MARKETING SCIENCE
Vol. 31, No. 3, May–June 2012, pp. 521–543
ISSN 0732-2399 (print) | ISSN 1526-548X (online)

http://dx.doi.org/10.1287/mksc.1120.0713

Mine Your Own Business: Market-Structure Surveillance Through Text Mining

Oded Netzer
Graduate School of Business, Columbia University, New York, New York 10027, on2110@columbia.edu

Ronen Feldman
School of Business Administration, Hebrew University of Jerusalem, Mount Scopus, Jerusalem, Israel 91905, ronen.feldman@huji.ac.il

Jacob Goldenberg
School of Business Administration, Hebrew University of Jerusalem, Mount Scopus, Jerusalem, Israel 91905; and Columbia Business School, New York, New York 10027, msgolden@huji.ac.il

Moshe Fresko
Jerusalem, Israel 91905, freskom@gmail.com

Web 2.0 provides gathering places for Internet users in blogs, forums, and chat rooms. These gathering places leave footprints in the form of colossal amounts of data regarding consumers' thoughts, beliefs, experiences, and even interactions. In this paper, we propose an approach for firms to explore online user-generated content and "listen" to what customers write about their and their competitors' products. Our objective is to convert the user-generated content to market structures and competitive landscape insights. The difficulty in obtaining such market-structure insights from online user-generated content is that consumers' postings are often not easy to syndicate. To address these issues, we employ a text-mining approach and combine it with semantic network analysis tools. We demonstrate this approach using two cases—sedan cars and diabetes drugs—generating market-structure perceptual maps and meaningful insights without interviewing a single consumer. We compare a market structure based on user-generated content data with a market structure derived from more traditional sales and survey-based data to establish validity and highlight meaningful differences.

Key words: text mining; user-generated content; market structure; marketing research
History: Received: January 30, 2010; accepted: January 20, 2012; Peter Fader served as the special issue editor and Alan Montgomery served as associate editor for this article.

1. Introduction

The spread of the Internet has led to a colossal quantity of information posted by consumers online through media such as forums, blogs, and product reviews. This type of consumer-generated content offers firms an opportunity to "listen in" on consumers in the market in general and on their own customers in particular (Urban and Hauser 2004). By observing what consumers write about products in a category, firms could, in principle, gain a better understanding of the online discussion and the marketing opportunities, the market structure, the competitive landscape, and the features of their own and their competitors' products that consumers discuss.

Recent years have seen an emergence of academic and commercial marketing research that taps into this abundant supply of data, but the utilization of these data sources remains in an early stage. Consumer-generated content on the Web is both a blessing and a curse. The wealth of data presents several difficulties: First, the amount of data provided is overwhelmingly large, making the information difficult to track and quantify. Second, this rich but unstructured set of consumer data is primarily qualitative in nature (much like data that can be elicited from focus groups or depth interviews but on a much larger scale), which makes it noisy—so much so that it has been nearly impractical to quantify and convert the data into usable information and knowledge. In this paper, we propose to use a combination of a *text-mining apparatus* and a *network analysis framework* to overcome these difficulties.

Our objective is to utilize the large-scale, consumer-generated data posted on the Web to allow firms to understand consumers' top-of-mind associative network of products and the implied market structure insights. We first mine these exploratory data and then convert them into quantifiable perceptual associations and similarities between brands. Because of the complexity involved in consumer forum mining, we apply a text-mining apparatus that is especially tailored to that venue. We combine an automatic conditional random field (CRF) approach (McCallum and Wellner 2005) with manually crafted rules. We

use network analysis techniques to convert the text-mined data into a semantic network that can, in turn, inform the firm or the researcher about the market structure and meaningful relationships therein.

The proposed approach provides the firm with a tool to monitor its market position over time at higher resolution and lower cost relative to more traditional market structure elicitation methods. We compare the insights about the market structure mined from consumer-generated content to those obtained from traditional market-structure approaches based on both large-scale sales and survey data sets. The comparison suggests that the market structure derived from the consumer-generated content is very similar to the market structure derived from the traditional data sources, providing external validity to our approach. At the same time, we identify important differences between the alternative approaches. For example, following a marketing campaign aimed at changing the position of Cadillac toward competing with the more luxurious import cars, the co-mention of Cadillac with the import luxury cars increased significantly and substantially. On the other hand, car switching between import luxury cars and Cadillac increased at a much slower pace. To the best of our knowledge, this is the first study that compares market-structure maps derived from consumer-generated content with market-structure maps derived from traditional approaches.

In what follows, we describe the current state of research with respect to applications of text mining to user-generated content and the market-structure literature. In §3, we briefly describe our proposed text-mining methodology. In §4, we demonstrate the use of the text-mining approach in the context of a sedan cars forum and pharmaceutical drugs forums. We conclude with a discussion of the potential of this approach, its limitations, and directions for future research.

2. Market Structure and Mining Consumer-Generated Content

2.1. Mining Consumer-Generated Content

One can think of consumer-generated content in venues such as forums and blogs as an online channel for word of mouth, or "word of mouse," which is one of the marketing operationalizations of the somewhat broader concept of social interaction. Numerous academic papers, industry market research, and a large body of anecdotal evidence point to the significant effect word of mouth has on consumer behavior and, in turn, on sales (e.g., Eliashberg et al. 2000, Reichheld and Teal 1996). Cyberspaces such as chat rooms, product review websites, blogs, and brand communities invite and encourage consumers to post their views and reviews. The level of activity in these channels of communication has grown exponentially in recent years.

In the past few years, academics and practitioners have begun to realize the potential in online consumer forums, blogs, and product reviews. Several studies have investigated the relationship between consumer-generated content and sales. One of the main difficulties in using such content for quantitative analysis is that the data are primarily qualitative in nature. In their conceptual paper on future directions for social interaction research, Godes et al. (2005) stated that one of the challenges inherent in tapping into user-generated content is the inability to analyze the communication content. To simplify the task, researchers often use moments of consumer-generated data, such as magnitude or valence, to represent the discussion. Alternatively, quantitative summaries of content, such as overall product ratings, can be used to represent consumers' opinions (Chevalier and Mayzlin 2006, Chintagunta et al. 2010, Dellarocas et al. 2007, Godes and Mayzlin 2004, Liu 2006). For example, Liu (2006) examined the volume and valence of messages posted on the Yahoo! movies message board to predict box office sales. Liu (2006, p. 80) reported that mechanically analyzing more than 12,000 movie reviews using human reviewers was "an extremely tedious task." Similarly, after manually coding (using judges) a sample of messages on television show ratings for valence and length, Godes and Mayzlin (2004) highlighted the potential of content analysis but concluded that the cost associated with their approach to data collection was prohibitively high and the data reliability was limited.

Unlike many product review sites, most online consumer forums do not include quantitative summaries of consumers' evaluations such as star ratings. Furthermore, evidence with respect to overall reliability and predictive validity of online product ratings is mixed (Chen et al. 2004, Godes and Mayzlin 2004). Archak et al. (2011) demonstrated the advantage of extracting a more multifaceted view of the content of product reviews via text mining to successfully predict product choices. Thus, although the aforementioned studies demonstrate that summary statistics about online word-of-mouth information can be useful in predicting outcomes such as sales and ratings, they also highlight the need to delve deeper into the content of online discussions.

Our objective is to explore the market structure and the brand-associative network derived from online discussions. To do so, we need to understand the co-mention of more than one brand within a linguistic unit, such as a sentence or paragraph, and the nature of the relationship between the brands. We leverage recent advances in text-mining techniques to achieve this goal.

2.2. Text Mining and Marketing

Text mining (sometimes called "knowledge discovery" in text) refers to the process of extracting useful, meaningful, and nontrivial information from unstructured text (Dörre et al. 1999, Feldman and Sanger 2006). For example, using what they call "undiscovered public knowledge," Swanson and colleagues found relationships between magnesium and migraines (Swanson 1988) and between biological viruses and weapons (Swanson and Smalheiser 2001) by merely text mining disjoint literatures and uncovering words common to both literature bases.

With the increasing availability of digitized data sources, the business world has begun to explore the opportunities offered by text-mining tools to collect competitive intelligence, to syndicate and meta-analyze the wealth of information consumers are posting online, and to automatically analyze the infinite stream of financial report data to search for patterns or irregularities (e.g., Feldman et al. 2010). Collaboration between computer scientists and business researchers has often facilitated the dissemination (albeit limited) of these tools to business research (e.g., Das and Chen 2007; Feldman et al. 2007, 2008; Lee and Bradlow 2011). These collaborations have led to fruitful research initiatives by opening opportunities to quantitatively explore new sources of business data.

The use of text-mining techniques to derive insights from user-generated content primarily originated in the computer science literature (e.g., Akiva et al. 2008, Dave et al. 2003, Feldman et al. 1998, Glance et al. 2005, Hu and Liu 2004, Liu et al. 2005; see Pang and Lee 2008 and Liu 2011 for a review). To handle the difficulties involved in extracting information from consumer forums, the text-mining approach we propose supplements machine-learning methods with handcrafted rules tailored to the particular domain to which the mining is applied. This hybrid approach is particularly useful for extracting relationships between brands and terms or brands and brands. Our paper aims at using text mining to assess consumers' associative network for multiple brands and the perceptual market structure derived from the discussion. We contrast the text-mining approach with traditional survey and sales-based approaches, providing external validation to the current as well as to previous approaches.

Recently, a handful of studies applied text mining to marketing applications. Archak et al. (2011) studied the relationship between product attributes and sales of electronic products. Ghose et al. (2012) used text mining together with crowdsourcing methods to estimate demand for hotels. Eliashberg et al. (2007) text mined movie scripts to predict their success. Seshadri and Tellis (2012) demonstrated that product "chatter," defined by the magnitude, sentiment, and star ratings of product reviews, can predict firms' stock performances. Decker and Trusov (2010) estimated consumer preferences for product attributes by text mining product reviews. Lee and Bradlow (2011) text mined semistructured product reviews to understand market structure based on the product attributes mentioned in the reviews. Similar to Lee and Bradlow, we are interested in utilizing text mining to understand market structure. However, unlike Lee and Bradlow, we define similarity between products based on their co-mention and top-of-mind association in the forum messages, as opposed to being based on the similarity of the products' mentions with various attributes. Such top-of-mind co-mention of products is more likely to appear in more unstructured text, which requires different text-mining approaches. Our results suggest that for the data used in this paper, direct co-mention association measures produce more sensible market structure maps than those produced based on the similarity in the mention, of products with terms used to describe these products (see §4.1.3).

We believe that these applications of text-mining techniques to marketing represent just the tip of the iceberg, and our research adds another dimension to these efforts. We focus on utilizing text mining to assess market structure (Rosa et al. 2004). Unlike most of the aforementioned research, our research focuses not on product reviews but on less structured consumer forums that discuss specific product categories (e.g., cars, pharmaceutical drugs). Such forums are more qualitative and less focused than product reviews. Furthermore, most of the earlier studies extracted well-structured information for a *single* entity, such as a product or product feature, and quantified its volume and/or valence. We are focused on extracting, analyzing, and visualizing information about a large number of entities. We then use that information to establish *relationships* between the entities and make comparisons between them to derive brand-associative networks and the market structure.

2.3. Brand-Associative Networks and Market Structure

The information that consumers voluntarily and willingly post on consumer forums and message boards opens a window into their associative and semantic networks, as reflected by co-occurrences of brand references and descriptions of those brands in the written text.

Saiz and Simonsohn (2012) provided compelling evidence for the face validity of using the frequency of occurrence of terms on the Web to reflect the "true" likelihood of a corresponding phenomenon. Going beyond the mere occurrence of terms, we propose assessing the proximity or similarity between several terms based on the frequency of their *co-occurrence* in the text.

The notion of using co-occurrence as a proxy for similarity has roots in the knowledge discovery and co-word analysis literature (He 1999). For example, co-occurrence of words (known as "co-word analysis") is frequently used to trace the development of a particular issue in science by tracking the frequency of co-occurrences of pairs of words in various research fields (Callon et al. 1986). One premise behind utilizing the co-occurrence of terms to analyze consumer forum discussions is that consumers indeed compare products quite often (Pang and Lee 2008). Schindler and Bickart (2005) found that direct comparisons of brands and products in consumer forums is one of the main information-seeking motives for content generators and readers of these forums.

But which brands are likely to be compared with each other? A rich literature in cognitive psychology suggests that individuals form mental associative networks that connect isolated items of stored knowledge (Anderson and Bower 1973). Spread of activation (Collins and Loftus 1975) suggests that activation of one node in the network (e.g., Toyota) is likely to spread to activation of other, closely connected nodes in the network (e.g., Lexus). The association strength reflects a semantic relatedness between the two nodes (Farquhar and Herr 1993). Accordingly, two brands that are closely connected in the associative network are more likely to be retrieved from long-term memory and used concurrently in a task. Similarly, attributes that most closely describe a product are likely to appear more frequently with that product in a sentence. John et al. (2006) developed a survey-based approach that applied the concept of a memory-associative network to brands and the concepts used to describe them. Henderson et al. (1998) demonstrated the use of brand-associative networks to understand relationships among brands such as competitiveness, complimentarity, segmentation, and market structure. In this research we propose an automatic text-mining approach to derive such market-structure insights.

We compare the text-mining-based market structure with traditional market-structure approaches based on consideration set (Urban et al. 1984) and on brand-switching data (Cooper and Inoue 1996, Grover and Srinivasan 1987). We elaborate on the existing market-structure methods in §4.1.2.

3. The Text-Mining Methodology

Our objective is to mine discussions contained in the user-generated content and look for relationships between the semantic components. To do so, we developed a text-mining apparatus specifically tailored to deal with the difficulties involved in mining consumer forums. In this section we provide the general framework. We delegate more technical details to the "Text-Mining Methodology" appendix in the electronic companion (at http://mktsci.journal.informs.org/).

3.1. The Text-Mining Apparatus

Extracting structured product (e.g., car brand or car model) and attribute data involves five main steps:

Step 1. *Downloading*: The Web pages are downloaded from a given forum site in HTML format.

Step 2. *Cleaning*: HTML tags and nontextual information such as images and commercials are cleaned from the downloaded files.

Step 3. *Information extraction:* Terms for products and product attributes are extracted from the messages.

Step 4. *Chunking*: The text is divided into informative units such as threads, messages, and sentences.

Step 5. *Identification of semantic relationships*: Two forms of product comparisons are computed: First, we generate a semantic network of co-occurrences of product mentions in the forum. This analysis can provide an overview of the overall market structure. Second, we extract the relationship between products and terms and the nature and sentiment of the relationship.

Figure 1 depicts a typical message downloaded from a forum that we use in our first empirical application. The first nontrivial step in the text-mining process is information extraction. The extraction of product names (e.g., Nissan Altima and Honda Accord in Figure 1) and the terms used to describe products (e.g., "paint" and "interior" in Figure 1) constitutes the process of converting unstructured textual data into a set of countable textual entities.

The computer science literature outlines a plethora of methods for information extraction (see Pang and Lee 2008 for a review). Unlike much of the extant literature, our focus is on information extraction methods

Figure 1 A Typical Message Downloaded from the Forum Edmunds.com

```
CarType: 2- Acura TL
MsgNumber: 2479
MsgTitle: r34
MsgAuther: r34
MsgDate: Jun 24, 2004 (11:38 am)
MsgRepliesTo:
That's strange. I heard many people complaint[sic] about
the Honda paint. I owned a 1995 Nissan Altima before and
its paint was much better than my neighbour's
Accord (1998+ model). I found the Altima interior was
quiet [sic] good at that time (not as strange as today's).
```

Source. http://townhall-talk.edmunds.com/direct/view/.ef0a892/2478#MSG2478; accessed April 13, 2012.

that find pairs of product names (e.g., companies, drugs, products, attributes) mentioned together, sometimes in the context of a phrase that describes the relationship between the terms (Feldman et al. 1998). Furthermore, the complexity of consumer forums and the informal style of the text require us to extend existing text-mining approaches. By combining *supervised machine learning* architectures such as CRFs with *rule-based* or *dictionary-based* text mining, we are able to extract meaningful and relatively accurate information from the text using as little human labor as possible. We extract the brand or product models primarily through a CRF machine learning approach (Lafferty et al. 2001) trained on a small, manually tagged training data set. We then use the rule-based approach primarily to fine-tune the terms extracted from the machine learning procedure. The rules are useful for more complex linguistic patterns (with deeper contextual information) that are specific to the domain studied and can be missed by the machine learning approach. Rules are also used to filter terms and to disambiguate certain product name instances. We describe the text-mining approach in detail in the "Text-Mining Methodology" appendix in the electronic companion.

We assess the accuracy of the information extraction procedure using human tagging of a random sample of validation messages that were not used to train the system. For the sedan car models (e.g., Honda Civic or Toyota Corolla) identified in our empirical application, we achieved recall (the proportion of entities in the original text that were identified and classified correctly) of 88.3% and precision (the proportion of entities identified that were classified correctly) of 95.2%, which led to an F1 $= 2 \times$ (recall $\times$ precision)/(recall + precision) $= 91.6\%$ (F1 is a harmonic mean between recall and precision commonly used to measure the accuracy of text-mining tools). For car brands (e.g., Honda or Toyota), we obtained even higher levels of accuracy: recall of 98.4%, precision of 98.0%, and an F1 of 98.2%. For the diabetes drugs application, we obtained recall of 88.9%, precision of 99.9%, and an F1 of 94.1% for the drugs; recall of 74.4%, precision of 90.3%, and an F1 of 81.6% for the adverse drugs reactions; and recall of 59.7%, precision of 95.8%, and an F1 of 73.6% for the more complex relationships between drugs and adverse reactions. By comparison, accuracy measures of 80% to 90% have often been achieved in prior simple (nonrelational) product entity extractions (Ding et al. 2009).

After extracting the information, we divided the records into chunks at three levels: threads, messages, and sentences. Threads often contain hundreds of messages, whereas messages are short, often composed of only one or a few sentences or sentence fragments. For the purposes of this study, we use messages as our primary unit of analysis. That is, we look for co-occurrences of pairs of products, brands, and terms in each message.

3.2. Occurrence, Co-Occurrence, and Lifts

The basis for much of the analysis that we will describe in §4 is the measure of co-occurrence of terms. We analyze co-occurrences to look for patterns of discussion in the text-mined data and to form semantic networks and market-structure perceptual maps. Comparisons are prevalent and helpful in the automatic analysis of sentences in the forums we mine. For example, given a sentence such as "Toyota is faster than Honda," we can automatically extract the two car manufacturers (Toyota and Honda) and the attribute(s) being compared (speed). We start by analyzing the context-free co-occurrence of products in the same message to build a perceptual map of the products. We then explore the topics discussed for each of the products.

One limitation of using simple co-occurrence as a measure of similarity between terms is that for any term that appears frequently in a forum, its co-occurrence with nearly any other term will be greater than that of a term that appears less frequently. For example, in the sedan car forum described later, the car model Toyota Camry appeared with safety-related words 379 times, whereas there were only 18 co-mentions of the car model Volvo S40 with safety-related words. However, consumers mentioned the Toyota Camry 34,559 times in the forum, whereas the Volvo S40 was mentioned only 580 times. Thus, once we *normalize* for the mere occurrence of each car model in the forum, we find that the likelihood of safety-related words appearing in a sentence that mentions "Volvo S40" is much greater than for such words appearing in a sentence that mentions "Toyota Camry." Such normalization is called lift (or pointwise mutual information; see Turney and Littman 2003). Lift is the ratio of the actual co-occurrence of two terms to the frequency with which we would expect to see them together.[1] The lift between terms A and B can be calculated as

$$\text{lift}(A, B) = \frac{P(A, B)}{P(A) \times P(B)}, \tag{1}$$

where $P(X)$ is the probability of occurrence of term X in a given message, and $P(X, Y)$ is the probability that both X and Y appear in a given message.

A lift ratio of less than (more than) 1 suggests that the two terms appear together less than (more than)

[1] In the context of brand switching, lift is sometimes referred to as "flow" (Rao and Sabavala 1981).

one would expect by the mere occurrence of each of the two terms in the forum separately. In the preceding example, the measure of lift between Toyota Camry and safety-related words (2.1) is smaller than the lift for those terms and the Volvo S40 (4.9), which suggests that, after normalization, the relationship between Volvo S40 and safety-related words is stronger than the relationship between Toyota Camry and such words. In §4.1.4 we test the sensitivity of our analyses to alternative co-occurrence and similarity measures.

Next, we describe two applications of the proposed approach to sedan car and diabetes drug forums.

4. Empirical Applications

The text-mining process described in §3 allows us to take a qualitative and unstructured data set consisting of millions of sentences and convert it into a *semantic network* of co-occurrence of terms in the forum's messages. In the semantic network, terms appear closer to one another if they are mentioned together more than one would expect from chance (a high lift ratio). An advantage of converting the text-mined co-occurrence data into a semantic network is the resulting ability to present the forum's discussion in a graphical manner. One can then "zoom in" on certain domains or subsets of the network and trace the relationship between terms in more detail. We demonstrate this process using two empirical applications: consumer forums about sedan automobiles and about diabetes drugs.

4.1. Sedan Automobile Forum

The first step in text mining involves downloading the data to be mined. We downloaded data on sedan cars from the Sedans Forum on Edmunds.com on February 13, 2007 (http://townhall-talk.edmunds.com/WebX/.ee9e22f/; see Table 1 for summary statistics about the forum).

The Sedans Forum included 868,174 consumer messages consisting of nearly 6 million sentences posted by 76,587 unique consumers between the years 2001 and 2007. From this repository, we extracted 30 car brands (e.g., Honda and Toyota), 169 car models[2] (e.g., Honda Civic and Toyota Corolla), and 1,200 common terms (mostly noun phrases and adjectives) used to describe these cars (e.g., "compact," "safe," "hybrid," "leg room"). We focus most of the analysis on the textual unit of a forum message. Within a message, we looked for co-occurrences between two car brands, two car models, and/or a car brand or model and a term used to describe it (e.g., Toyota Corolla and "engine"), sometimes with a term used to describe the relationship (e.g., "problem").

Table 1 Characteristics of the Sedans Forum on Edmunds.com

No. of threads	557,193
No. of messages	868,174
No. of sentences	5,972,699
No. of unique users	76,587
No. of brands	30
No. of car models	169
No. of describing terms	1,200

This data set involves a rich product category with a large number of products and multiple product dimensions. Furthermore, this category is popular; both enthusiasts and lay consumers are involved. These features make the discussion in this category interesting and challenging to map.

4.1.1. Co-Occurrence of Car Models. We begin with an analysis of the lifts of car models mentioned in the same message. The more frequently consumers mention two car models together, the closer those cars are in consumers' perceptual space. As mentioned previously, this line of reasoning stems from the notion of memory-associative networks (Anderson and Bower 1973) and has strong roots in the co-word analysis literature (He 1999). Whether a consumer highlights points of parity (e.g., "Toyota Corolla and Honda Civic have similar prices") or points of differentiation (e.g., "Toyota Corolla differs from Honda Civic in terms of mpg"), we postulate that the fact that the consumer consciously compares the two cars highlights a sense of proximity or relationship in her mind (i.e., the dissimilarities between Honda Civic and Lamborghini Murciélago would be too obvious to write about).

We constructed a 169 × 169 dyad matrix of lifts between each pair of sedan car models observed in the forum based on Equation (1). The relationship between each pair of nodes has a symmetric strength reflected by the magnitude of the lifts ($\text{lift}(i, j) = \text{lift}(j, i)$). Examining the characteristics of this network may help us understand the nature of the forum discussion and the centrality of different car models to the discussion.

Despite the large number of car models involved, the dyad matrix is relatively dense (74%) because of the richness of the data (nearly 900,000 messages). That is, on average, 74% of the car model pairs were co-mentioned at least once. The most compared-against car in the forum was the Honda Accord; it was compared at least once to each of the other 168 cars. We used network centrality measures

[2] In our data set and corresponding analyses, we included only brands and car models that were mentioned at least 100 times in the forum.

(betweenness centrality and eigenvector centrality) to determine the centrality of different cars to the discussion. The Honda Accord and Toyota Camry were not only the top two cars in terms of number of mentions in the forum but they also had the highest centrality measures among all cars. On the other hand, several cars that had a lower frequency of mention in the forum (e.g., Lexus ES and Chrysler 300, ranked only 16th and 27th, respectively, in terms of number of mentions in the forum) had very high centrality measures (Lexus ES and Chrysler 300 were ranked 4th and 9th, respectively, in terms of their centrality measures). These are cars that defy the typical car fragmentation of luxury and family cars, and therefore they tend to be compared against a large and heterogeneous set of (other) cars.

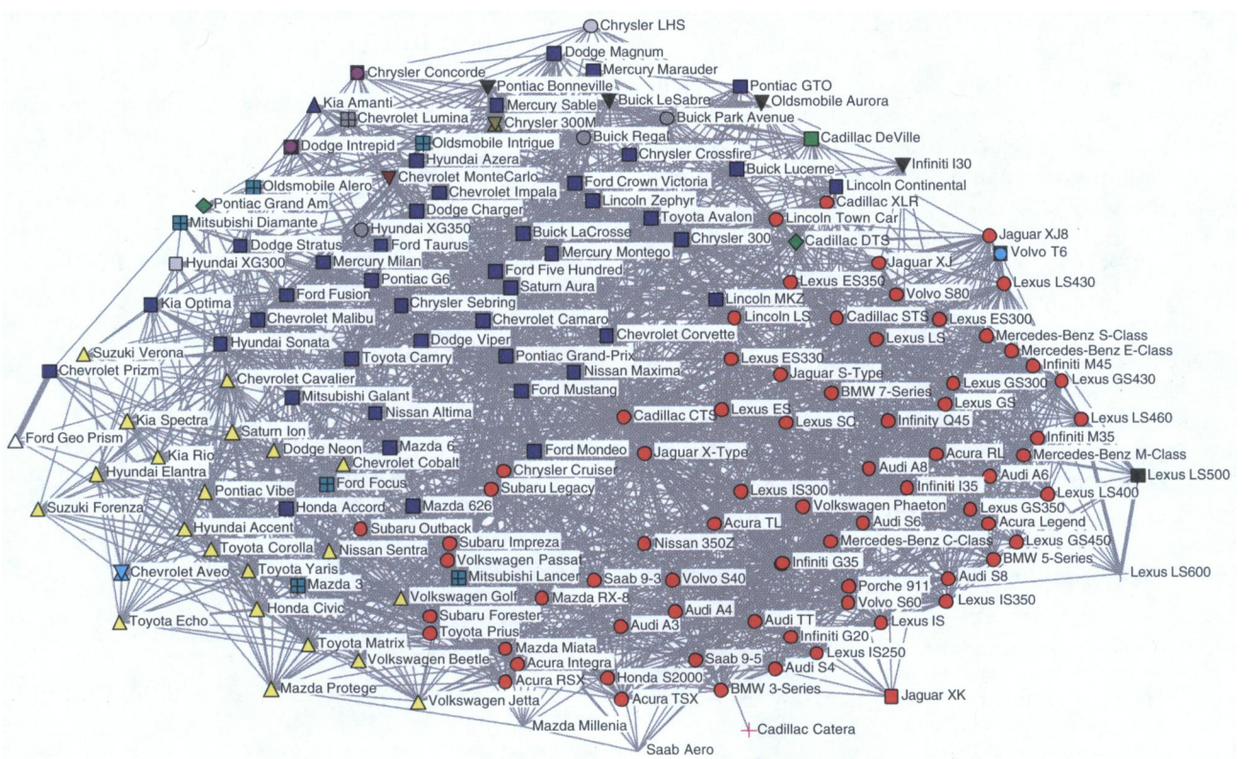
Figure 2 presents a visual depiction of the semantic network for the 169 car models mentioned in the forum using Kamada and Kawai's (1989) spring-embedded algorithm. This algorithm, much like multidimensional scaling, minimizes the stress of the spring system connecting the nodes in the network so that car models that are more similar (have higher lift) appear closer to one another in the graph. The width of the edges connecting the nodes (car models) in the graph represents the magnitude of the lift between two car models. Only lifts that are significantly greater than 1 at the 99% level based on the χ^2 test are depicted in the figure.

Although Figure 2 is somewhat crowded, it highlights some of the advantages of using the combination of text-mining and network analysis techniques to trace consumer perceptions and discussions. First, using the text-mining apparatus, we were able to simultaneously measure the discussions for 169 car models. Such an endeavor would be prohibitively difficult and costly using traditional marketing research methods. Second, network analysis permits us to analyze and visualize a large number of entities, thereby providing a comprehensive picture of the forum discussion.

The car models in the bottom left region of Figure 2 are the smallest sedan cars in the market—models such as Toyota Echo and Suzuki Forenza. As one moves right in the figure, the cars increase in size and luxuriousness, with the high-end Lexus models (LS500 and LS600) at the far right edge of the network. Cars of the same country of origin or the same make often appear close to one another (e.g., Audi A3, S4, and TT).

To further explore the sedan cars market structure, we looked for clusters of car models mentioned frequently together but less frequently with other groups of car models. We use the Girvan–Newman community clustering algorithm, which is commonly used to cluster networks (Girvan and Newman 2002). In the Girvan–Newman algorithm, clusters are defined by groups of nodes that are densely connected within

Figure 2 Spring-Embedded (Kamada and Kawai 1989) Network Graph of Sedan Car Models



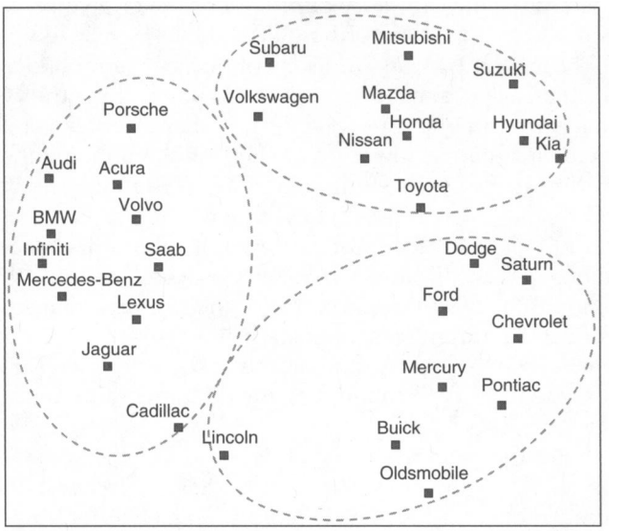
the cluster and less densely connected across clusters. Unlike typical social networks, in our semantic network, the communities consist of cars rather than people. The Girvan–Newman algorithm identified 26 clusters of car models. The shapes and color of the nodes (cars) in Figure 2 represent cluster membership. Although 26 seems like a large number of clusters, there were actually only 3 large clusters; the other 23 clusters consisted of only one car or a few cars each. Not surprisingly, the small clusters consist of cars at the edge of the network that are often outliers to the sedan cars market (e.g., the Jaguar XK) or older sedans that are not currently sold (e.g., Cadillac Catera). The clustering results provided high face validity. Cars that belong to the same family were grouped together. Specifically, the cluster of cars indicated by yellow triangles includes, for the most part, economy-class compact cars (e.g., Toyota's Corolla, Echo, Matrix, and Yaris). The cluster of cars represented by blue squares includes vehicles that fall primarily under the category of family cars (e.g., Toyota Camry and Toyota Avalon). The cluster of cars designated with red circles includes cars that belong predominantly to the luxury market (e.g., the Lexus models).[3]

Each of the sedan car models in Figure 2 is a member of a car brand. To get a somewhat cleaner picture of the associative network and resulting perceived market structure, we aggregated the discussion to co-occurrence of car brands. Overall, 30 car brands were mentioned in the forum. As in the car model analysis, we used lift between brands mentioned together in a message as a measure of similarity or association. With such a manageable number of brands, we were able to employ a more traditional market-structure analysis and visualization tool—multidimensional scaling (MDS). Figure 3 depicts the MDS map of car brands. We used the coordinates of the MDS to run a cluster analysis on the derived MDS. The *k*-means cluster analysis solution with three clusters fits the data best. The dashed ovals in Figure 3 reflect the three-cluster solution.

There are several insights that can be derived from Figure 3. Most American brands are clustered together in the bottom right section of the perceptual map. In the left part of the map, we see the high-end luxury European and Japanese brands such as BMW, Infiniti, Audi, Acura, Mercedes-Benz, and Lexus. Mainstream Japanese brands (i.e., Honda, Toyota, Mazda, and Nissan) are in the top section of the map.

[3] We also clustered car models based on traditional *k*-means cluster analysis. The results of the *k*-means three-cluster analysis were very similar to those of the Girvan–Newman algorithm, which suggests that the clustering solution found is robust to the clustering mechanism used (equivalence based or community based).

Figure 3 MDS Map of Discussion of Car Brands



However, Japan's Suzuki brand is mentioned more frequently with the Korean Kia and Hyundai brands. This result is consistent with the lower standing of the Suzuki brand in the U.S. market. The only American brands that are somewhat close to the luxury import brands in the MDS map are Cadillac and Lincoln. In fact, Cadillac is the only American brand that belongs to the luxury cars cluster. We discuss this finding in more detail later in §4.1.3.

4.1.2. Validating the Derived Market Structure. The perceptual and semantic network maps described thus far help us to map consumers' online discussions. A question that naturally arises is the extent to which consumers participating in online discussions and the topics discussed on the Web reflect the opinions of consumers at large and their off-line purchase behavior. Specifically, what are the similarities and differences between consumer forums-based market-structure maps and more traditional approaches to modeling market structure? In their review of market-structure models, Day et al. (1979) divided the analytical methods used to derive market structure into two groups: those that use purchase or usage data (e.g., brand switching) and those that use consumer judgment (e.g., consideration set or substitution). To examine the validity of the market-structure maps derived from the text-mining approach, we compare our results to market structures derived from brand-switching purchase data and from survey-based consideration set data.

Market Structure Based on a Consideration Set Survey. If the co-occurrence of product mentions in consumer forums reflects consumers' top-of-mind association between the two products, then such a relationship should be correlated with other measures and drivers

of product association. One of the strongest measures of co-association of products and of competitive market structure is based on products that consumers decide to place together in their consideration set (Urban et al. 1984).

To assess the external validity of the text-mining apparatus, we compared our results to consideration data elicited from a survey-based approach. In October 2006, BIGresearch conducted a Consumer Intentions and Actions Study involving a random sample of 7,623 respondents from BIGresearch's panel.[4] As part of the survey, respondents were asked if they were planning to buy or lease a car in the next six months. Respondents who answered "yes" were asked what type of car they were considering (sedan car, truck, minivan, sport-utility vehicle (SUV), crossover (between an SUV and a sedan car), or hybrid car). Respondents were then asked to select the make (brand) of the top two cars they were considering. Thus, this survey provided us with the two top brands in each consumer's consideration set. Of the 7,623 respondents, 426 (5.6%) mentioned that they were considering buying a sedan, crossover, or hybrid car in the next six months. Using the co-consideration data, we extracted the frequency of co-occurrence of any pair of car brands in consumers' top two brands considered. This resulted in a 26×26 matrix of brand co-occurrences.[5] We then converted the consideration-based co-occurrence matrix to a matrix of lifts between brands following Equation (1), and we normalized the matrix such that each row and column in the matrix summed to 1 to make the matrix comparable to the text-mining-based normalized lift matrix.

To make sure that the survey data were chronologically comparable to the text-mining data, we computed the table of co-mentions between the 26 car brands using only messages posted in our forum between January 1, 2006 and January 31, 2007 (around the time of the survey). We compared the aggregate survey-based co-consideration matrix of 26 brands with the text-mining-based co-mention lift matrix from the Edmunds.com forum. The correlation between the normalized matrices is relatively high ($r = 0.43$, pseudo-$p < 0.001$).[6] One of the reasons for the discrepancy between the two lift matrices is that the text-mining-based matrix is fully populated, whereas the 26×26 co-consideration matrix is relatively sparse (49% of the brand pairs are zero). To increase the statistical power of the comparison, we reran our analysis with car brands that were mentioned at least 15 times in the consideration set by the BIGresearch respondents. This resulted in 21 brands. The correlation between the text-mining-based and the survey-based 21×21 lift matrices is significantly higher ($r = 0.55$, pseudo-$p < 0.001$). We also calculated the correlation between the survey-based lift matrix and the text-mining-based matrix from messages posted in earlier years (2000–2005). As expected, the correlations are lower when the time period between the survey and the text-mining data did not match. This result provides additional evidence that the high correlation between the text-mining-based and the survey-based lifts is indeed a sign of the external validity of the text-mining approach.

The preceding analysis highlights some of the limitations of the survey-based approach. Even with a relatively large sample of more than 7,000 respondents, the ability to obtain a large enough sample of car buyers and a comprehensive set of car brands is limited. Furthermore, the BIGresearch survey did not ask consumers about the models of cars considered, which precludes an analysis of the market structure at the car model level. A survey-based approach to analyzing co-consideration at the car model level is likely to suffer from even more pronounced sample size limitations. Because of the sparsity of the survey-based data, we focus the remaining analysis on the possibly richer source of market-structure data based on actual purchase behavior.

Market Structure Based on Brand-Switching Data. One of the most common approaches to evaluating competitive market structure is the use of brand-switching data (Cooper and Inoue 1996, Grover and Srinivasan 1987, Harshman et al. 1982, Lehmann 1972). Like Cooper and Inoue (1996) and Harshman et al. (1982), we used a brand-switching matrix based on cars traded in during a new car purchase. Our data set includes 3,528,589 new car purchases, beginning with the first quarter of 2005 and extending through the second quarter of 2007. The data set was collected by the Power Information Network (PIN), an affiliate of J.D. Power and Associates. The PIN data set covers point-of-sale transaction data from 70% of the U.S. geographical markets and 30% of U.S. dealers (Silva-Risso et al. 2008). For each quarter and for each pair of car models (the car purchased and car traded in), we have the number of cars purchased or traded. Of the 470 car models and 54 car brands, which included sedans, SUVs, sport cars, and trucks, 108 models and 30 car brands (2,242,154 transactions)

[4] The BIGresearch survey included only households with income of $50,000 or more. We admit that this may lead to some discrepancy between the forum and the survey data.

[5] Only 26 car brands from the consideration set survey matched the 30 brands used in the text-mining analysis. Jaguar, Porsche, Saab, and Suzuki, which appeared in our text-mining data, did not appear or appeared less than five times in the consideration set data and were therefore removed from the analysis.

[6] The statistical significance for all correlations was assessed using the quadratic assignment procedure to account for the dependencies between rows in the co-occurrence matrix. See §4.1.5 for more details.

matched the car brand and car models mined in our sedan car forum.[7]

From the transaction data set, we created two car-switching matrices: a 30 × 30 car brands-switching matrix and a 108 × 108 car models-switching matrix. These switching matrices are similar in structure to the text-mining-based co-occurrence matrices previously described. To allow for comparisons between the two data sets, we counted brand similarities between cars i and j as the number of customers who switched either from car i to car j or from car j to car i (Lehmann 1972). As with the survey analysis described earlier, we compared normalized brand-switching and the text-mining lift matrices.

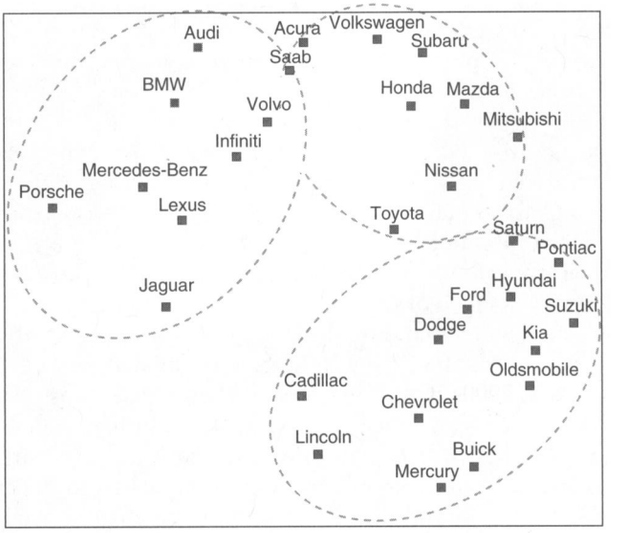
Model-switching matrix: We compared the switching matrix at the car model level with the text-mining lift matrix. The densities of the 108 × 108 text-mining and car model-switching lift matrices are 82% and 90%, respectively. The correlation between the normalized lifts of the two matrices is fairly high ($r = 0.470$, pseudo-$p < 0.001$). The difference between the matrices could be attributed to idiosyncratic differences for particular car models in the trade-in or forum data.[8]

Brand-switching matrix: Next, we compare the brand-level lift matrices. Both the text-mining and brand-switching 30 × 30 brand-level matrices are fully populated (trading in occurred in the data at least once between all 30 brands). The correlation between the brand-switching matrix and the text-mining normalized lift matrices is very high ($r = 0.753$, pseudo-$p < 0.001$) and provides supporting evidence for the external validity of the text-mining methodology. The strong correlation between the co-occurrence of brands in a message in the forum's messages and the likelihood of switching between the car brands suggests that consumers are more likely to discuss brands they are familiar with and are likely to trade between. Figure 4 depicts the market-structure MDS generated by the sales-based brand-switching data.

Consistent with the high correlation between the lift matrices, the perceptual maps of the two methods (Figures 3 and 4) are fairly similar. The correlation between the MDS coordinates of the two matrices is $r = 0.762$ (pseudo-$p < 0.001$). The k-means cluster analysis solution of the two figures also corresponds very well (83% hit rate).

Overall, the high correlations between the market-structure maps derived from the text-mining apparatus relative to the maps derived from the survey-based consideration set data, as well as the car sales data, provide sound evidence that top-of-mind co-mention of cars in forums is highly correlated with market outcome variables such as consumers' consideration-set structures and car-switching behaviors. Furthermore, this analysis provides supportive evidence for the ability of the text-mining apparatus to identify the relationship between car brands and models from unstructured text and for the overall relative representativeness of the forum's participants and discussion to the off-line world. No less interesting are the differences between the perceptual maps derived from the car-switching and consumer-generated content data. We discuss these next.

Figure 4 MDS of Car Brands Using Car-Switching Data



Differences Between the Brand-Switching and Text-Mining Market-Structure Maps. In comparing Figures 3 and 4, we also find some meaningful differences between the two maps. To examine the differences between the derived maps, we calculated the absolute difference between the Euclidean distances for each pair of brands in the MDS solution of the two maps.[9] Several brands have large differences between the maps. For example, in the brand-switching matrix, the Korean brands, Kia and Hyundai, as well as the Japanese Suzuki brand, are closer to the cluster of American brands than to the cluster of Japanese brand, whereas the reverse is true for the text-mining-based map. The difference possibly highlights that even though the Korean car brands were able to achieve top-of-mind association with the Japanese brands, such an association did not yet translate to car-switching behavior. As another example, Porsche is almost an outlier in the car trade-in map but is quite central in the forum discussion map. When it comes

[7] The 61 car models mentioned in the sedan car forum that were not traded in the PIN data set are mainly older car models such as the Cadillac DeVille.

[8] The PIN data set may be biased toward cars or car pairs that have better trade-in terms.

[9] The 30 × 30 matrix of the absolute differences between the maps is available from the authors upon request.

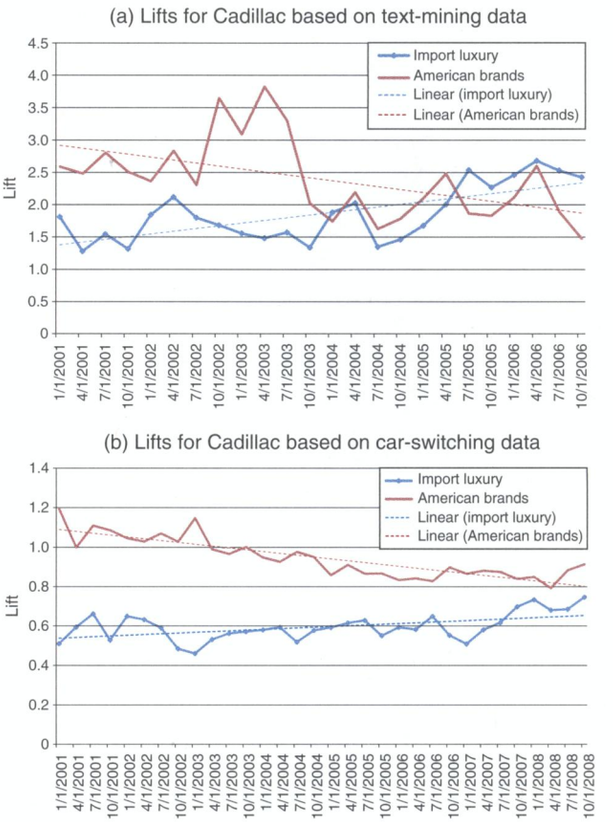
to Porsche, it may be easier to "talk the talk" than "walk the walk."

4.1.3. The Cadillac Case Study. Another interesting difference between the two market structure maps is the location of the Lincoln and Cadillac brands. In the text-mining-based map, Lincoln and Cadillac were the only American brands that appear close to the luxury import brands, with Cadillac even belonging to the import luxury brands cluster. This is not the case for the car-switching map. In the car-switching map, these brands clearly belong to the American cluster of brands. The position of Cadillac in the text-mining-based map may not have been the case all along. In the mid-2000s, Cadillac launched a $4.3 billion campaign to revamp the brand from "classic American" to "young luxury." In his *BusinessWeek* article, David Welch (2003, p. 79) wrote, "GM has taken significant strides toward making Cadillac a stronger rival to luxury import cars." The campaign included the introduction of new cars, such as the Cadillac XLR, which was intended to be a competitor for the sport models of Lexus and Mercedes (Csere 2003).

One of the advantages of text mining consumer forums is that each message comes with a time stamp. This allows us to analyze dynamics in the data. To test whether Cadillac was indeed able to change its positioning from a classic American brand to a rival for luxury import brands, we split the brand co-occurrence data into quarterly increments and analyzed the change in lifts between (1) Cadillac and the cluster of American brands and (2) Cadillac and the cluster of import luxury brands from Figure 3. As depicted in Figure 5(a), Cadillac was indeed able to change its positioning over time. Until the end of 2003, the average lift between Cadillac and the American brands was higher than its lift with the import luxury brands. However, starting in 2004—around the time of the repositioning campaign—Cadillac started to be mentioned at least as frequently, and often more frequently, with the import luxury brands than with the American brands. The positive slope for the lifts with the import luxury brands, the negative slope for the lifts with the American brands, as well as the difference between the slopes are statistically significant at the 0.05 level.

Next, we compared the trend observed in the text-mining data with the corresponding trend in lifts for the car-switching data. Figure 5(b) depicts the trade-ins (expressed in lifts) from import luxury and American brands to Cadillac between January 2001 and December 2008 by quarter. If the campaign was successful, not only in top-of-mind association but also in generating car switching to Cadillac, we would expect that over time, owners of import luxury brand

Figure 5 Trends in Lift Between Cadillac and Import Luxury and American Brands Over Time



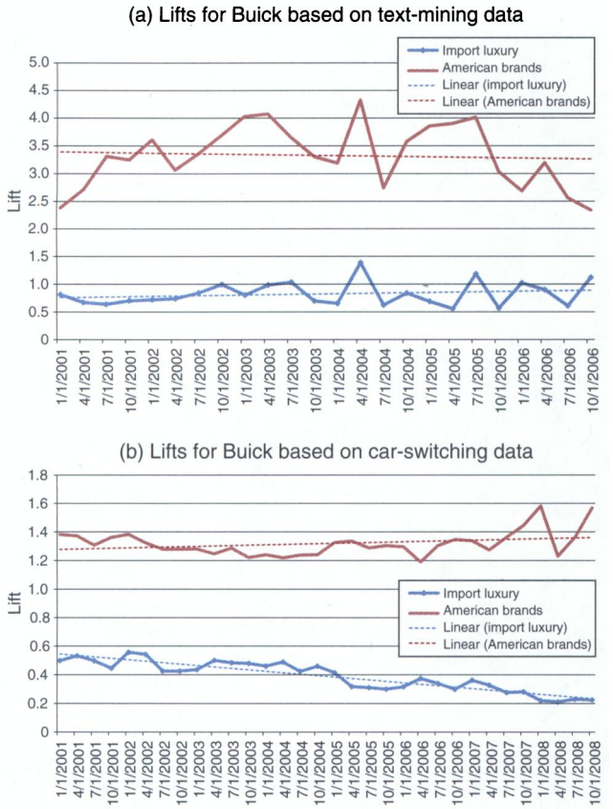
cars would be more likely to trade in their luxury cars for Cadillacs. The overall trend observed in the car-switching data is directionally similar to the one observed in text-mining data. Both the downward trend for trading in American cars for Cadillacs and the upward trend for trading from import luxury cars to Cadillacs are statistically significant ($P < 0.05$). However, whereas after the campaign the top-of-mind association between Cadillac and the import luxury brands was as strong as the association between Cadillac and the American brands, the trend is much weaker for the trade-in data. By the end of 2008, American car owners still had higher lifts for switching to Cadillac relative to that of import luxury cars owners.

A possible concern is that the observed trends are not specific to Cadillac and may be more general market or American brand-specific trends. To examine this concern, we repeated the lift trend analysis for the Buick brand (see Figure 6).[10]

Figure 6(a) suggests that, as expected, Buick is more likely than Cadillac to be associated with American

[10] We conducted similar analysis for other American brands such as Oldsmobile and Lincoln. These analyses revealed similar pattern to the Buick analysis and can be obtained from the authors upon request.

Figure 6 **Trends in Lift Between Buick and Import Luxury and American Brands Over Time**



brands and less likely than Cadillac to be associated with import luxury brands. In fact, Buick is less likely to be associated with import luxury brands than chance (average lift is less than 1). More importantly, the pattern of top-of-mind association of Buick with American and import luxury brands is fairly steady over time. Similar patterns were observed for the car-switching data (see Figure 6(b)) with an even downward trend for trading in import luxury cars for Buick over time. Thus, the trend observed for Cadillac is unlikely to be an overall market trend.

Several insights can be learned from the Cadillac case study. First, the case study anecdotally suggests that although the marketing campaign was able to "move the needle" in terms of top-of-mind association, the effect of the campaign on sales is much slower and possibly delayed in time. This result may suggest that market structure elicited from consumer forums can provide insights that cannot be obtained from sales data. Second, it demonstrates the opportunity to use the text-mining apparatus to measure the effectiveness of marketing campaigns. Third, it emphasizes the advantage that can be gained from dynamic and longitudinal market structure analysis using text-mined data. Finally, it highlights that mining consumer-generated content offers firms a tool to monitor their market positions over time at a higher resolution and often lower cost relative to traditional data sources.

Table 2 **QAP Correlation Between Car Brand Lift Matrices by Data Set Size**

1/16 of the messages	0.983
1/8 of the messages	0.989
1/4 of the messages	0.995
1/2 of the messages	0.997

4.1.4. Robustness Checks.

How Many Messages Are Needed? The sedan cars forum we used included nearly 900,000 messages. However, it is possible that in applying the proposed approach to other domains, the researcher may have sparser data. To test the sensitivity of the results to the size of the data set, we calculated the correlation between the car brands lift matrices generated using the full data set and matrices generated by randomly drawing only 1/16, 1/8, 1/4, and 1/2 of the messages. Table 2 describes these correlations.

Even with only 1/16 of the messages (less than 55,000 messages), the derived lift matrix is extremely similar ($r = 0.983$, pseudo-$p < 0.001$) to the matrix derived from the full data set. The MDS maps and the resulting insights derived from the data sets with varying sizes were also very similar. As expected, the positions of brands that appeared less frequently in the forum (e.g., Mitsubishi or Subaru) were more sensitive to sample size. This analysis suggests that the market-structure analysis is relatively robust to the number of forum messages studied. For more complex analyses that involve relationships between particular products and the terms used to describe them, the results may be more sensitive to the size of the data.

How Much Training Data Are Needed? The text-mining approach requires training data annotated by human experts to train the machine learning CRF algorithm. This task can be time consuming, though with the advent of crowdsourcing marketplaces (e.g., Amazon Mechanical Turk), this task can be crowdsourced at a fairly low cost. To test the sensitivity of the accuracy of the text-mining procedure to the size of the training data, we varied the size of the training data and calculated the recall, precision, and F1 measures (see Table 3) for identifying car brands in the validation sample as described in the "Text-Mining Methodology" appendix in the electronic companion.[11] For as few as 34 training messages, we obtained

[11] To separate the effect of training data on accuracy, in this sensitivity analysis we did not perform postprocessing after the CRF procedure (e.g., deleting terms that were identified as brands but are not car brands, such as Michelin). Postprocessing further increased accuracy in the results presented in the rest of the paper.

Table 3 Car Brand Recall, Precision, and F1 Accuracy Measures by Training Data Set Size

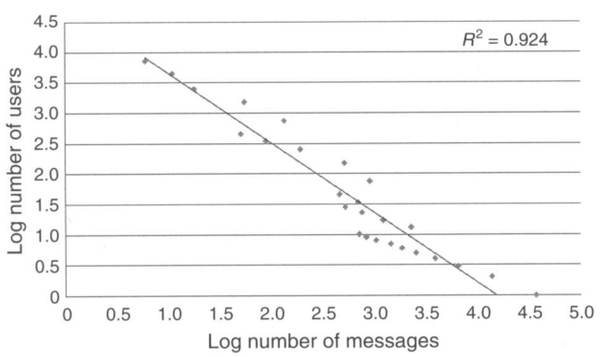
Size of training data	Recall (%)	Precision (%)	F1 (%)
276 messages	89	92	90
138 messages	86	90	87
69 messages	84	88	86
34 messages	81	86	83
17 messages	74	86	79

fairly good accuracy measures. This is attributed partially to the hybrid approach we take that combines machine learning with hand-crafted rules.

Are All Forum Participants Equal? In text mining the consumer forum, we conducted a census of all of the messages in the focal forum. Thus, from a sampling point of view, the analysis should effectively reflect the discussion in the forum. However, one possible concern with analyzing consumer forum data is the participation inequality in which a few "active" users contribute much of the content. Indeed, 10% of the users posted more than 80% of the messages in the sedan cars forum and 47% of the users posted only once. The log-log relationship between the number of users and number of messages they post is close to linear ($R^2 = 0.924$; see Figure 7). A similar relationship has been found in many user-generated content sites (e.g., Ochoa and Duval 2008).

To test whether the active users who generate the majority of the content in the forum differ from less active users, we calculated the correlation between the normalized lift matrix for content generated only by active users and the matrix generated by less active users. We defined active users as those who posted at least 10 messages in the forum (11% of the users who generated 82% of the content). The less active group consisted of the 89% who posted between one and nine messages each; this group accounted for the remaining 18% of the content. The correlation between the two matrices is quite high ($r = 0.79$, pseudo-$p < 0.001$), suggesting that, for the investigated forum, the active minority can well represent the entire forum population (Dwyer 2009). Future research could further explore heterogeneity in market structure based on participant and message characteristics (e.g., Ghose and Ipeirotis 2011).

Figure 7 Log Number of Users to Log Number of Messages in the Sedan Cars Forum



Short vs. Long Messages. The length of a message can relate to the type of user writing it or the state of mind of the user when he or she writes it. To test whether the market structure derived from longer messages is different from the market structure derived from shorter messages, we divided the forum messages into two groups based on a median split of the message length (the median length was 66 words). Longer messages naturally lead to greater chances for car references to co-occur. Indeed, 50% of the messages that contained more than 66 words accounted for 87% of the co-occurrences. The correlation between the car brand-normalized lift matrices of the short and long messages is extremely high (0.96, pseudo-$p < 0.001$), suggesting that the derived market structures is robust to message length.

Alternative Measures of Similarity. Throughout the analysis, we used the measure of lift to capture the association among products or between products and terms. We chose the lift measure because it is commonly used in the text-mining and co-word analysis literature and because of its intuitive appeal as a measure of co-occurrence relative to what would have been expected by chance. The lift measure captures the similarity between any pair of words by the direct co-occurrence between the words. An alternative approach to measuring the similarity between words is by measuring the similarity with which each pair of words is mentioned with all *other* words. In this section we compare the lift measure to other commonly used normalizations of the co-occurrence matrix (i.e., Salton's cosine, Jaccard index, and term frequency–inverse document frequency (tf-idf) weighting)[12] as well as to the Pearson correlation, which captures the similarity based on the correlation between the mention of each pair of words with all other words. Specifically, we compare the lift measure to the following measures.

1. Jaccard index:

$$\text{Jaccard}_{ij} = \frac{X_{ij}}{X_j + X_i - X_{ij}},$$

where X_{ij} is the co-occurrence between terms i and j, and X_i is the occurrence of term i.

[12] We note that cosine and tf-idf are sometimes used in the literatures as vector normalizations (similar to the correlation similarity measure) rather than normalizing the direct co-occurrences between words, as in our case. We also tested vector normalizations of these measures, but these led to less meaningful market-structure maps.

2. Salton cosine (Salton and McGill 1983):

$$\text{Cosine}_{ij} = \frac{X_{ij}}{\sqrt{X_j X_i}}.$$

3. Term frequency–inverse document frequency: tf-idf is used to weigh the occurrence of each term by its role in the document. The term frequency for term j in document m is defined by $tf_{jm} = X_{jm}/N_m$, where X_{jm} is the number of times term j appeared in document m, and N_m is the number of terms in document m. The inverse document frequency is defined by $idf_j = \log(D/M_j)$, where D is the total number of documents, and M_j is the total number of documents where term j appeared. tf-idf is given by $tf - idf_{jm} = tf_{jm} \times idf_j$. In our application, documents refer to forum messages and terms are cars. We calculated the tf-idf weighted co-occurrence between car brands i and j as

$$CO(tf - idf)_{ij} = \sum_{m \in D} (tf - idf_{jm} \times tf - idf_{im}).$$

4. Pearson correlation: $\rho_{ij} = \text{corr}(\mathbf{X}_i, \mathbf{X}_j)$ is the Pearson correlation between the co-occurrences of car brands i and j with all other car brands, where $\mathbf{X}_i$ is the vector of co-occurrences of car brand i with all other car brands.

Like the lift measure, measures 1–3 measure similarity based on the direct co-mention of any pair of words. Pearson correlation, on the other hand, measures the vector normalization of similarity of co-occurrence between every pair of words and all other words. Theoretically, we find the direct co-occurrences measures more appropriate to capture the associative network of car co-mentions in the forum. Table 4 describes the quadratic assignment procedure (QAP) correlations between the 30×30 car brand-normalized matrices based on the five similarity measures.

All the direct co-occurrence measures (lift, Jaccard index, Cosine, and tf-idf co-occurrence) are extremely highly correlated, producing virtually identical MDS maps. The correlation similarity measure, though still fairly highly correlated with the other measures, produces MDS maps that are far less meaningful (see the appendix).[13] Thus, for our application, association measures based on the direct co-occurrence between cars produce more meaningful market-structure maps relative to the vector normalization similarity measure.

Table 4 QAP Correlations Among Car Brand Matrices with Different Similarity Measures

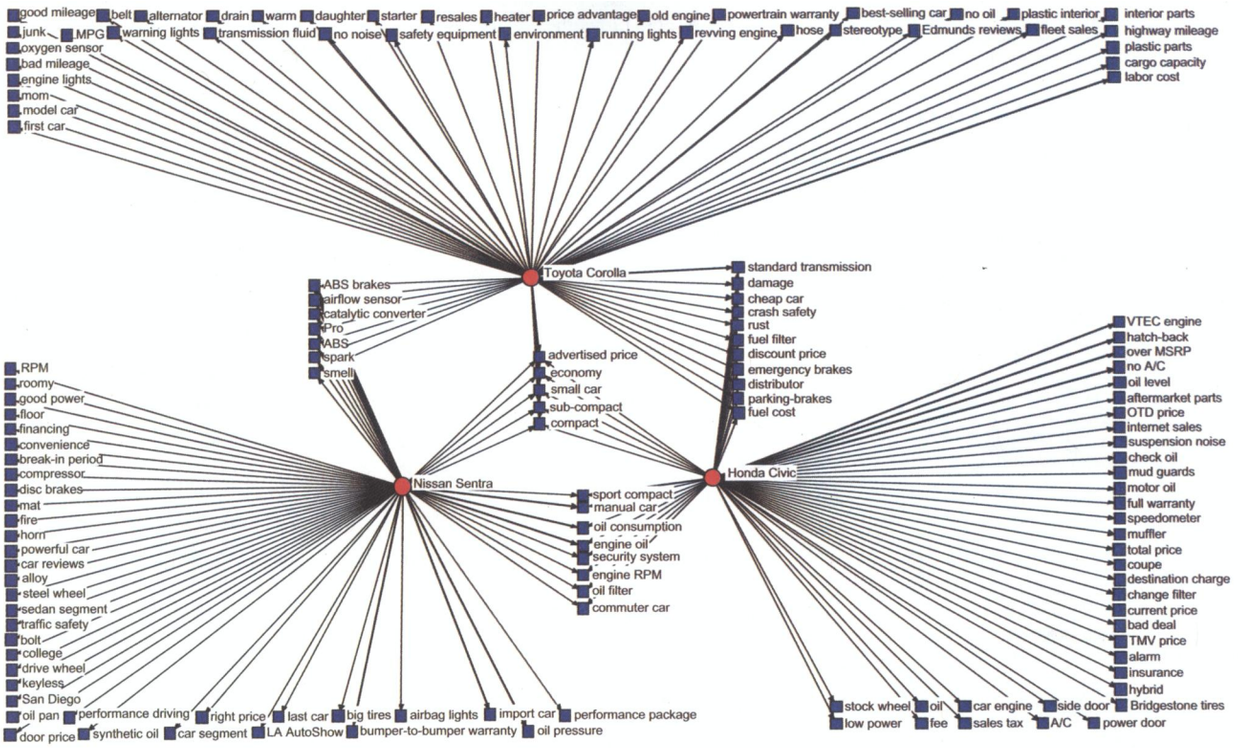
	Lift	Jaccard index	Cosine	tf-idf	Correlation
Lift	—				
Jaccard index	0.970	—			
Cosine	1.000	0.970	—		
tf-idf	0.961	0.919	0.961	—	
Correlation	0.623	0.575	0.623	0.473	—

4.1.5. "Zooming In" on the Discussion. Thus far, we have analyzed the co-occurrence of car models with one another in the forum. However, one of the most promising aspects of the text-mining methodology is the opportunity to quantify what consumers wrote about each of the cars. This type of analysis allows us to drill one level deeper into consumers' discussions. In addition to the 169 car models mentioned, we extracted 1,200 nouns and adjectives that consumers used to describe the cars. Therefore, we can investigate the frequency with which each term co-occurred with a car brand or model. Such analysis can help us to explore points of differentiation and points of parity between cars. As in the previous analyses, we focused on the lift measure to control for relative frequency of appearance of each term and the car brand or model in the forum. Figure 8 depicts the terms that exhibited strong lift (lifts > 2 and statistically significantly different from 1 at the 99% level) for three compact Japanese cars: the Honda Civic, Nissan Sentra, and Toyota Corolla. The width of the edge between the car and the term reflects the strength of the lift.

As Figure 8 shows, the terms mentioned with high lift with all three cars often describe the cars' category (e.g., "compact," "economy," and "small car"). More interesting are terms consumers mentioned frequently with one or two of the cars but not with other(s), suggesting a point of differentiation in top-of-mind associations. Although all three cars were frequently described as "compact," only the Nissan Sentra and Honda Civic were described as being "sport compact," differentiating these cars from the Toyota Corolla. Honda Civic was successful in differentiating itself from the other two cars based on terms such as "hatchback," and "hybrid." Indeed, during the period mined, only the Honda Civic offered hatchback and hybrid models among the three cars. The Nissan Sentra, on the other hand, was differentiated by a "performance package" and a "bumper-to-bumper warranty" that Nissan offered to consumers. Another interesting term associated with the Sentra is "college," which possibly suggests that the Sentra

[13] We also created a correlation similarity market-structure map based on the co-occurrence between cars and the terms used to describe them. This produced an even less meaningful market-structure map.

Figure 8 Terms Commonly Appearing with the Honda Civic, Nissan Sentra, and Toyota Corolla

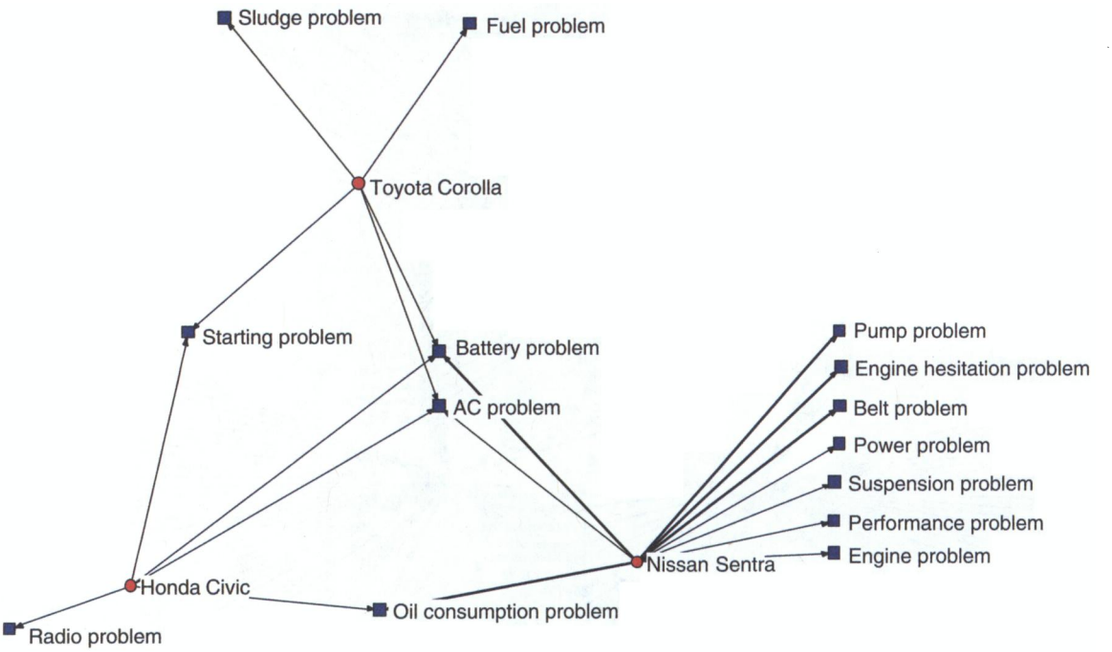


is perceived as a "college car." This market segment may not be obvious to Nissan through a simple demographic analysis of the Sentra's buyers because the buyers may be parents of college students. In general, the analysis of consumer-generated content opens a window for firms into a broader population than their own car buyers. For the Toyota Corolla, consumers frequently commented on its "plastic parts and interior." Furthermore, the car was mentioned with terms related to "good mileage." This may be more of a consumer perception than reality as the official gas mileage figures for all three cars are comparable (U.S. Department of Energy 2012).

The analysis presented in Figure 8 depicts lifts between cars and terms. One could zoom in one step further by analyzing the sentiment and context of the relationship between the car and the terms used to describe it. For example, both the Honda Civic and the Nissan Sentra were mentioned with "engine oil" and "oil consumption." A natural question then arises: Is oil consumption a problem the manufacturers should be worried about, or is it a positive or neutral claim? To investigate this issue further, we identified 71 terms that commonly appeared with a negative sentiment to terms such as "problem" or "issue." Figure 9 depicts the most common problems identified for the three cars (lift > 2 and statistically significantly different from 1 at the 99% level). The oil consumption terms identified in Figure 8 for the Honda Civic and Nissan Sentra did indeed relate to a problem that consumers frequently complained about for these two cars. Manufacturers can use such analyses to track common consumer complaints about problems in their cars and assess the relative position of their cars vis-à-vis the competition. We further explore more complex relationships that involve sentiment in our second text-mining application, which investigates adverse reactions to diabetes drugs.

4.1.6. Decomposing the Semantic Network. The network analyses and multivariate methods we have used thus far to analyze the text-mining data help us create perceptual maps that depict the similarities or dissimilarities between cars that emerge from forum discussions. Figures 8 and 9 go a step further in explaining the dimensions underlying the co-occurrences between three compact Japanese cars. However, one may wish to go beyond visual representations to a more systematic statistical analysis. Specifically, we investigated the characteristics of the cars and terms appearing in the forum that could explain the pattern of co-occurrences observed in the extracted semantic network. To do so, we related the lift measures among the 169 car models to the cars' characteristics (size, brand, manufacturer, country of origin, and price), independent mentions of the cars in the forum, and their co-occurrence with terms that

Figure 9 Problems Commonly Appearing with the Honda Civic, Nissan Sentra, and Toyota Corolla



consumers most commonly used to describe the cars in the forum.

We defined y_{ij} as the lift between car models i and j following Equation (1). We vectorized these lifts to create a vector of $(169 \times 168)/2$ lifts between pairs of car models ($\mathbf{Y}$). Similarly, each explanatory variable is a vector reflecting the match between a pair of associated car models and a variable of interest. We defined the following explanatory variables.

- **Brand:** $brand_{ij}$ equals 1 if both cars carry the same brand name (e.g., Honda Civic and Honda Accord) and 0 otherwise.
- **Manufacturer:** $manuf_{ij}$ equals 1 if both cars are manufactured by the same parent company (e.g., Honda Civic and Acura TL) and 0 otherwise.
- **Country of origin:** $country_{ij}$ equals 1 if the country of origin of both cars is the same (e.g., Honda Civic and Toyota Camry) and 0 otherwise.
- **Size:** $size_{ij}$ equals 1 if both cars belong to the same size category as defined in the forum website (the size categories used are compact, midsize, and large) and 0 otherwise.
- **Price difference:** $price\_difference_{ij} = |MSRP_i - MSRP_j|$. This number is calculated as the mean absolute deviation between the manufacturers' suggested retail prices (MSRP) of the two cars in thousands of dollars. Thus, the smaller the difference, the more similar the price of the two cars. The MSRP was elicited from the official MSRP listed on Edmunds.com, the website that hosts the forum we mined. For cars that did not have a new model in 2008, we replaced the MSRP with Edmunds.com's published True Market Value price for the most recent model of the car.
- **Occurrence:** Lift measures can be sensitive to the base frequency of occurrence of each of the terms. To control for the base occurrence of each car model in the forum, we included the product of the occurrences (in thousands) of the two car models: $occurrence_{ij} = occurrence_i \times occurrence_j$.

Additionally, we wish to relate the co-occurrence between cars to terms used to describe them in the forum. It would be impractical to include all 1,200 terms in the model directly. We used factor analysis to reduce the dimensionality of the term space and identify the underlying topics most relevant to the discussion. For the factor analysis, we focused on the 100 most frequently mentioned terms (see Table 1 in the "Factor Analysis of Terms" appendix in the electronic companion for the list of terms). We used factor analysis with varimax rotation to maximize the interpretability of the results. Based on the "elbow" in the scree plot (see Figure 1 in the "Factor Analysis of Terms" appendix in the electronic companion), the cumulative variance explained, and the ability to interpret the factors, we chose a six-factor solution. The first six factors explained 49% of the variance in the data. We named these six factors "upscale," "looks," "driving experience," "consumer value," "emotional sentiment," and "comfortable ride" (see Table 1 in the "Factor Analysis of Terms" appendix in the electronic companion for the rotated component matrix). We defined the similarity between the score of each pair of cars on these factors as follows.

• **Factor similarity**: For each factor k and car models i and j, the measure of similarity between the pair of cars is calculated by $factor(k)_{ij} = |score_{ki} - score_{kj}|$, where $score_{ki}$ is the score of car i on factor k.

From a statistical point of view, we need to address two issues before we can regress the vector of lifts between cars (**Y**) on the set of explanatory variables. First, to handle the ratio nature of lifts, we log-transform the ratio-based lifts by $y_{ij} = \log(1 + lift_{ij})$. Second, ordinary least square regression assumes that observations in the data are independent. This assumption is likely to be violated in our case because in the vector of lift dyads (**Y**), each car model appears 168 times. Although a simple regression would produce valid point estimates, the standard errors are likely to be incorrect. To solve this problem, we adopted the QAP method, which involves two steps. In the first step, a standard regression is performed to obtain the parameter estimates. In the second step, we permute the rows and columns of the dyad lift matrix and reestimate the model. We repeat the permutation 5,000 times to estimate the distribution of the regression parameters. This approach has been shown to yield unbiased estimates and standard errors (Krackhardt 1988). We report the result of the QAP regression in Table 5.

Model 1 regresses the lifts between car models on the similarity in car characteristics only; these are car characteristics that are exogenous to the forum discussion. Not surprisingly, cars sharing similar characteristics tend to be compared with one another more frequently in the forum. Looking at the standardized coefficients, we see that same brand has the strongest relationship to co-mention of the cars, followed by size, manufacturer, price tier, and country of origin.

Next, we looked at how information that is endogenous to the forum can help explain the lifts between cars. Specifically, we included in the regression the independent occurrence of each car in the forum and the similarity between the scores of each pair of cars on the six factors. We predict that cars that are mentioned with similar terms (low mean absolute deviation between their factor scores) are more likely to be mentioned together in the forum (high lifts). Thus we expect a negative relationship (coefficient) between factors' mean absolute deviations and lifts. Indeed, in Model 2 all factors except "comfortable ride" had a significant and negative effect on lifts. The factor that is most closely correlated with cars co-mentions is "upscale." Other factors contributing to the co-mention of cars are "consumer value" and, to some extent, "looks" and "emotional sentiment." In Model 3, we included car characteristics in addition to the occurrence and factor scores in a single model. The nature of the results did not change from the separate nested models. Each addition of a component further improved the adjusted R^2. Thus, similarity in the cars' characteristics and prices along with the discussion about the cars help to explain the degree of comparison between cars.

4.2. Diabetes Drug Forums

In the second application, we focus on pharmaceutical drugs. In this study we go into deeper textual relationships and sentiment analysis to investigate the mention of drugs in relation to adverse reactions associated with them. Specifically, we studied forums that discussed diabetes drugs because diabetes is a worldwide disease with multiple pharmaceutical treatments and an active and involved group of patients sharing their experiences over multiple forums.

Table 5 Parameter Estimates from the QAP Regression of Car Model Lifts

	Model 1—Car characteristics			Model 2—Car terms (factors)			Model 3—Car characteristics + Car terms		
	Coef.	Standardized coefficient	Pseudo-p-value	Coef.	Standardized coefficient	Pseudo-p-value	Coef.	Standardized coefficient	Pseudo-p-value
Intercept	0.2204	—	0.000	0.5104	—	0.000	0.3952	—	0.000
Brand	0.3423	0.2031	0.000	—	—	—	0.3197	0.1897	0.000
Manufacturer	0.1732	0.1708	0.000	—	—	—	0.1778	0.1754	0.000
Country of origin	0.0903	0.1259	0.000	—	—	—	0.0765	0.1067	0.000
Size	0.1325	0.1826	0.100	—	—	—	0.1332	0.1836	0.000
Price difference	−0.0054	−0.1629	0.000	—	—	—	−0.0029	−0.0866	0.000
Occurrence	—	—	—	−0.0002	−0.0477	0.001	−0.0003	−0.0609	0.000
Factor similarity									
Upscale	—	—	—	−0.0904	−0.2374	0.000	−0.0708	−0.1860	0.000
Looks	—	—	—	−0.0302	−0.0829	0.000	−0.0326	−0.0897	0.000
Driving experience	—	—	—	−0.0216	−0.0598	0.004	−0.0096	−0.0267	0.1259
Consumer value				−0.0530	−0.1490	0.000	−0.0396	−0.1114	0.000
Emotional sentiment				−0.0312	−0.0899	0.000	−0.0287	−0.0827	0.000
Comfortable ride				0.0022	0.0057	0.358	0.0061	0.0163	0.2324
Adj. R^2	0.230			0.113			0.297		

Note. The two-tailed pseudo-p-value is calculated based on the proportion of times the absolute value of the estimated coefficient was larger than the absolute value of the QAP permuted coefficient estimate across the 5,000 iterations.

Table 6 The Diabetes Drugs Forum Data

Forum	No. of threads	No. of messages	No. of sentences	No. of unique users	Dates
DiabetesForums.com	17,229	228,690	1,449,757	4,881	02/2002–05/2008
HealthBoards.com	4,418	24,934	216,220	3,723	11/2000–05/2008
Forum.lowcarber.org	22,092	325,592	3,106,362	7,172	10/2002–05/2008
Diabetes.Blog.com	61	29,359	227,878	3,922	07/2005–05/2008
DiabetesDaily.com	5,884	62,527	380,158	2,169	05/2006–05/2008
Total	**49,684**	**671,102**	**5,380,375**	**21,867**	

4.2.1. Diabetes Drug Data. We downloaded the entire forum discussions from five of the largest diabetes drug forums. Table 6 provides summary statistics of each forum. Overall, we mined more than 670,000 messages (more than 5 million sentences).

4.2.2. Analyzing Adverse Drug Reactions. We used the consumer forums to assess consumers' discussions about a phenomenon called "adverse drug reaction" (ADR), which is medical damage caused by taking a given medication at a normal dose. An ADR is more commonly referred to as a "side effect"; however, side effects can be both negative and positive, whereas ADRs refer only to negative effects. Recent estimates place ADRs as the cause of 3%–5% of all hospitalizations (around 300,000 annually in the United States). Prior to a drug's approval and market introduction, ADRs are examined in clinical trials on a sample of patients. Because of the relatively small number of patients studied, the short duration of the trials, and idiosyncratic conditions, clinical trials often miss ADRs. Accordingly, there are several mechanisms for surveillance of ADRs both pre- and post-marketing, such as cohort and case studies, population statistics, and anecdotal reporting from journals and doctors (see Table 2 in Edwards and Aronson 2000 for a comprehensive list). Additionally, the World Health Organization and the Food and Drug Administration collect information on post-marketing ADR events using channels such as the Adverse Event Reporting System (AERS). Patients are constantly searching for ADR information. Package inserts often consist of long checklists that make it difficult for patients to see the forest for the trees. The difficulty patients typically experience when trying to find information on the prevalence of ADRs likely explains much of the popularity of pharmaceutical and disease-focused forums, where patients can share common experiences with such drugs. The forums report patients' firsthand experiences with the drugs and act as a living environment that keeps updating itself over time. As we saw with the car forum, drug companies can, by tapping into these forums, gain a real-time window into emerging consumer views on the drugs they and their competitors manufacture. A text-mining approach is likely to cost less than traditional medical post-marketing research methods and produce results that are less sensitive to sample-size concerns.

To identify drugs and ADRs, we created a dictionary of drugs and ADRs. We extracted ADRs for all the drugs (diabetes and others) from Drugs.com. We then broke down the ADRs to their components (i.e., part of body, problem, and symptom) and then combined these components to create a universal set of all possible ADRs. The textual relationship between drugs and ADRs goes beyond the mere co-occurrence of a drug and an ADR. For example, in the following three sentences, the drug "Actos" co-occurs with the ADR "nausea"; however, only the first sentence refers to "nausea" being an ADR for "Actos": (1) "I had terrible nausea after taking Actos"; (2) "Unlike other drugs, Actos does not cause nausea"; and (3) "I switched from Actos to Lantus and had terrible nausea." We used a head-driven phrase structure grammar linguistic parser to identify the role of each part of speech in the sentence. This process helped us identify the exact relationship (including negation) between each drug and the ADR that was mentioned with it (see the "Text-Mining Methodology" appendix in the electronic companion for details).

We created a list of all ADRs that were frequently mentioned as having a negative relationship with each of the diabetes drugs. The first three columns in Table 7 list all of the drug–ADR relationships that had a lift significantly greater than 1 at the 95% level.

To evaluate the validity of the extracted drug–ADR relationships, we also collected ADR information for each diabetes drug from WebMD, the leading health portal in the United States (Keohane 2008). On WebMD, ADRs are rated by their frequency of occurrence and severity. The last two columns in Table 7 report whether the ADR was listed in WebMD for a particular drug and, if so, what rating WebMD gave for its frequency (common, infrequent, or rare) and severity (severe or less severe). In total, 86% of the ADRs identified as appearing frequently with each drug by the text-mining apparatus were listed in WebMD. Moreover, most of the ADRs (78%) identified by the text-mining approach were associated with known ADRs reported as frequent and/or severe. The severe ADRs had a significantly higher average lift

Table 7 Drug–ADR Relationships Extracted from the Forums

Forum mentions			WebMD	
Drug	ADR	Lift	Frequency	Severity
Actos	Fluid retention	6.51	Infrequent	Severe
Actos	Liver problems	4.89	Rare	Severe
Actos	Edema	4.54	Rare	Severe
Actos	Swelling	4.45	Infrequent	Severe
Actos	Weight gain	3.12	Rare	Severe
Amaryl	Low blood sugar	8.23	Infrequent	Severe
Amaryl	Weight gain	3.81	Doesn't exist	
Avandia	Heart problems	6.77	Rare	Severe
Avandia	Edema	6.42	Rare	Severe
Avandia	Swelling	4.25	Infrequent	Severe
Avandia	Fluid retention	3.31	Infrequent	Severe
Avandia	Weight gain	2.24	Rare	Severe
Byetta	Bad taste	2.87	Rare	Less severe
Byetta	Hair loss	2.86	Rare	Less severe
Byetta	Jitteriness	2.55	Infrequent	Less severe
Byetta	Nausea	2.46	Common	Less severe
Byetta	Loss of appetite	2.42	Infrequent	Less severe
Byetta	Cold symptoms	2.35	Doesn't exist	
Byetta	Constipation	2.22	Rare	Less severe
Byetta	Bloated feeling	1.83	Rare	Less severe
Byetta	Rash	1.72	Rare	Severe
Glucotrol	Low blood sugar	4.42	Common	Severe
Glyburide	Increased hunger	5.45	Common	Less severe
Glyburide	Weight gain	2.59	Common	Severe
Humalog	Allergic reaction	8.92	Common	Severe
Humalog	Rapid heartbeat	6.84	Rare	Severe
Humalog	Kidney problems	5.58	Doesn't exist	
Januvia	Respiratory problems	10.59	Infrequent	Severe
Januvia	Jitteriness	9.27	Rare	Severe
Januvia	Irritability	6.18	Rare	Severe
Januvia	Sinus problems	5.29	Infrequent	Severe
Januvia	Cold symptoms	3.13	Infrequent	Less severe
Lantus	Mood problems	9.78	Doesn't exist	
Lantus	Irritability	5.25	Rare	Severe
Lantus	Lower blood sugar	2.90	Common	Severe
Levemir	Anxiety problems	9.34	Doesn't exist	
Levemir	Sleep problems	8.49	Doesn't exist	
Levemir	Allergic reaction	6.14	Rare	Severe
Levemir	Rash	3.70	Infrequent	Severe
Metformin	Lactic acid	3.76	Rare	Severe
Metformin	Taste problems	3.76	Common	Less severe
Metformin	Muscle pain	2.88	Infrequent	Less severe
Metformin	Stomach cramps	2.76	Common	Less severe
Metformin	Diarrhea	2.49	Common	Less severe
Metformin	Digestive disorders	2.49	Common	Less severe
Metformin	Leg pain	2.07	Infrequent	Less severe
Symlin	Low blood sugar	5.78	Infrequent	Severe
Symlin	Bloated feeling	3.62	Doesn't exist	
Symlin	Nausea	1.80	Common	Less severe

than the less severe ADRs (average severe lift = 5.31, average less severe lift = 2.75; p-value < 0.01). The average lifts for the three frequency categories did not significantly differ from one category to another. Thus, the severity of an ADR seems to influence the mention of that ADR in the forum more than does its frequency. Overall, this analysis provides additional evidence of external validity of the user-generated content and the text-mining apparatus.

Many more ADRs were reported in WebMD than were identified in the forums. There were 215 ADRs mentioned on WebMD for the 12 drugs mined. Of these, 56% were also mentioned in the forum (20% were mentioned with a lift significantly greater than 1).[14] The low recall of ADRs by the forums' participants is likely to be caused by what the medical community calls "overwarning" (Duke et al. 2011), which refers to the tendency of the medical community to report "exhaustive lists of every reported adverse event, no matter how infrequent or minor" (p. 945). Mining consumer forums along the lines of our analysis may serve as a tool to prioritize ADRs that patients seem to be most concerned with.

Possibly more interesting are the seven ADRs that were frequently mentioned in the forums but not reported in WebMD: weight gain (Amaryl), cold symptoms (Byetta), kidney problems (Humalog), mood problems (Lantus), anxiety problems (Levemir), sleeping problems (Levemir), and bloated feeling (Symlin). Patients' mentions of these ADRs in the forums should, at the very least, raise a flag for health officials to track these possible drug reactions. Note that several of the ADRs not reported in WebMD are "softer" ADRs, such as mood problems, anxiety problems, sleeping problems, and cold symptoms, that may not be considered medically serious but are reactions to which patients are sensitive. For example, the relationship between diabetes and mood problems such as depression is well documented (Anderson et al. 2001). Thus, although a patient may be mistaken in associating her psychological condition with a particular drug, pharmaceutical firms should be aware of such common misattributions in marketing their drugs.

5. General Discussion

In this paper, we propose "a sonar" by which marketing researchers can listen to consumers' ongoing discussions over the Web with the goal of converting online discussions to market-structure insights. We use text mining to overcome the difficulties involved in extracting and quantifying the wealth of online data that consumers generate, and we use network analysis tools to convert the mined relationships into co-occurrence among brands or between brands and terms.

We demonstrate the value of the proposed apparatus in two empirical applications involving sedan cars and diabetes drugs forums. Because the structure and environment of the automotive market is relatively familiar, we use this application to test the

[14] Some of mismatches between WebMD and the forum may be due to mentions of related ADRs. For example, low blood sugar mentioned in WebMD with Amaryl may be associated with the weight gain problem mentioned in our forums. We refrain from making such judgments because of limited medical expertise on our part.

validity of the proposed market-structure approach. Indeed, the associative network derived from the car forum provided a high degree of internal and external validity. The sedan market structure mined from the top-of-mind co-mention of cars in the consumer forum was found to be highly correlated with the market structure derived from traditional data collection methods such as survey-based consideration set data and transaction-based, brand-switching data. The differences between the text-mining and sales-based maps can provide a window into top-of-mind associations, which may not have yet translated into action, as demonstrated with the Cadillac case study. The cars domain provided an appropriate testing ground to investigate the external validity of the proposed approach because sales brand-switching data are commonly available. Future research could also explore applications of the proposed approach in domains that are not as well established or in emerging domains in which alternative sources of market structure data are not readily available. Analyzing ADRs mentioned in diabetes drug forums and comparing them with ADRs reported in formal media venues provides additional evidence for the validity of the proposed approach.

The analysis of the adjectives and nouns commonly mentioned with each car model provided insightful information with respect to the content of the discussion. Our investigation of factors that drive the co-occurrence of cars in the forum reveals that cars that share similar characteristics and/or are similarly mentioned with respect to terms referring to the luxuriousness of the car, the value consumers receive from it, its appearance, and the emotional sentiment mentioned about the car are more likely to be mentioned together in a message. We further explore specific textual relationships and go beyond mere co-occurrence of terms to investigate textual sentiment associated with problems mentioned with different cars and drug–ADR relationships. These analyses demonstrate the ability of the text-mining approach to zoom in on a discussion to assess the competitive market structure through consumers' perceptions of the products' attributes (car problems or ADRs). Thus, in using a text-mining apparatus such as the one described here, firms can monitor their market position over time at a higher resolution and often lower cost relative to traditional data sources.

Using a case study of the Cadillac brand, we demonstrate how the proposed text-mining approach can track the *dynamics* in market structure using the real-time stream of data that consumer forums provide. This analysis also highlights the use of the text-mining approach to measure the effectiveness of a marketing campaign and to showcase how a marketing manager can affect brand position using marketing actions. The comparison to the trends in trade-in data suggests that although the campaign was useful in changing top-of-mind association, the impact on actual car switching is much slower. Future research could explore this opportunity further and utilize text mining to track the effectiveness of marketing campaigns. Additionally, for new (or repositioned) products, one can use text mining of consumer forums to study market-structure dynamics before, during, and after the launch of those products.

In analyzing market structure, the objective is often more descriptive than predictive. We therefore focus on utilizing text-mining and network analysis tools to describe the nature of discussions in a forum. The comparison against traditional data collection methods suggests that the proposed approach has the required external validity to reflect not only the opinions and views of forum members but also the views of the wider population of consumers. Future research might explore the potential for using the proposed approach as a predictive tool. Such an endeavor should consider further how well forum participants represent the population of consumers at large and the risk of firms manipulating the discussion (Dellarocas 2006). Although text mining allows us to minimize recall bias and demand effects, which are commonly found in survey-based data collection methods, views posted on forums may be biased because respondents aim for their views to be publicly available on the Web and because of online herding behavior (Huang and Chen 2006).

One could extend the application of the text-mining apparatus beyond consumer forums to the mining of blogs, product reviews, and more formal news articles. In fact, mining more formal corpora is often easier because the context of the discussion is more organized and the language used tends to follow grammatical standards and rules.

In summary, we hope the text-mining and derived market-structure analysis presented in this paper provides a first step in exploring the extremely large, rich, and useful body of consumer data readily available on Web 2.0.

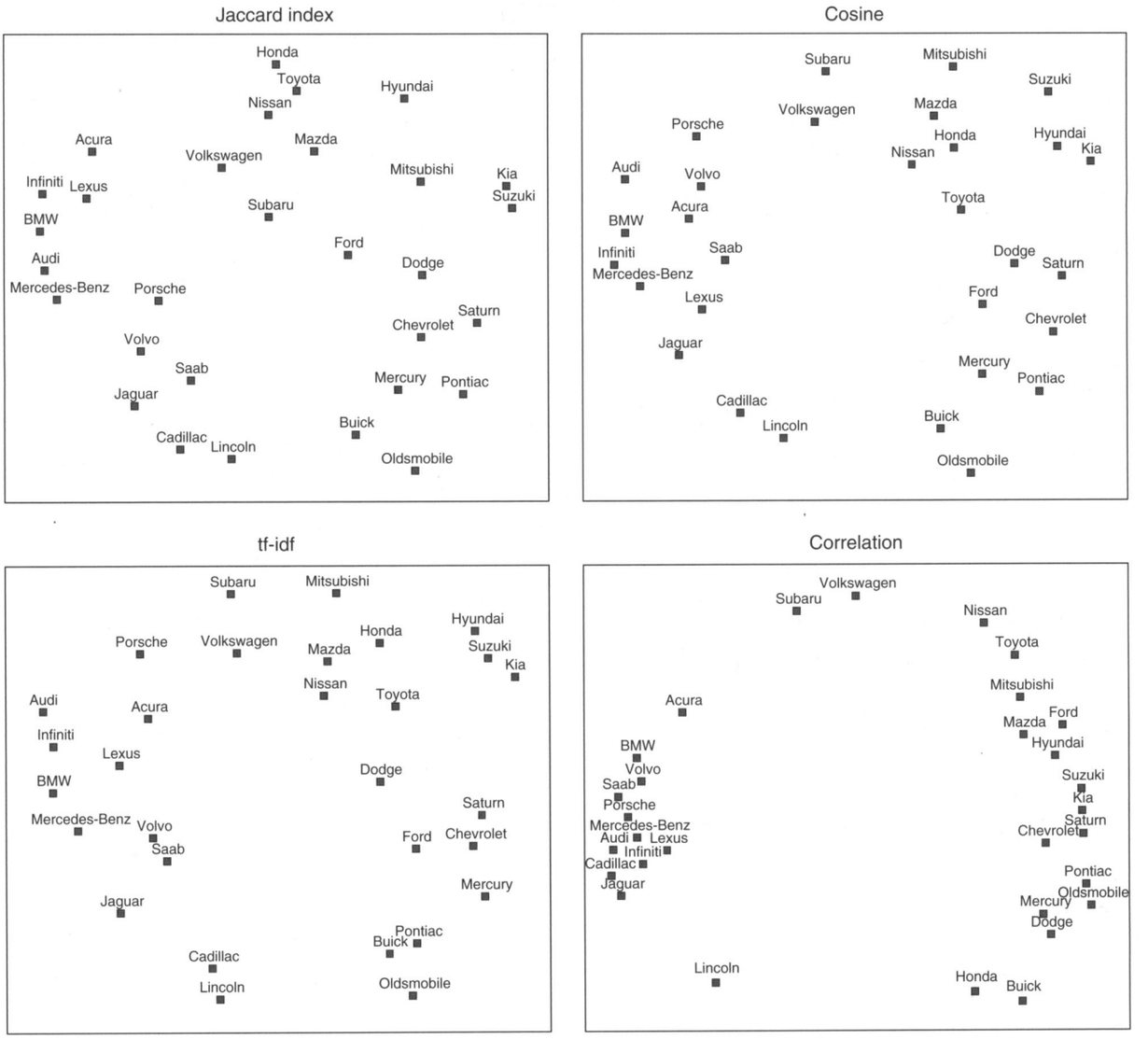
Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://mktsci.journal.informs.org/.

Acknowledgments

The authors thank the Marketing Science Institute, the Wharton Customer Analytics Initiative, and the Ogilvy Foundation for financial support. The authors thank Florian Stahl for data support; Niv Goldman, Yael Karlinsky, and Lauren Pully for research assistance; and Don Lehmann, Olivier Toubia, Kamel Jedidi, and seminar participants at Columbia University, the University of Delaware, and the Harvard Business School for comments on earlier versions of this research.

Appendix. MDS Maps Generated Using Different Similarity Measures



References

Akiva N, Greitzer E, Krichman Y, Schler J (2008) Mining and visualizing online Web content using BAM: Brand Association Map. *Proc. Second Internat. Conf. Weblogs Soc. Media 2008* (Association for the Advancement of Artificial Intelligence, Seattle), 170–171.

Anderson JR, Bower GH (1973) *Human Associative Memory* (V.H. Winston & Sons, Washington, DC).

Anderson RJ, Freedland KE, Clouse RE, Lustman PL (2001) The prevalence of comorbid depression in adults with diabetes: A meta-analysis. *Diabetes Care* 24(6):1069–1078.

Archak N, Ghose A, Ipeirotis PG (2011) Deriving the pricing power of product features by mining consumer reviews. *Management Sci.* 57(8):1485–1509.

Callon M, Law J, Rip A (1986) *Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World* (Macmillan, Houndmills, Basingstoke, UK).

Chen P-Y, Wu S-Y, Yoon J (2004) The impact of online recommendations and consumer feedback on sales. *Proc. Internat. Conf. Inform. Systems 2004* (Association for Information Systems, Washington, DC), 711–724.

Chevalier JA, Mayzlin D (2006) The effect of word of mouth on sales: Online book reviews. *J. Marketing Res.* 43(3):345–354.

Chintagunta PK, Gopinath S, Venkataraman S (2010) The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Sci.* 29(5):944–957.

Collins AM, Loftus EF (1975) A spreading-activation theory of semantic processing. *Psych. Rev.* 82(6):407–428.

Cooper LG, Inoue A (1996) Building market structures from consumer preferences. *J. Marketing Res.* 33(3):293–306.

Csere C (2003) Cadillac stakes a claim in the luxury-roadster arena. *Car Driver* (June) http://www.caranddriver.com/reviews/2004-cadillac-xlr-road-test.

Das SR, Chen MY (2007) Yahoo! for Amazon: Sentiment extraction from small talk on the Web. *Management Sci.* 53(9):1375–1388.

Dave K, Lawrence S, Pennock DM (2003) Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *Proc. 12th Internat. Conf. World Wide Web* (ACM, New York), 519–528.

Day G, Shocker AD, Srivastava RK (1979) Identifying competitive product markets: A review of customer oriented approaches. *J. Marketing* 43(Fall):8–19.

Decker R, Trusov M (2010) Estimating aggregate consumer preferences from online product reviews. *Internat. J. Res. Marketing* 27(4):293–307.

Dellarocas CN (2006) Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Management Sci.* 52(10):1577–1593.

Dellarocas C, Zhang XM, Awad NF (2007) Exploring the value of online product ratings in revenue forecasting: The case of motion pictures. *J. Interactive Marketing* 21(4):23–45.

Ding X, Liu B, Zhang L (2009) Entity discovery and assignment for opinion mining applications. *Proc. 15th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining 2009* (ACM, New York), 1125–1134.

Dörre J, Gerstl P, Seiffert R (1999) Text mining: Finding nuggets in mountains of textual data. *Proc. Fifth ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 398–401.

Duke J, Friedlin J, Ryan P (2011) A quantitative analysis of adverse events and "overwarning" in drug labeling. *Arch. Internal Medicine* 171(10):944–946.

Dwyer P (2009) Measuring interpersonal influence in online conversations. Working paper, Marketing Science Institute, Cambridge, MA.

Edwards RI, Aronson JK (2000) Adverse drug reactions: Definitions, diagnosis, and management. *Lancet* 356(9237):1255–1259.

Eliashberg J, Hui SK, Zhang JZ (2007) From story line to box office: A new approach for green-lighting movie scripts. *Management Sci.* 53(6):881–893.

Eliashberg J, Jonker J-J, Sawhney MS, Wierenga B (2000) MOVIEMOD: An implementable decision-support system for prerelease market evaluation of motion pictures. *Marketing Sci.* 19(3):226–243.

Farquhar PH, Herr PM (1993) The dual structure of brand associations. Aaker DA, Biel AL, eds. *Brand Equity and Advertising: Advertising's Role in Building Strong Brands* (Lawrence Erlbaum Associates, Hillsdale, NJ), 263–277.

Feldman R, Sanger J (2006) *The Text Mining Handbook* (Cambridge University Press, New York).

Feldman R, Govindaraj S, Livnat J, Segal B (2010) Management's tone change, post earnings announcement drift and accruals. *Rev. Accounting Stud. J.* 15(4):915–953.

Feldman R, Fresko M, Goldenberg J, Netzer O, Ungar L (2007) Extracting product comparisons from discussion boards. *Proc. Seventh IEEE Internat. Conf. Data Mining 2007* (IEEE, Piscataway, NJ), 469–474.

Feldman R, Fresko M, Goldenberg J, Netzer O, Ungar L (2008) Using text mining to analyze user forums. *Proc. Service Systems Service Management 2008 Internat. Conf.* (IEEE Systems, Man, and Cybernetics Society, Melbourne, VIC, Australia), 1–5.

Feldman R, Fresko M, Kinar Y, Lindell Y, Liphstat O, Rajman M, Schler Y, Zamir O (1998) Text mining at the term level. Zytkow JM, Quafafov M, eds. *Principles Data Mining Knowledge Discovery Proc. Second Eur. Sympos.* (Lecture Notes in Artificial Intelligence, Vol. 1510) (Springer-Verlag, Berlin), 65–73.

Ghose A, Ipeirotis PG (2011) Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Trans. Knowledge Data Engrg.* 23(10):1498–1512.

Ghose A, Ipeirotis PG, Li B (2012) Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Sci.* 31(3):493–520.

Girvan M, Newman MEJ (2002) Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* 99(12):8271–8276.

Glance N, Hurst M, Nigam K, Siegler M, Stockton R, Tomokiyo T (2005) Deriving market intelligence from online discussion. *Proc. Eleventh ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 419–428.

Godes D, Mayzlin D (2004) Using online conversations to study word-of-mouth communication. *Marketing Sci.* 23(4):545–560.

Godes D, Mayzlin D, Chen Y, Das S, Dellarocas C, Pfeiffer B, Libai B, Sen S, Shi M, Verlegh P (2005) The firm's management of social interactions. *Marketing Lett.* 16(3):415–428.

Grover R, Srinivasan V (1987) A simultaneous approach to market segmentation and market structuring. *J. Marketing Res.* 24(2):139–153.

Harshman RA, Green PE, Wind Y, Lundy ME (1982) A model for the analysis of asymmetric data in marketing research. *Marketing Sci.* 1(2):205–242.

He Q (1999) Knowledge discovery through co-word analysis. *Library Trends* 48(1):133–159.

Henderson GR, Iacobucci D, Calder BJ (1998) Brand diagnostics: Mapping branding effects using consumer associative networks. *Eur. J. Oper. Res.* 111(2):306–327.

Hu M, Liu B (2004) Mining and summarizing customer reviews. *Proc. Tenth ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 22–25.

Huang J-H, Chen Y-F (2006) Herding in online product choice. *Psych. Marketing* 23(5):413–428.

John DR, Loken B, Kim K, Monga AB (2006) Brand concept maps: A methodology for identifying brand association networks. *J. Marketing Res.* 43(4):549–563.

Kamada T, Kawai S (1989) An algorithm for drawing general undirected graphs. *Inform. Processing Lett.* 31(1):7–15.

Keohane E (2008) ComScore: Traffic to health information sites grows. *Direct Marketing News* (September 10) http://www.dmnews.com/comscore-traffic-to-health-information-sites-grows/article/116509/.

Krackhardt D (1988) Predicting with networks: Nonparametric multiple regression analysis of dyadic data. *Soc. Networks* 10(4):359–381.

Lafferty J, McCallum A, Pereira F (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proc. 18th Internat. Conf. Machine Learn.* (Morgan Kaufmann, San Francisco), 282–289.

Lee T, Bradlow E (2011) Automated marketing research using online customer reviews. *J. Marketing Res.* 48(5):881–894.

Lehmann DR (1972) Judged similarity and brand-switching data as similarity measures. *J. Marketing Res.* 9(3):331–334.

Liu Y (2006) Word-of-mouth for movies: Its dynamics and impact on box office revenue. *J. Marketing* 70(3):74–89.

Liu B (2011) Opinion mining and sentiment analysis. Liu B, ed. *Data Centric Systems and Application: Web Data Mining*, 2nd ed. (Springer-Verlag, Berlin), 459–526.

Liu B, Hu M, Cheng J (2005) Opinion observer: Analyzing and comparing opinions on the Web. *Proc. 14th Internat. Conf. World Wide Web* (Association for Computer Machinery, Chiba, Japan), 342–351.

Malouf R, Davidson B, Sherman A (2006) Mining web texts for brand associations. *Proc. AAAI-2006 Spring Sympos. Comput. Approaches Analyzing Weblogs* (AAAI Press, Stanford, CA), 125–126.

McCallum A, Wellner B (2005) Toward conditional models of identity uncertainty with application to noun coreference. *Adv. Neural Inform. Processing Systems* 17:905–912.

Ochoa X, Duval E (2008) Quantitative analysis of user-generated content on the web. Roure DD, Hall W, eds. *Proc. First Internat. Workshop Understanding Web Evolution '08 (Beijing, China)*, 19–26.

Pang B, Lee L (2008) Opinion mining and sentiment analysis. *Inform. Retrieval* 2(1–2):1–135.

Rao VR, Sabavala DJ (1981) Inference of hierarchical choice processes from panel data. *J. Consumer Res.* 8(1):85–96.

Reichheld FF, Teal T (1996) *The Loyalty Effect: The Hidden Force Behind Growth, Profits, and Lasting Value* (Harvard Business School Press, Boston).

Rosa JA, Spanjol J, Porac JF (2004) Text-based approaches to marketing strategy research. Moorman C, Lehmann DR, eds. *Assessing Marketing Strategy Performance* (Marketing Science Institute, Cambridge, MA), 185–211.

Saiz A, Simonsohn U (2012) Downloading wisdom from online crowds. *J. Eur. Econom. Assoc.* Forthcoming.

Salton G, McGill MJ (1983) *Introduction to Modern Information Retrieval* (McGraw Hill, New York).

Schindler RM, Bickart B (2005) Published word of mouth: Referable, consumer-generated information on the Internet. Haugtvedt CP, Machleit KA, Yalch RF, eds. *Online Consumer Psychology: Understanding and Influencing Consumer Behavior in the Virtual World* (Lawrence Erlbaum Associates, Hillsdale, NJ), 35–61.

Seshadri T, Tellis GJ (2012) Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Sci.* 31(2):198–215.

Silva-Risso J, Shearin WV, Ionova I, Khavaev A, Borrego D (2008) Chrysler and J.D. Power: Pioneering scientific price customization in the automobile industry. *Interfaces* 38(1):26–39.

Swanson DR (1988) Migraine and magnesium: Eleven neglected connections. *Perspect. Biol. Medicine* 31(4):526–557.

Swanson DR, Smalheiser NR (2001) Information discovery from complementary literatures: Categorizing viruses as potential weapons. *J. Amer. Soc. Inform. Sci. Tech.* 52(10):797–812.

Turney PD, Littman ML (2003) Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inform. Syst.* 21(4):315–346.

Urban GL, Hauser JR (2004) "Listening in" to find and explore new combinations of customer needs. *J. Marketing* 68(2):72–87.

Urban GL, Johnson PL, Hauser JR (1984) Testing competitive market structures. *Marketing Sci.* 3(2):83–112.

U.S. Department of Energy (2012) Fuel economy. http://www.fueleconomy.gov.

Welch D (2003) The second coming of Cadillac. *Bus. Week* (November 24), 79–80.