Empirical Analysis of Implicit Brand Networks on Social Media

Kunpeng Zhang
University of Illinois at Chicago
Department of IDS, College of Business Administration
Chicago, USA
kzhang6@uic.edu

Siddhartha Bhattacharyya
University of Illinois at Chicago
Department of IDS, College of Business Administration
Chicago, USA
sidb@uic.edu

Sudha Ram
University of Arizona
Department of MIS, Eller College of Management
Tucson, USA
ram@eller.arizona.edu

ABSTRACT

This paper investigates characteristics of implicit brand networks extracted from a large dataset of user historical activities on a social media platform. To our knowledge, this is one of the first studies to comprehensively examine brands by incorporating user-generated social content and information about user interactions. This paper makes several important contributions. We build and normalize a weighted, undirected network representing interactions among users and brands. We then explore the structure of this network using modified network measures to understand its characteristics and implications. As a part of this exploration, we address three important research questions: (1) What is the structure of a brand-brand network? (2) Does an influential brand have a large number of fans? (3) Does an influential brand receive more positive or more negative comments from social users? Experiments conducted with Facebook data show that the influence of a brand has (a) high positive correlation with the size of a brand, meaning that an influential brand can attract more fans, and, (b) low negative correlation with the sentiment of comments made by users on that brand, which means that negative comments have a more powerful ability to generate awareness of a brand than positive comments. To process the large-scale datasets and networks, we implement MapReduce-based algorithms.

Categories and Subject Descriptors

H.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information networks; E.1 [**DATA STRUCTURES**]: Graphs and networks; H.5.4 [**INFORMATION INTERFACES AND PRESENTATION**]: Hypertext/Hypermedia

General Terms

Algorithms, Experimentation

Keywords

Network analysis; sentiment identification; social media; marketing intelligence; MapReduce

1. INTRODUCTION.

Social media has become one of the most popular communication platforms allowing users to discuss and share topics of interest without necessarily having the same geo-location and time. Information is be generated and managed through computers or mobile devices by one person and consumed by many others. Different people express different opinions on the same topic, and some also express their opinions on multiple topics of interest. A wide variety of topics, ranging from current events and political debate, to sports and entertainment, are being actively discussed on these social forums. For example, Facebook users comment on or 'like' campaigns posted by a company; Twitter users send tweets with a maximum length of 140 characters to instantly share and deliver their opinions on politics, movies, sports, etc. Some e-commerce platforms, such as Amazon.com, allow users to leave their reviews on products. These actions generate a rich data source which can be analyzed to understand the interactions among entities on social media. These networks of interactions may be of two types: explicit or implicit. Explicit networks are formed via "friendship" relationships on Facebook or "following" relationships on Twitter; implicit networks are formed, for example, via reviewing actions of consumers on products from Amazon.com. The networks may also be established when people share common interests, or when brands have overlapping customers. Such networks can be constructed from large datasets of user-generated content, and analyzed to obtain actionable insights to help users or brands make informed decisions.

HT'14, September 1–4, 2014, Santiago, Chile.
Copyright 2014 ACM 978-1-4503-2954-5/14/09 ...$15.00.
http://dx.doi.org/10.1145/2631775.2631806.

For example, analysis of large brand-brand networks enables the identification of influential brands, facilitating targeted on-line advertising and eventually leading to product or service purchases.

Analysis of these networks helps in getting a better understanding of brand characteristics and is useful for making intelligent marketing decisions. For example, studying such kinds of networks by incorporating user-generated textual content can help identify influential brands and interactions among brands, which can lead to a better online brand advertising strategy. Most explicit networks are fairly easy to construct; However, they have some shortcomings. They are built based on explicit relationships among brands, which ignores activities between users and brands. In addition, current networking approaches do not consider textual sentiment of social content. In this paper, we attempt to overcome these shortcomings by leveraging user generated social media content including comments, "likes", and posts to build and analyze a new kind of implicit brand-brand network. Unlike regular network analysis ([28]) through users' friend networks, we leverage data on user interactions with brands' "fan" pages to extract networks that capture relationships between different brands. We then use a network analysis approach, together with sentiment analysis, to explore characteristics of the network. In addition, this brand-brand network obtained from a very large dataset of users and their interactions on a social platform requires efficient techniques for construction and analyses via distributed techniques based on Hadoop and MapReduce.

Our empirical study for brand-based networks built from large scale social data makes several contributions in the area of 'big' data on social interactions. The first major contribution is our new approach to build a weighted and undirected brand-brand network representing interactions among users and brands, based on a large amount of social content generated by users on a social media platform. To investigate network properties, we propose a technique for normalizing relationship weights in the network from a global perspective. In addition, we define new structural measures for analyzing the network by modifying traditional measures such as degree, diameter, clustering coefficient, and centrality to incorporate these weights. We then explore the structure of this network to understand its implications. As a part of this exploration, we address three research questions to develop a deeper understanding of brand characteristics from the network perspective coupled with sentiment analysis: (1) What is the structure of a brand network? (2) Does an influential brand have a large number of followers/fans? (3) Does an influential brand receive more positive or more negative comments from social users? Since our datasets and networks are very large, we implement MapReduce-based techniques in a distributed Hadoop[1] environment, for network generation. We collected data from the popular social platform Facebook through their Graph API[2] for an empirical analysis. In addition, we designed some simple but effective rules to filter out spam activities and spam users to improve the data quality.

The rest of this paper is organized as follows. Section 2 reviews all related work and Section 3 describes the overall framework. Section 4 describes the dataset and data cleansing process. Section 5 introduces the brand networks, describes how these are generated and normalized, and lays out important network measures used for analyzing the networks. Section 6 describes our empirical results from analyses of a large Facebook dataset to answer our research questions. We also describe a sentiment identification algorithm in this section. This is followed by conclusions and directions for future work in Section 7.

2. RELATED WORK.

In this section, we describe relevant work in three related areas: brand communities in social networks, network measures, and sentiment analysis on social media content.

Several studies in the marketing literature have examined the spread of influence and behavior across connected consumers; these arise, for example, from product reviews and recommendations, user interactions through comments and likes in social sites, or through participation in brand communities. The potential for word-of-mouth effects in promoting product adoptions, role and influence of early adopters, and broader issues of social contagion in consumer networks have seen much interest ([11, 12, 18]). Diffusion of information over consumer networks has been deemed effective for rapid reach over large online audiences through viral marketing approaches ([3, 1]). Various brands from across industries as well as non-profits are active on social networks for promoting brand image and building brand communities ([30]). 'Fan' pages on Facebook are an example of such communities, and form the basis for our study presented in this paper. Existing research has examined consumer interactions in online communities ([7]), social network based communities for promoting engagement with a brand ([22, 8]), consumer motivations for participation, and effective means for developing consumer-brand relationships ([30]). The topic of brand-brand relationships, however, has not received much attention in the literature - this is the focus of our study, using large-scale data, sentiment analysis and network analysis.

Many different measures exist in literature to quantify and understand network structures. For instance, degree is an important characteristic of a vertex in a network. Based on the degree of the vertices, it is possible to derive many other structural measures for the network. [2] found that a common property of many large networks is that the vertex connectivities follow a scale-free power-law distribution. Such scale free structures occur when networks expand by adding new vertices which attach preferentially to nodes that are already well connected. In addition, there are many studies related to centrality measures. For example, [10] assumed that the interactions in a network follow the shortest paths between two vertices; it is then possible to quantify the importance of a vertex or an edge in terms of its betweenness centrality. [20] proposed a betweenness measure that relaxes this assumption to include contributions from essentially all paths between nodes, although it gives more weight to short paths. It is based on random walks, counting how often a node is traversed by a random walk between two other nodes. [14] presented a novel formulation of centrality for dynamic networks that measures the number of paths in a network.

Sentiment identification has been widely studied in the past. These efforts mainly fall into three major categories. 1) Bag-of-Words approaches produce domain-specific lexi-

[1] Apache Hadoop: http://hadoop.apache.org/

[2] The Graph API: https://developers.facebook.com/docs/graph-api/

cons, and there is a vast body of research which attempts to incorporate them as features in machine learning models [32, 21, 9]. 2) Rule-based approaches have also been studied by many researchers. The authors in [4] proposed compositional semantics, based on the assumption that the meaning of a compound expression is a function of the meaning of its parts and of the syntactic rules by which they are combined. They have developed a set of compositional rules to assign sentiments to individual clauses, expressions and sentences. 3) Recently, there has been a wide range of machine learning techniques, which classify the whole opinion document (e.g., a product review) as positive or negative [21, 29, 6, 16]. In [4], the authors viewed such sub-sentential interactions in light of compositional semantics, and presented a novel learning-based approach that incorporates structural inference motivated by compositional semantics into the learning procedure. In [21], authors employed machine learning techniques to classify documents by overall sentiments and results on movie review data show that three machine learning methods they employed (Naïve Bayes, maximum entropy classification, and support vector machines) do not perform as well on sentiment classification as on traditional topic-based categorization. In [19], authors presented a linguistic analysis of conditional sentences, and built some supervised learning models to determine if sentiments expressed on different topics in a conditional sentence are positive, negative or neutral. Several researchers have also studied feature/topic-based sentiment analysis [25, 24, 17, 13, 9]. Their objective is to extract topics or product features in sentences and determine the associated sentiments. In [32], authors used feature-based opinion mining model to identify noun product features that imply opinions. In [15], authors proposed an approach to extract adverb-adjective-noun phrases based on clause structure obtained by parsing sentences into a hierarchical representation. They also proposed a robust general solution for modeling the contribution of adverbials and negation to the score for degree of sentiment. In [26], authors showed that information about social relationships can be used to improve user-level sentiment analysis. In [33, 27], authors considered network context to model the effect of emotions in sentiment. In this paper, we apply state-of-the-art sentiment engine to identify brand sentiment based on user's historical comments on that brand.

3. OVERALL FRAMEWORK.

To understand characteristics of brands on a social media platform, we examine three research questions. **RQ1:** What is the structure of a brand network in terms of various network measures? To answer this question, we build and normalize an implicit brand-brand network based on user historical activities, and then analyze the network using modified network measures incorporating weights. **RQ2:** We also examine if the influence of a brand correlates with the size of a brand. The influence of a brand can be identified by the eigenvector centrality calculated from the brand network. **RQ3:** Since there is a lot of user-generated text (e.g. comments made by users) on social media platforms, the third question we address incorporates such textual information to investigate whether an influential brand attracts more positive or negative comments. The sentiment of comments can be identified by state-of-the-art algorithms described in Section 6.1. Results shown in this paper are obtained from experiments based on a large Facebook dataset.

Table 1: Description and statistics of raw dataset.

Number of downloaded brands	$13,806$
Number of unique users	$286,862,823$
Number of unique countries	122
Number of categories defined by Facebook	172

4. DATA.

We collected a large (approximately 2 TB) dataset from Facebook using their Graph API. All analyses methods proposed and described in the following sections can also be applied to data from other social media platforms, such as Twitter. In this section, we describe the details of the data collection, pre-processing, and cleaning performed to generate a high quality dataset for network analysis.

4.1 Data Collection.

Facebook, the largest and most popular social network platform, has more than 1 billion accounts. Many organizations, and individuals build their own pages on Facebook to share and communicate with their fans. The extensive amount of textual and interaction information generated by users has made it a promising platform for brand analysis. In this work, our focus is on the top social brands as the object of analysis, i.e. the brands with a large number of fans. We used the Facebook Graph API to download all activities on a brand page such as posts initialized by the brand page administrator, as well as posts by users, such as comments, "likes" on posts, and public user profiles (e.g. gender and locale). Each brand may have a number of posts depending on the posting frequency. A post is any information that the brand wants to share and interact with users and may include text, photos, videos, links or a combination of these. For instance, posts may be about a new product release, company annual report filing announcement, special day greetings, surveys, or other important events and activities. Any Facebook user can respond to these posts by liking or making comments on them. While there is no 'dislike' action on Facebook, textual responses to posts can be used to indicate positive, neutral, as well as negative opinions. The dataset (shown in Table 1) used for this work was collected from January 1, 2009 thru January 1, 2013. It contains data from $13,806$ brand pages and approximately 280 million users. It covers data from brands in 122 countries in 172 categories as defined by Facebook's classification system.

4.2 Data Cleansing.

Data quality is of paramount importance in any analytics study as it can affect model performance and results. To ensure quality of the dataset we performed a number of cleansing operations. First we removed brands for which most of the posts and comments were not in English, because sentiment identification for non-English text is not well understood and accuracy is not high. To produce robust results we applied a spam filter to remove fake users and their corresponding activities. Our data shows that on average, a user comments on 4 to 5 pages and likes posts on 7 to 8 pages as shown in Figures 1 and 2, respectively.

Users connecting to an extremely large number of brands / pages are likely to be spam users or bots. For example, we found one spam user who appeared on 600+ different brand pages. We also detected one user who "liked" posts across 520 different brand pages. As most users are likely to be interested in a small number of brands, we discarded users making comments on more than 100 brands and those liking posts on more than 150 brands. In addition, we detected other kinds of spam users. For example, there was one user who liked $7,963$ posts out of $8,549$ posts for a brand. We assume that it is likely to be a spam user if this ratio is very high. We set this threshold to be 90% for every user except the page owner. Lastly, we also removed users who posted many duplicate comments containing URL links. A test on Barack Obama's page, found $209,864$ duplicate comments out of $2,987,505$ in total. The dataset for our analyses is from the top $2,000$ brands, selected as those having the largest numbers of fans on their Facebook page.

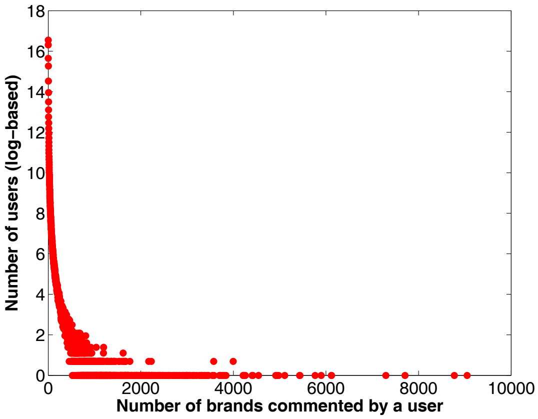


Figure 1: Distribution of individual brands / pages on which users comment. Y-axis is the *log* of number of users. $X-$axis is the number of brands(pages) on which a user makes comments.

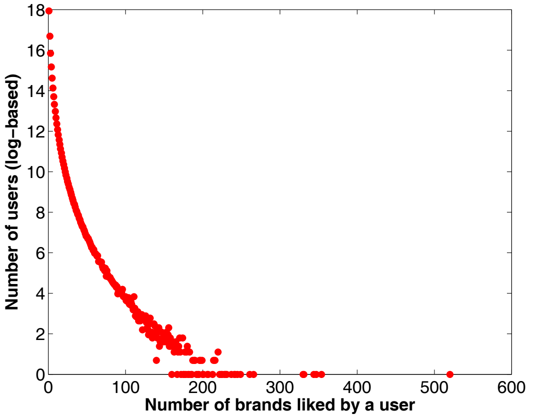


Figure 2: Distribution of individual brands / pages liked by a user. Y-axis is the *log* of number of users. $X-$axis is the number of brands(pages) liked by a Facebook user.

5. NETWORK ANALYSIS.

From the cleaned dataset, we constructed an implicit brand-brand network based on user historical activities. This is a weighted and undirected network. Link weights in the network were globally normalized. The data is very large, and hence we implemented a MapReduce algorithm (described in algorithm 1) to construct the network using Hadoop. We then modified the standard structural property measures in the network to incorporate weights, and used them to analyze the network. These structural properties include (weighted) degree, density, network diameter, clustering coefficient, and various centrality measures.

5.1 Weighted and Undirected Brand-brand Network.

Each brand has various properties such as a category as defined by the Facebook classification system, number of fans, number of people "talking about it", and a record of users' activities. This information can be used to capture the implicit relationships among brands and extract the brand-brand network. In this network, brands are designated as nodes, and a link between two brands is created if the same user commented on or liked posts made by both brands. Thus, two brands are bridged by common users. The larger the number of common users having activities on two brands, the higher the weight of their interconnecting link. This network represents brand-brand affinity. Formally, we define a weighted and undirected brand network ($\mathbb{B}$) as shown below.
$\mathbb{B} =< \mathbb{V}, \mathbb{E} >$, where $\mathbb{V} = \{b_i \mid b_i$ is a brand. Each b_i has f_i as the number of fans$\}$,

$\mathbb{E} = \{(b_i, b_j) \mid b_i$ has some common users with b_j, the corresponding weight is: $w_{ij} =$ the number of common users$\}$, where $1 \leq i, j \leq N$, N is the total number of brands, $N = 2,000$ in this study.

Alternatively, for the convenience of explaining network measures in the following section, we use the adjacency matrix A defined below to represent the network $\mathbb{B}$ as

$$A_{ij} = \begin{cases} w_{ij} & \text{if node } j \text{ connects to node } i \\ 0 & \text{otherwise} \end{cases}$$

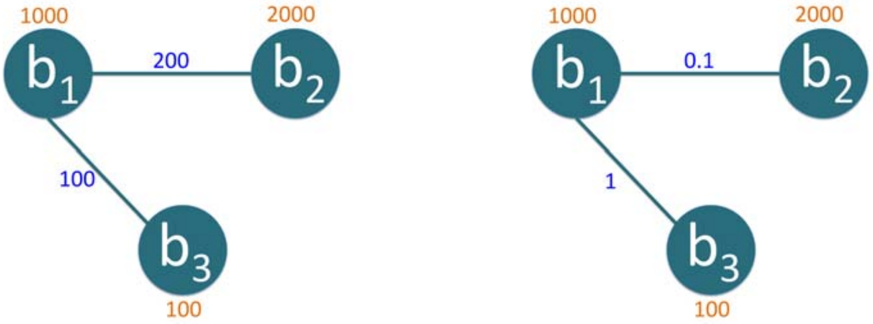
where w_{ij} is the weight between brand i and brand j, which is the number of common users between brand b_i and brand b_j.

Normalization of brand-brand network: ($\mathbb{B} \rightarrow \mathbb{B}_n$)
Well-known brands typically attract more fans and have more common users with other big brands. A comparison across brands in the network requires normalization of the link weights. However, if we normalize the network by using the global maximum weight in the network, we lose global network semantics such as the distribution of connection strength among links of a brand relative to the size of a brand. Consider the case shown in Figure 3a. The connection (b_1, b_3) can be considered relatively stronger than the connection (b_1, b_2), because all (100%) of b_3 users are connected to b_1, while only 10% of b_2 users are interested in b_1. We propose a two step normalization process to characterize the strength of a link in $\mathbb{B}_n$. See example shown in Figure 3b.
The normalization for network $\mathbb{B} \rightarrow \mathbb{B}_n$ is as follows.

- We first normalize each individual link between two brands b_i, b_j by setting $w'_{ij} = \frac{w_{ij}}{f_i * f_j}$.

Table 2: Description and statistics before and after data cleansing. Cleaned dataset containing top $2,000$ brands.

After cleaning		After selecting top brands
Number of brands	$7,580$	$2,000$
Number of unique users	$97,699,832$	$16,306,977$
Number of comments	$2,327,635,302$	$470,742,158$
Number of positive comments	$651,231,870$	$179,009,470$
Number of negative comments	$234,571,177$	$60,613,968$
Number of brand categories	150	118
Number of posts	$13,206,402$	$3,793,941$



(a) Before normalization

(b) After normalization

Figure 3: An example to show relationship between brands based on common users. Number in red indicates number of fans for that brand. Before the normalization, brand b_1 has 1000 fans. Brand b_2 has 2000 fans, and brand b_3 has 100 fans. The number of common users between b_1 and b_2 is 200. The number of common users between b_1 and b_3 is 100. (b) shows the relative weights after normalization from a global perspective.

- We then normalize all w'_{ij} by setting $w''_{ij} = \frac{w'_{ij}}{max_{\forall(i,j)}\{w'_{ij}\}}$.

where f_i here is the number of fans of brand i.

5.2 Network Generation Using Hadoop.

Having defined all the networks, i.e., the network without normalization $\mathbb{B}$ and the network with normalization $\mathbb{B}_n$, we now focus on the process used to generate a network containing common users between brands. The raw data downloaded from Facebook is in the following format for each brand: $< user_{id}$, comment> or $< user_{id}$, post like>. They are aggregated to generate a large text file consisting of triplets: $< brand_{id}$, $user_{id}$, # of activities[3]>. The size of the file is too large to be processed by a single machine. For example, to get common users between two brands b_i and b_j, we need to consider intersections between two sets S_i: $\{$all users having activities in brand $b_i\}$ and S_j: $\{$all users having activities in brand $b_j\}$. This consumes enormous processing time because each brand typically has millions of unique users who have activities on its page. We used Hadoop to efficiently generate our network file in the following format of $< b_i$, b_j, # of common users>. The basic map and reduce functions are shown in the algorithm 1. Without using Hadoop and other distributed computing techniques, it would have been impossible to even load such a large dataset (approximately 2 TB) into one single machine.

5.3 Network Measures.

Various structural properties have been defined in literature for networks as a whole and for individual nodes, including node degree, network diameter, network density, clustering coefficient, and centrality. Most of these have been defined in the context of unweighted graphs. In this work, we extend these metrics for weighted graphs ($\mathbb{B}_n$). We first provide formal definitions of each structural property and then report on an analysis of these measures for our extracted networks.

Weighted Node Degree. The simplest yet most frequently used property of a node is its degree, i.e. the number of connections it has to other nodes. The degree of node i (brand b_i) can be easily computed from the adjacency matrix A:

$$k_i = \sum_j A_{ji}$$

In our case, A is a weighted network. Figure 4 shows the degree distribution for our weighted network. The average degree for the weighted network is 0.662 and gives the average connection strength of node neighbors.

Network Density. The density of a network is the ratio of the number of links L that exist in the network to $N(N-1)/2$, to the maximum number of links possible (in an undirected network). Network density is thus determined as:

[3] Activity implies either making comments or liking posts. # of activities = # of comments + # of post likes.

Algorithm 1 Two MapReduce jobs are chained to generate the brand-brand network.

Input: A text file contains lines of $\langle brand_{id},\ user_{id},\ \#$ of activities$\rangle$

Output: A text file contains lines of $\langle b_i,\ b_j,\ \#$ of common users$\rangle$

```
/* The first job */
input: $\langle brand_{id}, user_{id}, \#$ of activities$\rangle$   ▷ //Each line in the text file
function MAPPER
    output $\langle user_{id}, brand_{id}\rangle$
end function
function REDUCER
    for all $v \in values$ do
        add v→list
    end for
    for all $\langle b_i, b_j\rangle$, $b_i, b_j \in list$ do   ▷ //$(b_i, b_j) = (b_j, b_i)$. Either one is used
        $\langle k2, v2\rangle \leftarrow \langle (b_i, b_j), 1\rangle$
    end for
    output $\langle k2, v2\rangle$
end function

/* The second job */
function IDENTITY MAPPER   ▷ //Output is the same as input
end function
function REDUCER
    for all $v \in values$ do
        sum += v   ▷ For the same key, sum over all values
    end for
    output $\langle key, sum\rangle$
end function
```

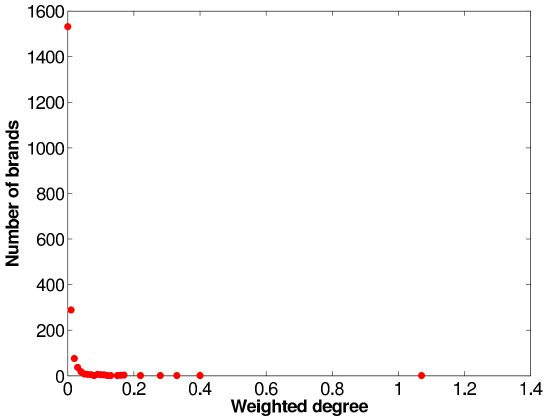


Figure 4: Degree distribution for the weighted brand-brand network. X−axis is the degree value. Y−axis is the number of brands. The weighted network shows the connection strength of each node to its neighbors. We eliminated brands with weighted degree less than 0.01. This is similar to common scale-free networks.

$$\rho = \frac{2L}{N(N-1)} = \frac{\sum\limits_i k_i}{N(N-1)} = \frac{\langle k\rangle}{N-1} \approx \frac{\langle k\rangle}{N},$$

where $\langle k\rangle = \frac{1}{N}\sum\limits_i k_i$ is the average node degree of the entire network.

Network density can also be interpreted as the fraction of links a node has on average normalized by the potential number of neighbors. It shows how densely nodes are connected to others. We consider all weights $w_{ij} > 0$ to be 1 when we compute the density of a network.

Network Diameter. Obviously, there are many paths between any two nodes i and j. The set of all such paths is ϕ_{ij}. We define a subset of these as shortest paths, i.e., those paths that have the minimal number of steps or geodesic distance. Geodesic distance is used to define the diameter of a network. The network diameter D_0 is defined as the longest of all the calculated shortest paths in the entire network. It is representative of the linear size of a network. It also reflects how fast information can be transmitted from one node to another in the network, and is expressed as $D_0 = max(\phi)$, where $\phi = \cup_{i,j}\phi_{ij}$ is the collection of all paths between all pairs of nodes.

Clustering Coefficient. The clustering coefficient can be interpreted as a measure of an “all-my-friends-know-each-other” property. It provides a mechanism for measuring transitivity of an undirected network by the fraction of triangles that exist in the network as compared to all combinations of triples. It is a measure of the extent to which nodes in a graph tend to cluster together. The clustering coefficient of a node is the ratio of existing links from a node's neighbors to each other to the maximum possible number of such links. The clustering coefficient for the entire network is the average of clustering coefficient of all the nodes. A high clustering coefficient for a network is another indication of a small world. Mathematically, it can be defined as below.

The clustering coefficient of the i^{th} node in a network N is:

$$CC_i = \frac{2e_i}{k_i(k_i-1)},$$

where k_i is the number of neighbors of the i^{th} node, and e_i is the number of connections between these neighbors. The maximum possible number of connections between neighbors is,

$$\begin{pmatrix} k_i \\ 2 \end{pmatrix} = \frac{k_i(k_i-1)}{2}$$

Thus, $CC_N = \frac{1}{n}\sum_{i=1}^{n} CC_i$, where n is the number of nodes in the network N. More detailed description of clustering coefficient can be found in Appendix A.

Centrality. In a network, there are four main measures of centrality: degree, betweenness, closeness, and eigenvector.

(1). *Degree centrality* is the same as the degree of a node. It measures the connectivity of a node.

(2). *Closeness centrality* is defined as the reverse of the length of the average of all shortest path from node i to the rest of the network. $x_i = \frac{1}{\langle l(i)\rangle}$, where $\langle l(i)\rangle = \frac{1}{N-1}\sum\limits_j l_{ij}$ and l_{ij} is a shortest path from i to j. The distance in a weighted network is defined as $d(i,j) = \frac{1}{w_{ij}}$. A small value of x_i indicates that the node is far away from the rest of the network; if it is large, then the node is close to the center.

(3). *Betweenness centrality* is the fraction of shortest paths that pass through a node. It quantifies the number

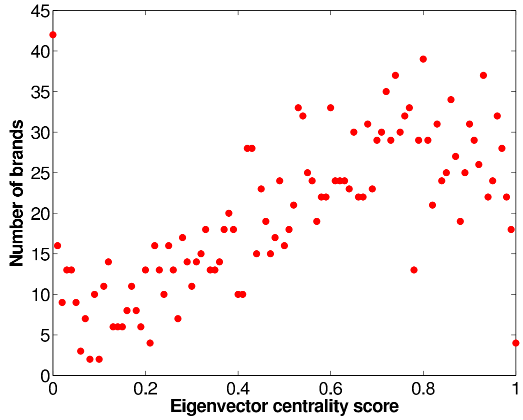


Figure 5: The eigenvector centrality distribution for the brand-brand network $\mathbb{B}_n$. $X-$axis is the eigenvector centrality score for each brand; $Y-$axis is the number of nodes in the network. Centrality score of 0 indicates isolated nodes i.e., without any connections to others. All eigenvector centrality scores are rounded to two decimal places.

of times a node acts as a conduit along the shortest path between two other nodes.

(4). *Eigenvector centrality* is widely used to measure the influence of a node in a network. It is based on topological features alone and takes into account only information in the neighborhood of a node. It assigns relative scores to all nodes in the network based on the idea that connections to more important nodes contribute more to the importance of the node in question, than connections to less important nodes. Since our brand-brand network $\mathbb{B}_n$ is weighted, we modify the original eigenvector centrality measure. For the given network $\mathbb{B}_n = (V, E)$ and adjacency matrix $A = (w_{ij})$, the eigenvector centrality score c_i of each brand i can be defined as:

$$c_i = \frac{1}{\lambda} \sum_{j \in N(i)} c_j = \frac{1}{\lambda} \sum_{j \in \mathbb{B}_n} w_{ij} c_j$$

where $N(i)$ is a set of the neighbors for brand i and λ is a constant. This calculation can be rewritten in vector notation with a small mathematical rearrangement as the eigenvector equation: $Ac = \lambda c$.

In general, there will be many different eigenvalues λ for which an eigenvector solution exists. Based on the **Perror-Frobenius theorem** ([23]), the requirement that all entries in the eigenvector be positive implies that only the greatest eigenvalue results in the desired centrality measure. Power iteration is one of the most commonly used eigenvalue algorithms to find the dominant eigenvector.

The eigenvector centrality distribution for our brand-brand network $\mathbb{B}_n$ as shown in Figure 5, reveals that brands in the network $\mathbb{B}_n$ have different influence scores and they are distributed widely in $(0, 1)$. There are around 30 isolated brands in the network, as indicated by the point at the upper left corner in the graph. The rest of the brands have eigenvector centrality scores between 0 and 1, meaning that they have either multiple strong connections or few weak connections to other brands. This centrality measure is useful for ranking brands in terms of influence.

Each brand has a category defined by Facebook, these include sports, politician, food/ beverages, clothing, and TV show. Table 3 shows top 10 influential brands and their associated categories. Among top 100 influential brands, **31 brands are categorized as TV shows, 12 as food beverages, and 9 as musician bands.** A detailed distribution is shown in Figure 6.

Table 3: Top 10 influential brands and their categories.

Rank	Top 10 influential brands	Category
1	Barack Obama	Politician
2	CNN	Media news publishing
3	Starbucks	Food beverages
4	Coca Cola	Food beverages
5	Victoria's Secret	Clothing
6	True Blood	TV show
7	Dexter	TV show
8	Taco Bell	Food beverages
9	Lady Gaga	Musician band
10	Pepsi	Food beverages

Table 4: Different properties of undirected and weighted normalized brand-brand network $\mathbb{B}_n$.

Property	Network $\mathbb{B}_n$
Number of nodes	$2,000$
Number of links	$965,605$
Average weighted degree	0.662
Network density	0.483
Network diameter	4
Average clustering coefficient	0.785
Average weighted clustering coefficient	0.882
Average path length	1.503

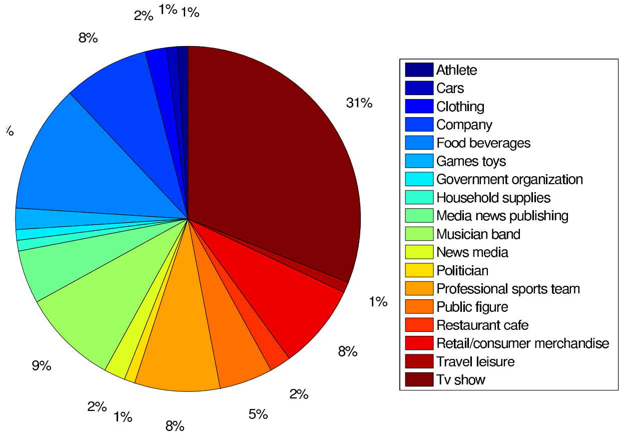


Figure 6: Category distribution of top 100 influential brands.

To summarize, we list the basic properties of our weighted and undirected normalized brand-brand network $\mathbb{B}_n$ in Table 4.

6. IMPLICATIONS OF BRAND-BRAND NETWORK.

Our network analysis provided insights into the properties of the brand-brand network. In this section, we will answer two remaining research questions: **RQ2** and **RQ3**. For this, we first introduce the notion of sentiment of a brand based on all comments made by users on the brand (after data cleansing).

6.1 Brand Sentiment Identification.

Users tend to express their opinions positively, neutrally, or negatively through their comments. Many easy-going or optimistic users tend to make non-negative comments or "like" other's posts, while some tough or pessimistic users like to leave non-positive comments. Previous researchers have developed ways to identify brand sentiment based on user activities on social media platforms [31]. In this work, we use a simple technique to identify brand sentiment, by calculating the positive ratio of all historical comments made by users on that brand. Since we already eliminated spam comments during the data cleansing process, this technique works well and generates good results. Sentiment of a brand b is defined as:

$$\text{SENT}_b = \frac{\#\text{ of positive comments}}{\#\text{ of positive comments} + \#\text{ of negative comments}}$$

We ignore neutral comments here because they do not express any opinions but just state facts.

The textual sentiment algorithm we use in this paper is explained here. We consider three types of values: positive, negative, and neutral. Our textual sentiment identification algorithm integrates the following three different individual components. The first is a rule-based method extended from the basic compositional semantic rules which include twelve semantic rules and two compose functions ([5]). For instance, Rule A is: If a sentence contains the key word "but", then consider only the sentiment of the "but" clause. According to this rule, the following statement is considered positive: "*I've never liked that director and major actors, but I loved the story shown in this movie.*" Compose functions generate integers from −5 to +5 as output to represent sentiment scores. The second component is a frequency-based method. We argue that rather than simply being classified as positive, negative, or objective, the sentiment should be given a continuous numerical score (e.g., −5 to +5) to reflect the sentiment strength. The strength of a sentiment is expressed by the adjective and adverb used in the sentence. We consider two kinds of phrases that derive numerical scores: the phrases in the forms of Adverb-Adjective-Noun (abbreviated as AAN) and Verb-Adverb (VA). Scores were calculated for key words based on a large collection of customer reviews, each of which is associated with a rating. The details of the score calculation can be found in our previous work [31]. Here, we present a few examples. "Easy" has a score of 4.1, "best" 5.0, "never" -2.0, and "a bit" 0.03. Furthermore, the third bag-of-word component considers special characters commonly used in social media text, such as emoticons, negation words and their corresponding positions, and domain-specific words. For example, ":-)" is a positive sentiment and ":-(" a negative sentiment. Some words and phrases express positive opinions like "1st", "Thank you, Obama", "Go bulls", "Thumbs up". Some domain specific words are also included, like "Yum, Yummy" for food related brands. Finally, a random forest machine learning model is applied to the features generated from the output of the three components. The out- puts are represented as three basic features (TS_1 , TS_2 , TS_3) and two derived features (TS_1 + TS_2, TS_1 - TS_2). Our sentiment identification algorithm is trained on manually labeled Facebook comments (2,000) and Twitter text(2,000) using 4 different learning algorithms (decision tree, neural network, logistic regression, and random forest). The random forest learning algorithm was found to achieve the best accuracy of 86%.

Table 5: Spearman rank correlation between the eigenvector centrality of a brand and the size of a brand, the sentiment of a brand, respectively. "EC": Eigenvector Centrality; "SRC": Spearman Rank Correlation; "sent": sentiment.

SRC(EC vs. size)	SRC(EC vs. sent)
0.676	−0.282

6.2 Network Analyses Implications.

Each brand has a number of followers (also called fans) on Facebook. The number of fans can represent the size of a brand, in that a big brand has a strong ability to attract more fans. However, the question that arises is, ÒIs a big brand also more influential/importantÓ? To answer this question, we calculate the Spearman rank correlation between eigenvector centrality and the number of fans. The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the ranked variables. For a sample of size n, the n raw scores X_i, Y_i are converted to ranks x_i, y_i, and ρ is computed using the following equation:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

The correlation value (SRC(EC vs. size)) between eigenvector centrality and the size of a brand in Table 5 tells us that **the size of a brand has a highly positive correlation with the influence of a brand, which means that a big brand is likely to influence other brands in the network.**

For each brand in the network, we collected all comments made by fans across various topics and calculated its sentiment using the sentiment identification algorithm described earlier. To determine if an influential brand receives more positive or negative comments we calculated the Spearman rank correlation between eigenvector centrality and the sentiment of a brand. Surprisingly, the value of SRC(EC vs. sent) in Table 5 demonstrates that **the influence/importance of a brand within the network has a low but negative correlation with its sentiment. This also implies that negative comments on brands are likely to propagate much faster and get more attention than positive comments.**

7. CONCLUSION AND FUTURE WORK.

This paper described a network analysis approach to analyze large dataset containing user historical activities on a social media platform to study characteristics of brands. It is based on an implicit brand-brand network extracted from user interests expressed in brand communities. To our knowledge, it is one of the first studies that develop networks showing relationships between brands based on a large social media dataset.

We proposed a framework for an empirical analysis of network characteristics which includes three components: the first is constructing and normalizing a brand-brand network. Second, we analyzed the network using modified network measures, including weighted degree, density, network diameter, clustering coefficient, and centrality to answer the first research question mentioned earlier; eigenvector centrality reveals brand importance/influence in the network. Third, we addressed two additional research questions. Our findings show that an influential brand has a highly positive correlation with the size of a brand, but a low negative correlation with the sentiment of a brand.

We conducted our experiments on a dataset collected from Facebook. Some simple but effective rules were designed to remove spam activities and users to improve data quality, which is an important consideration when using noisy social media datasets. Given the large data volume, we implemented the network generation algorithm using a MapReduce-based technique in a Hadoop environment; this ensures scalability needed for analysis of large networks.

The brand-brand network developed here is based on user historical activities, but not the content of these activities (except the sentiment of brands determined from comments). Incorporating content analysis can provide a deeper understanding of user activities and interactions and is a topic for continued research. The brand-brand network that our work develops provides a unique view of relationships between brands from the aggregation of consumers' overlapping interests. Such brand networks can be significant for exploring inter-relationships among brands, and related brand communities. Analyses of the undirected and weighted brand network using network measures, such as, centrality can help identify influential brands. Such brand networks can be of much interest for marketing and for obtaining broader understanding of social influence and communication through social media.

References

[1] S. Aral and D. Walker. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Manage. Sci.*, 57(9):1623–1639, Sept. 2011.

[2] A.-L. BarabÃąsi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[3] A. D. Bruyn and G. L. Lilien. A multi-stage model of word-of-mouth influence through viral marketing. *International Journal of Research in Marketing*, 25(3):151 – 163, 2008.

[4] Y. Choi and C. Cardie. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 793–801, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[5] Y. Choi and C. Cardie. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 793–801, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[6] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 519–528, New York, NY, USA, 2003. ACM.

[7] K. de Valck, G. H. van Bruggen, and B. Wierenga. Virtual communities: A marketing perspective. *Decis. Support Syst.*, 47(3):185–203, June 2009.

[8] L. de Vries, S. Gensler, and P. S. Leeflang. 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, 2012.

[9] X. Ding, B. Liu, and P. S. Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 231–240, New York, NY, USA, 2008. ACM.

[10] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, Mar. 1977.

[11] S. Hill, F. Provost, and C. Volinsky. Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science*, 22(2):256–275, 2006.

[12] R. Iyengar, C. Van den Bulte, and T. W. Valente. Opinion leadership and social contagion in new product diffusion. *Marketing Science*, 30(2):195–212, Mar. 2011.

[13] L. Ku, Y. Liang, and H. Chen. Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs'06*, pages 100–107, 2006.

[14] K. Lerman, R. Ghosh, and J. H. Kang. Centrality metric for dynamic networks. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, MLG '10, pages 70–77, New York, NY, USA, 2010. ACM.

[15] J. Liu and S. Seneff. Review sentiment scoring via a parse-and-paraphrase paradigm. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, EMNLP '09, pages 161–169, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[16] R. Mcdonald, K. Hannan, T. Neylon, M. Wells, and J. Reynar. Structured models for fine-to-coarse sentiment analysis. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007.

[17] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 171–180, New York, NY, USA, 2007. ACM.

[18] S. Nam, P. Manchanda, and P. K. Chintagunta. The effect of signal quality and contiguous word of mouth on customer acquisition for a video-on-demand service. *Marketing Science*, 29(4):690–700, 2010.

[19] R. Narayanan, B. Liu, and A. Choudhary. Sentiment analysis of conditional sentences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, EMNLP '09, pages 180–189, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[20] M. J. Newman. A measure of betweenness centrality based on random walks. *Social Networks*, 27(1):39 – 54, 2005.

[21] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[22] N. J. Patterson, M. J. Sadler, and J. M. Cooper. Consumer understanding of sugars claims on food and drink products. *Nutrition Bulletin*, 37(2):121–130, 2012.

[23] O. Perron. Zur theorie der matrices. *Mathematische Annalen*, 64(2):248–263, 1907.

[24] A.-M. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 339–346, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[25] V. Stoyanov and C. Cardie. Topic identification for fine-grained opinion analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 817–824, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[26] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 1397–1405, New York, NY, USA, 2011. ACM.

[27] J. Tang, Y. Zhang, J. Sun, J. Rao, W. Yu, Y. Chen, and A. C. M. Fong. Quantitative study of individual emotional states in social networks. *Affective Computing, IEEE Transactions on*, 3(2):132–144, April 2012.

[28] C. Tucker. Social advertising. *SSRN: http://ssrn.com/abstract=1975897*, 2012.

[29] P. D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[30] A. M. Turri, K. H. Smith, and E. Kemp. Developing Affective Brand Commitment Through Social Media. *Journal of Electronic Commerce Research*, 14(3):201–214, 2013.

[31] K. Zhang, Y. Cheng, Y. Xie, D. Honbo, A. Agrawal, D. Palsetia, K. Lee, W. keng Liao, and A. N. Choudhary. Ses: Sentiment elicitation system for social media data. In *ICDM Workshops'11*, pages 129–136, 2011.

[32] L. Zhang and B. Liu. Identifying noun product features that imply opinions. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 575–580, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[33] Y. Zhang, J. Tang, J. Sun, Y. Chen, and J. Rao. Moodcast: Emotion prediction via dynamic continuous factor graph model. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 1193–1198, Dec 2010.

APPENDIX

A. CLUSTERING COEFFICIENT.

The basic idea behind transitivity is how reliably we can say that if nodes i and j and node i and k are connected whether j and k are connected as well. If that is the case then we have a triangle. In general, if $A_{ij} = 1$ and $A_{ik} = 1$, then the nodes i, j, k form a triplet. There are only two types of triplets: triangles and non-triangles. Triangles contain 6 paths of length 3 whereas non-triangles, regular triplets, contain two paths of length 2. Particularly in social networks a large fraction of triplets are triangles, which means if X is friends with Y and Z then with a high probability Y and Z are also friends. It can be also applied and explained in our brand-brand network. Thus, one way of measuring the strength of transitivity of an undirected unweighted network is by the fraction of triangles with respect to the entire set of triplets.

$$C = \frac{3 * \#\ triangles}{\#\ triplets} = \frac{6 * \#\ triangles}{\#\ paths\ of\ length\ 2}$$

The number of triangles is given by:

$$n_{\triangle} = \frac{1}{6} TrA^3$$

The number of paths of length 2 between two given nodes is given by:

$$n_2(i, j) = \sum_k A_{ik} A_{kj} = (A^2)_{ij}$$

The total amount of paths of length 2 is:

$$n_2 = \parallel A^2 \parallel - TrA^2$$

where $\parallel \cdot \parallel$ means summing over all matrix elements. The trace of matrix A means $TrA = A_{11} = A_{22} + \cdots + A_{nn} = \sum_i^n A_{ii}$. Finally, we obtain:

$$C = \frac{TrA^3}{\parallel A^2 \parallel - TrA^2}$$

This also called clustering coefficient.