A Probabilistic Graphical Model for Brand Reputation Assessment in Social Networks

Kunpeng Zhang Doug Downey Zhengzhang Chen Yusheng Xie Yu Cheng

Ankit Agrawal Wei-keng Liao Alok Choudhary

EECS Department
Northwestern University
Evanston, Illinois 60208
Email: {kzh980, ddowney, zzc472, yxi389, ych133, ankitg, wkliao, choudhar}@eecs.northwestern.edu

***Abstract*—Social media has become a popular platform that connects people who share information, in particular personal opinions. Through such a fast information exchange mechanism, reputation of individuals, consumer products, or business companies can be quickly built up within a social network. Recently, applications mining social network data start emerging to find the communities sharing the same interests for marketing purposes. Knowing the reputation of social network entities, such as celebrities or business companies, can help develop better strategies for election campaigns or new product advertisements. In this paper, we propose a probabilistic graphical model to collectively measure reputations of entities in social networks. By collecting and analyzing large amount of user activities on Facebook, our model can effectively and efficiently rank entities, such as presidential candidates, professional sport teams, musician bands, and companies, based on their social reputation. The proposed model produces results largely consistent with the two publicly available systems - movie ranking in Internet Movie Database and business school ranking by the US news & World Report - with the correlation coefficients of 0.75 and -0.71, respectively.**

I. Introduction

Social networks, such as Facebook, Twitter, and Flicker, are becoming ubiquitous that change the way the modern world operates. They make it convenient to keep up with friends, family, and colleagues, discover great contents, connect to causes, share photos, drum up business, and learn about fun events. Online shopping companies, such as Amazon, eBay, and Bestbuy, provide e-commerce social network platforms that allow shoppers to leave comments for the purchased products. Such consumer reviews can immediately help potential customers to make purchase decisions. They also greatly benefit manufacturers for improving their next generation products.

Many data mining methods have recently been proposed to study the online consumer reviews. Hu *et al*. [10] developed a set of sentiment analysis algorithms and opinion mining techniques to summarize the customer reviews based on product features. Pang *et al*. [17] applied machine learning and rule-based techniques to classify text sentiments and achieved results with a satisfactory accuracy for movie reviews. Liu *et al*. [27], authors exploited social relations for sentiment analysis in microblogging proposing a Sociological Approach to handling Noisy and short Texts (SANT). Zhang *et al*. [25] made use of graph-based mining methods to rank online products by considering both overall and a set of features.

There is also a growing need to evaluate social brands. This will help not only business managers monitor the marketing growth of their brands, but also consumers make informed purchase decisions. Similar to the e-commerce, user opinions on other social networks are as important to cause a significant impact toward the decision making, political elections, travel plans, school applications, and so on. However, there are no previous efforts on reputation assessment through analyzing social media data. Most related work on mining the sentiments do not consider the fact that people have different evaluation standards. In other words, comments made by different users should not be considered to carry the same. A user's tendency to make positive or negative comments made across all topic domains can be used to help normalize the sentiment weights among the group of users toward the same subject. Incorporating this network information will generate results with more confidence due to the reduction in various biases.

In this paper, we propose a probabilistic graphical model to measure social brand reputation. This model adopts a block-based Markov Chain Monte Carlo (MCMC) sampling method to infer the probability of hidden variables, social brand reputations and user positivities. Direct calculating the joint probability of hidden variables is very expensive because of the large state space. The block-based MCMC sampling method considers users and brands as two separate blocks. The property of conditional independence enables us to implement variable sampling in parallel. The model is evaluated by using a large amount of Facebook data. The experiments show that social brand reputation ranking produced by our system is significantly consistent with some existing publicly available ranking systems: IMDB movie ranking and top business schools ranking by the US News & World Report. In addition, we also explain that some other measurements such as the number of fans, the number of likes, the number of positive comments, and the percentage of positive comments are not good indicators for measuring social brand reputation. The general framework of our system could be pictorially depicted in the Fig. 1. Users make either positive or negative comments across brands. Brands can receive comments from different

users.

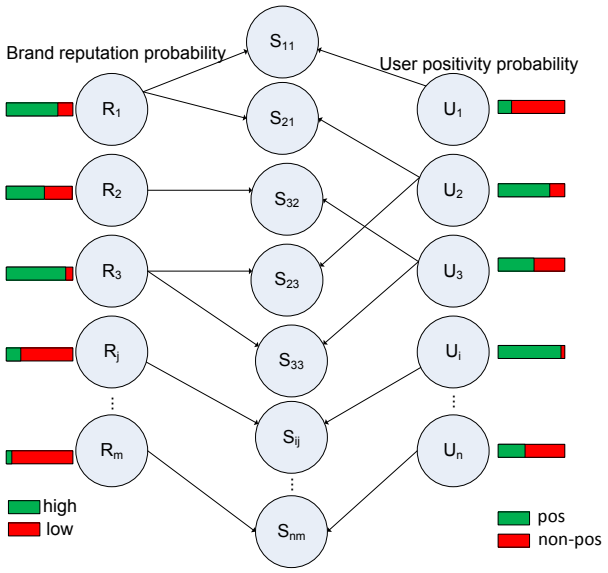


Fig. 1: A collective inference framework of our system. U_i: positivity of a i-th user; R_j: reputation of a j-th brand; S_{ij}: sentiment of comments made by U_i on brand R_j. Green means the high reputation for brands and more positive for users. Red has the opposite meaning. (Best viewed in color)

The task of measuring social brand reputation from large scale social media data is faced with the following challenges: 1) There are many ways to represent the relations among users and social brands. Choosing a reasonable model to bridge this relationship becomes a fundamental step for this task, which accordingly would affect the performance of the system. 2) The accuracy of sentiment identification of social media data is of course very crucial. 3) Due to the volume of the data (described in detail in the experimental results section), the data quality and the efficiency of inference is another big challenge. To address these challenges, 1) A probabilistic graphical model is developed to bridge the relationship between users and social brands. 2) A parallelized block MCMC technique is applied here to infer the probability of hidden variables. 3) We filter out some spam data by designing a very effective strategy. To the best of our knowledge, this is the first work which takes such a large amount of social media data and applies graphical modeling techniques to infer social brand reputation.

The rest of this paper is organized as follows: we study some related work in section 2. In section 3, we focus on our proposed model and collective inference. Section 4 describes the data we use for experiments, presents experimental results, and is followed by conclusion and future work in section 5.

II. Related Work

In this section, we describe related work on three aspects: 1) Recently, a number of papers have begun to investigate how to apply modern statistical machine learning methods on user generated social media text and social network information for online advertising. 2) Different recommendation algorithms and systems have been proposed and developed in many areas. 3) With the social media data growing incredibly fast, analyzing their sentiments and summarizing their content become important to help make informed decisions.

Behavioral Targeting: Most related works are called social targeting or behavioral targeting which learns from past user behaviors, especially feedbacks (i.e., comments, clicks) to match the best advertisements to users. This has resulted in a spike of interest in user data analysis and profile generation as published in [4], [13]. In the area of audience selection, Provost *et al*. [20] have recently shown that user profiles can be built based on co-visitation patterns of social network pages. These profiles are used to predict the performance of brand display advertisements over 15 campaigns. In [2], Ahmed *et al*. presented a time-varying hierarchical user model that captures both the user's long term and short term interests. A dynamic topic model was employed. They also showed a streaming distributed inference algorithm that both scales to tens of millions of users and adapts the inferred user's interest as it gets to know more about the user. Our work is different in that we focus on collective inference of social brand reputation based on large scale user behaviors.

Recommender Systems: In [9], authors proposed that Twitter users can be usefully modeled by the tweets and relationships of their Twitter social graphs. They developed a recommender system called Twittomender which demonstrated how user profiles can be used as the basis for recommendation. This work showed the potential of real-time web, and micro-blogging services like Twitter, to serve as a useful source of recommendation information. In [29], Liu et al. modeled the topic-level social influence in heterogeneous networks. In [30], authors predicted links and made recommendations across heterogenous social networks. In [11], Jin *et al*. applied a set of new patent features to capture important textual and time-evolving properties for patent maintenance. They also proposed a network-based refinement approach utilizing the patent information network for prediction, smoothing and optimization. Abundant online news gave researchers opportunities to build some online product recommendation systems. In [24], Wu *et al*. presented the recommendation and summarization components of personalized news filtering and summarization system. They designed a content-based news recommender that automatically obtains World Wide Web news from the Google news website and recommends news to users according to their preference. K-nearest neighbor and Naïve Bayes methods were used to model user interest preference in their paper. Our work is different from those recommender systems in that we estimate reputation probabilities of social brands based on user historical behaviors instead of recommending any brand to social users.

Sentiment Analysis: Recently, there has been a wide range of research done on sentiment analysis, from rule-based, bag-of-words approaches to machine learning techniques which classifies the whole opinion document as positive or negative [17], [21], [23]. In [14], [15], authors analyzed sentiments from a sentence level. Researchers have also studied feature/topic-based sentiment analysis [6], [12], [16], [18]. In [28], authors aimed to identify social sentiments by incorporating user-lever in social networks.

In addition, anomaly detection and dynamical evolution for social network data is also related to our work, because we

also designed some rules to filter out noisy data and spam users. In [3], authors provided a comprehensive and structured overview of the research on anomaly detection techniques. In [19], authors presented an unsupervised framework for detecting anomalous nodes in bipartite graphs. In [22], Osmar *et al*. presented a framework for modeling and detecting community evolution in social networks, where a series of significant events is defined for each community. A community matching algorithm is also proposed to efficiently identify and track similar communities over time. In [8], authors proposed an efficient solution by modeling networked data as a mixture model composed of multiple normal communities and a set of randomly generated outliers. The probabilistic model characterizes both data and links simultaneously by defining their joint distribution based on hidden Markov random fields (HMRF). Maximizing the data likelihood and the posterior of the model gives the solution to the outlier inference problem.

III. Methodology

In this section, we describe our proposed probabilistic graphical model by first defining several terminologies referred in this paper. Then, we identify sentiments of all comments made by users are identified by applying our ensemble learning technique. The overlapped graphical model to represent the relationship among brands and users bridged on comments is built under some assumptions. Based on the probability conditional independences, we propose a blocked-based MCMC sampling method to collectively infer probabilities of brand reputation and user positivity. To efficiently do the inference, we implement a paralleled version.

A. Definitions

Social Brand - A social brand is an entity in the social network that allows other users to leave comments on it (e.g. on Facebook its page). Examples are companies, organizations, individuals, or consumer products.

Brand Reputation - Brand reputation means how good a given brand is perceived in the market, especially how it is evaluated by users on the social media platform. A brand with higher reputation is likely to attract more attentions and positive comments from their fans. In contrast, a lower reputation brand is likely to receive more non-positive comments.

User Positivity - Different users may have different evaluation standards. The contributions from different users shall not be considered equally. For example, a tough user tends to make non-positive comments on the brands and the opposite for an easy-going user.

B. Sentiment Identification

Sentiment analysis for social texts (comments) is the key component of our model. Our sentiment identification algorithm integrates the following three different individual components [26]. The first is a rule-based method extended from the basic compositional semantic rules [5] which include twelve semantic rules and two compose functions. Compose functions generate integers from -5 to $+5$ as output to represent sentiment scores. Here gives an example. Rule A: 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 (score is $+3$): *"I've never liked that director, but I loved this movie."* The second component is a frequency-based method. We argue that the sentiment should not be simply classified as positive, negative, or objective but 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). The scores of key words were used are calculated based on a large collection of customer reviews, each of which is associated with a rating. The details of score calculation can be found in [26]. Here are 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 Internet language expresses positive opinions like "1st!", "Thank you, Obama", "Go bulls!". 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 outputs of the three components. The outputs from three individual components are represented as three basic features (S_1, S_2, S_3) and we also have two derived features $(S_1 + S_2, S_1 - S_2)$. Our sentiment identification algorithm is trained on manually labeled Facebook comments and Twitter tweets and achieved an accuracy of 86%. In this paper, we only consider binary sentiment values for comments produced by the trained model: positive for scores larger than a threshold ($\gamma = 0.7$) and non-positive otherwise.

C. Graphical Model

Many social media platforms allow users to leave comments on public pages. For instance, Facebook fans can express their opinions on consumer product campaigns of business brands. It is very common to see that the same users make comments on campaigns in different domains. Therefore, extracting such user activities across domains and analyzing the sentiments are important to justify the degree of contributions toward a brand for a given individual. We propose a probabilistic graphical model to incorporate information about the networked structure and semantics of the text. One fact is that a positive comment is made by a positive user to a higher reputation brand with a high probability. If a brand has a low reputation, it obviously attracts more non-positive comments. In this case, we shall observe some weakly negative comments made by easy-going users who usually write positive comments. If a brand has a lower reputation and most of the comments came from tough users, then those comments are most likely negative. Based on these assumptions, we construct a Bayesian model which is depicted in Fig. 2. This plate model has a similar but succincter representation of variable relationships than Fig. 1. It describes the fact that a user makes comments on multiple brands and a brand receives comments from different users.

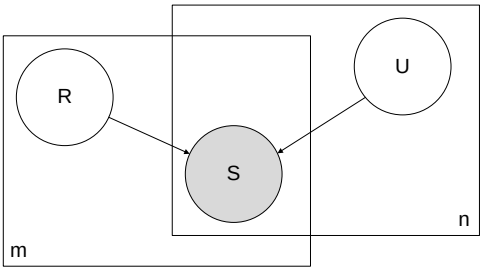


Fig. 2: Plate model. The shaded S is an observed variable, representing the sentiment of comments made by a user on a brand. R and U are hidden variables representing brand reputation and user positivity, respectively. All these variables in this model have binary values. m: number of brands, n: number of users.

D. Collective Inference

The goal of this task is to infer reputation of social brands (R) and positivity of social users (U) from the sentiments of observed comments (S). It is equivalent to infer the joint probability of R's and U's from S, as represented by the following equation (Eq. (1)).

$$
\begin{aligned}
&P(R_i \mid S_{11}, \cdots, S_{ij}, \cdots, S_{mn}) \\
&= \sum_{R_{-i},U} P(R_1, \cdots, R_m, U_1, \cdots, U_n \mid S_{11}, \cdots, S_{ij}, \cdots, S_{mn}) \\
&= \sum_{R_{-i},U} \frac{P(R_1,R_2,\cdots,R_m,U_1,U_2,\cdots,U_n,S_{11},\cdots,S_{ij},\cdots,S_{mn})}{\sum_{i,j} P(S_{ij})} \qquad (1)
\end{aligned}
$$

where $1 \leq i \leq m, 1 \leq j \leq n$, S_{ij} is the sentiment value of comments made by user i on brand j ($S_{ij} = 1$ if the sentiment score is greater than a designated threshold, $S_{ij} = 0$ otherwise.). Once we have this probability, we can get the probability of R_i and U_j by summing out some other variables. However, it is difficult to calculate the denominator (partition function) due to a large discrete state space, which often arises in statistical physics. We apply a method developed by physicists and statisticians to sample from the target distribution with the Markov Chain Monte Carlo method (MCMC). In order to calculate the posterior distribution, we define the conditional probability distribution (CPD): $P(S \mid R, U)$. Table I presents the CPD values. Note that a noise factor δ is introduced to make our Bayesian model more realistic. Parameters of α and β for $P(S \mid R = 0, U = 1), P(S \mid R = 1, U = 0)$ are chosen based on our prior knowledge of this domain. We assume that users with lower positivity give more positive comments on brands with higher reputation. Similarly, users with higher positivity give less positive comments to the brands with lower reputation. Thus, we obtain $\alpha < \beta$.

In the MCMC method, a Markov chain is first constructed to converge the target distribution, and samples are then taken from the Markov chain. The state of each chain is an assigned value to the variables being sampled, and the transitions between states follow a rule based on MCMC method, known as the heat bath algorithm in statistical physics. The rule asserts that the next state of a chain is reached by sequentially sampling all variables from their distribution when conditioned on the current values of all other variables and the data. To apply this algorithm, we define the full conditional distribution $P(R_i \mid R_{-i}, U, S)$ for brands and $P(U_j \mid U_{-j}, R, S)$ for users. The distribution uses the probabilistic arguments from Table I by canceling out some terms due to the properties of Bayesian network, which yields:

TABLE I: Conditional probability distribution for R, U, S. S^1: positive sentiment, S^0: non-positive sentiment. α, β, and δ are parameters based on prior domain knowledge.

R	U	S^1	S^0
0	0	δ^2	$1-\delta^2$
0	1	α	$1-\alpha$
1	0	β	$1-\beta$
1	1	$1-\delta$	δ

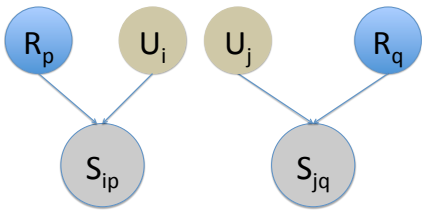
$$
\begin{aligned}
&P(R_i \mid R_{-i}, U, S) \\
&= \frac{P(R,U,S)}{P(R_{-i},U,S)} \\
&= \frac{P(R,U,S)}{\sum_{R_i} P(R,U,S)} \\
&= \frac{P(R_1)\cdots P(R_m)\cdot P(U_1)\cdots P(U_n)\cdot \prod_{i,j} P(S_{ij}|U_i,R_j)}{\sum_{R_i} P(R_1)\cdots P(R_m)\cdot P(U_1)\cdots P(U_n)\cdot \prod_{i,j} P(S_{ij}|U_i,R_j)} \\
&= \frac{P(R_i)\prod_k P(S_{ki}|U_k,R_i)}{\sum_{R_i} P(R_i)\prod_k P(S_{ki}|U_k,R_i)} \qquad (2)
\end{aligned}
$$

$$
\begin{aligned}
&P(U_j \mid U_{-j}, R, S) \\
&= \frac{P(U,R,S)}{P(U_{-j},R,S)} \\
&= \frac{P(U,R,S)}{\sum_{U_j} P(U,R,S)} \\
&= \frac{P(R_1)\cdots P(R_m)\cdot P(U_1)\cdots P(U_n)\cdot \prod_{i,j} P(S_{ij}|U_i,R_j)}{\sum_{U_j} P(R_1)\cdots P(R_m)\cdot P(U_1)\cdots P(U_n)\cdot \prod_{i,j} P(S_{ij}|U_i,R_j)} \\
&= \frac{P(U_j)\prod_k P(S_{jk}|U_j,R_k,)}{\sum_{U_j} P(U_j)\prod_k P(S_{jk}|U_j,R_k)} \qquad (3)
\end{aligned}
$$

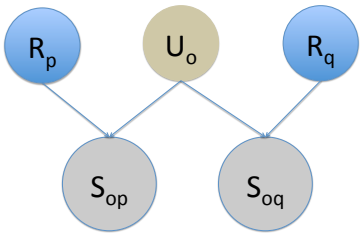
where R_{-i} represents $\{R_1, R_2, \cdots, R_{i-1}, R_{i+1}, \cdots, R_m\}$, while U_{-j} represents $\{U_1, U_2, \cdots, U_{j-1}, U_{j+1}, \cdots, U_n\}$. With further investigation of our model, we realize that there are lots of conditional independencies that can make the inference calculation more efficient. First, $R_1, R_2, \cdots, R_m$ are independent of each other given all $U_1, U_2, \cdots, U_n$ and all observed variables S_{ij}. Similarly, all $U_1, U_2, \cdots, U_n$ are independent of each other given $R_1, R_2, \cdots, R_n$ and all S_{ij}. There are two cases both of which show the conditional independence of R_p and R_q given all U and S, similar cases for U.

- R_p and R_q do not have any common users as shown in Fig. 3a. It is obviously that R_p and R_q are independent given all $U_1, U_2, \ldots, U_n$ and S_{ij};
- R_p and R_q have common users U_o as shown in Fig. 3b. They are still conditional independent because U_o blocks the path from R_p to R_q given U_o, $S_{op}, and S_{oq}$ are known.

These conditional independencies give us opportunities to



(a) Case 1: no common users



(b) Case 2: some common users

Fig. 3: Two cases for the conditional independence of R_p and R_q, given all U and S.

sample brands and users in parallel. We implement a block-based MCMC method that processes users and brands as two separate blocks. We alternately sample all R_i's and U_j's in each sampling round. The detailed algorithm is depicted in Algorithm 1. Performance comparison between our parallelized block-based MCMC and sequential MCMC is presented in the experimental results section.

Algorithm 1 Parallelized block-based MCMC inference

Require: $1 \leq i \leq m$ and $1 \leq j \leq n$
Require: noise factor: $\delta = 0.1; \alpha = 0.3; \beta = 0.6$; sentiment threshold: $\gamma = 0.7$
Initialization: $P(R_i) \leftarrow 0.5$; $P(U_j) \leftarrow 0.5$
for all (i, j) **do**
  $S_{ij} = \frac{\#PC}{\#PC+\#NC}$[1]
  if $S_{ij} > \gamma$ **then**
    $S_{ij} = 1$;
  else
    $S_{ij} = 0$;
  end if
end for
repeat
  In the k'th round:
  Parallelize sampling R_i based on equation (2);
  if $P(R_i) \geq rand(0,1)$[2] **then**
    $R_i^{(k)} \leftarrow 1$
  else
    $R_i^{(k)} \leftarrow 0$
  end if
  Parallelize sampling U_j based on equation (3);
  if $P(U_j) \geq rand(0,1)$ **then**
    $U_j^{(k)} \leftarrow 1$
  else
    $U_j^{(k)} \leftarrow 0$
  end if
  $P(R_i) = \frac{\sum_k R_i^{(k)}}{k}, P(U_j) = \frac{\sum_k U_j^{(k)}}{k}$;
until The target distribution converges (mixing time)
return $P(R_i), P(U_j)$

IV. Experimental Results

The data used in our experiments is collected from Facebook through its Graph API [7]. Once the data is downloaded, we first remove spam activities. After the data is cleaned, our block-based MCMC inference algorithm is applied to obtain the reputation probabilities of social brands and positivity probability of users. We compare our results with two existing ranking systems: movie rankings from the Internet Movie Database (IMDB), and top business schools ranked by the US News & World Report. Comparison will also be presented between our parallelized inference method and the sequential sampling algorithm. Performance analysis will be given to explain why other measurements, like the number of fans, post, comments, the percentage of positive comments, etc. are not sufficient to measure brand reputation.

A. Data Description

Facebook, the largest and most popular social network platform, has attracted a lot of attention from markets. Many companies, organizations, and individuals build their own pages to communicate with social users (fans). The extensive amount of network and text information also shifted researchers' focus to this emerging field, social network data analysis. In this paper, we mainly consider social brands as our target objects. We use Facebook Graph API to download the available activities made on brand side such as posts and user side, such as comments on posts, likes on posts, and public profiles.

Data Cleaning: We consider pages of top brands, i.e. the brands with a large number of fans. By May 1, 2012, we have collected $11,140$ pages and approximately 270 million users in our database. The data quality is important as it can affect our model performance. Our first step is to remove brands in which most of posts and comments are not written in English, because the sentiment identification for non-English texts has not been well studied. To produce results of more robust and reliable, we ignore the brand pages receiving very few comments, as user opinions drive the measurement of brand reputation. After the two initial filtering steps, there are $7,523$ brand pages left. We then applied a spam filter to remove spam users. Our data shows that on average, a user comments on 4 to 5 brands. Users making comments on an extremely large number of brands are likely spam users or bots. For example, we found one spam user appeared on $600+$ different brand pages. Fig. 4 shows the distribution of user activities on brands. As the most users are interested in a handful of brands, in our experiments we set the threshold of 100 to discard users making comments on more than 100 brands. We also ignore users who make comments on only one brand, because in most of the cases these users

[1] One user could have multiple comments on one brand. PC: positive comments made by user i on brand j, NC: Non-positive comments made by user i on brand j

[2] $rand(a, b)$ function generates a random number between a and b

like/dislike excessively the brand and can potentially bias our model. To fairly determine the positivity of users, we set the threshold of minimum number of brands on which a user must comment to be 2. Similarly, we also ignore users who have very few total comments across all brands. The threshold of the minimum number of comments is set to be 5. We also remove users who leave many duplicated comments on the same brand and the duplicated comments contain URL links. A test on a very popular brand/page, Barack Obama's page, found $209,864$ duplicated comments from same users out of $2,987,505$ in total. Our data cleaning process significantly and effectively improves the data quality. Table II describes the cleaned data used in our experiments.

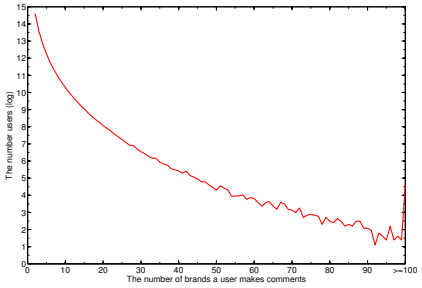


Fig. 4: The distribution of user activity is presented as the number of social brands on which users make their comments. The Y-axis is in log scale.

TABLE II: The stats of data after cleaning.

The number of unique users	$15,528,173$
The number of social brands	$7,523$
The number of comments	$126,613,072$
The number of positive comments	$93,233,898$
The number of negative comments	$33,379,174$
The number of different categories declared on Facebook	150
The number of different countries brands from	71
The number of total posts	$8,186,454$

B. Performance Evaluation

Two common important aspects related to MCMC-based inference algorithm are: 1) the convergence speed, 2) the time complexity of sampling. To investigate the first aspect, we randomly pick 5 brands from different categories and plot their reputation probabilities as the sampling proceeds. Fig. 5 clearly shows that they converge after we collect about approximately 250 samples (we also call this mixing time).

To address the second aspect, we parallelize block-based MCMC due to conditional independency and compare to the sequential version in terms of time taken. Speedup is a very common metric used in the field of parallelization. In this experiment, we use 8 processors in parallel. Fig. 6 shows that we achieve near-linear speedup (close to 7). It is the speedup with respect to the cumulative time until $50^{th}, 100^{th}, \cdots, 750^{th}$ sampling round. We believe that the reason for slightly uneven speedup at some sampling rounds is that MCMC is a stochastically approximate sampling technique based on the

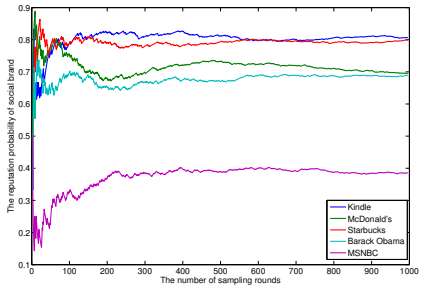


Fig. 5: The probability convergence of the block MCMC algorithm for five different brands. "Kindle" - Product/service, "McDonald's" - Food/beverages, "Starbucks" - Food/beverage, "Barack Obama" - Politician, "MSNBC" - News/media.

target distribution as shown in Eq. (2) and Eq. (3). Experiments were run on a machine with 256 GB memory and 24 cores.

$$S_p = \frac{T_1}{T_p}$$

where p is the number of processors ($p = 8$); T_1 is the execution time of sequential algorithm; T_p is the execution time of parallel algorithm with p processors.

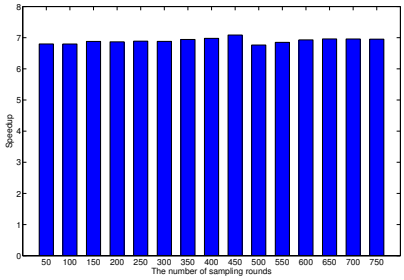


Fig. 6: Speedup of the parallelized block MCMC algorithm over the sequential algorithm on 8 computing cores.

Model Evaluation: We compare our results with two existing publicly available ranking systems: IMDB movie ranking and top business schools ranked by the US News & World Report. We first extract all movies existing in our dataset and then manually collect their corresponding ratings, the number of votes/reviews from reviewers, and the box office values from IMDB. The higher the rating score, the better the movie is. The total number of movies is 73. Table III shows that the IMDB rating score is consistently correlated with the reputation probability generated from our algorithm. These two different rating systems have different population of users. The IMDB rating evaluates movie reputations by capturing complicated relationships among many variables, while our system is based on user activities on the Facebook social platform. An obvious question that arises here is whether we can use the number of votes/reviews or the box office revenue to rank movie reputations. The results from Table III shows that these two metrics do not have significant correlations with

the reputation and consequently they can not be considered as good indicators to measure movie reputation. The larger number of votes does not mean the better reputation, because some voters might give biased votes even without watching the movies. Similarly, a movie might attract a large number of audience due to a large amount of money spent in advertising. Most of the box office revenue is from the first couple of months since their release. But this can not guarantee better reputation because some people may regret watching it and leave negative comments later.

To demonstrate the independence of our model to specific categories, we compare with another popular and authoritative data - top business school overall rankings based on many factors, including faculty, students, funding, research, graduates, academics, endowment, etc. We collected information for about 35 business schools which also exist in our dataset. Table III shows that the rank correlation between reputation probability and the ranking by the US News & World Report is relatively high. The lower the ranking is, the better the school is, which leads to a negative correlation. P-value here for both datasets is set to be 0.05.

TABLE III: Correlations between social brand reputation and IMDB / top business school ranking.

Reputation $\leftrightarrow$ IMDB rating score	**0.757**
Reputation $\leftrightarrow$ the number of IMDB votes	0.440
Reputation $\leftrightarrow$ the box office revenue	0.283
Reputation $\leftrightarrow$ the US News & World Report ranking of top business school	**-0.715**

C. Discussions

In addition to making comments on Facebook brand pages, brand page administrators also allow users ("fans") to like their posts (we call it "post like") and like their brand. Then the obvious question that arises is: can we use metrics related to these (the number of fans, the number of post likes, the number of comments, the percentage of positive comments) to rank brands in terms of reputations? To investigate this, we calculate the correlation between our computed reputation and these measurements respectively. Table IV shows that these are not very well correlated with our brand reputation probability and the two publicly available ranking systems. Consequently, these are not good metrics for the purpose of predicting IMDB rankings and Business school rankings. The general reasons for this could be: 1) Some brands have longer history since their foundation earlier than others on Facebook. Thus, it is quite likely that they might have more "fans" in general. 2) Tough users and easy-going users are considered as equally weighted which is not fair. Other reasons might include: 1) Different brands might have different posting frequencies resulting in different number of posts and post likes. 2) The most important reason according to us is that all these social actions performed by users are less indicative of their intent than what they wrote (comments).

Another obvious question coming out is that can we mine some patterns based on the combination of all these metrics. We want to use these metrics as features to build a learning model based on our collected and labeled data. In this work, we employ some existing linear and non-linear regression techniques, such as least absolute deviation, Poisson regression, logistic regression, and SVM regression algorithms to train our model on the movie rating data and test on the business school ranking data. We calculate the rank correlation between predicted value and existing labeled ranking. The absolute value of the best rank correlation we obtained is 0.52 through SVM regression. This shows that combination of all these parameters might also be not a good indicator to measure social brand reputation.

TABLE IV: Correlations between other metrics and social brand reputation, IMDB rating, Business school ranking. '#': number; '%': percentage.

Social brand reputation VS. other metrics	
Reputation $\leftrightarrow$ # of fans	-0.129
Reputation $\leftrightarrow$ # of post likes	-0.174
Reputation $\leftrightarrow$ # of post likes per post	-0.117
Reputation $\leftrightarrow$ # of comments	0.153
Reputation $\leftrightarrow$ # of positive comments	0.472
Reputation $\leftrightarrow$ % of positive comments	0.424
IMDB rating VS. other metrics	
IMDB rating $\leftrightarrow$ # of fans	0.093
IMDB rating $\leftrightarrow$ # of post likes	-0.129
IMDB rating $\leftrightarrow$ # of post likes per post	0.228
IMDB rating $\leftrightarrow$ # of comments	0.142
IMDB rating $\leftrightarrow$ # of positive comments	0.251
IMDB rating $\leftrightarrow$ % of positive comments	0.257
Bschool ranking VS. other metrics	
Bschool ranking $\leftrightarrow$ # of fans	-0.151
Bschool ranking $\leftrightarrow$ # of post likes	-0.107
Bschool ranking $\leftrightarrow$ # of post likes per post	-0.138
Bschool ranking $\leftrightarrow$ # of comments	-0.255
Bschool ranking $\leftrightarrow$ # of positive comments	-0.440
Bschool ranking $\leftrightarrow$ % of positive comments	-0.479

Parameter Settings: Parameters are chosen based on prior knowledge. We tried different parameter settings in our experiments. Table V lists some combinations of parameters. It shows that different parameter settings do not affect the ranking results a lot, but they have some impact on the reputation probability. All these probabilities are collected after the mixing time when the probabilities are already converged.

V. Conclusion and Future Work

In this work, we proposed a probabilistic graphical model to represent the relationship between social brands and users. This model not only captures the network information but also includes the semantic information from users in terms of the comments they make. One of the biggest advantages of this model is that it reduces the biased effect from a single user and a single comment. It collectively infers the brand reputation and user positivity. To efficiently perform the inference, we implemented a parallelized block-based MCMC algorithm due to the existence of many conditional probability independencies in our model. We conducted our experiments on a large amount of data from Facebook, and compared our brand reputations with two existing ranking systems: IMDB movie ranking and top business school ranking by US News & World Report. We also explained why other measurements like number of fans, number of post likes, the percentage of positive comments, etc. are not good indicators for brand ranking. We also empirically studied the impact of parameter settings on measuring brand reputation.

TABLE V: Different parameter settings for calculating probability of brand reputation ($*$: the default setting). CI: correlation with IMDB, CB: correlation with Business school ranking.

	Probability of brand reputation					CI	CB
Parameter settings	Kindle	McDonald's	Starbucks	Barack Obama	MSNBC		
$\delta = 0.1, \alpha = 0.3, \beta = 0.6, \gamma = 0.7^*$	0.828	0.720	0.783	0.675	0.398	**0.757**	**-0.715**
$\delta = 0.05, \alpha = 0.3, \beta = 0.6, \gamma = 0.7$	0.880	0.694	0.829	0.689	0.358	0.711	-0.692
$\delta = 0.1, \alpha = 0.2, \beta = 0.4, \gamma = 0.7$	0.840	0.640	0.804	0.658	0.389	0.726	-0.700
$\delta = 0.1, \alpha = 0.25, \beta = 0.45, \gamma = 0.7$	0.807	0.710	0.784	0.642	0.345	0.743	-0.705
$\delta = 0.1, \alpha = 0.3, \beta = 0.6, \gamma = 0.8$	0.769	0.700	0.789	0.623	0.322	0.708	-0.686

In the future, we will make our graphical model more comprehensive. Although more complicated model could make the model more realistic, the computational complexity of corresponding inference will also become more challenging. Incorporating network information from other social platforms like Twitter, Google+, LinkedIn, etc. could make brand reputation assessment more reliable.

ACKNOWLEDGMENT

This work is supported in part by the following grants: NSF awards CCF-0833131, CNS-0830927, IIS-0905205, CCF-0938000, CCF-1029166, and OCI-1144061; DOE awards DE-FG02-08ER25848, DE-SC0001283, DE-SC0005309, DESC0005340, and DESC0007456; AFOSR award FA9550-12-1-0458.

REFERENCES

[1] C. Aggarwal, and P. Yu, *Outlier detection with uncertain data*. In SDM, 483493.

[2] A. Ahmed, and Y. Low, and M. Aly, and V. Josifovski, and A. Smola, *Scalable distributed inference of dynamic user interests for behavioral targeting*. In Proceedings of the 17th SIGKDD, 2011, 114122. New York, NY, USA.

[3] V. Chandola, and A. Banerjee, and V. Kumar, *Anomaly detection: A survey*. ACM Comput. Surv. 41(3):15:115:58.

[4] Y. Chen, and D. Pavlov, and J. Canny, *Large-scale behavioral targeting*. In Proceedings of the 15th SIGKDD, 2009, 209218. New York, NY, USA.

[5] Y. Choi, and C. Cardie, *Learning with compositional semantics as structural inference for subsentential sentiment analysis*. In Proceedings of the EMNLP, 2008, 793801. Stroudsburg, PA, USA.

[6] X. Ding, and B. Liu, and P. Yu, *A holistic lexicon-based approach to opinion mining*. In Proceedings of the WSDM, 2008, 231240. New York, NY, USA: ACM.

[7] Facebook Graph API. https://developers.facebook.com/docs/reference/api/.

[8] J. Gao, and F. Liang, and W. Fan, and C. Wang, and Y. Sun, and J. Han, *On community outliers and their efficient detection in information networks*. In Proceedings of the 16th SIGKDD, 2010, 813822. New York, NY, USA.

[9] J. Hannon, and M. Bennett, and B. Smyth, *Recommending twitter users to follow using content and collaborative filtering approaches*. In Proceedings of the 4th RecSys, 2010, 199206. New York, NY, USA.

[10] M. Hu, and B. Liu, *Mining and summarizing customer reviews*. In Proceedings of the 10th SIGKDD, 2004, 168177. New York, NY, USA.

[11] X. Jin, and S. Spangler, and Y. Chen, and K. Cai, R. Ma, and L. Zhang, X. Wu, and J. Han, *Patent maintenance recommendation with patent information network model*. In Proceedings of 11th ICDM, 2011, 280289. Washington, DC, USA.

[12] L. Ku, and Y. Liang, and H. Chen, *Opinion extraction, summarization and tracking in news and blog corpora*. In AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, 2006, 100107.

[13] R. Kumar, and A. Tomkins, *A characterization of online browsing behavior*. In Proceedings of the 19th WWW, 2010, 561570. New York, NY, USA.

[14] J. Liu, and S. Seneff, *Review sentiment scoring via a parse-and-paraphrase paradigm*. In Proceedings of the EMNLP, 2009, 161169. Stroudsburg, PA, USA.

[15] R. McDonald, and K. Hannan, and T. Neylon, and M. Wells, and J. Reynar, *Structured models for fine-to-coarse sentiment analysis*. In Proceedings of the 45th ACL, 2007, 432439. Prague, Czech Republic.

[16] Q. Mei, and X. Ling, and M. Wondra, and H. Su, and C. Zhai, *Topic sentiment mixture: modeling facets and opinions in weblogs*. In Proceedings of the 16th WWW, 2007, 171180. New York, NY, USA.

[17] B. Pang, and L. Lee, and S. Vaithyanathan, *Thumbs up?: sentiment classification using machine learning techniques*. In Proceedings of the EMNLP, 2002, 7986. Stroudsburg, PA, USA.

[18] A. Popescu, and O. Etzioni, *Extracting product features and opinions from reviews*. In Proceedings of the EMNLP, 2005, 339346. Stroudsburg, PA, USA.

[19] H. Dai, and F. Zhu, and E. P. LIM, and H. H. PANG, *Detecting Anomalies in Bipartite Graphs with Mutual Dependency Principles*, The 12th ICDM, 2012, Brussels, Belgium.

[20] F. Provost, and B. Dalessandro, and R. Hook, and X. Zhang, and A. Murray, *Audience selection for on-line brand advertising: privacy-friendly social network targeting*. In Proceedings of the 15th SIGKDD, 2009, 707716. New York, NY, USA.

[21] E. Riloff, and J. Wiebe, *Learning extraction patterns for subjective expressions*. In Proceedings of the EMNLP, 2003, 105112. Stroudsburg, PA, USA.

[22] M. Takaffoli, and F. Sangi, and J. Fagnan, and O. Zane, *Community evolution mining in dynamic social networks*. Procedia - Social and Behavioral Sciences 22(0):4958.

[23] P. Turney, *Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews*. In Proceedings of the 40th ACL, 2002, 417424. Stroudsburg, PA, USA.

[24] X. Wu, and F. Xie, and G. Wu, and W. Ding, *Personalized news filtering and summarization on the web*. In Proceedings of the 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence, 2011, 414421. Washington, DC, USA.

[25] K. Zhang, and R. Narayanan, and A. Choudhary, 2010. *Voice of the customers: mining online customer reviews for product feature-based ranking*. WOSN, 2010. Berke-ley, CA, USA: USENIX Association.

[26] K. Zhang, and Y. Chen, and Y. Xie, and D. Honbo, and A. Agrawal, and D. Palsetia, and K. Lee, and W. Liao, and A. Choudhary, *SES: Sentiment elicitation system for social media data*. IEEE 11th workshop on ICDM, 2011.

[27] X. Hu, and L. Tang, and J. Tang, and H. Liu, *Exploiting social relations for sentiment analysis in microblogging*. WSDM'2013. ACM, New York, NY, USA, 537-546.

[28] C. Tan, and L. Lee, and J. Tang, and L. Jiang, and M. Zhou, and P. Li, *User-level sentiment analysis incorporating social networks*. In Proceedings of the Seventeenth ACM SIGKDD, 2011, 1397-1405.

[29] L. Liu, and J. Tang, and J. Han, and M. Jiang, and S. Yang, *Mining topic-level influence in heterogeneous networks*. In Proceedings of the 19th ACM CIKM. 2010, 199–208.

[30] Y. Dong, and J. Tang, and S. Wu, and J. Tian, and N. Chawla, and J. Rao, and H. Cao, *Link Prediction and Recommendation across Heterogenous Social Networks*. In ICDM'12. 2012, 181-190.