PepsiCo acquired its two major bottling companies, Pepsi Bottling Group and PepsiAmericas, and formed a wholly owned subsidiary called Pepsi Beverages Company. Through the mergers of the two businesses, PepsiCo controlled about 80% of its bottling network (Kaplan, 2010), making it a highly vertically integrated firm.

Insert Figure 1 about here

Next, in Figure 2, we plot the three major soft drink manufacturers' average distance against their rivals within the beverage industry over time. As stated earlier, the cosine distance of the text vectors captures the dissimilarity between business descriptions and, thus, the average cosine distance of a firm against its rivals reveal how a firm differentiated its business against its industry rivals.

Two historical patterns can be noticed from the figure. First, the negative slope of the median splines in the late 90s and early 2000s for Coca-Cola and PepsiCo indicate that the two firms were converging towards each other. During this period, the two firms added various lines of drink products other than carbonated soft drinks. For instance, Coca-Cola attempted to

increase sales for its sports drink brand Powerade and fruit drink brand Oasis. In the meantime, PepsiCo purchased Tropicana in 1998, adding juice products to its beverage portfolio. PepsiCo's purchase of Quaker Oats Company also brought Gatorade under their control, which led the firm to directly compete against Coca-Cola in the sports drink market.

Second, the positive slope of the lines after early-mid 2000s correspond to the differentiation strategies undertaken by the three major soft drink manufacturers. Around this period, Coca-Cola started to reduce the level of sugar in its soft drink products as consumers increasingly concerned about sugar-related health problems. In addition, in 2010, Coca-Cola started its major bottling consolidation operations. Recently, it has also started buying up non-carbonated drink companies in international markets. In the meantime, as discussed above, PepsiCo has set up an agenda to expand their food and snack business and include healthier ingredients in the entire portfolio of products, also with an international focus. The Dr. Pepper Snapple Group was spun off from U.K. based confectionary company, Cadbury, in 2008 and has been focusing on the carbonated soft drink market in the U.S. The differences in the strategic orientations of the three firms show why the distance scores have trended upwards for all the firms since 2010.

Insert Figure 2 about here

To better understand how strategic change and differentiation work in tandem, we map out, in Figure 3, the relative distance between the text vectors using a dimension reduction technique called the principal component analysis (PCA)[4]. In brief, the PCA plot shows a two-dimensional projection of the given text vectors which originally belongs to the high-dimensional tf-idf vector space. In the plot, the dots represent a text vector for a given year and the lines show the temporal progression between the text vectors. For simplicity, we labeled only the earliest and latest text vectors for each firm. One noticeable pattern from Figure 3 is that, in

[4] Although the PCA plot provides an intuitive visual of the relative position of the text vectors, we caution that PCA does not consistently correspond to cosine distance measures since information is lost in the dimension reduction process.

correspondence to Figure 1, PepsiCo's text vectors are located further apart than the other firms' text vectors, which reflects the firm's increased level of change during the early and late 2000s. Another interesting pattern is that the three groups of text vectors are rather located equally distant from each other. This pattern suggests that the three major beverage firms were rather distinct from each other, with no one firm particularly close to the other firms. Such a pattern corresponds to the firms showing similar distance values in Figure 2.

Insert Figure 3 about here

As an example of our strategic focus measure, we display, in Figure 4, the historical trend of the soft drink manufacturers' strategic focus from 1995 to 2016. A higher value in the measure means that the firm's business description has been more focused in terms of the topics present in the business descriptions. Among the three players, we find that PepsiCo has been the least focused player among the three majors. As evidenced in their large-scale acquisitions during the 2000s, PepsiCo expanded its operations beyond carbonated soft drinks (e.g., Pepsi and Mountain Dew) and snacks (e.g., Frito-Lay) into various food (e.g., Quaker Oats) and non-carbonated beverage (e.g., Tropicana) markets. In the meantime, Coca-Cola remained focused on its carbonated beverages until 2007, when it then diversified its brand portfolio and acquired several companies in the still drink segment[5]. While the Dr. Pepper Snapple Group has been more of a focused player in the beverage industry, since its spin-off from Cadbury, the firm has been a heavy diversifier within the industry, holding more than 50 carbonated and non-carbonated drink brands.

Insert Figure 4 about here

5.5 The airline industry

As another example to show how our measures perform, we apply the three measures to the scheduled air transportation industry (SIC = 4512). We selected the airline industry since

[5] These were Honest Tea (an organic tea company), Innocent Drinks (a London-based manufacturer of fruit smoothies), and Glacéau (the maker of VitaminWater and SmartWater).

prior research has found that it is an attractive empirical setting to study competitive dynamics of market players due to its single-business characteristic (e.g., Guo et al., 2017; Baum & Korn, 1996). We picked six major airlines (Southwest Airlines, American Airlines, Delta Air Lines, United Airlines, JetBlue Airways, and Spirit Airlines[6]) for comparison. Figure 5 shows the strategic change of these airline firms from 1996 to 2016.

As the graph shows, the major airline firms exhibited distinctive periods of strategic change in the past two decades. Again, we find that the periods of strategic change correspond remarkably well to the individual histories of the firm. Below, we focus the discussion on Southwest Airlines and United Airlines. First, the results show that Southwest Airlines have gone through some substantial strategic change around 2010. The most important event during this period is likely to be the AirTran acquisition. Southwest announced its plan to acquire AirTran Airways in September, 2010 and finalized the transaction in May, 2011. The merger significantly increased Southwest's presence in the international markets (Central America and northern parts of South America) as well as the East Coast region. Most notably, AirTran gave Southwest an access to the Atlanta airport (200 daily departures). In addition, 140 AirTran aircrafts were added to the existing 550 Southwest aircrafts, marking a 25% increase in fleet size.

Second, we find that United Airlines has undergone two distinct periods of major strategic change in the early and the late 2000s with the distance value around 0.6. In 2001, two of the United Airlines' aircrafts were hijacked and crashed in the September 11 attack. Moreover, coupled with low demand (post-dotcom and post-9/11) and increased costs (increased oil prices and pilot pay raise), United Airlines folded its dual brand strategy and discontinued its low-cost carrier, Shuttle by United. Meanwhile, in 2010, after a four year-long merger discussion and regulatory approval, Continental Airlines and United merged and formed a new company called United Continental Holdings. The combined airline network covered 370 airports in 59 countries around the world, making it the largest airline in terms of revenue passenger miles. The peak in 2010 reflects the merger between United and Continental.

[6] Founded in 1980 as Charted One, Spirit Airlines started to submit their annual reports to the SEC in 2012.

Insert Figure 5 about here

In Figure 6, we plot the average distance of the airline firms to the rest of its competitors. An interesting point about Southwest Airlines is that, unlike prior reports that find the firm as a highly differentiated airline during the pre-2000 era (e.g., Porter, 1996), our measure of differentiation shows that its distance against its rivals has been relatively decreasing in the post-2000 era. Such a trend may have resulted from the fact that Southwest has become a major domestic airline over the years by serving numerous geographical markets across the whole nation. In addition, Southwest's entry to the international markets, through its acquisition of AirTran in 2010, may have contributed to the loss in differentiation. Essentially, the move represented an entry into the territories of the other major airlines, such United Airlines or American Airlines, which already had significant operations in the international markets. A different reason that may have affected Southwest's relative position is the rise of other low-cost airlines, such as JetBlue and Spirit. These firms explicitly adopted various aspects of Southwest's strategy, such as maintaining a limited number of airports, a single type of aircraft, and a single class of service.

Insert Figure 6 about here

Furthermore, to get a better understanding of the relationship between strategic change and positioning, we show the PCA result of the airline firms' text vectors in Figure 7. In the plot, the relative distances between the dots do not perfectly correspond with the cosine distance shown in Figure 5.[7] While the distance between the text vectors for Southwest Airlines and Delta Air Lines matches their pattern of strategic change, the PCA result does not reflect the major peaks shown in United's (2001 and 2010) and American Airlines' (2013) subplots. Despite the partial correlation between the strategic change and the relative distance in the PCA plot, the PCA result still reveals some information about the similarity across the text vectors. In

[7] As mentioned before, this is due to the loss of information when reducing the number of dimensions during a PCA.

particular, though Southwest lost its extent of differentiation in the later years, as evidenced in Figure 6, their positioning was rather distinct to the rest of the competitors. According to the PCA result, American Airlines, United Airlines, and Spirit Airlines were located in a similar space. While JetBlue was initially close to this group of firms, the firm slowly moved away from the group over time. When viewed in conjunction with the results in Figure 5 an Figure 6, we can understand that JetBlue was trying to carve out its own market position over the past decade through a moderate level of strategic change.

Insert Figure 7 about here

Lastly, in Figure 8, we show the strategic focus of the six major airline firms from 1995 to 2016. Most notably, Delta Air Lines exhibits the highest level of focus among the six major airlines throughout the whole period. However, despite its high level of focus, Delta has also been the firm that has lost a significant degree of focus since the early 2000s. To get a relative sense of what the scale of the focus measure implies, in Figure 9, we compared the strategic focus of Delta Air Lines with General Electronics, which is a multi-business firm known for its broad business portfolio. Similar to the bottom panel in Figure 8, we overlay the calculated focus measures for Delta and GE in the upper panel of Figure 9. As one would expect, the plot intuitively shows that GE's focus is significantly lower than Delta's throughout the past two decades.

To get a better sense of how the focus measure was calculated, in the bottom side of Figure 9, we show two PCA plots of the LDA topics extracted for Delta and GE in 2015. In this PCA plot, a circle denotes a topic present in a given business description, the size of the circle refers to the relative topic weight, and the relative position of the circle shows how similar a topic was to the others. For instance, the large "0" circle in Delta's PCA plot (bottom left panel) refers to a topic with keywords related to the air transportation industry[8]. As represented by its

[8] The top 10 keywords of the topic "0" were "aircraft", "airline", "service", "travel", "passenger", "airport", "regulation", "cost", "operation", and "addition."

size, Delta's business description mostly concerns the airline industry topic, and this is the reason why Delta had a higher focus score in the upper panel. In the meanwhile, GE's PCA plot (bottom right panel) shows that the firm's business description had a greater number of topics with relatively equal and small weights. Since our focus measure takes the sum of the squared topic weights, the distribution of the topic weights tells why GE's focus value was significantly lower than Delta. Interestingly, we find that the circle "35" referred to a topic with keywords related to the aerospace business[9]. The presence of this topic and its relative size reflects GE's presence in the aerospace sector through its aircraft engine division, GE Aviation.

Insert Figure 8 about here

Insert Figure 9 about here
---

6. RELATIONS TO PERFORMANCE

Using the methods described in the previous section, we empirically analyzed how each measure of the strategy constructs relates to firm performance. In the regression analyses, we used a one-year leading Tobin's *q* as the dependent variable since we are focusing on the future performance implications of these constructs as predicted by theory. Following previous work (Villalonga, 2004), Tobin's *q* was calculated as the ratio of market value to the book value of total assets. We included three control variables across all regressions: the current ratio, the selling, general, and administrative expenses (SG&A) intensity, and the logged total sales. The current ratio reflects the short-term cash constraints and is defined as the firm's current assets over current liabilities. The SG&A intensity addresses a firm's overhead efficiency and was calculated as the ratio of SG&A expenses over total sales. The log of total sales was used as a proxy for firm size. We ran pooled OLS regressions with year fixed effects and industry fixed

[9] The top 10 keywords of the topic "35" were "aviation", "drone", "defense", "aerospace", "overhaul service", "aircraft", "helicopter", "part", "otp", "u.s. government."

effects at the four-digit SIC level. The fact that the findings are robust to the inclusion of industry fixed effects demonstrates the cross-industry validity of the constructs. Robust standard errors were used to account for heteroscedasticity in all regression models.

Insert Table 1 about here

6.1 The relationship between strategic change and performance

Table 1 reports the results of the pooled OLS regressions. Model (1) reports the baseline regression result without the key independent variable. The coefficients for the control variables are all statistically significant. The current ratio and the SG&A intensity were positively related to Tobin's *q*, suggesting that the lesser the firm is constrained in cash and the more resources allocated to overhead activities, the higher the firm performance. The proxy for firm size is negatively related to Tobin's *q*, indicating that the larger the firm, the lower the firm performance.

Model (2) in Table 1 reports the results of the pooled OLS regressions in which the independent variable is the strategic change variable, calculated in terms of the cosine distance between a firm's current and previous text vectors. The result indicate that strategic change has a statistically significant ($p < 0.05$) and negative effect on Tobin's *q*. This result suggests that the more the firm changes its strategy from its previous year, the lesser the firm performance in the subsequent year. Thus, the given result lends support to the studies which suggest that strategic change is associated with a negative firm performance (e.g., Sing et al., 1986).

6.2 The relationship between strategic positioning and performance

Next, Models (3) and (4) displays the results of the pooled OLS regressions for the strategic positioning measure. The strategic positioning measure was also calculated in terms of the cosine distance between text vectors. In Model (3), the coefficient on the positioning variable based on cosine distance is positive, but not significant. When including the quadratic terms of each variable in Model (4), however, we find that there is an inverted U-shape relationship between positioning and firm performance (statistically significant at $p < 0.01$ for both the linear

and quadratic terms). Increasing differentiation against industry players is first associated with performance increase, but later associated with performance degradation. This is in line with the standard argument about differentiation that differentiated firms achieve competitive advantage up to a moderate degree (e.g., Porter, 1985).

6.3 The relationship between strategic focus and performance

In Model (5) and (6) of Table 1, we report the results of the pooled OLS regressions for the strategic focus measure and its squared term. As stated earlier, the strategic focus measure is a HHI-like measure of the topic weights in a firm's annual report and thus is bound by 0 and 1. Model (5) estimates the linear effect of strategic focus on Tobin's Q. The coefficient on Focus is positive and significant. Model (6) reports the estimation of the curvilinear effect of strategic focus on performance. The coefficient on the quadratic term is negative and significant ($p<0.01$), while the coefficient on the linear term is positive and significant ($p<0.01$). The inflection point is 0.47 and it is within the range of the focus variable [0,1]. Thus, the results suggest that there is an inverted-U shaped relationship between strategic focus and firm performance. Such result is in line with the argument that firm performance is maximized when firms achieve an optimal level of related diversification (e.g., Zahavi & Lavie, 2013).

6.4 Robustness tests

To test the robustness of the regressions results, we ran regressions with return on assets (ROA) as the dependent variable. The results are qualitatively similar to the results reported in Table 1. We also ran several sensitivity tests of the sample of firms used in the regression. We tested whether limiting our sample to the firms listed in the Russell 3000 and the S&P 500 or whether including only the business descriptions with a sizeable length (i.e., more than 5,000 words) changed the results. Again, the results were qualitatively similar to the results reported in Table 1. Thus, we find additional evidence that our text-based measures of strategic change and positioning reflect the underlying constructs, by showing that the measures support the performance implications of change, differentiation, and focus as discussed in the literature.

7. DISCUSSION

7.1 Measuring strategy

In this paper, we demonstrated that we can utilize the qualitative information in unstructured textual data and create novel measures of strategy using two NLP techniques, called vector space models and topic modeling. By applying new text analytic techniques to the annual reports of U.S. firms, we created text-based measures of three core strategy concepts: strategic change, positioning, and focus. Examples and statistical results show that our measures of the three strategy constructs reflect relatively well the purported characteristics of firm strategy. In particular, we find evidence that examples of our measures closely correspond to descriptive accounts of several firm histories. We also statistically test the relationship between our measures and firm performance, and find support that each measure is significantly associated with future firm profitability as predicted by prior theoretical arguments.

While a nascent stream of studies in the management literature has started to use NLP tools to utilize textual data (e.g., Kaplan & Vakili, 2015; Arts, Cassiman, & Gomez, 2018), its usage has been mostly limited to the study of technology by analyzing patent abstracts. To our knowledge, no study has yet applied the tools to study the core questions in strategy, such as those related to strategic change, positioning, and focus. To be clear, we do not intend to suggest that what we have done here is a robust empirical testing of prior theories, but rather to suggest that our text-based measures reflect the underlying strategy concepts. We expect that the use of our methods can serve as a complement, not a substitute, for the existing methods found in prior empirical studies of strategy.

More generally, we believe that the approach presented here can be seen as the first steps towards creating constructs that can “measure” some crucial aspects of strategy in a way that can cut across industries and geographies. Hopefully, this will trigger further such attempts to translate rich constructs that have heretofore remained in the conceptual-theoretical domain to the empirical one. This should in turn allow us to empirically test some of the core claims in the field of strategic management, as well as explore novel ones.

7.2 Links to other streams of management research

We believe that the methods introduced in this paper would also be useful for various other inquiries in management research. Recently, strategy researchers have started to actively examine textual data to investigate questions such as how firms strategically communicate with external constituents (Guo, Yu, & Gimeno, 2017), how market agents engage in a discourse around new technologies (Kahl & Grodal, 2016), and how firm executives impress investors through verbal presentations (Whittington, Yakis-Douglas, & Ahn, 2016). As evidenced in these efforts, NLP techniques can be a powerful tool to the study of communication, cognition, representation, culture, and institutions.

Using the two methods introduced in this paper, researchers may examine other types of texts (e.g., press releases, conference call transcripts, and news or magazine articles) to identify the linguistic contents in the respective text data and create measures of constructs to answer different kinds of questions. For instance, given a corpus of texts that records a continuous flow of events (e.g., a diary of project development or a set of news articles), one may utilize the vector space model and create similarity measures of texts to examine how an event has progressed over time.

7.3 Template for introducing machine learning into strategy

Although we have here introduced two NLP techniques that efficiently and effectively analyze large textual data sets, other NLP techniques can also be explored and applied to analyze unstructured textual data. In fact, the two techniques introduced here are some of the simplest from the NLP domain. Many more techniques exist, such as named entity recognition, relationship extraction, text classification, sentiment analysis, and semantic networks, which extract other types of information from texts. For example, using semantic networks, researchers can automatically extract relationships between concepts underlying a text and create network representations of the concepts. Such techniques can be applied after the preprocessing steps we have covered in this paper.

Moreover, in recent years, new NLP applications that incorporate machine learning frameworks have been introduced. Though many applications are focused on addressing traditional NLP tasks, such as text classification, using neural network models, vastly original methods that extract new types of textual information also exist. For instance, word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) creates vector representations of words that capture a syntactic and semantic relationships between words. Such methods provide opportunities for researchers to extract even deeper relational information within a text. We concur with the statement that new methods "allow scholars to address old questions in new ways and to investigate questions that were not tractable using existing methods" (Arora, Gittelman, Kaplan, Lynch, Mitchell, & Siggelkow, 2016: p.3).

Our hope is that this paper can serve as a template for future work on how increasingly sophisticated machine learning techniques can be introduced into strategic management research, and how it can help shed light on some of the most fundamental questions in the field.

REFERENCES

Amburgey, T. L., D. Kelly, W. P. Barnett. 1993. Resetting the clock: The dynamics of organizational change and failure. ***Administrative Science Quarterly***. 38: 51–73.

Amihud, Y., & Lev, B. 1981. Risk reduction as a managerial motive for conglomerate mergers. ***The Bell Journal of Economics***, 605-617.

Arora, A., Gittelman, M., Kaplan, S., Lynch, J., Mitchell, W., & Siggelkow, N. 2016. Question-based innovations in strategy research methods. ***Strategic Management Journal***, 37(1): 3-9.

Arts, S., Cassiman, B., & Gomez, J. C. 2018. Text matching to measure patent similarity. ***Strategic Management Journal***, 39(1): 62-84.

Axelrod, R. 1976. ***Structure of decision: The cognitive maps of political elites***. Princeton university press.

Barr, P. S., Stimpert, J. L., & Huff, A. S. 1992. Cognitive change, strategic action, and organizational renewal. ***Strategic Management Journal***, 13(S1): 15-36.

Baum, J. A., & Korn, H. J. 1996) Competitive dynamics of interfirm rivalry. ***Academy of Management Journal***, 39(2): 255-291.

Bettman, J. R., & Weitz, B. A. 1983. Attributions in the board room: Causal reasoning in corporate annual reports. ***Administrative Science Quarterly***, 165-183.

Blei, D. M. 2012. Probabilistic topic models. ***Communications of the ACM***, 55(4), 77-84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. 2003. Latent Dirichlet allocation. ***Journal of Machine Learning Research***, 3(Jan): 993-1022.

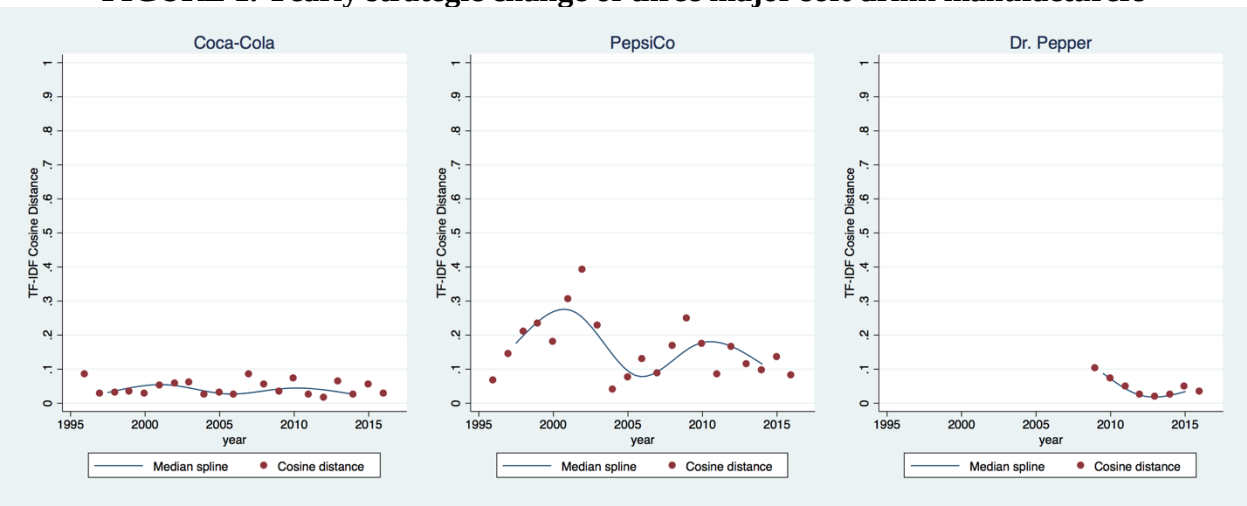
Bowman, E. H. 1978. Strategy, annual reports, and alchemy. ***California Management Review***, 20(3): 64-71.
Bowman, E. H. 1984. Content analysis of annual reports for corporate strategy and risk. ***Interfaces***, 14: 61–71.
Boyd, B. K., Dess, G. G., & Rasheed, A. M. 1993. Divergence between archival and perceptual measures of the environment: Causes and consequences. ***Academy of Management Review***, 18(2): 204-226.
Brandenburger, A. M., & Stuart, H. W. 1996. Value-based business strategy. ***Journal of Economics & Management Strategy***, 5(1): 5-24.
Carley K, Palmquist M. 1992. Extracting, representing and analyzing mental models. ***Social Forces***, 70: 601–636.
Carley, K. 1993. Coding choices for textual analysis: A comparison of content analysis and map analysis. ***Sociological Methodology***, 75-126.
Caves, R. E., & Porter, M. E. 1977. From entry barriers to mobility barriers: Conjectural decisions and contrived deterrence to new competition. ***The Quarterly Journal of Economics***, 241-261.
Chatterjee, S., & Blocher, J. D. 1992. Measurement of firm diversification: Is it robust? ***Academy of Management Journal***, 35(4): 874-888.
Chen, M. J. 1996. Competitor analysis and interfirm rivalry: Toward a theoretical integration. ***Academy of Management Review***, 21(1): 100-134.
Cho, T. S., & Hambrick, D. C. 2006. Attention as the mediator between top management team characteristics and strategic change: The case of airline deregulation. ***Organization Science***, 17(4): 453-469.
Deephouse, D. L. (1999). To be different, or to be the same? It's a question (and theory) of strategic balance. ***Strategic Management Journal***, 20(2): 147-166.
Denny, M. J., & Spirling, A. 2016. **Assessing the consequences of text preprocessing decisions**. Unpublished manuscript, Stanford University.
Duriau, V. J., Reger, R. K., & Pfarrer, M. D. 2007. A content analysis of the content analysis literature in organization studies: Research themes, data sources, and methodological refinements. ***Organizational Research Methods***, 10(1): 5-34.
Dutton, J. E., & Duncan, R. B. 1987. The creation of momentum for change through the process of strategic issue diagnosis. ***Strategic Management Journal***, 8(3): 279-295.
Edwards, C. D. 1955. Conglomerate bigness as a source of power. In Rosenbluth, G. (Ed). ***Business concentration and price policy,*** 331-359. Princeton University Press.
Finkelstein, S., & Hambrick, D. C. 1990. Top-management-team tenure and organizational outcomes: The moderating role of managerial discretion. ***Administrative Science Quarterly***, 484-503.
Fiol, C. M. 1989. A semiotic analysis of corporate language: Organizational boundaries and joint venturing. ***Administrative Science Quarterly***, 277-303.
Gavetti, G., & Rivkin, J.W. 2007. On the origin of strategy: Action and cognition over time. ***Organization Science***, 18(3): 420–439.
Ghemawat, P. 1991. ***Commitment: The Dynamic of Strategy***. Free Press, New York.
Ginsberg, A. 1988. Measuring and modelling changes in strategy: Theoretical foundations and empirical directions. ***Strategic Management Journal***, 9(6): 559-575.
Gioia, D. A., & Chittipeddi, K. 1991. Sensemaking and sensegiving in strategic change initiation. ***Strategic Management Journal***, 12(6): 433-448.

Greve, H. R. 1998. Performance, aspirations, and risky organizational change. ***Administrative Science Quarterly***, 58-86.
Grimmer, J., & Stewart, B. M. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. ***Political Analysis***, 21(3): 267-297.
Guo, W., Yu, T., & Gimeno, J. 2017. Language and Competition: Communication Vagueness, Interpretation Difficulty, and Market Entry. ***Academy of Management Journal***, 60(6): 2073-2098.
Hannan, M. T., & Freeman, J. 1984. Structural inertia and organizational change. ***American Sociological Review***, 149-164.
Hasan, S., Ferguson, J. P., & Koning, R. 2015. The lives and deaths of jobs: Technical interdependence and survival in a job structure. ***Organization Science***, 26(6): 1665-1681.
Hatten, K. J., & Hatten, M. L. 1987. Strategic groups, asymmetrical mobility barriers and contestability. ***Strategic Management Journal***, 8(4): 329-342.
Haveman, H. A. 1992. Between a rock and a hard place: Organizational change and performance under conditions of fundamental environmental transformation. ***Administrative Science Quarterly***, 48-75.
Hoberg, G., & Phillips, G. 2010. Product market synergies and competition in mergers and acquisitions: A text-based analysis. ***The Review of Financial Studies***, 23(10): 3773-3811.
Hoberg, G., & Phillips, G. 2016. Text-based network industries and endogenous product differentiation. ***Journal of Political Economy***, 124(5): 1423-1465.
Hoskisson, R. E., Hitt, M. A., Johnson, R. A., & Moesel, D. D. 1993. Construct validity of an objective (entropy) categorical measure of diversification strategy. ***Strategic Management Journal***, 14(3): 215-235.
Hotelling, H. 1929. Stability in competition. ***The Economic Journal***, 39(153): 41-57.
Huff, A. S. 1990. ***Mapping strategic thought***, John Wiley & Sons.
Huff, A. S., Narapareddy, V., & Fletcher, K. E. 1990. Coding the causal association of concepts. In Huff, A. S. (Ed), ***Mapping strategic thought***, John Wiley & Sons, 311-325.
Jensen, M. C. 1986. Agency costs of free cash flow, corporate finance, and takeovers. ***The American Economic Review***, 76(2): 323-329.
Jurafsky, D., & Martin, J. H. 2014. ***Speech and language processing*** (Vol. 3). London: Pearson.
Kabanoff, B. 1997. Introduction: computers can read as well as count: computer-aided text analysis in organizational research. ***Journal of Organizational Behavior***, 507-511.
Kahl, S. J., & Grodal, S. 2016. Discursive strategies and radical technological change: Multilevel discourse analysis of the early computer (1947–1958). ***Strategic Management Journal***, *37*(1): 149-166.
Kaplan, Andrew (2010). What PepsiCo hopes to gain from the merger with its two largest bottlers. ***Beverage World***. April: 20-24.
Kaplan, S. 2008. Cognition, capabilities, and incentives: Assessing firm response to the fiber-optic revolution. ***Academy of Management Journal***, 51(4): 672-695.
Kaplan, S., & Vakili, K. 2015. The double-edged sword of recombination in breakthrough innovation. ***Strategic Management Journal***, 36(10): 1435-1457.
Kelly, D., & Amburgey, T. L. 1991. Organizational inertia and momentum: A dynamic model of strategic change. ***Academy of Management Journal***, 34(3): 591-612.
Knoke, D., & Kuklinski, J. H. 1982. ***Network analysis***. Sage: Newbury Park, CA.

Kraatz, M. S., & Zajac, E. J. 1996. Exploring the limits of the new institutionalism: The causes and consequences of illegitimate organizational change. ***American Sociological Review***, 812-836.
Lang, L. H., & Stulz, R. M. 1994. Tobin's q, corporate diversification, and firm performance. ***Journal of Political Economy***, 102(6): 1248-1280.
Lant, T. K., & Mezias, S. J. 1992. An organizational learning model of convergence and reorientation. ***Organization Science***, 3(1): 47-71.
Lee, J., Lee, K., & Rho, S. 2002. An evolutionary perspective on strategic group emergence: a genetic algorithm‐based model. ***Strategic Management Journal***, 23(8): 727-746.
Manning, C. D., & Schütze, H. 1999. ***Foundations of statistical natural language processing***. MIT press.
Manning, C.D., Raghavan, P., & Schütze, H. 2008. ***An introduction to information retrieval***. Cambridge University Press.
March, J. G. 1981. Footnotes on organizational change. ***Administrative Science Quarterly***. 26: 563–577.
March, J. G., & Simon, H. A. 1958. ***Organizations***. John Wiley & Sons: New York.
Markides, C. C., & Williamson, P. J. 1994. Related diversification, core competences and corporate performance. ***Strategic Management Journal***, 15(S2): 149-165.
Menon, A. R., & Yao, D. A. 2017. Elevating Repositioning Costs: Strategy Dynamics and Competitive Interactions. ***Strategic Management Journal***. 38(10): 1953-1963.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. 2013. Distributed representations of words and phrases and their compositionality. ***Advances in Neural Information Processing Systems***. 26: 3111-3119.
Montgomery, C. A. 1985. Product-market diversification and market power. ***Academy of Management Journal***, 28(4): 789-798.
Montgomery, C. A. 1994. Corporate diversification. ***The Journal of Economic Perspectives***, 8(3): 163-178.
Palich, L. E., Cardinal, L. B., & Miller, C. C. 2000. Curvilinearity in the diversification–performance linkage: an examination of over three decades of research. ***Strategic Management Journal***, 21(2): 155-174.
Penrose, E. T. 1959. ***The theory of the growth of the firm***. Wiley, New York.
Peteraf, M. A. 1993. Intra‐industry structure and the response toward rivals. ***Managerial and Decision Economics***, 14(6): 519-528.
Porter, M. E. 1980. ***Competitive Strategy: Techniques for Analyzing Industries and Competitors***. Free Press, New York.
Porter, M. E. 1985. ***Competitive Advantage: Creating and Sustaining Superior Performance***. Free Press, New York.
Porter, M. E. 1991. Towards a dynamic theory of strategy. ***Strategic Management Journal***, 12(S2): 95-117.
Porter, M. E. 1996. What is strategy? ***Harvard Business Review***. 74(6): 61–78.
Rajagopalan, N., & Spreitzer, G. M. 1997. Toward a theory of strategic change: A multi-lens perspective and integrative framework. ***Academy of Management Review***, 22(1): 48-79.
Ramanujam, V., & Varadarajan, P. 1989. Research on corporate diversification: A synthesis. ***Strategic Management Journal***, 10(6): 523-551.
Ravenscraft, D. J., & Scherer, F. M. 1987. Life after takeover. ***The Journal of Industrial Economics***, 147-156.

Rumelt, R. P. 1982. Diversification strategy and profitability. ***Strategic Management Journal***, 3(4): 359-369.
Rumelt, R. P., Schendel, D., & Teece, D. J. 1994. ***Fundamental issues in strategy: A research agenda***. Harvard Business Press.
Salton, G. & Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. ***Information Processing and Management***. 24(5): 513–523
Siggelkow, N. 2001. Change in the presence of fit: The rise, the fall, and the renaissance of Liz Claiborne. ***Academy of Management Journal***, 44(4): 838-857.
Siggelkow, N. 2002. Evolution toward fit. ***Administrative Science Quarterly***, 47(1): 125-159.
Siggelkow, N. 2003. Why Focus? A study of intra- industry focus effects. ***The Journal of Industrial Economics***, 51(2): 121-150.
Singh, J. V., House, R. J., & Tucker, D. J. 1986. Organizational change and organizational mortality. ***Administrative Science Quarterly***, 587-611.
Singh, J. V., House, R. J., & Tucker, D. J. 1986. Organizational change and organizational mortality. ***Administrative Science Quarterly***, 587-611.
Smith, K. G., Grimm, C. M., Wally, S., & Young, G. 1997. Strategic groups and rivalrous firm behavior: Towards a reconciliation. ***Strategic Management Journal***, 149-157.
Tushman, M. L., & Romanelli, E. 1985. Organizational Evolution: A Metamorphosis Model of Conveyance and Reorientation," In L. Cummings & B. Staw (Eds.), ***Research in Organizational Behavior***. Vol. 7, Greenwich, CT: JAI Press, 177-222.
Villalonga, B. 2004. Diversification discount or premium? New evidence from the business information tracking series. ***The Journal of Finance***, *59*(2), 479-506.
Virany, B., Tushman, M. L., & Romanelli, E. 1992. Executive succession and organization outcomes in turbulent environments: An organization learning approach. ***Organization Science***, 3(1): 72-91.
Weber, R. P. 1990. ***Basic content analysis*** (No. 49). Sage.
Wernerfelt, B. 1984. A resource-based view of the firm. ***Strategic Management Journal***, 5(2): 171-180.
Whittington, R., Yakis-Douglas, B., & Ahn, K. 2016. Cheap talk? Strategy pr esentations as a form of chief executive officer impression management. ***Strategic Management Journal***, *37*(12): 2413-2424.
Wernerfelt, B., & Montgomery, C. A. 1988. Tobin's q and the importance of focus in firm performance. ***The American Economic Review***, 78(1): 246-250.
Zahavi, T., & Lavie, D. 2013. Intra- industry diversification and firm performance. ***Strategic Management Journal***, 34(8): 978-998.
Zajac, E. J., & Kraatz, M. S. 1993. A diametric forces model of strategic change: Assessing the antecedents and consequences of restructuring in the higher education industry. ***Strategic Management Journal***, 14(S1): 83-102.

FIGURE 1. Yearly strategic change of three major soft drink manufacturers



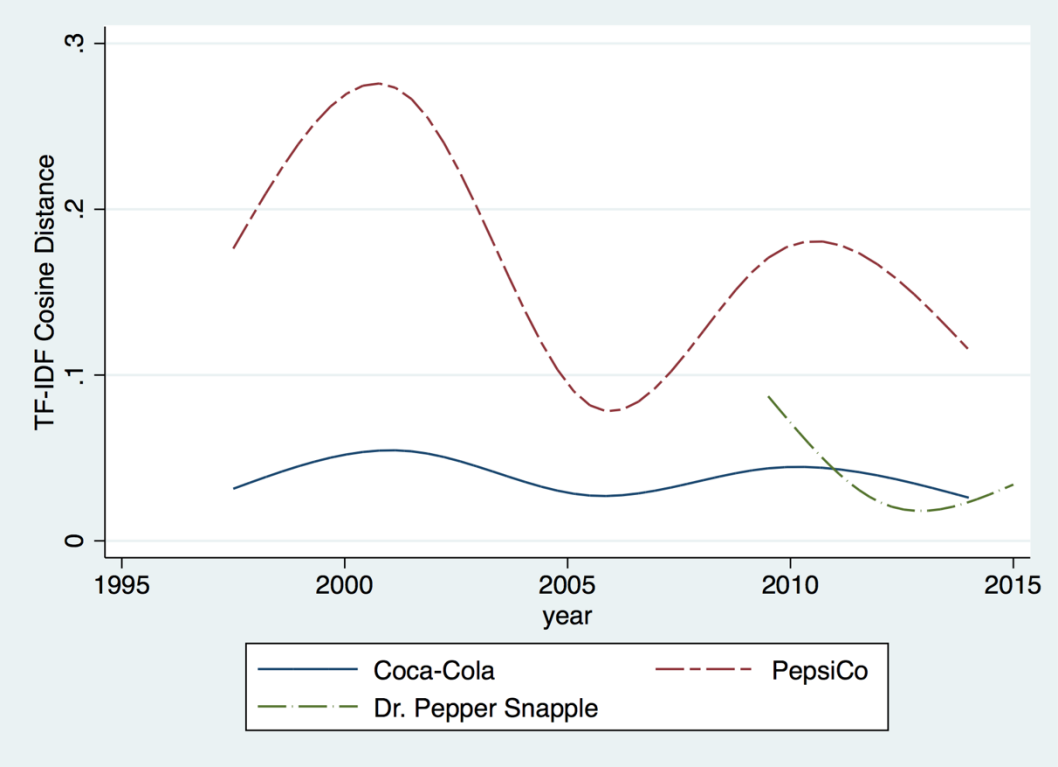
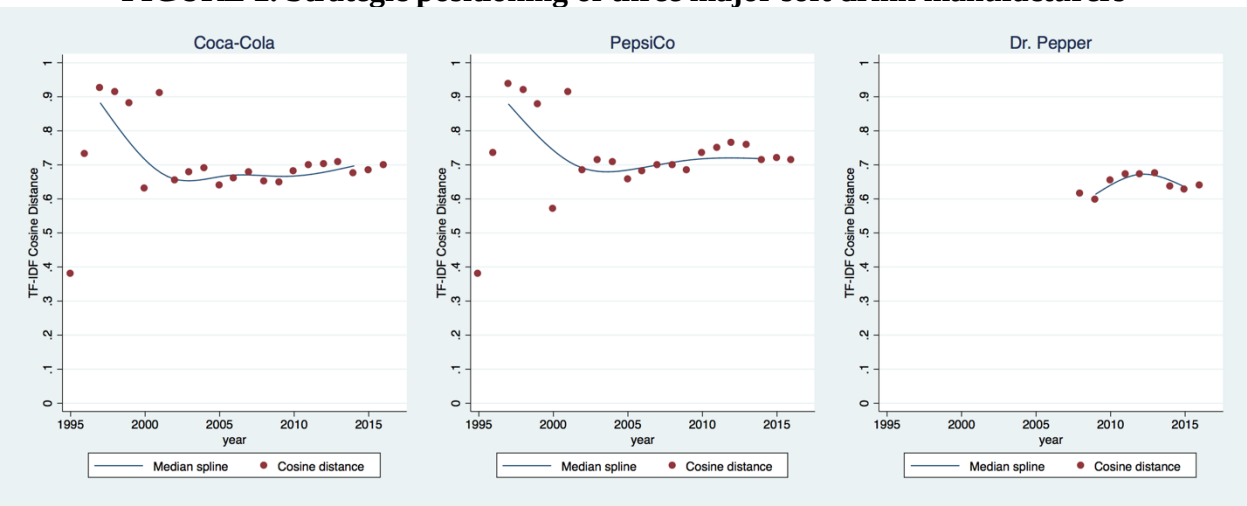


FIGURE 2. Strategic positioning of three major soft drink manufacturers



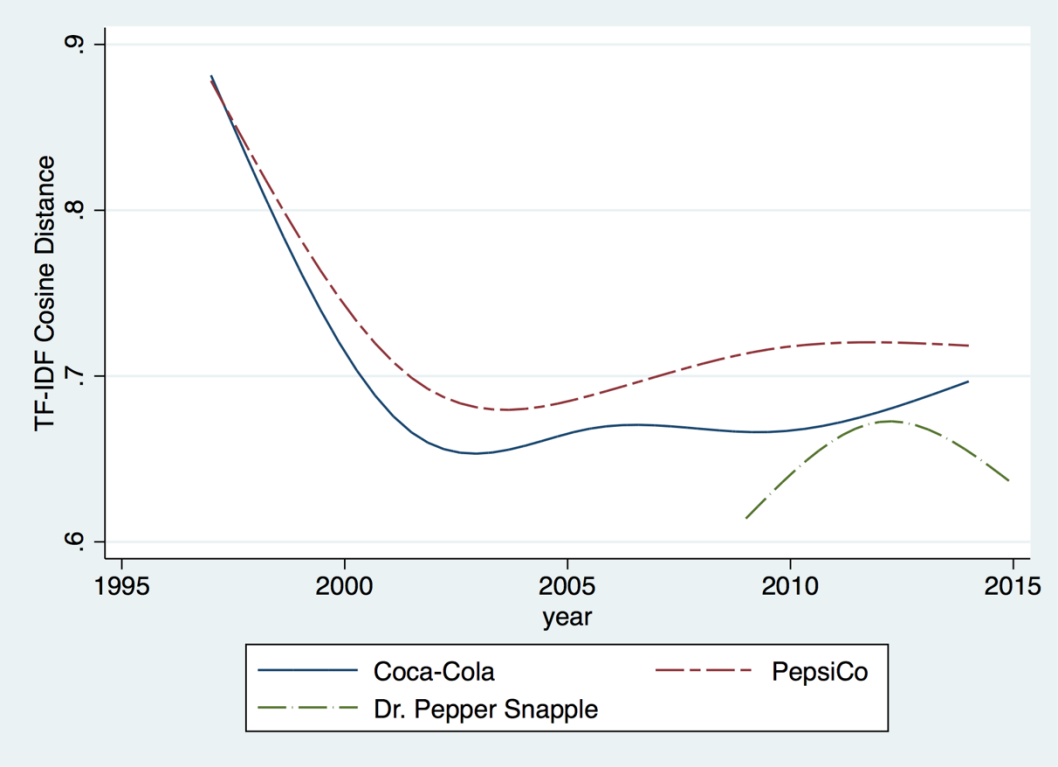
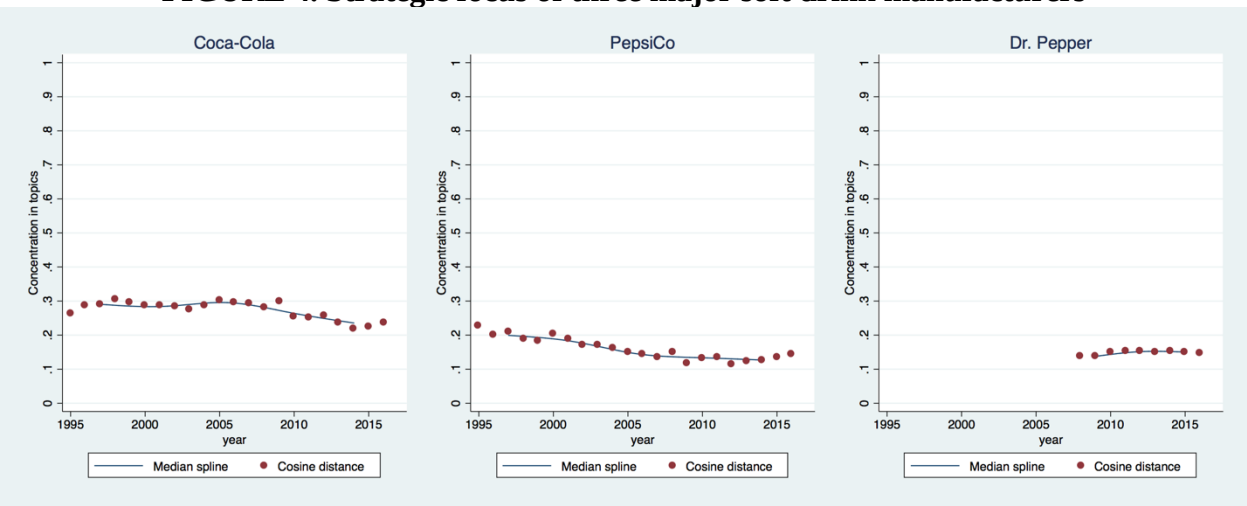


FIGURE 3. Principal component analysis of three major soft drink manufacturers' tf-idf text vectors

FIGURE 4. Strategic focus of three major soft drink manufacturers



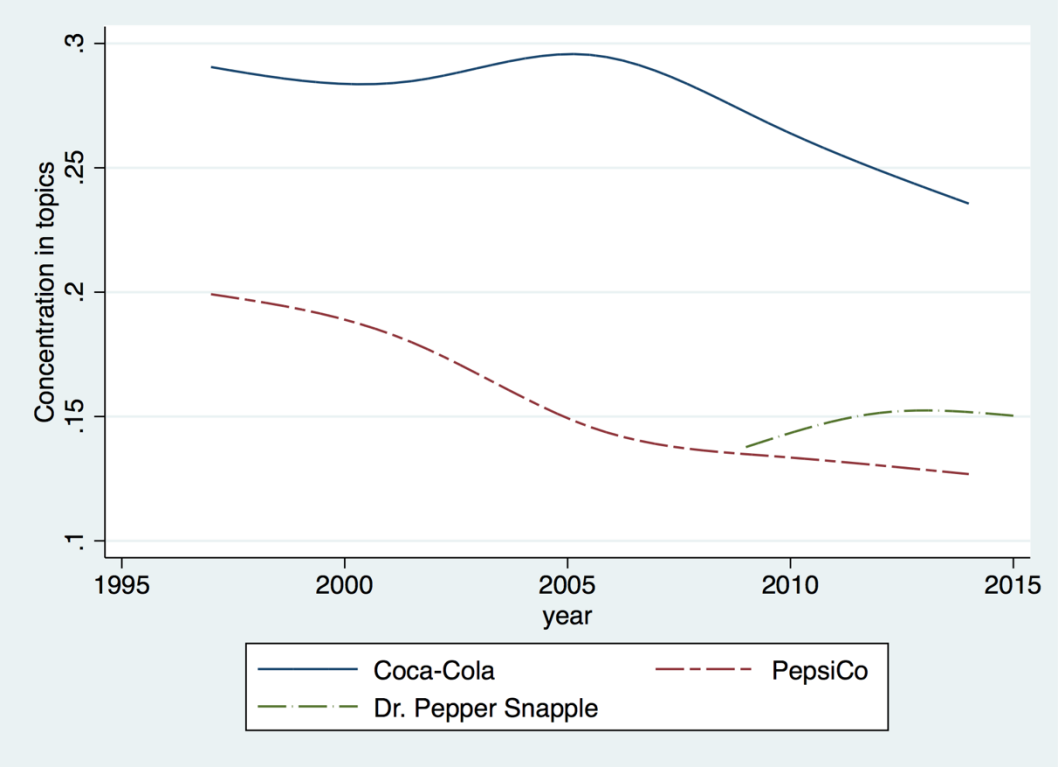


FIGURE 5. Yearly strategic change of six major airline firms

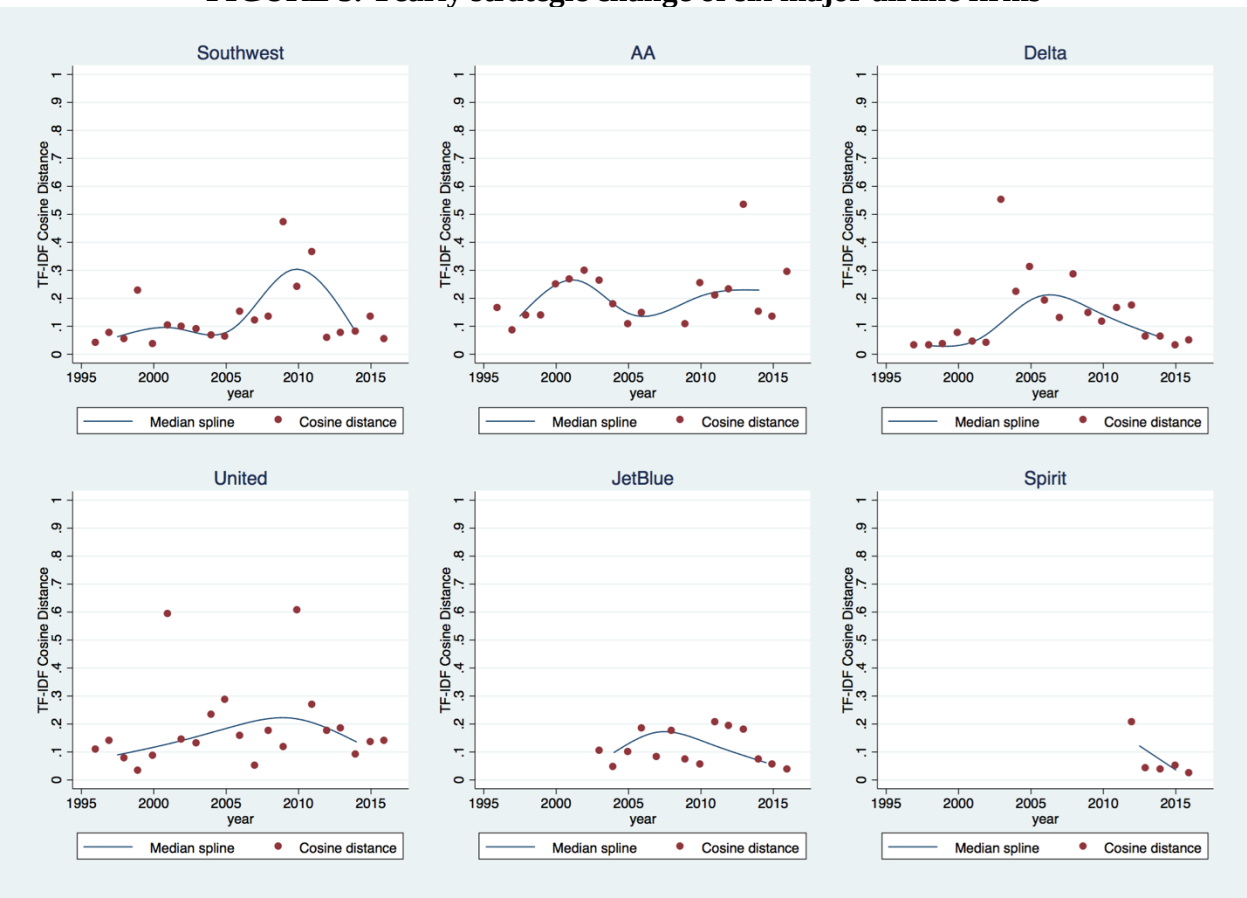
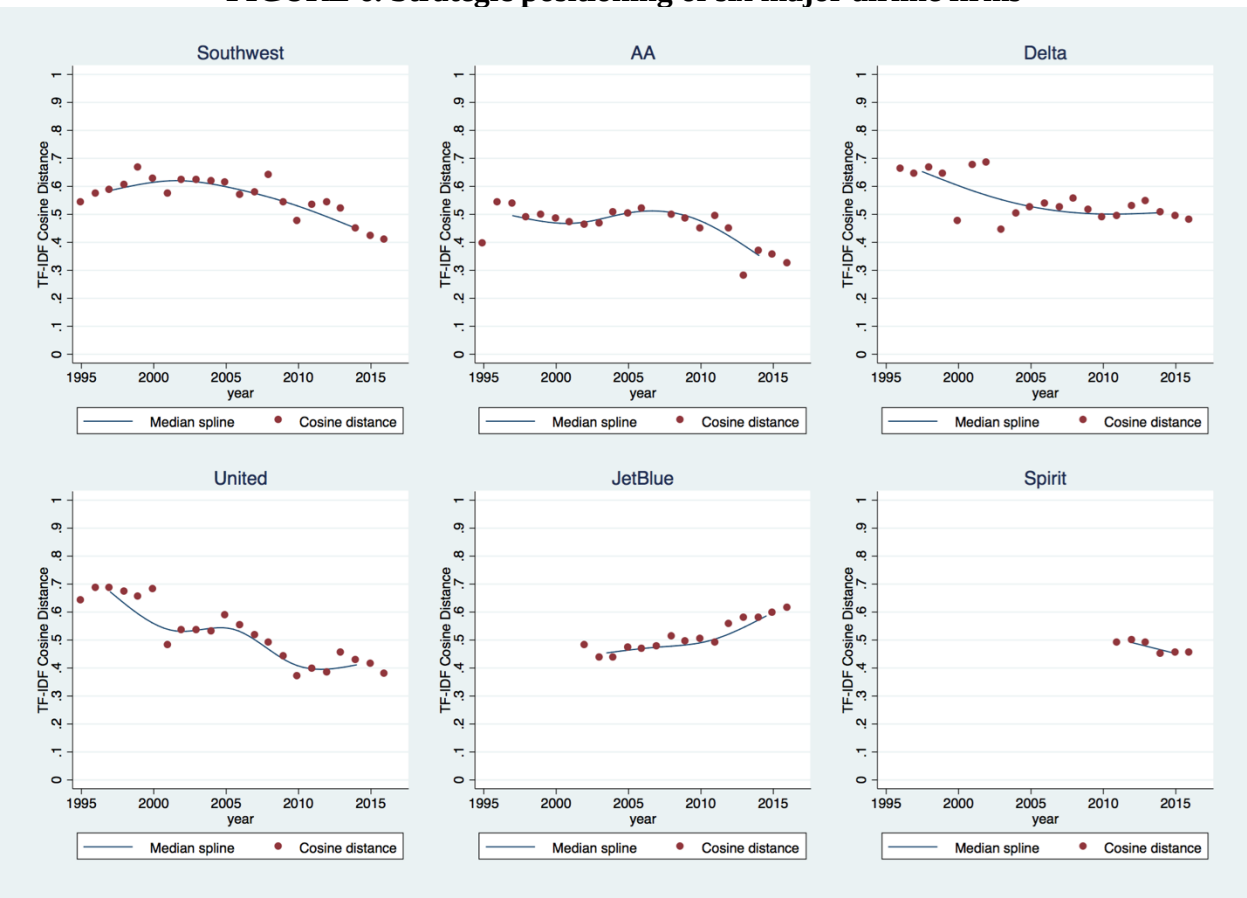


FIGURE 6. Strategic positioning of six major airline firms



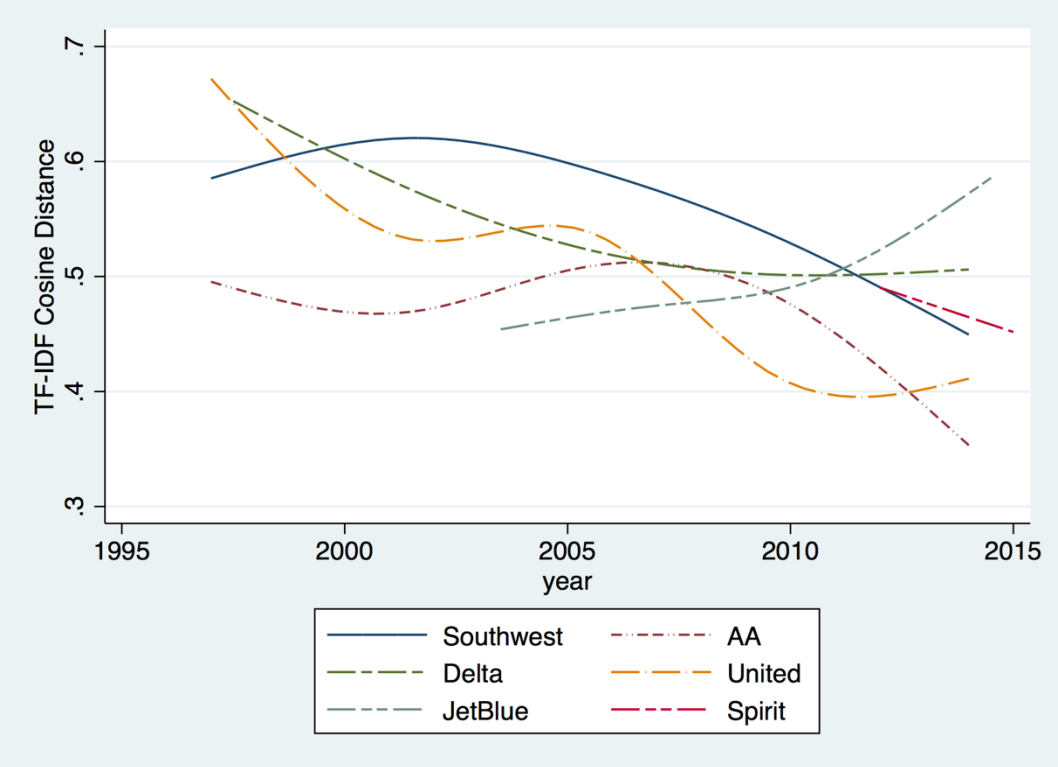


FIGURE 7. Principal component analysis of six major airline firms' tf-idf text vectors

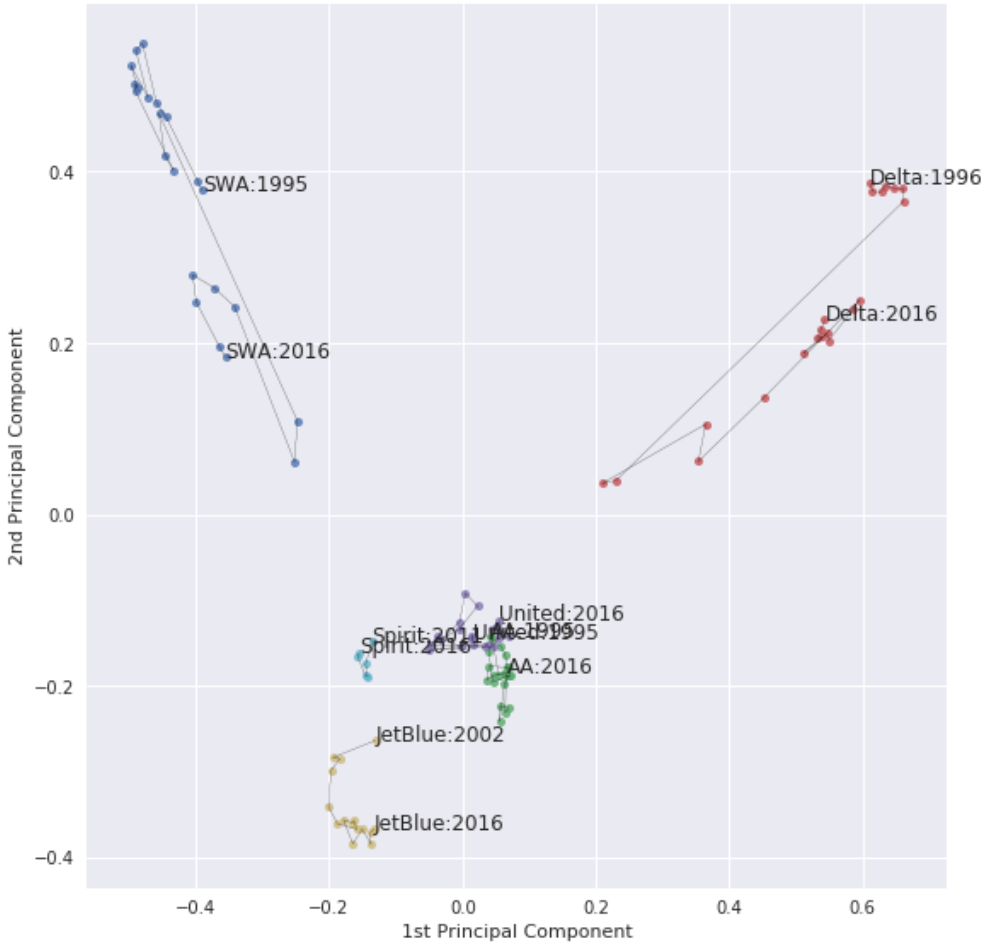
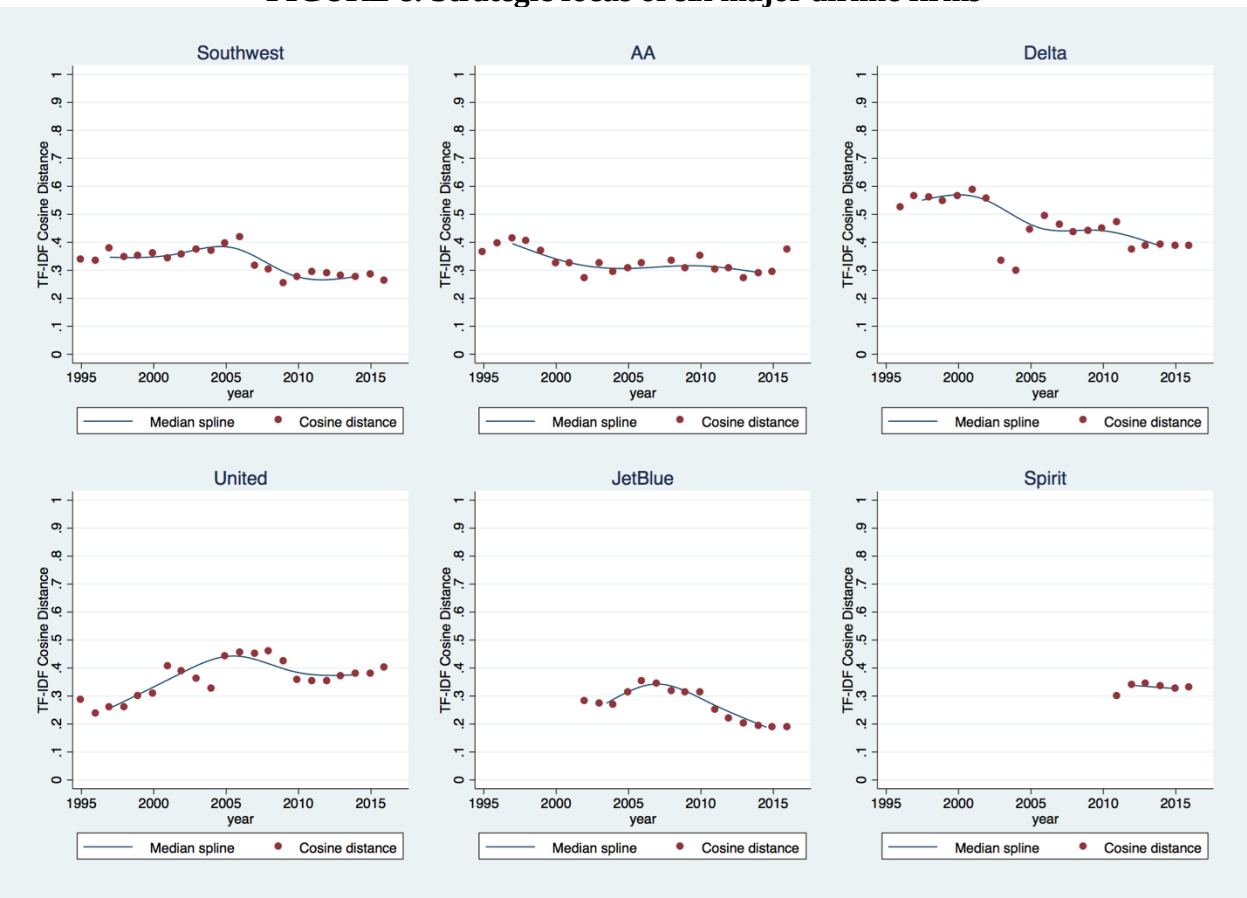


FIGURE 8. Strategic focus of six major airline firms



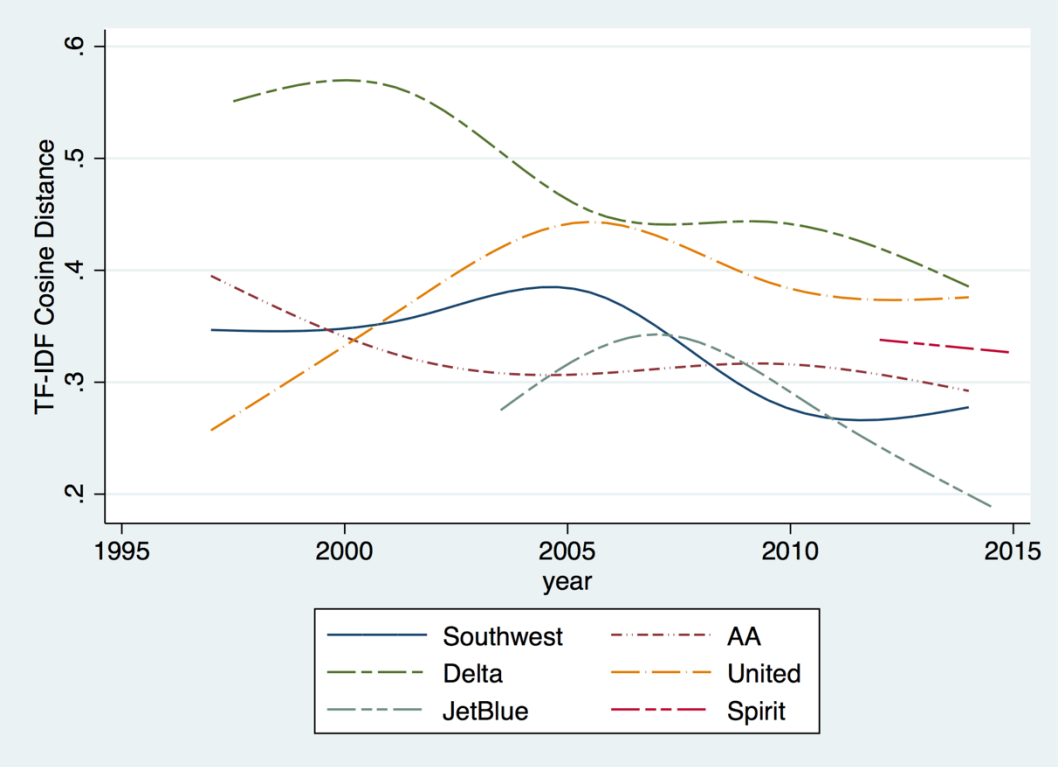
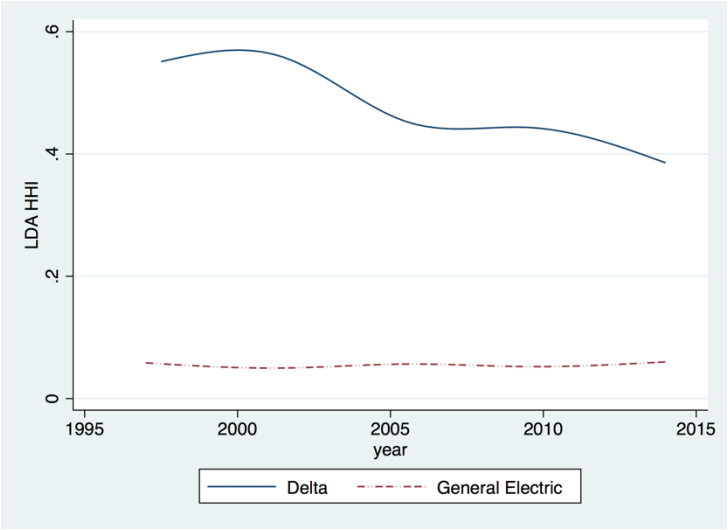


FIGURE 9. Topical distribution of Delta Air Lines and General Electrics



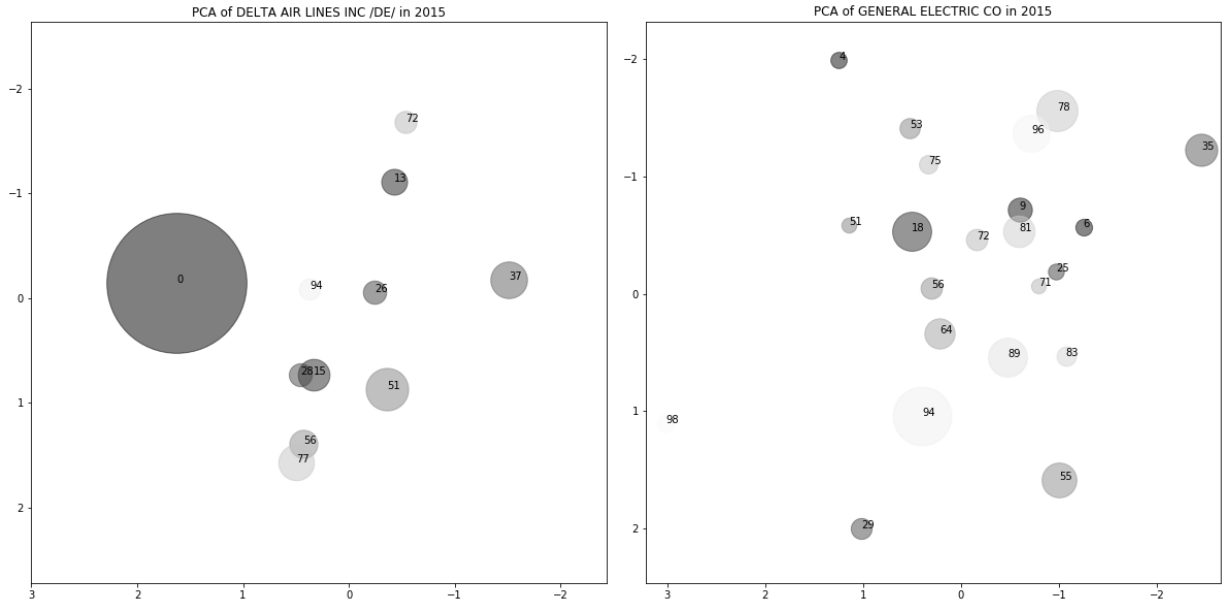


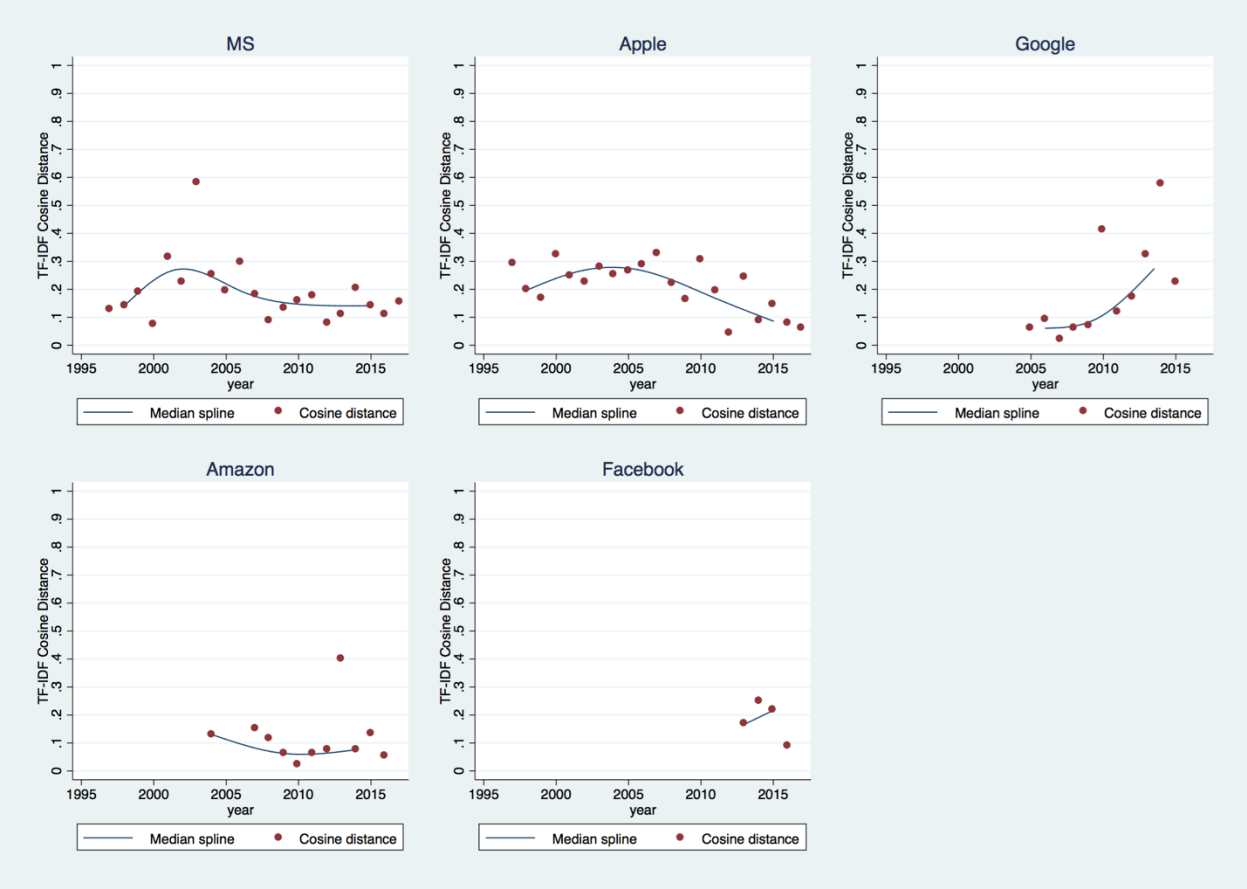
TABLE 1. Pooled OLS regression of strategic change, differentiation, and focus on Tobin's Q

	(1)	(2)	(3)	(4)	(5)	(6)
Change $_t$ (Cosine)		-0.086**				
		(0.039)				
Differentiation $_t$ (Cosine)			0.072	0.561***		
			(0.053)	(0.169)		
Differentiation $_t^2$ (Cosine)				-0.418***		
				(0.156)		
Focus $_t$ (Cosine)					0.877***	1.559***
					(0.095)	(0.246)
Focus $_t^2$ (Cosine)						-1.478***
						(0.499)
Current Ratio $_t$	0.005**	0.005**	0.005**	0.005**	0.005*	0.005*
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
ln(Total sales t)	-0.012***	-0.012***	-0.012***	-0.012***	-0.013***	-0.013***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
SGA intensity t	0.003***	0.003***	0.003***	0.003***	0.003***	0.003***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Constant	2.143***	2.159***	2.092***	1.962***	2.000***	1.939***
	(0.045)	(0.045)	(0.058)	(0.067)	(0.047)	(0.051)
Observations	52,392	52,392	52,392	52,392	52,392	52,392
SIC4 FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Appendix

FIGURE A1. Yearly strategic change of five major tech firms



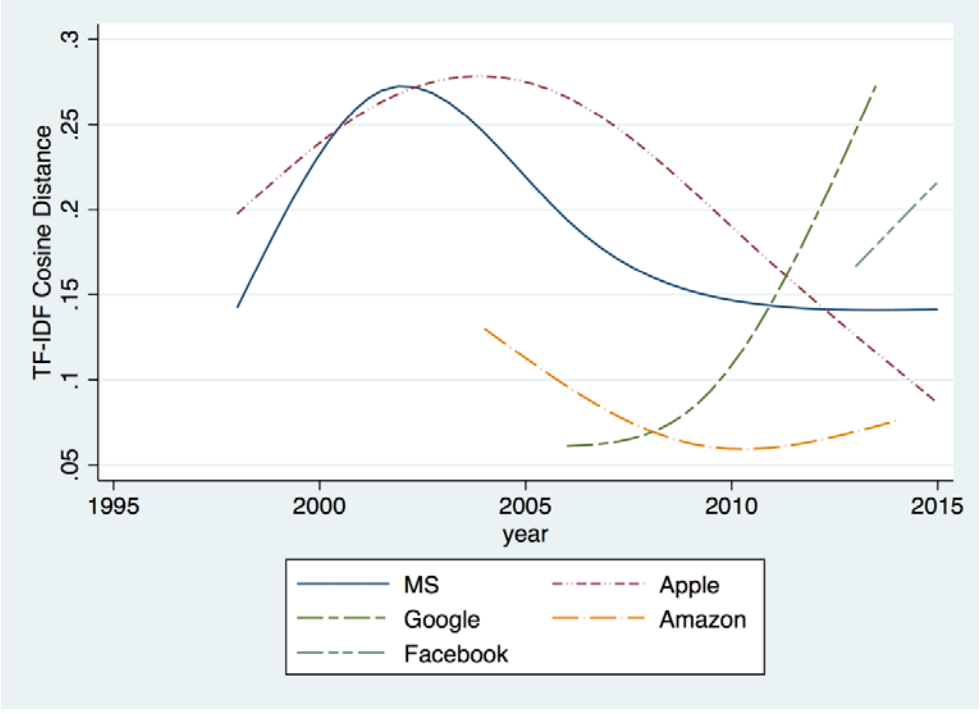
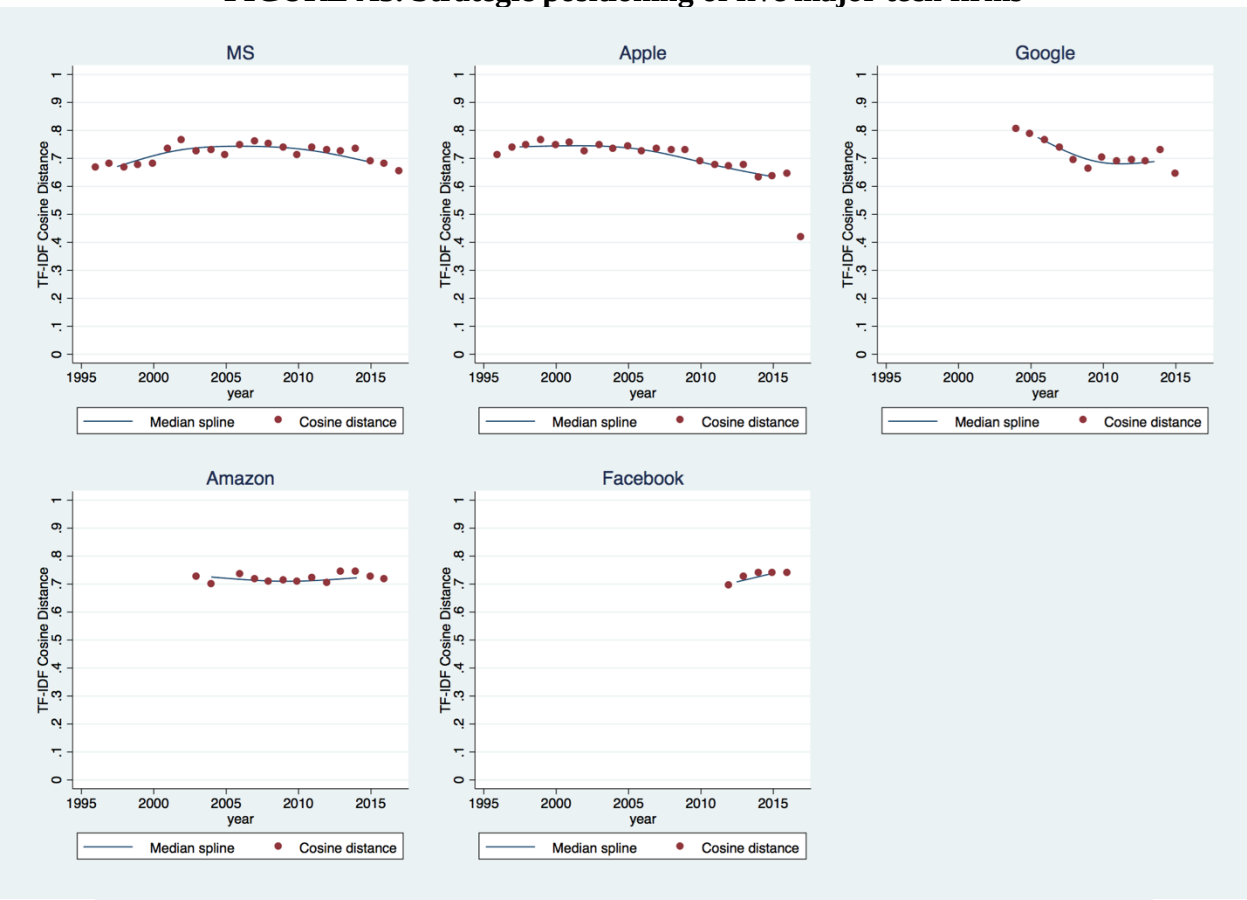


FIGURE A3. Strategic positioning of five major tech firms



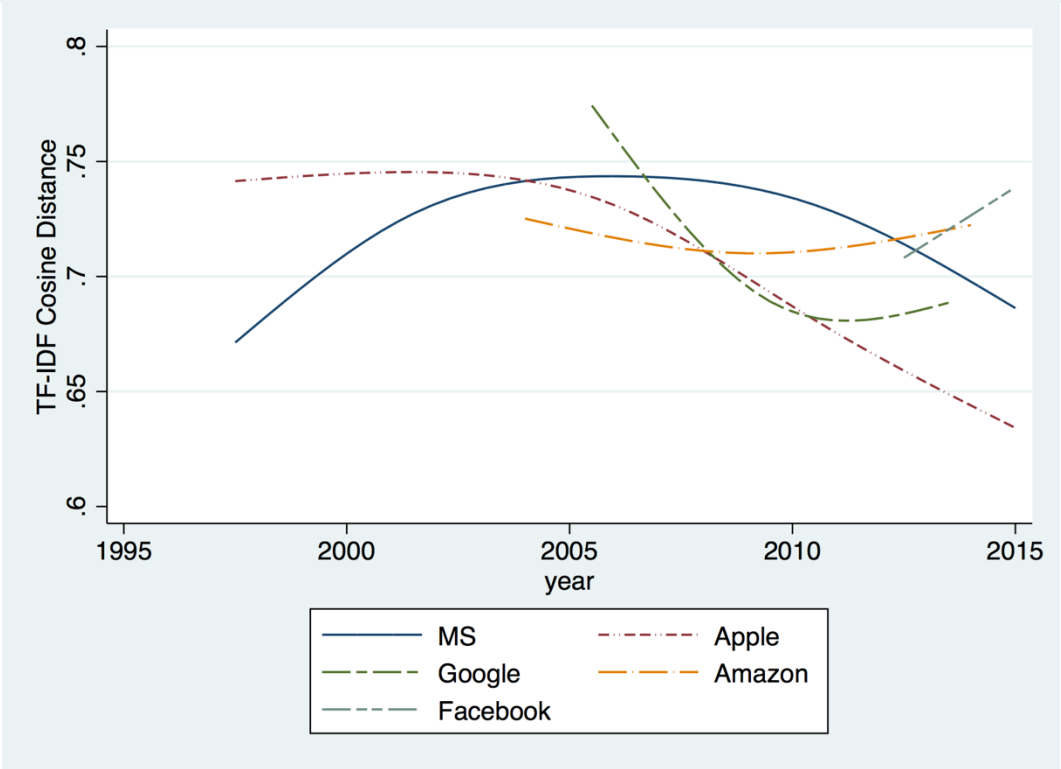


FIGURE A3. Principal component analysis of five major tech firms' tf-idf text vectors

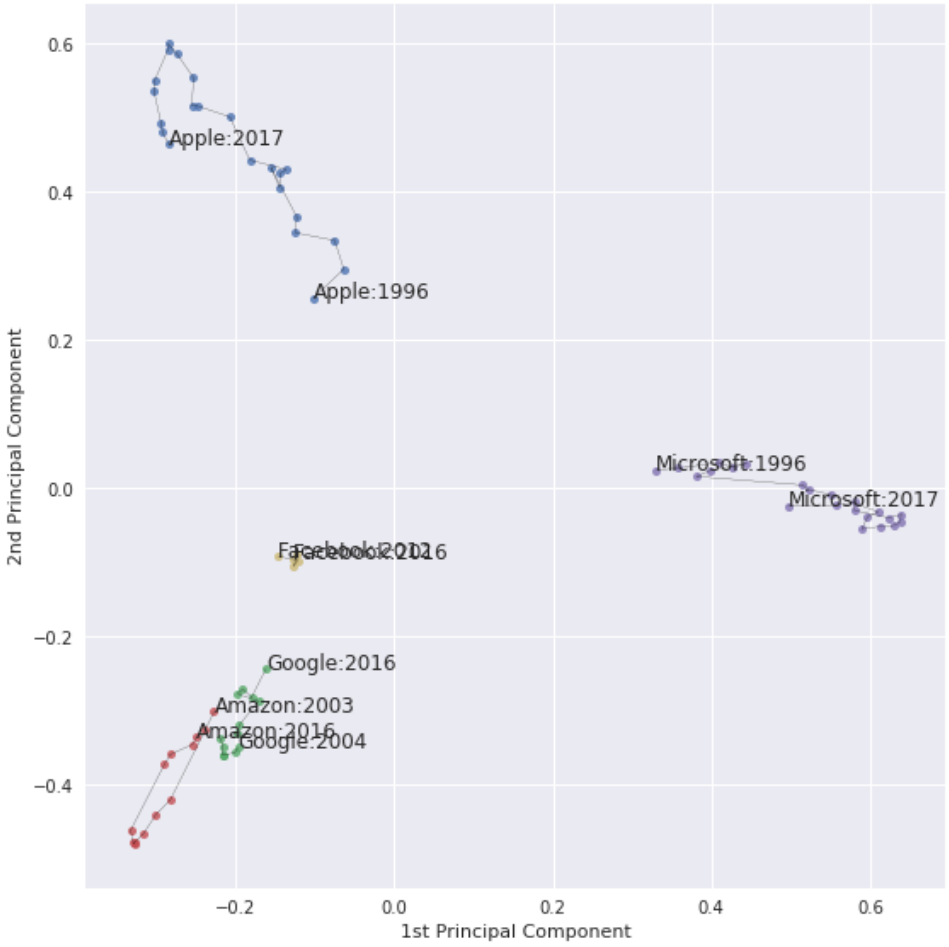


FIGURE A4. Strategic focus of five major tech firms