Identifying Market Structure: A Deep Network Representation Learning of Social Engagement

Yi Yang
Assistant Professor
Hong Kong University of Science and Technology
LSK Business Building, RM 4041
Clear Water Bay, Sai Kung, HK
+852-23586359
imyiyang@ust.hk

Kunpeng Zhang
Assistant Professor
University of Maryland, College Park
4316 Van Munching Hall
7699 Mowatt Ln
College Park, MD, USA 20742
301-405-0702
kzhang@rhsmith.umd.edu

P.K. Kannan
Professor of Marketing
University of Maryland, College Park
3445 Van Munching Hall
7699 Mowatt Ln
College Park, MD, USA 20742
301-405-9187
pkannan@rhsmith.umd.edu

November 15, 2019

Identifying Market Structure: A Deep Network Representation Learning of Social Engagement

Abstract

With rapid technological developments, product-market boundaries have become more dynamic. Consequently, competition for products and services is emerging outside the product-market boundaries traditionally based on SIC and NAICS classification codes. Identifying these fluid product-market boundaries is critical for firms not only to compete effectively within a market, but also to identify lurking threats and latent opportunities. Extant methods using surveys on consumer perceptions or purchase data will be unable to identify the impact that a brand from outside the boundary may have on brands within a product-market. Newly available big data on social media engagement presents such an opportunity. We propose a deep network representation learning framework to capture latent relationships among thousands of brands and across many categories, using millions of social media users' brand engagement data. We build a heterogeneous brand-user network and then compress the network into a lower dimensional space using a deep Autoencoder technique. We validate our technique using a novel link-prediction method and visualize the learned representations pictorially. We illustrate how our method can capture the dynamic changes of product market boundaries using two well-known events: the acquisition of Whole Foods by Amazon and the introduction of the Model 3 by Tesla.

Keywords: AI, Machine Learning, Deep Representation Learning, Social Media, Competitive Market Structure, Big Data

Introduction

Identifying product-market boundaries and examining the strength of competition between brands within a product-market has long been an important issue with strategic implications for next-generation product design, product positioning, new customer acquisition, and pricing and promotion decisions (Bergen and Peteraf 2002; Kannan and Wright 1991; Urban, Johnson, and Hauser 1984). Over the years, academics and practitioners have contributed significantly to developing various methods to define and identify market structure (see review by Shugan 2014. These include survey-based methods such as brand concept maps (BCM) (John et al. 2006) and ZMET (Zaltman and Coulter 1995), methodologies based on observational purchase data (e.g., brand switching) (Kannan and Sanchez 1994), consideration sets (Ringel and Skiera 2016), and scanner-based purchase data (Erdem 1996; Novak 1993; Shugan 1987). Within the online context, researchers have used unstructured user click streams (Moe 2006), online search logs (Kim, Albuquerque, and Bronnenberg 2011; Ringel and Skiera 2016), and customer reviews (Lee and Bradlow 2011). Some of these methods use data from the bottom of the purchase funnel, such as evaluation and purchase stage data, and thus assume the product-market boundaries are pre-specified. Even if some of these methods use data from the top of the funnel at the awareness or pre-evaluation stage, such as forum discussions (Netzer et al. 2012) and hashtags (Nam, Joshi, and Kannan 2017), the data volume and scalability issues force these methods to define a product-market boundary first and then examine the competition ***within*** the pre-specified product-market to make these methods implementable. Thus, many of the methods will not be able to capture the changes that occur to the product-market boundaries and/or the impact that a brand from outside the boundary may have on brands within a product-market.

However, with the rapidly changing competitive environment ushered in by technological developments, firms need to be aware of how the product-market boundaries themselves are changing and how competitive threats and opportunities may be emerging outside of the narrowly defined product-market boundaries. For example, the digital camera product-market has been upended by developments in smartphone categories: large e-commerce platforms are entering into many retail markets and firms are integrating vertically and horizontally. In such situations, market boundaries based on SIC and NAICS codes may be unable to spot new threats and opportunities. Instead of using data captured at the lower end of the customer journey, managers now need to view the product-market boundaries from a much higher level and derive insights from data that spans categories and brands. This is precisely what our paper seeks to do using large-scale (over a hundred million) social media user engagement data. We provide valuable insights into overlapping product-market boundaries and changes in those boundaries that help managers spot competitors and complements, identify cross-promotion strategies, and develop firm-level strategies.

Our research generates important insights regarding the product-market boundaries by applying deep-learning techniques on user-brand engagement data generated on social media platforms. Recently, owing to its ability to handle big data, deep learning-empowered Artificial Intelligence (AI) (Agrawal, Gans, and Goldfarb 2018; LeCun, Bengio, and Hinton 2015) has made important contributions in many applications such as healthcare (Esteva et al. 2019) and marketing (Timoshenko and Hauser 2019). At the same time, social media platforms such as Facebook, Twitter, and Instagram, have generated a large volume of user data on brand engagement activities such as following brand public pages, liking, commenting, and sharing brand posts. The goal of our study is to analyze this informative brand-user engagement data to identify latent relations among a large number of brands and reveal the product-market boundaries. Previous research has documented that a user connecting to a brand online can be interpreted as an expression of affinity (Kuksov, Shachar, and Wang 2013; Naylor, Lamberton, and West 2012), and that most online fans are offline customers (Pereira, de Fátima Salgueiro, and Mateus 2014). We make the minimal assumption that if a user interacts with two brands, for example, Samsung and an HTC phone, it indicates the user has some level of interest -- greater than awareness -- in both brands. Simply by sharing interested users, the two brands are related either as a substitute or complement to one another. If such patterns exist after observing activities on various brands from a large group of users (which could be millions of users on a social media platform), we could argue that such a pair of brands have latent relations on some dimension, as learned from user-brand data.

Based on the above premise, we first construct a large-scale heterogeneous user-brand network based on user engagement on brands' social media public fan pages. Then, we propose a deep network representation learning method to discover relationships within the data. Specifically, we develop a deep learning method suitable for (1) handling large data efficiently and (2) learning complex patterns from data effectively. The process leads to a low-dimensional representation (i.e., a vector) for each brand and each user by training a deep Autoencoder on the network data. The deep Autoencoder is similar to traditional dimensionality reduction methods such as Principal Component Analysis in capturing latent factors in data with few dimensions. It is however very different from those methods in that it uses a non-linear transformation function to understand the latent patterns in data and at the same time reduce the noise in the data. In our context, the deep Autoencoder can preserve the first-order (user-brand direct connection) and the second-order (two users connecting to the same brand, or one user connecting to two different brands) network topology so that brands with network structural equivalence are located closer in the representation space, while brands with dissimilar network structures are located further away. This method also projects users and brands onto the same dimensional space, which can be used for many different follow-up analyses. For example, in this study we take learned brand vectors and apply state-of-the-art visualization tools such as *t-SNE* (Maaten and Hinton 2008) to visualize the product-market boundaries characterizing the brands. Moreover, the proposed framework can capture changes in product-market boundaries by constructing a sequence of networks across different time frames to understand the dynamics of market structure.

We validate the product-market boundaries identified through our methodology using a link prediction method, where using a calibration sample we learn the network representation and use that to predict the network structure in a validation sample. The results show that our proposed approach significantly outperforms several baselines on two standard metrics of predicting user-brand engagement on out-of-sample data. We also establish the face validity of the results through the identification of product-market boundaries. Our analysis of the user-brand engagement data of over five thousand brands and nearly 26 million users reveals product-market boundaries with high face validity – grouping of specific categories, high-end brands, and overlaps. Our event studies on Amazon's acquisition of Whole Foods and Tesla introducing the Model 3 illustrate how our methodology captures the changes in product-markets associated with these events. We also discuss how the market structure maps can reveal opportunities and threats facing a brand.

The key contribution of our paper lies in leveraging the information embedded in big data of user-brand engagement networks to identify product-market boundaries. Using deep learning techniques to examine over five thousand brands, we study market structure from a network analysis perspective that is very different from extant studies. Prior studies in competitive market structure were forced to pre-define the product-market boundaries due to scalability issues or through using data lower in the purchase funnel. We overcome such constraints using user-brand engagement network data at a much higher level (interest phase) and deep learning techniques that provide us with insights no one existing technique can accomplish at this scale and level of detail. Our paper is among the first to apply deep network representation learning on social media data and show its usefulness for market structure discovery. Specifically, we implement a deep Autoencoder to capture complex network structural equivalence to learn latent brand relations. We are able to pin a large number of brands on the market structure map to precisely visualize brand relationships using the learned vector representations. We can view the global market structure as well as the zoomed-in sub-market structure. Finally, we showcase how a new technology, Artificial Intelligence (AI) can be used to better tackle a traditional marketing problem. It is well known that three elements render AI techniques possible for-life applications: data, algorithm, and computing power (Agrawal, Gans, and Goldfarb 2018). In this paper, we leverage deep learning and a network representation learning (algorithm) to understand market structure using a large-scale social media data (data). The model implementation is efficient under Nvidia P100 GPU, with Tensorflow as the backend framework (computing power).

The remainder of this paper is organized as follows. We first give a brief background review of research on competitive market structures and discuss the theoretical foundation for using social media user engagement data to derive brand associations. We then describe our deep network representation learning approach, compare it with state-of-the-art baselines, and demonstrate its effectiveness. Based on the application, we discuss the key findings, insights, and the managerial implications. We conclude with a discussion on the limitations and future research direction.

Background and Theoretical Foundation

Extant work in identifying competitive market structures dates to the 1970s (e.g., Kalwani and Morrison 1977; Day, Shocker and Srivastava 1979) when diary-panel based brand-switching purchase data and survey-based consumer judgments of substitution-in-use or similarities were used to construct market structure maps. Developments since then have been based on the availability of the volume and variety of data and methodology in terms of their sophistication and capability to handle large volumes of data. We briefly review them from the two perspectives of data and methodology.

Data

Early studies depended on customer data generated either at a late stage of the customer journey or the very beginning of the journey. For example, purchase data collected using diary-panels or survey of brand perceptions – judgements of similarities or substitution-in-use – were commonly used for constructing brand-switching data or perceptual maps of brand relationships. The increased availability of scanner-panel data of purchases, market structure models with marketing mix (e.g., Carpenter and Lehmann 1985; Kannan and Wright 1991), and dynamic market structure models (e.g., Erdem 1996) provided more detailed insights into inter-brand relationships and competition. Focusing on the early stages of the customer journey, approaches such brand concept maps (BCM) (John et al. 2006) and ZMET (Zaltman and Coulter 1995) relied on data collected using surveys and, therefore, were effort intensive. Given the scaling issues with the MLE-based models and the limitations with survey data, the market definition problem was ignored, and product-market boundaries were pre-specified generally at the category level so that a smaller number of brands within a category could be analyzed.

The advent of online sources, such as review platforms, social media platforms, and clickstream data, has dramatically increased the volume and variety of data for market structure studies, especially at the awareness, search, and consideration stages of the customer journey. For example, the study by Kim et al. (2011) relies on Amazon's customer search logs on camcorders to derive market structure. Lee and Bradlow (2011) visualize competitive market structure in the digital camera category using user-generated online customer reviews that mainly comment on product attributes and brands' relative positions. The study by Netzer et al. (2012) relies on data from online discussion forums to build a market structure of the automobile industry using a hybrid text mining and network analysis method. Similarly, Ringel and Skiera (2016) use search history from a product and price comparison site to derive customers' consideration sets that reflect competition among LED-TVs at the SKU level. It is important to note that even with a large volume of data, these studies pre-define the product-market boundaries at the category level to make the analyses viable.

There are other studies where the product-market boundaries are not pre-defined. For example, France and Ghose (2016) introduce a method for identifying, analyzing, and visualizing sub-markets in product categories from online reviews. Nam, Joshi, and Kannan (2017) use hashtags from a social tagging website to infer brand relations across categories. Similarly, Culotta and Cutler (2016) propose to extract brand-related attributes and build brand conception maps using hashtags from Twitter, where such pre-defined boundaries are not necessary. Our proposed methodology utilizes user-brand engagement data from social media platforms and focuses on the early stages of the customer journey, but does not restrict the analysis using pre-defined product-markets, rather it allows the boundaries to emerge from the data.

More importantly, the scale at which we analyze that data, which is much larger than any of the extant method, is key to analyzing the relationships spanning multiple categories. Our proposed method is capable of handling thousands of brands and millions of users. Analysis conducted on this order of magnitude has not been seen in previous research. Table 1 summarizes the studies based on the type of data used.

<Insert Table 1 Here>

Methodologies

Much of the online data generated in the early stages of the customer journey tend to be unstructured (online reviews, social-tags, hashtags, etc.). In order to extract product names or attributes included in data, researchers develop and apply various text mining-based technologies to reviews, discussions, and summaries in online forums. For example, Netzer et al. (2012) propose a model combining conditional random fields (CRF) and predefined linguistic rules to extract product keywords. Pant and Sheng (2015) focus on firm-generated content in websites, compute firm similarity based on textual descriptions (from the first few pages of firm websites), and use TF-IDF (term frequency-inverse document frequency) and website structures link analysis (in-links and out-links) to derive firm competing relationship. Ringel and Skiera (2016) construct consideration sets from consumer co-search data. Our method is significantly different from these methodologies. We use social media engagement data to construct an explicit heterogeneous user-brand network and develop a deep network representation learning approach to jointly learn low-dimensional representation of brands and users while preserving the original network topology. Recently, the application of such neural representation learning to very large-scale data has attracted attention. Timoshenko and Hauser (2019) employ neural representation learning of words to represent user-generated content and develop deep learning models to extract customer needs. Our work instead focuses on representations of a social network.

Extant studies have also used different methodologies to visualize market structure. Multidimensional scaling (MDS), a popular dimension-reduction technique, has been used to visualize a set of products with high dimensional features (DeSarbo et al. 2006, Kim et al. 2011, Netzer et al. 2012). As MDS is sensitive to the number of products, Ringel and Skiera (2016) propose a new model called DRMABS to visualize market structure, which can parsimoniously represent more than thousand SKUs in one map. In our research, we apply a state-of-the-art visualization tool called *t-SNE*[1] for visualizing market structure among brands. *t-SNE* has been largely adopted in many high-dimensional data visualization tasks and is not sensitive to the number of data points. Table 2 summarizes selected extent studies on market structure analysis in multiple dimensions and highlights the positives of our proposed methodology.

<Insert Table 2 Here>

Social Media Engagement

Our proposed methodology analyzes social media engagement data in the form of user-brand links. Many social media platforms such as Facebook, Twitter, and Instagram host public fan pages created by firms to facilitate the communication with customers and promote products. The user-brand engagement could be in the form of a user "liking" a brand or a post by the brand, "sharing" the brand post, or "commenting" on a brand post. Since each of these "likes," "shares," and "comments/posts" is a user-brand link in our study, it is important to understand what they represent. Surveys of fans of brands have revealed many reasons as to why users "like" a brand or post/share comments. These include: to support a brand they like, to get a coupon or discount, to get regular updates from the brands they like, to participate in contests, to share personal experiences, to share their interests/lifestyles with others, to research brands, to imitate a friend who likes the brand, to act on a recommendation from another fan, etc. (Syncapse 2013; Pellitier and Horky 2015; Pereira, de Fátima Salgueiro, and Mateus 2014).

[1] *t-SNE* is a nonlinear dimensionality reduction technique that is particularly well-suited for visualizing high-dimensional data (Maaten and Hinton 2008).

Prior research has examined the importance of such a marketing channel and studied the impact of user engagement on brand image and customer purchase intentions with mixed results (De Vries, Gensler, and Leeflang 2012; Lipsman et al. 2012; Naylor, Lamberton, and West 2012; Goh, Heng, and Lin 2013; Hoffman, Novak, and Kang 2017). For example, (Goh, Heng, and Lin 2013) find that user engagement in social media brand communities leads to a positive increase in purchase expenditures. Mochon et al. (2017) use a field experiment to find that users who liked a gym brand online were likely to become members of that gym offline. It is not the organic liking by the user that lead to the offline purchase; more often "liking" a Facebook page is used as a platform for firm-initiated promotional communications. In another field experiment setting, John et al. (2017) find that "liking" is simply a symptom of a positive brand attitude and does not imply the fan is any more loyal to the brand or any more likely to purchase the brand. Additionally, it is only when users who liked the brand are targeted using promotional communication by the firm that purchase probabilities increase. Thus, for our research purposes we will treat a "like," "share," or a "comment/post" as exhibiting positive attitude or interest towards the brand at the beginning of the customer journey. Such a tendency for users to connect to brands is generally interpreted as an expression of affinity (Culotta and Cutler 2016; Kuksov, Shachar, and Wang 2013; Naylor, Lamberton, and West 2012), which is consistent with our treatment.

Our work takes a data-driven angle to examine the user-brand links which reflect a minimal level of affinity of users to the brands, and thus form a large heterogeneous user-brand network. Our proposed approach is motivated by research in social network analysis suggesting that social network structure equivalence reflects value/interest homophily and can be used to measure social proximity (McPherson, Smith-Lovin, and Cook 2001). These proximities are measured by the user-brand links which indicate how close are users to brands, brands to other brands, and users to other users, and become input to deriving market structure. In order to capture complex and latent brand associations at a scale and frequency that is not possible using extant methods, we aggregate millions of user engagement links and examine these relationships using deep learning.

Methodology

Many social network platforms, such as Facebook, Instagram, and Twitter, can be abstracted as a network containing business (firm) accounts and individual user accounts. The public fan pages of business accounts are used by firms to communicate with their customers and fans. Users interact with brands and with each other in different ways, such as commenting, liking, sharing, and following. Our challenge in this paper is to analyze the large social network formed by such links with multiple types of nodes (i.e., brands and users) and their associated activities to effectively and efficiently discover latent relationships among brands. We propose a deep network representation learning framework to address this problem. There are several steps in the overall framework as summarized as in Figure 1.

<Insert Figure 1 Here>

Step 1: Data Collection. We specify a set of brands that is of interest in the social network platform. We then download all available user engagement data from the brands' public fan page during the appropriate time window depending on managerial interest. A user engagement is defined as either liking or commenting on a firm's post on its public fan page. We only obtain user IDs for the sake of preserving privacy of users. The details of data collected are described in the following section on Data.

Step 2: Network Construction. We start with a cleansing operation to remove spurious users. We then construct a heterogeneous user-brand network including all selected brands and all users engaging with them. A brand node and a user node are connected if the user engages with the brand. The strength of an edge between a brand node and a user node is the engagement frequency.

Step 3: Deep Network Representation Learning. The deep network representation learning algorithm represents each node (brand or user) as a low-dimensional vector, also known as node embedding. Representation learning is essential to data-driven analysis, and the learned low-dimensional embeddings are useful for the downstream task of identifying and visualizing the product-market boundaries. Timoshenko and Hauser (2019) adopt pre-trained word embeddings, where each word is represented as a low-dimensional vector, to extract insights from textual reviews. However, our node embeddings are trained via an unsupervised Deep Autoencoder.

The objective in using an Autoencoder is to learn the representation of the data and encode the user-brand link structure so that the node can be represented in a lower dimensional space while the structure is preserved. The dimensionality reduction is achieved by training the network to ignore the “noise” in the data and focus on the primary latent structure. The Autoencoder reduces the dimensionality of the input data to a “bottleneck” (the reduced encoding), and using the reduced encoding as input, reconstructs a representation of the original data. Learning occurs through backpropagation of the error to get the reconstructed representation as close as possible to the original representation while eliminating noise. It is the bottleneck reduced encoding we are interested in for developing market structure. In essence, we can compare the dimensionality reduction functionality of the Autoencoder with that of Principal Component Analysis (PCA). While in PCA the reduced dimensions are linear combinations of the input variables, the reduced dimensions in Autoencoder are non-linear and non-orthogonal achieved through non-linear activations of the neurons allowing the model to learn more powerful generalizations than what PCA can.

In our application, the Autoencoder works on the large heterogeneous network in an attempt to preserve the network structure such that (i) nodes directly connected have similar vectors (closer to each other) in the reduced embedding space, and (ii) nodes that are not directly connected but share structural equivalence (such as many common neighbors) are also similar in the embedding space. These two types of similarity are referred to as the first-order (direct connection) similarity and the second-order (network structural equivalence) similarity. Formally, we denote an aforementioned network as $G = (V^b, V^u, E)$, where $V^b = (v_1^b, v_2^b, \dots, v_n^b)$ represents a set of n brand nodes, $V^u = (v_1^u, v_2^u, \dots, v_m^u)$ represents m user nodes, and $E = \{e_{i,j}\}, i \leq m, j \leq n$ represents all links between users and brands. $e_{i,j}$ indicates an engagement between user i and brand j. Given such network G, network representation aims to learn a mapping function $f: v_i^b, v_j^u \longmapsto w_i^b, w_j^u \in R^d$, where $d \ll \min(m, n)$. w_i^b, w_j^u are called brand embedding and user embedding, respectively. A commonly used embedding dimensionality d is 300 (Mikolov, Chen, et al. 2013; Mikolov, Sutskever, et al. 2013). The objective of the mapping function is to develop appropriate embeddings so that the brand proximities, brand-user proximities, and user proximities exhibited in the original data are preserved as much as possible in the reduced embedding space. Technical details of the Autoencoder methodology are discussed in Appendix A1.

Prior research of network analysis relies on network adjacent matrix representation, that is, a brand node is represented as a $|V^u|$-dimension vector where $|V^u|$ is the number of unique user nodes in the network. Each element in the vector corresponds to a user. If the user at a particular index has a connection with the brand in the network, that element is marked as 1 and 0 otherwise. The brand vector is usually very sparse given the fact that each brand only engages with a small subset of users. Similarly, the user vector is also very sparse since each user only engages with a small subset of brands. Using this representation to measure similarity is inaccurate -- not mention inefficient -- for such an order of magnitude. In contrast, representing brands as dense low-dimensional vectors allows us to capture brand relations from multiple facets, and we use a toy example (shown in Figure 2) to illustrate how network representation learning works.

<Insert Figure 2 Here>

Suppose we have three brands (B1, B2, B3) and five users (U1, U2, U3, U4, U5) in a network. Representation learning aims to find a mapping function so that each node is represented as a low-dimensional vector (for the sake of this illustration let us assume that it is 3-dimensional). The mapping function is optimal when nodes exhibiting similar structures (first-order and/(or) second order) are projected onto similar vectors in the reduced embedding space (assumed to be 3-dimensional space in this illustration). Since U1 engaged with B1, we expect the vector representation of B1 and U1 to be close. Similarly, B2's representation is closer to B1's representation than to B3's as B2 shares more users with

B1 than with B3. Since B2 has some additional network structure, such as connections with U4 and U5, its representation leans towards U4 and U5. All user representations are jointly learned in a similar way.

Step 4: Market Structure Discovery. Once we obtain vector representation for brands and users, we can use learned embeddings to efficiently compute similarity among brands and to visualize natural clusters of related brands in a low-dimensional space. In this study, we primarily focus on identifying brand relationships. Finding similar brands to a focal brand can be achieved by a nearest neighbor search based on the widely-used cosine similarity. Cosine similarity measures the cosine of the angle between two vectors and has a range [-1, 1][2]. Visualizing natural clusters of related brands can be achieved by a dimension reduction method, such as *t-SNE* (Maaten and Hinton 2008), which projects high-dimensional data into a low-dimensional space (e.g., two or three dimensions). It has been used for visualization in a wide range of applications and is especially well-suited for visualizing high-dimensional representations learned from deep neural networks. *t-SNE* preserves the distance of data points well such that data points nearby in the high-dimensional space would be close in the lower dimensional space, while distant data points would be further apart in the lower dimensional space. Specifically, the input of *t-SNE* is the vectors in the reduced dimension space with d=300 and the output is the vectors with 3 dimensions. Thus, we shall observe that related brands are surrounding each other in the reduced 3-dimensional space after *t-SNE*.

Data

In this study, we use Facebook as our empirical benchmark, as it is one of the largest and most representative online social network platforms. Note that our model can be generalized to other similar social network platforms.

[2] Note that we do not use Jaccard similarity because the brand vector space is a continuous space and Jaccard similarity is specifically designed for a discrete space, nor Euclidian distance due to its poor performance in a high dimensional space.

To collect Facebook data, we first obtain a list of U.S. brands with the most followers from the social media marketing website Socialbakers[3]. Facebook public fan pages are categorized into several groups on Socialbakers, such as Brands, Celebrities, Community, Entertainment, Media, Place, Society, and Sport. Without loss of generality, we focus on the “Brands” category as it covers a wide range of different industries and is more interesting to marketers. In total, we obtain 5,478 different brands, covering 25 different categories. On Facebook, firms post on their public fan pages and allow users to comment, like, and share. The posts become an important marketing channel for businesses to interact with their customers. An example of a post on Walmart’s fan page is shown in Figure 3. We use Facebook Graph API[4] to download all activities visible on a brand page such as posts by the brand administrator, as well as posts by users, including comments and likes on brand posts. Facebook added more reaction emotions such as ‘love,’ ‘haha,’ ‘wow,’ ‘sad,’ and ‘angry’ in 2016. Our dataset does not include these reaction emotions because the Graph API returns only the ‘likes.’ Moreover, users can ‘share’ brands’ posts, but the Graph API does not provide individual level data regarding who shares which post, rather it provides an overall share count for each post. Therefore, our dataset does not include ‘share’ engagement. It is worth emphasizing that to ensure privacy protection, we do not download any user profile information nor examine the content of user comments. The dataset collected for this study covers the duration from January 1, 2017, through January 1, 2018.

Data cleaning. To ensure data quality and robust results, we design a set of rules to remove fake users and their corresponding activities. As fake accounts and fraudulent activities have become more pervasive, researchers and social media firms are increasingly paying more attention to these problems (Mukherjee, Liu, and Glance 2012; Van Vlasselaer et al. 2016; Zahedi, Abbasi, and Chen 2015). In August 2012, Facebook admitted in a regulatory filing that 8.7% of its 955 million active accounts were

---

[3] Socialbakers is a global AI-powered social media marketing company offering a marketing software-as-a-service platform called the Socialbakers Suite. It includes data from Facebook, Twitter, and YouTube. https://www.socialbakers.com/

[4] https://developers.facebook.com/docs/graph-api/

fake[5]. For comments on Facebook brand pages to reflect genuine user experiences, opinions, and interactions with brands, such fraudulent activities should be detected and removed. Following prior work on Facebook (Zhang, Bhattacharyya, and Ram 2016), we replicate a set of similar rules to remove fake users and their posts. Table 3 describes the resulting data using a heterogeneous user-brand network. The brands' degree distribution (number of connections) exhibits a scale-free distribution (shown in Figure 3), a well-documented phenomenon in most social networks.

<Insert Table 3 Here>

<Insert Figure 3 Here>

Evaluation and Results

In this section, we first introduce our design to quantitatively evaluate our proposed methodology and compare its performance with several baselines. Then we present the market structure derived from our learned brand representation. Specifically, we visualize the market structure using a sophisticated dimension reduction technique *t-SNE*, and identify the top similar brands for several focal brands in different industries.

Evaluation using Link Prediction

In studying market structure, there is lack of ground truth about the identified structure, that is, knowledge of what the "true" structure is. As a result, demonstrating the performance of various proposed methods is challenging. One may argue that the industry classification (e.g., SIC or NAICS) of brands can be used for evaluation and face validity for the results. However, these classification systems are static, do not re-classify firms as the product market evolves, and, therefore, are unable to accommodate innovations that create entirely new product markets (Bhojraj, Lee, and Oler 2003; Hoberg and Phillips 2016; Jacobs and O'Neill 2003). To address the challenge, we propose an alternative and novel way to evaluate the

[5] Kelly, Heather (2012), "83 Million Facebook accounts are fakes and dupes," CNN, August 3, http://www.cnn.com/2012/08/02/tech/social-media/facebook-fake-accounts/

identified market structure. An identified market structure is a function of the brand representation, meaning that an accurate representation -- able to capture latent and complex semantic and structural relationships among brands and users -- is more likely to identify valid market structures. This approach is supported by prior research showing a strong relationship between brand image and the characteristics of brand's supporters and followers (Naylor, Lamberton, and West 2012; Kuksov, Shachar, and Wang 2013; Culotta and Cutler 2016). If a network learning method is capable of accurately representing network nodes accounting for these relationships between brands and users, then it would be able to predict the future links between brands and users accurately. Therefore, we use a cross-validation procedure using a novel *link prediction research design*, where we predict the most likely formed links of user-brand engagement in an out-of-sample network given the brand vectors and user vectors learned from a training network. The likelihood of a link formation is measured by the proximity of a brand vector and a user vector. Note that link prediction performance is significantly correlated with the quality of learned vectors, given the assumption that a better network representation learning can predict new interactions between users and brands with a high accuracy. To demonstrate the superiority of our proposed method, we compile a set of representative baselines (see details below). Our analysis shows that (i) link prediction using representation learned from our heterogeneous user-brand network performs better than a reduced homogenous network – a widely used method by extant approaches; (ii) deep learning-based methods learn better representation than shallow machine learning methods such as matrix factorization; and (iii) our deep learning-based model is robust and able to handle sparse networks as compared to baselines.

To demonstrate the effectiveness of our framework, we compile a 2x2 research design with two different network structures (homogeneous vs. heterogeneous) and two different algorithms (shallow model vs. deep model) (see the algorithm in Appendix A2). The link prediction process follows the seminal work (Liben-Nowell and Kleinberg 2007). Let $G_{0,2} = (V^b_{0,2}, V^u_{0,2}, E_{0,2})$ denote a network snapshot during a time period (t_0, t_2). The network $G_{0,2}$ can be chronologically split into two non-overlapping subnetworks $G_{0,1} = (V_{0,1}^b, V_{0,1}^u, E_{0,1})$ and $G_{1,2} = (V_{1,2}^b, V_{1,2}^u, E_{1,2})$. Conventionally, we call $G_{0,1}$ and $G_{1,2}$ training network and testing network, respectively. The overall evaluation process is as follows. First, we train on $G_{0,1}$ to obtain brand representation (and user representation). Second, we randomly select N users in the period of (t_0, t_1). For each user, we calculate its proximity to all non-connected brands. We sort all proximity scores for all N users and choose top k pairs (denoted as L) as predicted links. Finally, we evaluate the performance using two standard metrics: $precision@k$ and $recall@k$, defined below. Precision indicates the accuracy of the link prediction algorithm while recall is referred to as the true positive rate or sensitivity. The larger value for both metrics indicates the better performance.

$$precision@k = \frac{|L \cap E_{1,2}|}{k}, \quad recall@k = \frac{|L \cap E_{1,2}|}{|E_{1,2}^T|},$$

where $E_{1,2}$ is the set of all newly formed links in $G_{1,2}$. $precision@1$ checks whether a non-connected brand-user pair with the highest proximity in the training period forms a link in the testing network. Note that this evaluation process might be slightly different when it comes to a brand-brand homogenous network where we only have vector representation for brands. To obtain the proximity scores to all non-connected brands for N randomly selected users, we employ a weighted average strategy, a similar idea used in the item-based collaborative filtering framework. For each user u_i, we have all brands that u_i has connected (i.e., b_1, b_2, …, b_m that u_i connects to in $G_{0,1}$). The similarity score between u_i and each non-connected brand b_j is S_{ij}: $S_{ij} = \frac{\sum_{k=1}^{m} S_{kj}}{m}$, where S_{kj} is the similarity between brand b_j and brand b_k that u_i connects to in $G_{0,1}$, m is the number of brands u_i connects to in $G_{0,1}$.

Prior research on network analysis commonly constructs an implicit brand-brand network where two brands are connected if there is a common user (or keywords). The advantage of analyzing such a homogeneous network is an increase in computational efficiency because the network size is dramatically reduced. However, such a simplified operation that converts an original heterogeneous user-brand network into an implicit homogeneous network usually results in decreased performance because some important information encoded in user-brand interactions is completely ignored. In contrast, our deep learning-based approach performs well because it jointly learns optimal representation for both brands and users while preserving latent relationships among brands and users. For the shallow model, we use matrix factorization, denoted as *MF*. *MF* is commonly used to learn latent low-dimensional factors by decomposing the matrix of user-brand interactions. Factorization algorithms such as singular-value decomposition and nonnegative matrix factorization have spurred the development of many different applications in other domains, especially in recommender systems. For our *MF* benchmark method, we use the singular-value decomposition (SVD) to decompose the user-brand interaction matrix M into a lower rank approximation: $M = U\Sigma V^T$, where U conceptually represents how much each user "likes" an underlying dimension, V^Tconceptually represents how relevant each underlying dimension is to each brand, and Σ is a diagonal matrix of singular values, which are essentially weights. For the purpose of prediction, we first approximate the original matrix through U, Σ, and $V^T$and then predict a link to a brand with the highest predicted preferences that the user has not connected.

We use the user-brand interactions from the first half of time span in our data to build a training network ($G_{0,1}$) and use the second half to build a testing network ($G_{1,2}$). $G_{1,2}$ has 7,247,410 links (1,996,354 new links), formed by 1,547,762 users and 1,511 brands (also in the training network). Note that the accuracy of a link prediction under a random strategy is approximately 0.085%.

For our link prediction procedure, we randomly select 100 and 1,000 users and summarize the performance in Table 4 and Table 5. Note that all *p*-values in parentheses are obtained under a t-test from 10 runs of every model. We can see that our method significantly outperforms baselines in both *precision@k* and *recall@k* at all different *k*'s. As an illustration, consider Column 4 (with k = 1,000) in Table 4, the shallow model on the homogeneous brand-brand network has precision of 0.078 which means that only 7.8% of the predicted links are actually formed during the testing period. In contrast, our deep model on the same network brings a slight improvement to 0.082. This suggests that shallow and deep models have comparable performance when data is small and homogeneous. We then apply our deep model to the heterogeneous network, which significantly improves the precision by approximately 58.9% over traditional methods (0.124 vs. 0.078). We observe similar trends for metric *recall@k*. For *k*=1,000, shallow and deep models on the homogeneous network are able to retrieve 60.2% and 68.6% links, respectively (Table 4, Column 4). Further, *precision@k* decreases as *k* becomes larger, while *recall@k* increases. When *k* is large, many false links are predicted as well as true links. This reflects the famous precision-recall tradeoff that any model can be adjusted to improve precision at the expense of recall, or vice versa.

By further investigating Table 4 (*N*=100) with Table 5 (*N*=1,000), we find that the *precision* is higher and the *recall* is lower when the number of selected users is large. This is expected because we have higher chances to select true positive links when the number of users increases. On the other hand, the total number of true links in the testing network also increases by a magnitude, and thus the *recall* decreases.

<Insert Table 4 Here>

<Insert Table 5 Here>

To study the impact of training size on performance, we vary the training size with different network sparsity. We randomly remove a certain percentage of links from the training network and learn representation of users and brands. Then we predict user-brand links and measure the *precision* and *recall* using the out-of-sample testing network. As we can see in Table 6, our method still significantly outperforms baselines, especially when network sparsity is extremely high. For example, our method improves the *precision*@1000 by 77.7% (0.183 vs. 0.103) and *recall*@1000 by 55.0% (0.080 vs. 0.124), when only 10% links are kept in the training network. This suggests that our method handles sparsity better than baselines, which is very important since most real-world networks are extremely sparse.

<Insert Table 6 Here>

Visualization of Market Structure

With the learned brand representation vectors, we can visualize how the brands are clustered, and even zoom-in to examine sub-markets. We use *t-SNE* to obtain market structure visualization. *t-SNE* (Maaten and Hinton 2008) is a popular dimension reduction technique largely used by the machine learning community to visualize representations, and it is shown to preserve global structure better than the multi-dimensional scaling (MDS) that has been used in prior marketing research (e.g. Kim, Albuquerque, and Bronnenberg 2011). In our context, we use *t-SNE* on the learned 300-dimensional brand representations to obtain the associated two-dimensional visualization map. Figure 4 presents the global structure of the brands in our Facebook data. Each data point in the figure denotes a brand belonging to one of the 25 categories, and each category is indicated by a different color. We can interpret the visualization as follows: the closer any two brands are in the figure, the more similar their brand representations are in the 300-dimensional space (see Figure 4). The color codes in the map indicate brands in the same industry/category, with the industry/category label self-identified by the brands themselves on Facebook.

<Insert Figure 4 Here>

There are several observations from the global Facebook brand market structure map. First, there are clear clustering patterns, particularly between brands in the same industry (points with same color tend to be in a cluster). For example, Cluster 1 in Figure 4 includes non-luxury domestic and imported automobile brands such as Toyota, Nissan, Mazda, as well as some automobile accessories brands such as Michelin, DENSO, and Auto Parts. Note that in our data we have several luxury car brands such as BMW, Mercedes-Benz, Audi, Tesla, and Maserati, which are not close to the brands in Cluster 1. In fact, they are clustered in a different region of the map with other luxury brands such as Channel, Gucci, Cartier, and others. Such a separation between luxury car brands and non-luxury car brands further confirms that brand representation learned from our approach captures latent semantics in multiple dimensions, not only on the industry dimension but also on the price dimension. Second, some brands appear in categories which are different than what one would normally expect. There are two explanations for this category mixture (in addition to the one related to the latent semantics captured in the luxury-non-luxury automobile case). First, our category label is obtained from the self-identified category label indicated by each brand on Facebook. For example, Amazon lists itself in the retail category instead of the e-commerce category while Apple's chosen category is service rather than electronics. Second, each brand might have several businesses across different categories. For example, Amazon's business is related to the high-tech, shipping and delivery industry, as well as retail and supermarket (after acquiring Whole Foods) industries. The strength of our methodology lies in capturing these relationships into a single map given the ease with which it locates thousands of brands in the market structure map, thereby highlighting the complex product-market boundaries characterizing these brands.

To further examine the specifics of the product-market boundaries, we zoom-in on the four areas in the figure to examine the sub-market structures, which are displayed in Figure 5. Subfigure 1 which we already discussed displays automobile brands along with automobile accessories and motorcycle brands at the top. Subfigure 2 displays premium vacation resort brands, such as The Signature at MGM Grand and the Coconut Bay Beach Resort & Spa. Subfigure 3 and Subfigure 4 contain airline brands and cosmetic brands, respectively. Taken together these maps provide face validity to our methodology in terms of core brands making up a category and the overlaps among product-markets.

<Insert Figure 5 Here>

Additionally, since each brand in our data is associated with a category, we can visualize how the 25 Facebook categories are related. We take a weighted average on vectors of all brands within each category to obtain a category vector. It can be considered as a "centroid" of all brands that belong to that category. More formally, given a category C that includes a set of brands $\{b_1, b_2, \ldots, b_k\}$, where $\boldsymbol{v_i}$ is the vector representation of each brand b_i, f_i is the number of users who have engaged with b_i. Then, we use a weighted average to obtain the category representation of C as: $\boldsymbol{v_c} = \sum_{i=1}^{k} \log(f_i)\, \boldsymbol{v}_i$.

Once we obtain a vector representation for all 25 categories, we visualize them on a two-dimensional space using *t-SNE*, as shown in Figure 6. We can see that travel is next to airlines, and not surprisingly, alcohol is close to sporting-goods and gambling.

<Insert Figure 6 Here>

Identifying Proximal Brands

While visual mapping is sufficient to provide a gestalt picture of all the five-thousand plus brands in the aggregate, it does not provide the actual distance between the brand vectors in the reduced dimension space. Since identifying proximal brands for competitor or complement analysis is a critical task in marketing decisions (Day, Shocker, and Srivastava 1979), we focus on identifying proximal brands from the perspective of a focal brand. In doing so, we offer a new perspective that reflects the competitive or complementary relationships in the social network space.

In this study, we choose United Airlines and Southwest Airlines from the airlines category and Audi USA and Nissan from the automobile category, as these brands are generally regarded as having different consumer bases and belonging to different sub-markets. Each of the four brands is referred to as a focal brand, and we find their top-10 proximal brands based on cosine similarity. From the results in Table 7 we can obtain several interesting insights. First, our method is able to capture specific brand latent characteristics. For example, Southwest Airlines is generally considered as a low-budget airline compared to United. The brands most proximal to Southwest Airlines and United reflect this difference. The proximal brands for Southwest Airlines are JetBlue, Frontier Airline, and Allegiant, while the most proximal brands for United are major domestic and international airlines, such as American Airlines, Delta, Lufthansa, All Nippon, Air China, LATAM Airlines, and Air New Zealand. Similar results also are identified in the automobile industry. Top proximal brands to Audi USA are Mercedes-Benz USA and BMW USA, which are generally high-end luxury car brands. In contrast, Nissan is closer to Mazda, Toyota, and Volkswagen, all of which produce more affordable cars.

Second, we also observe the asymmetric competition (cf., Ringel and Skiera 2016). Given a brand A and its top proximal brand set S_A, we can identify a set of brands S_B that is proximal to a brand B in S_A. S_B need not necessarily contain A. And even if $A \in S_B$, the order of A in S_B might be very different from the order of B in S_A. For example, Southwest Airlines is the fourth most proximal brand to United while United ranks sixth in the set of top proximal brands to Southwest Airlines.

Third, unlike prior market structure analysis where proximal brands are usually from the same category as the focal brand, the top most proximal brands derived from our analysis are from different categories. For example, a brand called "Airfarewatchdog" is proximal to both United and Southwest Airlines. Airfarewatchdog is a deal-finder for flight tickets and has a large follower base (over 1 million) on Facebook. Traditional market analysis may simply ignore this brand, as it is not an airline. Further, it is also interesting to see that Southwest Airlines is closer to Airfarewatchdog than to United which may indicate that the fans of Southwest Airlines are more likely to use a deal finder before purchasing flight tickets; thus, Airfarewatchdog could be a complement to Southwest when customers look for cheap flights at that site and end up at Southwest, or it could potentially compete with Southwest. In either case, Southwest could focus more on this site and examine the nature of the relationship. Southwest Airlines could also examine its relationship with the hotel brand Hyatt as it is proximal to Southwest. Similarly, we see Kawasaki USA, an innovative motorsport vehicle manufacturer, is proximal to Audi USA. This cross-industry brand proximity very well demonstrates that representation learning can capture latent characteristics of brands and explore brand relationship from different perspectives. We believe this advantage can provide new insights to marketing analysis.

<Insert Table 7 Here>

Case Studies on Market Structure Dynamics

Market structure evolves over time and can change dramatically especially under an unexpected industry shock. Whether our proposed method can be adaptively learned is also of interest as it could provide useful insights to marketing practitioners. In this section, we analyze how market structure changes under exogenous shocks by analyzing two case studies: (1) Amazon acquiring Whole Foods, and (2) Tesla introducing the Model 3. We take a before-after strategy where we use data for 3-month pre- and 3-month post the event announcement day and calculate the change in distance from the focal brand (e.g., Amazon and Tesla) to other representative brands that are selected from the same category, as shown in Figure 8 and Figure 9. The purpose of the event study is to examine how a focal brand relationship with other brands change as a major event happens. Specifically, for Amazon-Whole Foods, we select several brands from the retail and e-commerce category, and for Tesla, we select several brands from the auto category. This demonstrates that our proposed approach is able to learn effective representation; as a result, the dynamics in market structure are well captured. We calculate the change between focal brand i's representation w_i^b and target brand j's representation w_j^b before and after the specific event using cosine similarity: $cossim\left(w_i^b{}_{after}, w_j^b{}_{after}\right) - cossim\left(w_i^b{}_{before}, w_j^b{}_{before}\right)$. Therefore, positive numbers indicate similarity increase while negative numbers mean the decrease in similarity.

Amazon acquires Whole Foods

Amazon acquired Whole Foods in August 2017. This acquisition has had significant impacts on the grocery and retail industries. It is widely believed that Amazon plans to use its acquisition of Whole Foods to enter into the online grocery delivery business. Amazon and Whole Foods run separate Facebook pages. After the merger of the two firms, we see from Figure 8 that Amazon is more proximal to retail brands as measured by cosine similarity, while the proximity to other relevant brands decreases slightly. For example, the cosine similarity between Amazon and Loews Home Improvement decreases by 0.184. In contrast, the cosine similarity between Amazon and other super-market brands increases. Among them, proximity of Amazon to Whole Foods increases by 0.202, and increases between Amazon and Kroger by 0.165. As inferred from our data-driven model, Amazon even becomes more proximal to Walmart indicating that Amazon's competitive market structure landscape has shifted. By further examining our data, we find that after the Whole Foods acquisition the number of common users who interact with both Amazon and Whole Foods on their Facebook public pages increases. Some Amazon users posted comments on Whole Foods fan page mentioning Amazon. For example, in a Whole Foods post "Here are 6 New Healthy Products Coming to Whole Foods in March," a user, who had liked an Amazon post earlier, commented "You mean AMAZON… as they bought Whole Foods…right?" This direct link between Amazon and Whole Foods leads the deep Autoencoder to strength the proximity between the two brands. Moreover, in another Whole Foods post, a user who had liked a Kroger post earlier posted "The quality has gone downhill and prices have soared…. You've made Kroger look appealing….'' Although we do not find this user has ever interacted with Amazon before, her interaction with Whole Foods leaves an implicit connection between Amazon and Kroger which could be captured by the deep Autoencoder. In short, after Amazon acquired Whole Foods, online social media users who are Amazon's fans pay more attention to Whole Foods, and users who are fans of other supermarket brands engage more with Whole Foods due to the acquisition event. As a result, the deep Autoencoder captures the dynamics and updates the brand representation accordingly.

<Insert Figure 7 Here>

The acquisition by Amazon has an impact on the market structure of Whole Foods too. In Figure 9, we consider Whole Foods as the focal brand and calculate the change in proximities to other brands before and after the acquisition. Based on the results, we observe that Whole Foods' proximity to other retail brands such as Target, Walmart, and Best Buy increases. Among them, the proximity to Amazon increases the most due to the increased common users between them. In contrast, Whole Foods' proximity to supermarket brands such as Goya Foods, Enjoy Life Foods, and HelloFresh slightly decreases. Second, the magnitude of change in proximity values is smaller than that of Amazon to other brands. This seems to indicate that the acquisition has less impact on Whole Foods as it is still positioned around other supermarket brands, while Amazon is expanding closer to the grocery retail category.

<Insert Figure 8 Here>

Tesla announces the Model 3

Tesla sells two types of sedans, the Model S and the Model 3. The Model S is a luxury premium sedan with a larger range of acceleration and customization options, while the Model 3 is designed and built as a mass-market affordable electric vehicle. The Model S can cost over $100,000 depending on the configuration, while the Model 3 costs approximately $35,000. After the announcement of the new Model 3, we see that Tesla becomes more distant from luxury car brands and closer to non-luxury car brands. Examining data from the Auto Gallery, a Southern California premiere luxury and exotic dealership, we can see in Figure 10 that the cosine similarity between Tesla and luxury car brand Maserati decreases by 0.209. Similar trends exist between Tesla and other high-end or luxury car brands such as BMW, Mercedes-Benz, Audi, and so on. Meanwhile, Tesla becomes more proximal to Kia, Mazda, and other more affordable car brands.

<Insert Figure 9 Here>

Testing for Significance

In the above analysis, we compute the distance change between the focal brand (e.g., Amazon or Whole Foods) and other brands, before and after the acquisition. We can see that there is a significant increase in similarity between Amazon and Whole Foods after the acquisition. However, whether this distance change is caused by the acquisition or other unobserved factors, such as the difference of data split and noise of data, still remains unclear. Therefore, we conduct a further analysis by randomly splitting all data before the acquisition into two parts (i.e., d1 and d2, with d1 before d2). We then measure the distance between Amazon and Whole Foods using d1 and d2 separately. We repeat this process 30 times using different data cuts in the pre-acquisition data. The average distances between the two brands across using all d1s and d2s are 0.228 and 0.232, respectively. The two-tailed t-test on the distance is 0.055, which indicates there is no statistically significant difference between the distances between Amazon and Whole

Foods before and after the acquisition in different cuts of the pre-acquisition data. Accordingly, the substantial increase in similarity between Amazon and Whole Foods is not attributed to sample differences.

We perform a similar process on Tesla's introduction of the Model 3. In particular, we choose one non-luxury brand, Mazda, and compute its distances to Tesla before the event using various data splits. The average distances between Mazda and Tesla across using all d1s and d2s are 0.185 and 0.191, respectively, with a p-value of 0.076. This seems to indicate there is no statistically significant difference between Mazda and Tesla when the cutting point of data varies before the event. Therefore, we conclude that after Model 3's announcement, Tesla becomes more similar to non-luxury automobile brands on the social media platform. Note that we also conduct analyses on Tesla and other automobile brands and the results are consistent.

Implications and Conclusion

As our proposed approach handles a larger number of brands and millions of user engagement data across these brands, the results are very useful for brand managers to get a gestalt view of the relationships across thousands of brands. The visualization of potentially overlapping product-market boundaries across many categories helps managers to identify latent threats and potential opportunities which cannot be done with extant methods. For example, for Southwest, is Airfarewatchdog a potential competitor who might draw visitors away from Southwest or is it a complementor who would increase visits to Southwest? Having identified the overlapping market with Airfarewatchdog, Southwest could invest more attention to evaluate the exact nature of this relationship. If Airfarewatchdog is a competitor, then Southwest might focus on developing strategies to differentiate itself and channel visitors to its website exclusively. If it is a complementor, then Southwest might run display ad campaigns on Airfarewatchdog's website. Also, given Hyatt is closely associated with Southwest with common users who "like" both brands, Southwest could run mutually beneficial joint promotions with Hyatt. Identifying such unusual or unexpected insights is the greatest advantage of our approach.

Another important strategic use of our market structure maps is to identify competitors and complementors across categories and track how these relationships change over time. While Hoberg and Phillip (2016) use text analysis to 10-K statements to identify such grouping based on product descriptions that the firms provide, we provide a more dynamic structure based on actual customer/user social media activities. Moreover, our market structure map is more forward-looking and predictive of emerging competition and complementors and more proactive than those based on 10-K statements, which can be viewed as reactive. Since Hoberg and Phillips (2010) show that merging firms with more similar product descriptions in their 10-Ks results in more successful outcomes, using our market structure maps to identify merger and acquisitions targets (firms sharing common users) may have similar benefits. For example, given that Kawasaki USA, a motorsport vehicle manufacturer, is proximal to Audi USA, is there a benefit for Audi USA to acquire Kawasaki USA?

The power of our method lies in its ability to capture the dynamic changes in market structure. Since the maps are based on the analysis of big data which can be collected in a relatively short window of time, our methodology can be used to track changes in their relative position when firms introduce new products, new promotions, and new marketing initiatives. The case studies that we highlighted provide good illustrations of this. Additionally, although we have not analyzed this in the paper, firms can deploy our method to enhance their social network-based marketing efforts by better targeting specific potential customers, since user nodes in the network are also learned and represented as vectors in the same multi-dimensional space as brands. Our link prediction design demonstrates a possible utilization for targeting. Lastly, our proposed method is generalizable to other similar platforms if we can construct a heterogeneous user-brand network from public fan pages' engagement data.

Our research has some limitations. First, our analysis is conducted on one social network, Facebook. Even though Facebook is one of the largest online social networks that has billions of users and thousands of brands, it is likely that users on different platforms may exhibit different engagement behavior and some of the research findings may not be generalized to other platforms. For example, it is reported that Instagram users and Facebook users have different age groups[6]. We could apply the same technique to other social media platforms and compare findings. When it comes to the dynamic market structure analysis, we generate a series of networks at each give time window. Our current analysis treats these networks as equally important. In fact, the networks at an earlier stage might affect subsequent networks, because user-brand interactions might be dependent on their prior activities. Incorporating dependency among networks in a temporal way into the representation learning algorithm can improve market structure analysis, which we leave as our future work. Finally, each link in the user-brand network is created when the user engages with the brand on the public page. Facebook has introduced various reaction emotions to the platform to allow users interact with brands in different ways, such as Like, Love, Haha, Wow, Sad and Angry. Future work can build a multi-relation network to deeply capture user-brand engagement heterogeneity.

[6] https://www.statista.com

References

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb (2018), *Prediction machines: the simple economics of artificial intelligence*, Harvard Business Press.

Baesens, Bart (2014), *Analytics in a big data world: The essential guide to data science and its applications*, John Wiley & Sons.

Bergen, Mark and Margaret A. Peteraf (2002), "Competitor identification and competitor analysis: a broad-based managerial approach," *Managerial and Decision Economics*, 23 (4–5), 157–69.

Bhojraj, Sanjeev, Charles M.C. Lee, and Derek K. Oler (2003), "What's my line? A comparison of industry classification schemes for capital market research," *Journal of Accounting Research*, 41 (5), 745–774.

Carpenter GS, Lehmann DR (1985) A model of marketing mix, brand switching, and competition. Journal of Marketing Research 22(3):318–329.

Chintagunta, Pradeep, Dominique M. Hanssens, and John R. Hauser (2016), "Marketing Science and Big Data," *Marketing Science*, 35 (3), 341–342.

Culotta, Aron and Jennifer Cutler (2016), "Mining Brand Perceptions from Twitter Social Networks," *Marketing Science*, 35 (3), 343–62.

Day, George S., Allan D. Shocker, and Rajendra K. Srivastava (1979), "Customer-Oriented Approaches to Identifying Product-Markets," *Journal of Marketing*, 43 (4), 8–19.

De Vries, Lisette, Sonja Gensler, and Peter S.H. Leeflang (2012), "Popularity of brand posts on brand fan pages: An investigation of the effects of social media marketing," *Journal of Interactive Marketing*, 26 (2), 83–91.

DeSarbo, Wayne S., Rajdeep Grewal, and Jerry Wind (2006), "Who competes with whom? A demand-based perspective for identifying and representing asymmetric competition," *Strategic Management Journal*, 27 (2), 101–29.

Erdem, Tülin (1996), "A Dynamic Analysis of Market Structure Based on Panel Data," *Marketing Science*, 15 (4), 359–78.

Esteva, Andre, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean (2019), "A guide to deep learning in healthcare," *Nature Medicine*, 25 (1), 24.

Goh, Khim-Yong, Cheng-Suang Heng, and Zhijie Lin (2013), "Social media brand community and consumer behavior: Quantifying the relative impact of user-and marketer-generated content," *Information Systems Research*, 24 (1), 88–107.

Hoberg, Gerard, Gordon Phillips (2010), "Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis, " The Review of Financial Studies, Volume 23, Issue 10, October 2010, Pages 3773–3811.

——— and Gordon Phillips (2016), "Text-based network industries and endogenous product differentiation," *Journal of Political Economy*, 124 (5), 1423–1465.

Hoffman, Donna L., Thomas P. Novak, and Hyunjin Kang (2017), "Let's get closer: Feelings of connectedness from using social media, with implications for brand outcomes," *Journal of the Association for Consumer Research*, 2 (2), 216–228.

Jacobs, Gabriel and Cliona O'Neill (2003), "On the reliability (or otherwise) of SIC codes," *European Business Review*, 15 (3), 164–169.

John, Deborah Roedder, Barbara Loken, Kyeongheui Kim, and Alokparna Basu Monga (2006), "Brand concept maps: A methodology for identifying brand association networks," *Journal of marketing research*, 43 (4), 549–563.

John, L. K., Emrich, O., Gupta, S., & Norton, M. I. (2017), "Does "Liking" Lead to Loving? The Impact of Joining a Brand's Social Network on Marketing Outcomes," *Journal of Marketing Research*, 54(1), 144–155.

Kalwani MU, Morrison DG (1977) A parsimonious description of the Hendry system. Management Science 23(5):467–477.

Kannan, P. K. and Susan M. Sanchez (1994), "Competitive Market Structures: A Subset Selection Analysis," *Management Science*, 40 (11), 1484–99.

———, Gordon P. Wright, and N. Dan Worobetz (1991), "Testing for competitive submarkets," *International Journal of Research in Marketing*, 8 (3), 187–203.

Kim, Jun B., Paulo Albuquerque, and Bart J. Bronnenberg (2011), "Mapping online consumer search," *Journal of Marketing Research*, 48 (1), 13–27.

Kuksov, Dmitri, Ron Shachar, and Kangkang Wang (2013), "Advertising and consumers' communications," *Marketing Science*, 32 (2), 294–309.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015), "Deep learning," *Nature*, 521 (7553), 436–44.

Lee, Thomas Y. and Eric T. Bradlow (2011), "Automated marketing research using online customer reviews," *Journal of Marketing Research*, 48 (5), 881–894.

Liben-Nowell, David and Jon Kleinberg (2007), "The link-prediction problem for social networks," *Journal of the American Society for Information Science and Technology*, 58 (7), 1019–1031.

Lipsman, Andrew, Graham Mudd, Mike Rich, and Sean Bruich (2012), "The power of 'like': How brands reach (and influence) fans through social-media marketing," *Journal of Advertising Research*, 52 (1), 40–52.

Maaten, Laurens van der and Geoffrey Hinton (2008), "Visualizing data using t-SNE," *Journal of Machine Learning Research*, 9 (Nov), 2579–2605.

Martin, David, Hailin Wu, and Adil Alsaid (2003), "Hidden surveillance by Web sites: Web bugs in contemporary use," *Communications of the ACM*, 46 (12), 258–264.

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook (2001), "Birds of a feather: Homophily in social networks," *Annual review of sociology*, 27 (1), 415–444.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013), "Efficient Estimation of Word Representations in Vector Space," *arXiv:1301.3781 [cs]*.

———, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean (2013), "Distributed Representations of Words and Phrases and their Compositionality," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds., Curran Associates, Inc., 3111–3119.

Mochon, Daniel, Karen Johnson, Janet Schwartz, and Dan Ariely (2017), "What Are Likes Worth? A Facebook Page Field Experiment," *Journal of Marketing Research*, 54 (2), 306–17.

Moe, Wendy W. (2006), "An empirical two-stage choice model with varying decision rules applied to internet clickstream data," *Journal of Marketing Research*, 43 (4), 680–692.

Mukherjee, Arjun, Bing Liu, and Natalie Glance (2012), "Spotting fake reviewer groups in consumer reviews," in *Proceedings of the 21st international conference on World Wide Web*, ACM, 191–200.

Nam, Hyoryung, Yogesh V. Joshi, and P.K. Kannan (2017), "Harvesting Brand Information from Social Tags," *Journal of Marketing*, 81 (4), 88–108.

Naylor, Rebecca Walker, Cait Poynor Lamberton, and Patricia M. West (2012), "Beyond the 'like' button: The impact of mere virtual presence on brand evaluations and purchase intentions in social media settings," *Journal of Marketing*, 76 (6), 105–120.

Netzer, Oded, Ronen Feldman, Jacob Goldenberg, and Moshe Fresko (2012), "Mine Your Own Business: Market-Structure Surveillance Through Text Mining," *Marketing Science*, 31 (3), 521–43.

Novak, Thomas P. (1993), "Log-linear trees: models of market structure in brand switching data," *Journal of Marketing Research*, 30 (3), 267–287.

Pant, Gautam and Olivia R. L. Sheng (2015), "Web Footprints of Firms: Using Online Isomorphism for Competitor Identification," *Information Systems Research*, 26 (1), 188–209.

Pereira, Hélia Gonçalves, Maria de Fátima Salgueiro, and Inês Mateus (2014), "Say yes to Facebook and get your customers involved! Relationships in a world of social networks," *Business Horizons*, 57 (6), 695–702.

Ringel, Daniel M. and Bernd Skiera (2016), "Visualizing Asymmetric Competition Among More Than 1,000 Products Using Big Search Data," *Marketing Science*, 35 (3), 511–34.

Shugan, Steven M. (1987), "Estimating brand positioning maps using supermarket scanning data," *Journal of Marketing Research*, 24 (1), 1–18.

Shugan, Steven M. (2014), "Market Structure Ressearch," in *The History of Marketing Science*, eds. Russ Winer and Scott Neslin, World Scientific Publishing Co., pp. 129-164.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014), "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, 15 (1), 1929–1958.

Timoshenko, Artem and John R. Hauser (2019), "Identifying customer needs from user-generated content," *Marketing Science*, 38 (1), 1–20.

Urban, Glen L., Philip L. Johnson, and John R. Hauser (1984), "Testing competitive market structures," *Marketing Science*, 3 (2), 83–112.

Van Vlasselaer, Véronique, Tina Eliassi-Rad, Leman Akoglu, Monique Snoeck, and Bart Baesens (2016), "Gotcha! Network-based fraud detection for social security fraud," *Management Science*, 63 (9), 3090–3110.

Wedel, Michel and P. K. Kannan (2016), "Marketing analytics for data-rich environments," *Journal of Marketing*, 80 (6), 97–121.

Xu, Haitao, Shuai Hao, Alparslan Sari, and Haining Wang (2018), "Privacy risk assessment on email tracking," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, IEEE, 2519–2527.

Zahedi, Fatemeh Mariam, Ahmed Abbasi, and Yan Chen (2015), "Fake-website detection tools: Identifying elements that promote individuals' use and enhance their performance," *Journal of the Association for Information Systems*, 16 (6), 448.

Zaltman, Gerald and Robin Higie Coulter (1995), "Seeing the voice of the customer: Metaphor-based advertising research," *Journal of Advertising Research*, 35 (4), 35–51.

Zhang, Kunpeng, Siddhartha Bhattacharyya, and Sudha Ram (2016), "Large-Scale Network Analysis for Online Social Brand Advertising.," *MIS Quarterly*, 40 (4).

Appendixes

A1: Deep Autoencoder

In this study, the network representation, also known as network embedding, is learned through a deep Autoencoder, an unsupervised learning model consisting of two joint components, an encoder, and a decoder. The encoder, implemented by a deep fully-connected feed forward neural network, is a compressor that transforms the input data into a latent representation (e.g., a low-dimensional vector), while the decoder is a reverter that reconstructs the latent representation back to the original input data. As the input data is often high dimensional (in a magnitude of millions), learning effective low dimensional representation (several hundred) in an efficient way while preserving information in the input data as much as possible is not trivial. In our case we study a large user-brand network where each brand (or user) node is originally represented as a one-hot encoding vector that is fed into the Autoencoder for learning a latent low-dimensional vector. To illustrate the Autoencoder in details, we first formally define user-brand heterogeneous network and network representation learning.

Definition 1: user-brand Heterogeneous Network A user-brand heterogeneous network is denoted as $G = (V^b, V^u, E)$, where $V^b = (v_1^b, v_2^b, \ldots, v_n^b)$ represents n brand nodes, $V^u = (v_1^u, v_2^u, \ldots, v_m^u)$ represents m user nodes, and $E = \{e_{i,j}\}, i \leq m, j \leq n$ represents all links between users and brands. $e_{i,j}$ is the link weight that indicates the frequency of engagement between user i and brand j. Engagement is defined as liking or commenting by a user on a brand's Facebook fan page.

Definition 2: Network Representation Learning Given a user-brand heterogeneous network G, network representation learning aims to learn a mapping function $f\colon v_i^b, v_j^u \longmapsto w_i^b, w_j^u \in R^d$, where $d \ll \min(m, n)$. w_i^b, w_j^u are called brand embeddings and user embeddings, respectively. Following previous research, we set the embedding dimensionality d to 300.

The objective of the mapping function is to learn good embeddings so that the brand proximities, brand-user proximities, and user proximities are preserved at maximum. More specifically, given network-like inputs, we tend to preserve the following two network structures into the learned representations.

1. Similarity to neighbors. Our user-brand network is a bipartite network where brand nodes and user nodes are neighbors. A user node and a brand node are connected with a large weight, indicating a strong relationship between them. The similarity from this one-hop connection for all links between users and brands are measured by first-order loss function, denoted as L_{1st}. L_{1st} with weights incorporated incurs a penalty if neighboring nodes are projected far apart, similarly to the idea of Laplacian Eigenmaps (Belkin and Niyogi 2003). Therefore, minimizing L_{1st} is an attempt to preserve local distances; if v_i^b and v_j^u are similar, then w_i^b and w_j^u are close in the embedding space.

$$L_{1st} = \sum_{j=1}^{n}\sum_{i=1}^{m} e_{i,j}\left(w_i^b - w_j^u\right)^2$$

2. Similarity to neighbors of neighbors. In our user-brand network, neighbors' neighbors of a brand node are other brand nodes. Neighbors' neighbors of a user node are other user nodes. If two brands share many common users, their similarity should be high. Similarly, if two users are fans for many common brands, their similarity should be high too. The objective of network representation learning is designed in such a way that a network structure similarity should be well captured. Therefore, to minimize the reconstruction error (denoted as L_{2st}) by compressing the latent information in hidden layers, the Autoencoder has the following objective function measured by second- order loss function, denoted as L_{2nd}.

$$L_{2nd} = \sum_{i=1}^{m}(x_i^{b'} - x_i^b)^2 + \sum_{j=1}^{n}(x_j^{u'} - x_j^u)^2$$

where x_i^b and x_j^u are the input of brand v_i^b and user v_j^u for the deep Autoencoder, respectively. They are represented as an adjacent one-hot encoding vector by all other nodes. The dimensionality of x_i^b and x_j^u

equals the total number of brands and users $(m+n)$ in the network. Each element in the vector corresponds to a node in the network. If the node at a particular index connects the brand node v_i^b (or user node v_j^u), the corresponding element is marked as the engagement frequency, and as 0 otherwise. This adjacent representation is a very common way for representing nodes in a network (Liben-Nowell and Kleinberg 2007).

Therefore, our overall objective function is to minimize the sum of first-order loss and second-order loss, as below.

$$L = L_{1st} + L_{2nd} = \sum_{j=1}^{n}\sum_{i=1}^{m} e_{i,j}\left(w_i^b - w_j^u\right)^2 + \lambda(\sum_{i=1}^{m}\left(x_i^{b'} - x_i^b\right)^2 + \sum_{j=1}^{n}\left(x_j^{u'} - x_j^u\right)^2)$$

$x_i^{b\prime}$ and $x_j^{u\prime}$ are the output of the deep Autoencoder, which are the reconstructed representation of the input x_i^b and x_j^u, respectively. The hyper-parameter λ plays a role to balance the first-order loss and second-order loss, and its value are tuned using grid search via the link prediction experiment. The essence of a deep Autoencoder is to minimize the reconstruction error between the input and output via deep neural networks. In particular, given input x_i^b, parameters of the intermediate representation for each encoder layer are as follows:

$$w_i^1 = \sigma(W^1 x_i + b^1)$$

$$w_i^k = \sigma\left(W^k w_i^{k-1} + b^k\right), k = 2, \ldots, K$$

After we obtain the intermediate representation w_i^K, the output x_i' can be generated via a reversing operation of the encoder. That is, the network parameters of the k-th layer are shared between the encoder and decoder. The reconstruction process for the decoder layers is as follows:

$$w_i^{K\prime} = \sigma(W^K x_i + b^K)$$

$$x_i' = \sigma(W^1 w^{1\prime} + b^{1\prime})$$

We implement the above deep learning model using the Tensorflow library on Nvidia P100 GPU. Gradient descent is used in optimization and parameter estimation. We also adopt dropout training (Srivastava et al. 2014), a common practice in neural network, to avoid overfitting. In our experiments, we use sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$ as the activation function to capture the non-linearity. For the deep Autoencoder, we use three hidden layers in the encoder, i.e., K=3. Each hidden layer has 5,000, 1,000, and 300 neurons respectively. The decoder uses the exact same number of hidden layers and neurons, i.e., 300, 1,000, and 5,000. More hidden layers increase the overall learning time and performance worsens due to potential overfitting.

A2: Link prediction algorithm

Algorithm: LINK PREDICTION ALGORITHM

Input: user-brand networks in training and testing; number of randomly selected users: N
k: *precision@k*, *recall@k*
Output: *precision@k* and *recall@k*

1. Obtain node representation V_i via deep autoencoder for i in $1, \ldots, m$ (m is the total number of users in training)
2. Select N users at random $U = \{u_1, u_2, \ldots, u_N\}$
3. $S \longleftarrow \Phi$ (initialization)
4. **foreach** user $u_i \in U$ **do**
   foreach brand b_j in training **do**
   $p_{ij} \longleftarrow$ proximity score between u_i and b_j
   S += $(u_i \longleftrightarrow b_j, p_{ij})$
   end
   end
5. $L = \{l_1, \ldots, l_k \mid l_i$ is a user-brand pair$\}$ (top k predicted links based on their proximity scores)
6. $precision@k = \frac{|L \cap E_{1,2}|}{k}$, $recall@k = \frac{|L \cap E_{1,2}|}{|E^T_{1,2}|}$ (see the definition of $E_{1,2}$ and $E^T_{1,2}$ in the Evaluation and Results Section)

A3: Comment Network and Like Network

On Facebook, users engage with brands in multiple ways, such as liking or commenting. In this study, we aggregate user comments and likes as engagements and build a user-brand network. However, it is also interesting to know what would happen if we construct a user-brand network with only comments or likes. Prior research shows that Facebook likes affect offline customer behavior (Mochon et al. 2017). To have a deeper understanding of learned network representation, we conduct two complimentary link-prediction experiments based on the comment network and the like network. The comment network is constructed between a user and a brand if the user leaves comments on the brand's public page. Similarly, the like network is constructed between a user and a brand if the user likes posts on the brand's public page.

Similar to our previous experiments, we measure the link prediction performance on two metrics *precision@k* and *recall@k*.

We can observe several findings from the results shown in Table A1 and Table A2. First, the network representations learned from the like network or the comment network have less predictive power than those learned from the network constructed using both likes and comments. For example, the *precision*@1000 of the comment network, the like network and the like+comment network is 0.168, 0.314 and 0.355, respectively. Similarly, results for the *recall* metric show that the deep network learning is better at capturing the hidden relationships among brands and users with more volume and variety of data. Second, we see that deep network learning approach consistently performs better than linear models (e.g., matrix factorization) in the heterogeneous user-brand network setting, where the performance gain is limited in the homogeneous brand-brand network setting. This indicates that a common practice of reducing heterogeneous networks to homogeneous networks loses important information for learning good representation. Third, we can see that link prediction performance is better for the like network than the comment network. The reasons are two-fold: (1) the like network has more data than the comment network, which facilitates better network representation learning, and (2) the like engagement is more meaningful than the comment engagement in the market structure discovery. A user liking a brand signals a preference for the brand, while a user commenting on a brand can be a complex signal as the comment may be positive or negative.

Table A1: Performance comparison for different models on Like network. The number of randomly selected users is N=1,000.

precision@k		k=10	k=100	k=500	k=1,000	k=5,000	k=10,000	k=100,000
Homogeneous brand-brand network	Linear model	0.320	0.279	0.258	0.233	0.127	0.067	0.011
		(0.094)	(0.056)	(0.008)	(0.008)	(0.004)	(0.001)	(0.001)
	Deep model	0.323	0.284	0.258	0.235	0.135	0.069	0.011
		(0.147)	(0.082)	(0.017)	(0.009)	(0.014)	(0.034)	(0.002)
Heterogenous brand-user network	Linear model	0.424	0.365	0.312	0.287	0.152	0.087	0.011
		(0.035)	(0.042)	(0.039)	(0.008)	(0.032)	(0.003)	(0.000)
	Deep model	**0.486*****	**0.398*****	**0.354*****	**0.314*****	**0.178*****	**0.091*****	**0.011**
		(0.026)	(0.032)	(0.023)	(0.009)	(0.037)	(0.004)	(0.001)
recall@k		k=10	k=100	k=500	k=1,000	k=5,000	k=10,000	k=100,000
Homogeneous brand-brand network	Linear model	0.002	0.024	0.111	0.201	0.458	0.563	0.896
		(0.001)	(0.005)	(0.003)	(0.006)	(0.015)	(0.010)	(0.006)
	Deep model	0.002	0.025	0.124	0.204	0.476	0.560	0.882
		(0.002)	(0.002)	(0.011)	(0.018)	(0.052)	(0.023)	(0.034)
Heterogenous brand-user network	Linear model	0.041	0.056	0.332	0.350	0.521	0.635	0.911
		(0.003)	(0.004)	(0.029)	(0.029)	(0.075)	(0.079)	(0.009)
	Deep model	**0.049*****	**0.068*****	**0.350*****	**0.404*****	**0.562*****	**0.663*****	**0.929*****
		(0.005)	(0.006)	(0.021)	(0.043)	(0.037)	(0.063)	(0.028)

Table A2: Performance comparison for different models on Comment network. The number of randomly selected users is N=1,000.

precision@k		k=10	k=100	k=500	k=1,000	k=5,000	k=10,000	k=100,000
Homogeneous brand-brand network	Linear model	0.189	0.179	0.156	0.134	0.067	0.045	0.010
		(0.169)	(0.041)	(0.014)	(0.008)	(0.005)	(0.003)	(0.000)
	Deep model	0.189	0.168	0.162	0.137	0.062	0.044	0.010
		(0.097)	(0.019)	(0.052)	(0.010)	(0.032)	(0.002)	(0.001)
Heterogenous brand-user network	Linear model	0.213	0.192	0.167	0.154	0.122	0.080	0.010
		(0.025)	(0.087)	(0.029)	(0.024)	(0.052)	(0.020)	(0.001)
	Deep model	**0.234*****	**0.210*****	**0.173*****	**0.168*****	**0.126*****	**0.088*****	**0.011***
		(0.045)	(0.023)	(0.067)	(0.019)	(0.033)	(0.002)	(0.002)
recall@k		k=10	k=100	k=500	k=1,000	k=5,000	k=10,000	k=100,000
Homogeneous brand-brand network	Linear model	0.002	0.017	0.068	0.117	0.291	0.393	0.834
		(0.002)	(0.003)	(0.006)	(0.008)	(0.017)	(0.018)	(0.008)
	Deep model	0.002	0.019	0.068	0.114	0.295	0.393	0.842
		(0.001)	(0.012)	(0.022)	(0.032)	(0.042)	(0.053)	(0.012)
Heterogenous brand-user network	Linear model	0.019	0.042	0.077	0.162	0.333	0.442	0.885
		(0.003)	(0.019)	(0.045)	(0.029)	(0.029)	(0.056)	(0.034)
	Deep model	**0.018**	**0.044****	**0.082*****	**0.182*****	**0.352*****	**0.453*****	**0.894*****
		(0.004)	(0.012)	(0.051)	(0.037)	(0.026)	(0.033)	(0.046)

Figure 1: The overall framework of the proposed deep network representation learning

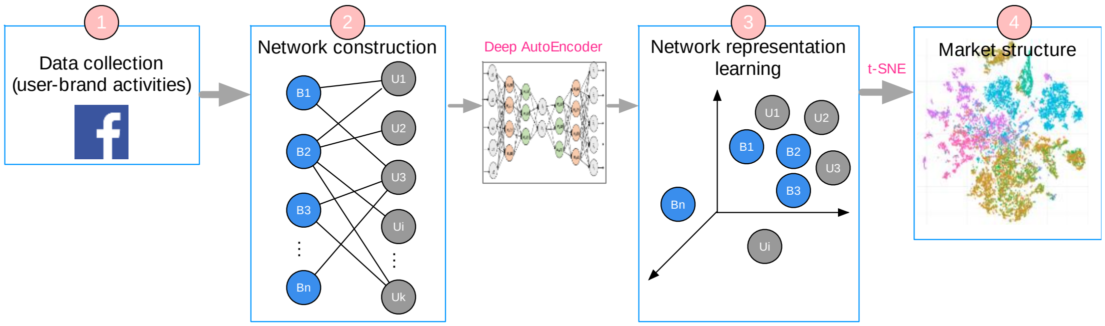

Figure 2: An illustration of deep network representation learning

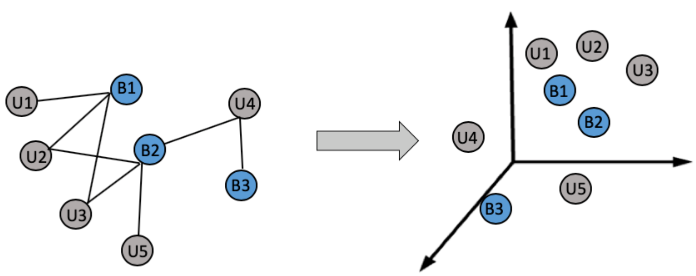

Figure 3: Degree distribution of brands in the user-brand network

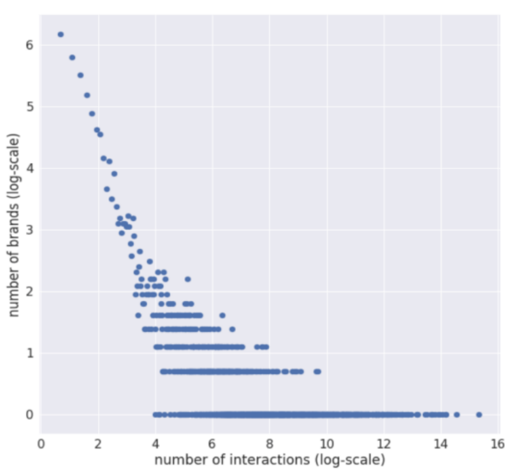

Figure 4: The global structure among brands in our Facebook data

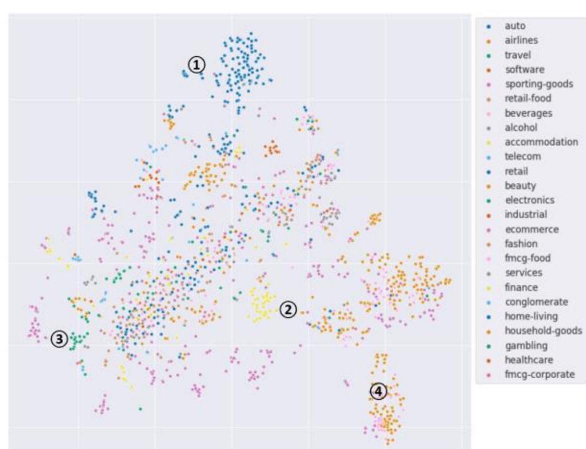

Figure 5: Zooming-in on Clusters 1, 2, 3, and 4

Figure 6: Visualization of category structure

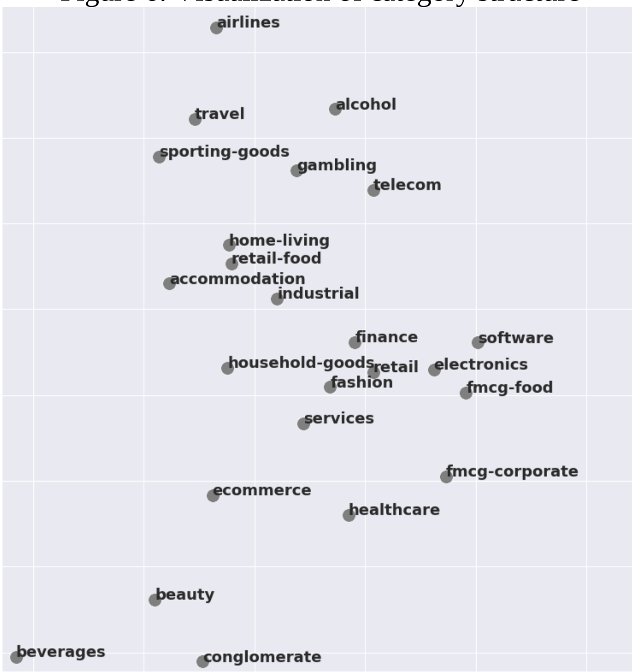

Figure 7: Proximity change of Amazon to other brands in retail and e-commerce industry

Figure 8: Proximity change of Whole Foods to other brands in retail and e-commerce industry

Figure 9: Proximity change of Tesla to other selected brands in the auto industry

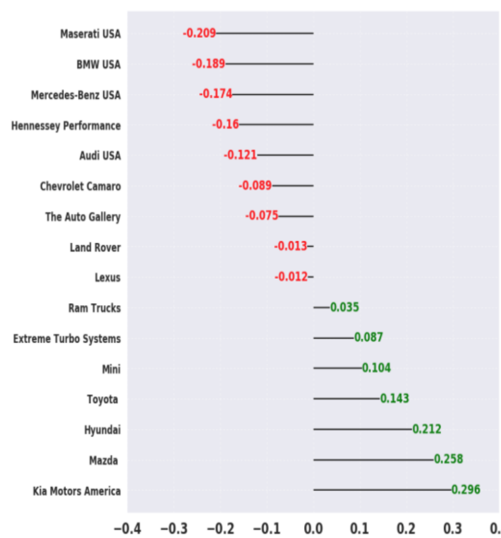

Table 1: Comparison of different types of work on market structure discovery

	Primary/Survey Data	**Text Mining**	**Social Tag-based**	**Search Data**	**Social Engagement**
Data Volume	Small	Large	Large	Large	Very large
Data Veracity	Authentic	Noisy	Moderately noisy	Moderately noisy	Moderately noisy
Privacy Preserving	Yes	Yes	Yes	No (need to insert a tracking pixel)	Yes
Data Availability	Low (need to do survey)	High (publicly available)	High (publicly available)	Low (need to insert a tracking pixel)	High (publicly available)
Data pre-processing cost	Low (use consideration set directly)	High (text mining is error-prone)	High (text mining is error-prone)	Low (use consideration set directly)	Low (use network raw data)

Table 2: Summary of difference among extant literature on market structure discovery

	Kim et al. 2011	**Lee and Bradlow 2011**	**Netzer et al. 2012**	**Ringel and Skiera 2016**	**Culotta and Cutler 2016**	**Nam, Joshi and Kannan 2017**	**Our study**
Objective	To visualize user search behavior and understand market structure	To visualize competitive market structure using text mining on customer review	To visualize competitive market structure using text mining on forum discussion	To understand asymmetric competition in the product categories	To infer attribute-specific brand ratings	To analyze user generated tags for marketing research	To propose a novel deep network representation learning framework for marketing research
Brands/Products	62 products, 4 brands	9 brands	169 products, 30 brands	1,124 products	200 brands	7 brands	5,478 brands

Consumers/Users	N.A.	N.A.	76,587	100,000+	14.6 million	N.A.	25,992,832
Data sources	Amazon	Customer review at Epinions	Online discussion forum	Product comparison website	Twitter	Social tagging platform Delicious	Facebook public fan page
Data type	Consumer search	Text	Text	Consumer search	Network	Social tags	Network
Brand association methodology	Consideration set	Text-mining	Text-mining	Consideration set	Network learning	Network learning	Network learning
Asymmetry	Yes	No	No	Yes	No	No	Yes
Dynamic	No	No	No	No	No	Yes	Yes
Dimension reduction	Yes	Yes	No	No	No	Yes	Yes
External validation	N.A.	N.A.	Purchase data, survey	Survey	Survey	Brand concept map (survey)	Event study, link prediction
Privacy preserve	Yes	Yes	Yes	No (need to insert a tracking pixel)	Yes	Yes	Yes
Data availability	Low (need to collect data daily)	High (publicly available)	High (publicly available)	Low (need to insert a tracking pixel)	High (publicly available)	High (publicly available)	High (publicly available)
Data preprocessing cost	Low (use consideration set directly)	High (text mining is error-prone)	High (text mining is error-prone)	Low (use consideration set directly)	Low (use network raw data)	Low (tags are well defined)	Low (use network raw data)

Table 3: Data description and statistics

Number of brands	5,478
Number of users	25,992,832
Number of unique user-brand interactions	36,927,613
Number of like interactions	87,876,623
Number of unique user-brand like interactions	29,611,805
Number of comment interactions	18,703,549
Number of unique user-brand comment interactions	7,612,358
Total number of user-brand interactions	106,580,172

Table 4: Performance comparison for different models. The number of randomly selected users is N=100

precision@k		k=10	k=100	k=500	k=1,000	k=5,000	k=10,000	k=100,000
Homogeneous brand-brand network	Shallow model	0.400	0.262	0.132	0.078	0.022	0.012	0.001
		(0.109)	(0.023)	(0.018)	(0.008)	(0.002)	(0.000)	(0.000)
	Deep model	0.410	0.271	0.139	0.082	0.023	0.014	0.001
		(0.092)	(0.027)	(0.020)	(0.009)	(0.003)	(0.001)	(0.000)
Heterogenous brand-user network	Shallow model	0.430	0.291	0.157	0.095	0.028	0.018	0.001
		(0.102)	(0.030)	(0.024)	(0.008)	(0.005)	(0.002)	(0.000)
	Deep model	**0.52*** **	**0.322*** **	**0.173*** **	**0.124*** **	**0.034*** **	**0.028*** **	**0.001*** **
		(0.092)	(0.022)	(0.051)	(0.011)	(0.008)	(0.001)	(0.000)
recall@k		k=10	k=100	k=500	k=1,000	k=5,000	k=10,000	k=100,000
Homogeneous brand-brand	Shallow model	0.031	0.260	0.488	0.602	0.828	0.918	0.996
		(0.008)	(0.002)	(0.060)	(0.050)	(0.036)	(0.016)	(0.005)

network		0.032	0.275	0.505	0.621	0.832	0.912	0.997
	Deep model	(0.013)	(0.032)	(0.054)	(0.047)	(0.049)	(0.032)	(0.003)
Heterogenous brand-user network	Shallow model	0.037 (0.015)	0.287 (0.065)	0.521 (0.074)	0.637 (0.045)	0.870 (0.023)	0.935 (0.047)	0.998 (0.000)
	Deep model	**0.056****** (0.013)	**0.311****** (0.035)	**0.582****** (0.077)	**0.686****** (0.054)	**0.897****** (0.078)	**0.967****** (0.024)	**0.999**** (0.002)

Table 5: Performance comparison for different models. The number of randomly selected users is N=1,000

precision@k		k=10	k=100	k=500	k=1,000	k=5,000	k=10,000	k=100,000
Homogeneous brand-brand network	Shallow model	0.460 (0.132)	0.387 (0.112)	0.331 (0.021)	0.291 (0.012)	0.130 (0.004)	0.078 (0.003)	0.012 (0.000)
	Deep model	0.490 (0.020)	0.393 (0.003)	0.332 (0.018)	0.295 (0.017)	0.131 (0.003)	0.078 (0.003)	0.012 (0.000)
Heterogenous brand-user network	Shallow model	0.500 (0.102)	0.422 (0.060)	0.344 (0.022)	0.320 (0.072)	0.162 (0.010)	0.087 (0.017)	0.012 (0.000)
	Deep model	**0.522***** (0.092)	**0.436***** (0.040)	**0.365***** (0.012)	**0.355***** (0.035)	**0.187***** (0.014)	**0.091***** (0.047)	**0.013***** (0.000)
recall@k		k=10	k=100	k=500	k=1,000	k=5,000	k=10,000	k=100,000
Homogeneous brand-brand network	Shallow model	0.031 (0.008)	0.033 (0.021)	0.128 (0.008)	0.223 (0.008)	0.509 (0.013)	0.607 (0.013)	0.915 (0.008)
	Deep model	0.032 (0.005)	0.035 (0.047)	0.131 (0.018)	0.226 (0.011)	0.510 (0.010)	0.605 (0.015)	0.921 (0.007)
Heterogenous brand-user network	Shallow model	0.049 (0.022)	0.056 (0.009)	0.241 (0.012)	0.365 (0.010)	0.549 (0.012)	0.658 (0.024)	0.981 (0.015)
	Deep model	**0.049***** (0.009)	**0.076***** (0.003)	**0.352***** (0.010)	**0.412***** (0.007)	**0.584***** (0.009)	**0.743***** (0.008)	**0.990***** (0.002)

Table 6: Performance comparison for different sizes of training set. The number of randomly selected users is N=1,000

precision@1000		10%	30%	50%	70%	90%	100%
Homogeneous brand-brand network	Shallow model	0.103 (0.012)	0.195 (0.008)	0.248 (0.008)	0.263 (0.012)	0.282 (0.015)	0.291 (0.012)
	Deep model	0.097 (0.042)	0.190 (0.010)	0.248 (0.021)	0.267 (0.031)	0.284 (0.023)	0.295 (0.017)
Heterogenous brand-user network	Shallow model	0.143 (0.015)	0.225 (0.031)	0.256 (0.042)	0.283 (0.008)	0.312 (0.052)	0.320 (0.072)
	Deep model	**0.183***** (0.024)	**0.242***** (0.032)	**0.273***** (0.037)	**0.301***** (0.012)	**0.337***** (0.032)	**0.355***** (0.035)
recall@1000		10%	30%	50%	70%	90%	100%
Homogeneous brand-brand network	Shallow model	0.080 (0.009)	0.153 (0.006)	0.193 (0.006)	0.203 (0.007)	0.219 (0.011)	0.223 (0.008)
	Deep model	0.075 (0.005)	0.150 (0.010)	0.194 (0.007)	0.204 (0.003)	0.220 (0.005)	0.226 (0.011)
Heterogenous brand-user network	Shallow model	0.108 (0.031)	0.179 (0.018)	0.223 (0.013)	0.257 (0.026)	0.271 (0.017)	0.241 (0.010)
	Deep model	**0.124***** (0.009)	**0.198***** (0.008)	**0.24***** (0.019)	**0.289***** (0.029)	**0.314***** (0.008)	**0.352***** (0.007)

Table 7: Top 10 proximal brands to each focal brand

Focal brand		**United**	**Southwest Airlines**	**Audi USA**	**Nissan**
Rank	1	American	JetBlue	Mercedes-Benz USA	Mazda
	2	Delta	Frontier	BMW USA	Toyota
	3	Lufthansa	Allegiant	Land Rover	Volkswagen
	4	Southwest	Delta	Lexus	Kia Motors America
	5	Alaska	Alaska	Chevrolet Camaro	Subaru of America
	6	All Nippon	United	Maserati USA	Chrysler
	7	Air China	Airfarewatchdog	Kawasaki USA	FIAT
	8	LATAM	American	Firestone Tires	Jaguar
	9	Air New Zealand	Virgin America	Tesla	Alfa Romeo
	10	Airfarewatchdog	Hyatt	Ram Trucks	KLIM