Modeling a Retweet Network via an Adaptive Bayesian Approach

Bin Bi
Microsoft Research
One Microsoft Way
Redmond, WA 98052
bibi@microsoft.com

Junghoo Cho
University of California, Los Angeles
405 Hilgard Avenue
Los Angeles, CA 90095
cho@cs.ucla.edu

ABSTRACT

Twitter (and similar microblogging services) has become a central nexus for discussion of the topics of the day. Twitter data contains rich content and structured information on users' topics of interest and behavior patterns. Correctly analyzing and modeling Twitter data enables the prediction of the user behavior and preference in a variety of practical applications, such as tweet recommendation and followee recommendation. Although a number of models have been developed on Twitter data in prior work, most of these only model the tweets from users, while neglecting their valuable retweet information in the data. Models would enhance their predictive power by incorporating users' retweet content as well as their retweet behavior.

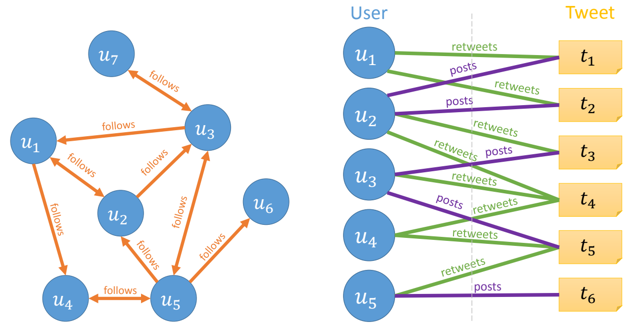
In this paper, we propose two novel Bayesian nonparametric models, URM and UCM, on retweet data. Both of them are able to integrate the analysis of tweet text and users' retweet behavior in the same probabilistic framework. Moreover, they both jointly model users' interest in tweet and retweet. As nonparametric models, URM and UCM can automatically determine the parameters of the models based on input data, avoiding arbitrary parameter settings. Extensive experiments on real-world Twitter data show that both URM and UCM are superior to all the baselines, while UCM further outperforms URM, confirming the appropriateness of our models in retweet modeling.

General Terms

Algorithms, Human Factors, Experimentation

1. INTRODUCTION

Microblogging services like Twitter have become important platforms for Web users to share interesting stories, breaking news, and rich media content. Twitter data has become a valuable resource for user modeling to predict the user behavior and preference in various applications on Twitter, such as tweet recommendation and followee recommendation.



(a) **Follow network**: The graph has one type of nodes (i.e., users) and one type of action edges (i.e., follow links).

(b) **Retweet network**: This is a bipartite graph that has two types of nodes (i.e., users and tweets) and two types of action edges (i.e., posting and retweeting).

Figure 1: Comparison of a follow network and a retweet network

There are different kinds of information contained in Twitter data. Tweet text in the data plays an important role in characterizing users' topical interest in tweeting. A number of models have been developed on tweet text to identify and understand users' personal interest and their tweeting behavior in prior work [27, 4, 14]. Such tweet analysis methods faced a challenge of extremely noisy tweet data [36].

To remedy the noisy data problem, a number of studies complemented noisy tweet text with extra information of users' follow relationship, which is an additional signal reflecting users' topical interest, since a user tends to follow another user with similar interest [19, 11]. There also exists prior work that developed models on both tweet text and the follow network [5, 32, 17]. However, most of the existing work neglected the valuable *retweet* information contained in Twitter data. Missing retweet information in a model can lead to inferior performance, because a large number of Twitter users retweet messages from others much more frequently than they tweet new messages. A user's profile can be enriched by properly incorporating the messages he/she retweeted into his/her own tweets. This strategy remedies the data sparsity problem for the users with few tweets.

On the other hand, the retweet relationship is a cleaner signal than the signal of the follow relationship in terms of specifying users' topical interest. This is because existing models essentially make the assumption that the follow relation from follower A to followee B indicates that A is interested in every single tweet from B, but this assumption

WWW 2016, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4143-1/16/04.
http://dx.doi.org/10.1145/2872427.2882985 .

is actually not valid in reality. Instead, user A follows user B typically because a portion of tweets, but not all, from B interest A. As a result, the performance of the models can be harmed by the invalid assumption.

By contrast, retweet is a relationship between a user and a tweet. Figure 1 visualizes the difference between a follow network and a retweet network. The fact that user A retweets tweet T clearly indicates that A is interested in this particular tweet T. As a result, explicitly utilizing the retweet relationship enables the model to accurately identify users' personal interest, and thus enhances its predictive power. There have been some existing work on retweet analysis, but these studies mostly used retweet information by aggregating retweets in various ways, rather than explicitly modeling each individual retweet relation. Such aggregative methods lose the valuable information of binary retweet relations.

In this paper, we propose two novel hierarchical mixture models, User-Retweet Model (URM) and User-centric Model (UCM), which jointly exploit tweet text and the retweet relationship for user behavior analysis. Both models are able to incorporate the analysis of users' retweet behavior into that of their tweet text. Furthermore, URM and UCM both make use of both tweet and retweet to derive the topics of interest to individual users. To avoid arbitrary parameter settings (e.g., the number of mixture components) in typical mixture models, both URM and UCM are designed not to require the number of topics as an input parameter. Instead, they are nonparametric and fully driven by data in the sense that the best number of topics in the models can be automatically determined based on the characteristics of observed data.

In a nutshell, URM uses a three-layer Dirichlet process hierarchy to jointly model tweet text and the retweet relationship. It characterizes each user and each retweet as a unique mixture model. As another DP hierarchy, on the other hand, UCM differs from URM in that UCM further differentiates the personal interest in tweet and retweet for each user by introducing two random measures for each user's tweet interest and retweet interest, respectively. Our empirical study on real-world Twitter data shows that URM and UCM both significantly outperform all the baselines while UCM further improves over URM.

In particular, the major contributions of our work are summarized as follows:

- We propose two novel Bayesian models, URM and UCM, on retweet data for user behavior analysis on Twitter. Such modeling allows the prediction of the user preference and future behavior in a wide range of practical applications.
- Both URM and UCM are designed to leverage the signals from tweet text and the retweet relationship in Twitter data. Both models explicitly exploit individual retweet relations to identify users' interest. Moreover, the two models enable tight coupling of the analysis of text and the network structure based on the solid probabilistic foundation in Bayesian modeling.
- To avoid manually parameter tuning, URM and UCM are both able to let the retweet data speak for itself by automatically figuring out the best number of topics based on the data characteristics. We also propose Bayesian inference algorithms based on collapsed Gibbs sampling to learn from the data the optimal values of various parameters in the models.
- Through extensive experimentation on a real-world data set collected from Twitter, we demonstrate (a) the substantial better accuracy achieved by both URM and UCM than all the baselines in terms of the quality of distilled topics, model precision and predictive power, (b) the further improvement of UCM over URM, and (c) various interesting insights gained from the experiments.

The rest of this paper is organized as follows. In Section 2, we describe the prior work related to ours. The background and preliminaries of Bayesian nonparametric modeling are given in Section 3. Section 4 introduces our two novel Bayesian nonparametric models, URM and UCM. In Section 5, we present the details of our empirical evaluation and experimental results. Finally, we conclude the paper in Section 6.

2. RELATED WORK

2.1 Microblog data analysis

Many research efforts have been devoted to analyzing and modeling text content for a variety of microblogging applications. Cheong *et al.* [12] discovered trend patterns in tweet data to identify users who contribute towards the discussions on specific trends. A parallel effort has been devoted to studying the emergent topics from tweet text. For instance, TwitterMonitor [22] identifies emerging topics on Twitter in real time and provides meaningful analytics that synthesize an accurate description of each topic.

On Twitter, tweets are accompanied by the follow network. Given the valuable social information contained in the follow graph, a number of studies have been conducted on incorporating the follow network in their models. Topic-Sensitive PageRank (TSPR) [17] was the first attempt to combine analysis of text content and a network structure. In the context of Twitter, Weng *et al.* proposed TwitterRank [32] to find topic-level key influencers on Twitter by leveraging both tweet text and the follow network. In TwitterRank, a set of topics is first produced by LDA [8] on the tweets. Then TwitterRank applies a method similar to TSPR to compute the per-topic influence rank. Unlike TSPR and TwitterRank, the FLDA model [5] was proposed to integrate both content topic discovery and social influence analysis in the same generative model.

Despite utilizing the follow network, all these prior studies neglect the more valuable signal of the retweet network. As discussed above, the retweet network is a social structure which contains a cleaner signal in terms of identifying users' topical interest. Therefore, in this work, we propose new models that jointly exploit tweet text and the retweet network to analyze users' topics of interest and behavior on Twitter.

2.2 Retweet study

There have been many efforts on studying the different aspects of retweeting. The research work on this specific topic can be divided into retweeting behavioral analysis and predicting retweets.

Boyd *et al.* [9] studied some basic issues about retweet behavior: how people retweet, why people retweet and what people retweet. Bild *et al.* [6] instead analyzed aggregate user behavior and the retweet graph with a focus on quantitative descriptions, and found that the retweet graph is small-world and scale-free, like the social graph, but is less disassortative and has much stronger clustering. Yang *et al.* [34] investigated how retweeting behavior was influenced by factors like posting time. In [21], Macskassy and Michelson studied a set of Twitter users and sought to explain the individual information diffusion behavior, as represented by retweets. They found that content-based propagation models could explain the majority of retweet behavior seen in their data. Comarela *et al.* [13] identified factors that influence a users' response or retweet probability, and found that previous response to the user, the user's sending rate, the freshness of information, the length of tweet could affect the user's response.

Moving away from retweet behavioral analysis, researchers have also studied the retweet prediction problem. Hong *et al.* [18] proposed a method to predict the popularity of tweets, and estimated the number of times a tweet would be retweeted. In their work, content features, temporal information, as well as metadata of tweets and tweeters were explored. Zaman *et al.* [35] used a collaborative filtering approach to predict for a pair of users whether a tweet posted by one would be retweeted by the other user. They found that the identity of the source of the tweet and retweeter were the most important features for predicting future retweets. In [33], Yang and Counts studied how to predict the speed, scale, and range of information diffusion by analyzing how tweets on the same topic spreaded. Artzi *et al.* [3] proposed a model for predicting the likelihood of responding which includes retweeting and replying.

Unlike all these prior retweet studies, we aim to design and to evaluate descriptive models that probabilistically express hypotheses about the way in which retweet data may have been generated. By uncovering the hidden structure inherent in the data, we are able to conduct exploratory analysis of the retweet network, and to gain valuable insight into the underlying properties of the retweet data.

2.3 Bayesian nonparametric modeling

Because of the nonparametric nature, our proposed models can automatically figure out the optimal values of the parameters, e.g., the number of topics, based on input data. There exist a number of studies on Bayesian nonparametric modeling for a wide range of practical applications. Orbanz and Teh [24] presented an overview of how Bayesian nonparametric models work for a variety of machine learning problems, and provided a few examples where the models can be employed. [29] reviewed prior research works on the specific topic of hierarchical Bayesian nonparametric modeling, and gave a series of its successful applications ranging from problems in biology to computational vision to natural language processing.

Ahmed and Xing [1] introduced the temporal nonparametric mixture model as a framework for evolutionary clustering. They provided an intuitive construction of this framework using the recurrent Chinese restaurant process (RCRP) metaphor, as well as a Gibbs sampling algorithm to carry out posterior inference in order to determine the optimal cluster evolution. In the context of Twitter modeling, Lim [20] proposed the Twitter-Network (TN) topic model to jointly model the tweet text and the follow network in a Bayesian nonparametric way. The TN model employs the hierarchical Poisson-Dirichlet processes for text modeling and a Gaussian process random function model for follow network modeling. Therefore, TN differs from our models in modeling information of different nature.

3. PRELIMINARIES

The rich content and structural information contained in a retweet network presents an exciting opportunity for statistical modeling. Given the dynamic nature of Twitter data, we choose to build Bayesian nonparametric models, which allow the representation of data to grow structurally as more data are observed. As opposed to a parametric model, it is capable of letting the data speak for itself to automatically determine the complexity of the nonparametric model.

There are two main constituents of our models: DPM and HDP. The Dirichlet Process Mixture (DPM) model [2] is the key building block in Bayesian nonparametric models for a broad range of applications. The DPM model has been extended to Hierarchical Dirichlet Processes (HDP) [30] to cluster grouped data. For the purpose of clarity, in this section, we describe the two components of our models, and explain the notations used throughout the paper.

3.1 Dirichlet Process Mixture

There are three different views on the DPM model: (1) a distribution of a random probability measure, (2) intuitive Chinese Restaurant Process (CRP), and (3) a limit of a finite mixture model. All of these perspectives are equivalent, but each one provides a different view of the same process, and some of them might be easier to follow.

A *Dirichlet process* (DP) is defined as a distribution of a random probability measure G [15]. A DP, denoted by $DP(\lambda, G_0)$, is parameterized by a concentration parameter λ, and a base measure G_0. $G \sim DP(\lambda, G_0)$ denotes a draw of a random probability measure G from the Dirichlet process. G is technically a distribution over a given parameter space θ, so one can draw parameters $\theta_1, \dots, \theta_n$ from G. Previously drawn values of θ_i have strictly positive probability of being redrawn again, which makes the underlying probability measure G discrete [7]. Using a DP at the top of a hierarchical model leads to the Dirichlet Process Mixture model for Bayesian nonparametric modeling [2].

Sampling from the DPM model is conducted by the following generative process:

$$\begin{aligned} G &\sim DP(\lambda, G_0), \\ \theta_i &\sim G, \\ w_i &\sim F(.|\theta_i) \end{aligned} \tag{1}$$

where F is a given likelihood function parameterized by θ. The clustering property of a DP prefers to use fewer than n distinct θ. An equivalent Chinese Restaurant Process metaphor exhibits the clustering property. In particular, consider a Chinese restaurant with an unbounded number of tables. Each θ_i corresponds to a customer who enters the restaurant. The i-th customer θ_i sits at table k that already has n_k customers with probability $\frac{n_k}{i-1+\lambda}$, and shares the dish (parameter) ψ_k served there, or sits at a new table with probability $\frac{\lambda}{i-1+\lambda}$, and orders a new dish sampled from

G_0. This process can be expressed as:

$$\theta_i|\theta_1,\ldots,\theta_{i-1},\lambda,G_0 \sim \sum_{k=1}^{i-1}\frac{n_k}{i-1+\lambda}\delta_{\psi_k}+\frac{\lambda}{i-1+\lambda}G_0. \tag{2}$$

where δ_ψ is a probability measure concentrated at ψ.

Finally, a DPM model can be derived as the limit of a sequence of finite mixture models, where the number of mixture components is taken to infinity. Therefore, a DPM can be used to build an infinite-dimensional mixture model, and has the desirable property of extending the number of clusters with the arrival of new data. This flexibility enables the DPM to conduct model selection automatically.

3.2 Hierarchical Dirichlet Processes

The DPM is widely used to build a model with a discrete random variable of unknown cardinality (i.e., a cluster indicator). The HDP, on the other hand, applies to the problems in which multiple different groups of data would share the same settings of partitions. In such applications, the model for each of the groups incorporates a discrete variable of unknown cardinality. The HDP model is able to share clusters across multiple clustering problems.

The key building block of the HDP model is a recursion where the base measure G_0 for a DP: $G \sim DP(\lambda, G_0)$ is itself a draw from another DP: $G_0 \sim DP(\alpha, H)$. By this recursive construction, the random measure G are constrained to place its atoms at the discrete locations determined by G_0. Such a construction is commonly used for conditionally independent hierarchical models of grouped data.

More formally, in HDP, we model each of the groups as a DP, which is gathered into an indexed collection of DPs $\{G_j\}$. In order to be tied probabilistically, the random measures share their base measure, which is defined to be random as well, as follows:

$$\begin{aligned} G_0 &\sim DP(\alpha, H) \\ G_j &\sim DP(\lambda, G_0). \end{aligned} \tag{3}$$

This means that we first draw G_0 from the base measure H. The random measure G_0 is then, in turn, used as a reference measure to obtain the measures G_j. As a result, each random measure G_j inherits its set of atoms from the same G_0. Therefore, this conditionally independent hierarchical model induces sharing of atoms among these random measures G_j.

Integrating out all random measures, we obtain the equivalent Chinese Restaurant Franchise processes (CRF) [30]. In the CRF, the metaphor of a Chinese restaurant is extended to a set of restaurants which share a set of dishes. The customers in the j-th restaurant sit at tables in the same manner as the CRP, and this is done independently in the restaurants. The coupling among restaurants is achieved via a franchise-wise menu. The first customer to sit at a table in a restaurant chooses a dish from the menu and all subsequent customers who sit at that table inherit that dish. Dishes are chosen with probability proportional to the number of tables (franchise-wide) which have previously served that dish.

HDP is a building block of our proposed models specifically designed for retweet modeling, which will be described later. Also, it serves as a baseline for the evaluation of the quality of models in our experiments.

4. RETWEET MODELING

In this section, we present two different Bayesian nonparametric models on retweet data. Both of them are able to integrate the analysis of tweet text and users' retweet behavior in the same probabilistic framework. Moreover, they both jointly model users' interest in tweet and retweet.

4.1 User-Retweet Model (URM)

Identifying users' interest in tweet and retweet is key for user modeling to predict the user behavior and preference in various applications on Twitter, such as tweet recommendation and followee recommendation. Therefore, a Bayesian model, which properly captures the great diversity of user interests on Twitter, is clearly needed. We refer to the first model as User-Retweet Model (URM).

Twitter has become a central nexus for discussion of the topics of the day. On Twitter, users from all over the world tweet a variety of topics of interest. Naturally, each user has distinct preference and topical interest. To characterize the heterogeneity among all users, we model each user as a unique mixture of a set of topics, where the mixing proportion governs his or her personal interest. In detail, each user possesses a distinct probability distribution over the topics, indicating the probability that he or she is interested in tweeting each individual topic. For example, consider a mini set of two topics: *politics* and *food*. One user may tweet the politics topic with a higher probability than the food topic, while another may be more interested in tweeting food than tweeting politics. Given a set of topics, a Twitter user generates each word in their tweets from one of the topics based on the distribution specific to this topic.

In addition to tweets, retweets convey useful clues about the users' interest and preference. If multiple users retweet a certain message, they are likely to have common topical interest reflected by this message. In order to capture the diversity of topics exhibited by retweets, we further model each retweet as a mixture of a set of topics. Specifically, each retweet is represented as a probability distribution over the topics, quantifying the probability of covering each individual topic.

In a typical mixture model, the number of mixture components is usually manually specified and empirically tuned to determine the granularity of the model. However, given the dynamic nature and large scale of retweet data, it is infeasible to manually exhaust the optimal number of topics in a retweet model. To address this limitation, we resort to a fully data-driven approach, i.e., imposing Dirichlet process priors over the mixture components [15], which allows the number of topics to be automatically determined based on the data characteristics.

4.1.1 Generative Process for URM

The problem of retweet modeling is to specify a probabilistic process by which the observed data, i.e., all the words in tweets, denoted by $\mathbf{w}$, and all the words in retweets, denoted by $\mathbf{x}$, may have been generated. In URM, we assume that in tweeting, to choose a word, a user would first select a topic of interest according to his or her unique topic distribution, from which he or she would then pick a word w based on its generative probability in this selected topic. This stochastic process repeats for every word in the tweets of every user.

On the other hand, unlike a tweet created by one single user, a retweet may be forwarded by multiple users, and thus the retweet should exhibit the topics of interest to these forwarders. Therefore, to generate a word in a retweet, a topic would be first picked based on the topic distributions of all the users who forwarded this retweet. A word x would then be chosen from the word distribution specific to this picked topic.

Let us formally describe the URM model. Let y index each topic exhibited by words in tweets $\mathbf{w}$. As a result, there is a word distribution, denoted by ϕ_y, for each tweet topic y. To avoid manually setting the number of tweet topics, we assume ϕ_y itself to be a random variable drawn from a Dirichlet process. As discussed before, draws from a DP often share common values and thus naturally form clusters. Instead of being pre-specified, the number of clusters, which is often smaller than the total number of draws, varies with respect to data.

As a result, the global probability of generating tweets $p(\mathbf{w})$ is distributed as a DP, which can be expressed with a stick-breaking representation [28]:

$$p(\mathbf{w}) = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}, \qquad (4)$$

where ϕ_k follows the prior H over multinomial distributions: $\phi_k \sim H$; δ_ϕ is a probability measure concentrated at ϕ; and $\beta = (\beta_k)_{k=1}^{\infty} \sim \mathrm{GEM}(\alpha)$ is an infinite sequence defined as:

$$\beta'_k \sim Beta(1, \alpha), \qquad \beta_k = \beta'_k \prod_{l=1}^{k-1} (1 - \beta'_l)$$

The global distribution defined in Equation (4) captures the homogeneity for the tweet behavior of users for the global population, but it does not reflect each individual user's behavior. As stated earlier, to capture the heterogeneity among all users, we characterize each user by a mixture model. These mixture models of all users are linked together via the global distribution defined in Equation (4). Linking these mixture models is significant and useful in that it allows the tweet topics to be shared among all users. For instance, consider a user who is interested in the food topic and the politics topic, and another user who likes the food topic and the technology topic. It would be helpful for a model to relate the food topic discovered in the analysis of the former user to that detected from the latter user.

Specifically, the probability of generating user u's tweets can be written as:

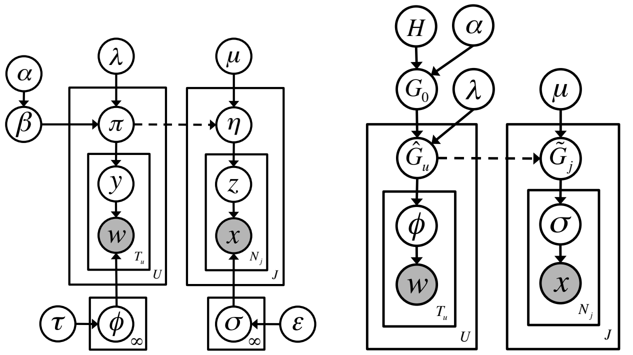
$$p(\mathbf{w}_u) = \sum_{k=1}^{\infty} \pi_{uk} \delta_{\phi_k}, \qquad (5)$$

where the mixing proportion $\pi_u = (\pi_{uk})_{k=1}^{\infty} \sim DP(\lambda, \beta)$. In this way, we introduce another layer of DP for the mixture of tweet topics in each user.

Moreover, as discussed before, each retweet is modeled as a mixture of a set of topics as well. Let z index each topic exhibited by words in retweets $\mathbf{x}$. σ_z denotes the word distribution for retweet topic z. R_j denotes the set of all the users who forwarded the j-th retweet message. The probability of generating the j-th retweet is thus given as:

$$p(\mathbf{x}_j) = \sum_{k=1}^{\infty} \eta_{jk} \delta_{\sigma_k}. \qquad (6)$$

where $\eta_j = (\eta_{jk})_{k=1}^{\infty} \sim DP(\mu, \frac{1}{|R_j|} \sum_{u \in R_j} \pi_u)$. As a result, the generation of a retweet is attributable to the topics of interest to all of its forwarders. The stick-breaking representation of the URM model is depicted in Figure 2(a).



(a) Stick-breaking representation of URM

(b) URM as a DP hierarchy

Figure 2: Graphical models for URM

4.1.2 URM as a Three-layer DP Hierarchy

In a way, URM generalizes HDP by using a three-layer Dirichlet process hierarchy for retweet modeling. The URM model defines a set of random probability measures in each layer of the DP hierarchy. In particular, first we draw a global probability measure G_0 from a DP with base measure H and concentration parameter α influencing the sparsity of the global topic distribution:

$$G_0 \sim DP(\alpha, H). \qquad (7)$$

To characterize personal topical interest in tweeting, we then draw a topic distribution $\hat{G}_u$ from the global probability measure over the topic space G_0 for each user:

$$\hat{G}_u \sim DP(\lambda, G_0) \qquad (8)$$

with concentration parameter λ.

To model the generation of retweets, since a single message can be retweeted by multiple users, for each tweet we draw a probability measure $\widetilde{G}_j$ from a set of multiple topic probability measures, $\{\hat{G}_u | u \in R_j\}$, corresponding to all the forwarders of this tweet, R_j.

Here we introduce a novel notion of drawing a probability measure from a set of probability measures. An equivalent representation of the set of probability measures $\{\hat{G}_u | u \in R_j\}$ is given by a DP with base measure $\frac{1}{|R_j|} \sum_{u \in R_j} \hat{G}_u$ which averages the probability measures in this set. We show that a DP with an average of multiple probability measures as its base measure is equivalent to a standard DP in the following. Suppose $\sigma_1, \ldots, \sigma_{i-1}$ are observed samples from $\widetilde{G}_j$. The probability of the i-th draw σ_i to be sampled from $\widetilde{G}_j$ can then be given by integrating out $\widetilde{G}_j$ using the properties of the Dirichlet distributed partitions [23] and replacing the base measure with the average of multiple probability measures:

$$\sigma_i | \sigma_1, \ldots, \sigma_{i-1}, \mu, \widetilde{G}_j \sim \frac{1}{i-1+\mu} \sum_{k=1}^{i-1} \delta_{\sigma_k} + \frac{\mu}{|R_j|(i-1+\mu)} \sum_{u \in R_j} \hat{G}_u, \qquad (9)$$

which gives a standard Dirichlet process. The URM model as a DP hierarchy is illustrated in Figure 2(b).

4.1.3 Bayesian Inference for URM

To estimate the latent topic structures in URM, we perform posterior inference to "invert" the generative process described above. In particular, we develop an efficient Markov chain Monte Carlo (MCMC) algorithm [26], or more precisely a Gibbs sampler, to approximate the posterior for URM. In a Gibbs sampler, each latent variable is iteratively sampled conditioned on the observations and all the other latent variables, so the key to Gibbs sampling is to derive a full conditional distribution for each latent variable, which is given in the following.

Sampling y:
Let w_{ua} denote the a-th word in user u's tweets. Given the current values of the remainder of the variables, denoted by $\bullet$, the probability of word w_{ua} assigned to an existing topic k can be derived as:

$$p(y_{ua} = k|\bullet) \propto (c_{uk}^{-(ua)} + \lambda\beta_k)\frac{e_{kw_{ua}}^{-(ua)} + \tau_{w_{ua}}}{e_{k*}^{-(ua)} + \tau_*}, \tag{10}$$

whereas the probability that the topic assignment y_{ua} takes on a new value k_{new} is given by:

$$p(y_{ua} = k_{\text{new}}|\bullet) \propto \frac{\lambda\beta_{k_{\text{new}}}}{V}, \tag{11}$$

where $c_{uk}^{-(ua)}$ denotes the number of words in user u's tweets assigned to topic k, excluding the current assignment y_{ua}. $e_{kw}^{-(ua)}$ denotes the number of times word w is assigned to topic k across all tweets, excluding the current assignment. V is the total number of unique words in the vocabulary.

During the sampling process, if a topic assignment takes on a new value k_{new}, we include this new topic $\phi_{k_{\text{new}}}$ into the set of tweet topics, for which we draw a new global proportion $\beta_{k_{\text{new}}}$. On the other hand, if, as a result of updating topic assignments, none of words is assigned to some topic, we delete this unallocated topic from the set of tweet topics, and update the global proportions β accordingly.

Sampling z:
Gibbs sampling for retweet topics $\mathbf{z}$ is similar to that for tweet topics $\mathbf{y}$. Let x_{jb} denote the b-th word in the j-th retweet. The probability of word x_{jb} assigned to a previously used topic k can then be given by:

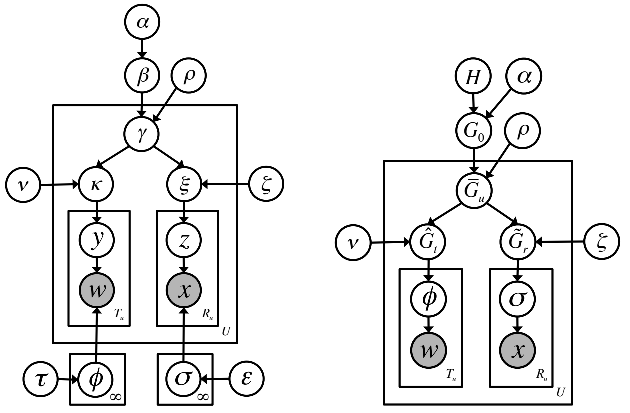
$$p(z_{jb} = k|\bullet) \propto (d_{jk}^{-(jb)} + \sum_{u \in R_j} \mu\pi_{uk})\frac{g_{kx_{jb}}^{-(jb)} + \epsilon_{x_{jb}}}{g_{k*}^{-(jb)} + \epsilon_*}, \tag{12}$$

while the probability that the topic assignment z_{jb} takes on a new value k_{new} is as follows:

$$p(z_{jb} = k_{\text{new}}|\bullet) \propto \frac{\sum_{u \in R_j} \mu\pi_{uk_{\text{new}}}}{V}, \tag{13}$$

where $d_{jk}^{-(jb)}$ denotes the number of words in the j-th retweet assigned to topic k, excluding the current assignment z_{jb}. $g_{kx}^{-(jb)}$ denotes the number of times word x is assigned to topic k across all retweets, excluding the current assignment. R_j denotes the set of all the users who forwarded the j-th retweet.

Sampling β:
Following the simulation of new tables in the CRF introduced in [25], the prior global proportions β can be sampled by simulating how new topics are created for c_{uk} draws from the DP with precision $\lambda\beta_k$ (dishes in the CRF), which is a sequence of Bernoulli trials for each u and k:



(a) Stick-breaking representation of UCM

(b) UCM as a DP hierarchy

Figure 3: Graphical models for UCM

$$p(m_{ukr} = 1) = \frac{\lambda\beta_k}{\lambda\beta_k + r - 1} \qquad \forall r \in [1, c_{uk}]. \tag{14}$$

A posterior sample of β is then obtained by:

$$\beta \sim \text{Dirichlet}(m_1, \ldots, m_K, \alpha), \tag{15}$$

where $m_k = \sum_u \sum_r m_{ukr}$, and K is the number of active topics with which there exist words associated. β has dimension $K+1$ because the mass for α in the Dirichlet distribution corresponds to generating a new topic out of an infinite set of empty topics. If a topic has lost all its words, it is merged with the unknown topics in the mass associated with α. Iterative sampling based on Equations (14) and (15) gives the posterior samples of β, which are needed by sampling tweet topics $\mathbf{y}$.

Sampling π:
Similarly, Equations (12) and (13) for sampling retweet topics $\mathbf{z}$ require the posterior samples of π. The posterior proportion π_u for user u is given by:

$$\pi_u \sim \text{Dirichlet}(n_{1u}, \ldots, n_{Ku}, \lambda), \tag{16}$$

where $n_{ku} = \sum_j \sum_r n_{jkur}$. The auxiliary Bernoulli variable n_{jkur} for retweet j, topic k and user u is defined as:

$$p(n_{jkur} = 1) = \frac{\mu\pi_{uk}}{\mu\pi_{uk} + r - 1} \quad \forall r \in [1, d_{jk}]. \tag{17}$$

All the above posterior distributions create a Markov chain for Gibbs sampling. The Gibbs sampler for URM iteratively samples $\mathbf{y}$, $\mathbf{z}$, β, and π as described above in turn.

4.2 User-centric Model (UCM)

4.2.1 Generative Process for UCM

The User-Retweet Model characterizes each user and each retweet as a unique mixture model. In other words, it constructs a separate mixture model for each retweet in addition to user modeling. Given that users' behavior of both tweet and retweet reflects their distinct preference and topical interest, an alternative to user modeling would be introducing a random measure specific to each user that captures his or her unique interest. There often exist differences between the tweet interest and the retweet interest of a user. For instance, a user may be interested in retweeting jokes, but he or she could never tweet anything joking. To differentiate a user's interest in tweet and retweet, we should introduce two random measures which capture his or her tweet interest and retweet interest, respectively. This alternative model is referred to as User-centric Model (UCM).

Formally, in the UCM model, we introduce a probability measure $\bar{G}_u$ specific to any user u, which is distributed as a DP:

$$\bar{G}_u \sim DP(\rho, G_0), \tag{18}$$

where $G_0 \sim DP(\alpha, H)$. Equation (18) can be represented with a stick-breaking process as:

$$\bar{G}_u = \sum_{k=1}^{\infty} \gamma_{uk} \delta_{\phi_k}, \tag{19}$$

where $\gamma_u = (\gamma_{uk})_{k=1}^{\infty} \sim DP(\rho, \beta)$. The mixing proportion γ_u quantifies the user u's common interest in each different topic, which reflects the homogeneity of u's behavior of tweet and retweet. To separate the modeling of tweet interest and that of retweet interest, we draw from $\bar{G}_u$ a probability measure $\hat{G}_t$ for tweet generation and a probability measure $\tilde{G}_r$ for retweet generation:

$$\begin{aligned} \hat{G}_t &\sim DP(\nu, \bar{G}_u), & (20)\\ \tilde{G}_r &\sim DP(\zeta, \bar{G}_u). & (21) \end{aligned}$$

Using the stick-breaking representation, Equations (20) and (21) can be expressed as:

$$\begin{aligned} \hat{G}_t &= \sum_{k=1}^{\infty} \kappa_{uk} \delta_{\phi_k}, & (22)\\ \tilde{G}_r &= \sum_{k=1}^{\infty} \xi_{uk} \delta_{\sigma_k}, & (23) \end{aligned}$$

where $\kappa_{uk} = (\kappa_{uk})_{k=1}^{\infty} \sim DP(\nu, \gamma)$, which measures the user u's topical interest in tweet, and $\xi_{uk} = (\xi_{uk})_{k=1}^{\infty} \sim DP(\zeta, \gamma)$, which quantifies u's retweet interest over the topics. The stick-breaking representation of UCM is illustrated in Figure 3(a). Figure 3(b) depicts the graphical model for UCM as a DP hierarchy.

4.2.2 Bayesian Inference for UCM

We develop a Gibbs sampler specifically for Bayesian inference for UCM, which is similar to the sampler for URM. In this section, we describe the posterior distributions for topic assignments $\mathbf{y}$ and $\mathbf{z}$, conditioned on the values of all the other variables.

Sampling $\mathbf{y}$:
The Gibbs sampling equation for topic assignment y_{ua} of the a-th word in user u's tweets is:

$$p(y_{ua} = k|\bullet) \propto (c_{uk}^{-(ua)} + \nu\gamma_{uk}) \frac{e_{kw_{ua}}^{-(ua)} + \tau_{w_{ua}}}{e_{k*}^{-(ua)} + \tau_*}, \tag{24}$$

whereas a new value k_{new} is sampled for y_{ua} based on the following probability:

$$p(y_{ua} = k_{\text{new}}|\bullet) \propto \frac{\nu\gamma_{uk_{\text{new}}}}{V}, \tag{25}$$

where $c_{uk}^{-(ua)}$ denotes the number of words in user u's tweets assigned to topic k, excluding the current assignment y_{ua}, and $e_{kw}^{-(ua)}$ denotes the number of times word w is assigned to topic k across all tweets, excluding the current assignment.

Sampling $\mathbf{z}$:
For the b-th word in user u's retweets, a previously seen topic k is sampled from the distribution given by:

$$p(z_{ub} = k|\bullet) \propto (f_{uk}^{-(ub)} + \zeta\gamma_{uk}) \frac{g_{kx_{ub}}^{-(ub)} + \epsilon_{x_{ub}}}{g_{k*}^{-(ub)} + \epsilon_*}, \tag{26}$$

whereas the probability that the topic assignment z_{ub} takes on a new value k_{new} is:

$$p(z_{ub} = k_{\text{new}}|\bullet) \propto \frac{\zeta\gamma_{uk_{\text{new}}}}{V} \tag{27}$$

where $f_{uk}^{-(ub)}$ denotes the number of words in user u's retweets assigned to topic k, excluding the current assignment z_{ub}, and $g_{kx}^{-(ub)}$ denotes the number of times word x is assigned to topic k across all retweets, excluding the current assignment.

5. EMPIRICAL EVALUATION

To evaluate the quality of our proposed models, URM and UCM, we conducted experiments on a real-world dataset crawled from Twitter. First, we demonstrate the latent topics discovered by both models, which qualitatively reflect the effectiveness of the models. Then, we quantitatively measure the quality of the topics discovered by our proposed models and baselines. Finally, we assess and compare the predictive power and generalizability of these models to objectively evaluate their effectiveness.

5.1 Dataset and Experiment Settings

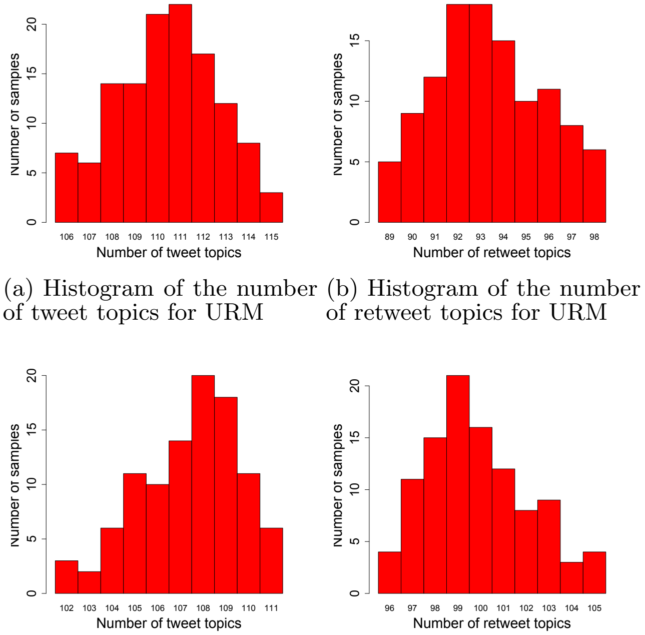
Our experiments were conducted on a Twitter dataset collected between October 2009 and January 2010. We crawled the dataset based on the follow network in a breadth-first search manner. The crawler began with an initial seed set of the top 1000 users in `twitterholic.com`, and traversed the follow links in a forward direction. Users' tweet content and retweet activities were collected during the crawling process. The dataset includes over 1.1 million Twitter users, with more than 273 million follow links and over 2.9 million retweet links. An in-depth analysis of this dataset can be found in [31].

We used the tokenizer from the TweetNLP project [16] in order to improve the accuracy of the recognized terms in the noisy text. Furthermore, we reduced the inherent noise of tweets, by removing terms that appear in less than 20 tweets.

URM and UCM require a set of hyper-parameters to be determined a priori. In our experiments, we set the hyper-parameters: $\alpha = 1, \lambda = 0.5, \mu = 0.5, \tau = 0.1, \epsilon = 0.1, \rho = 0.5, \nu = 0.5, \zeta = 0.5$. We ran the Gibbs sampling algorithms for 1000 iterations. Since the nearly samples from a Markov chain are usually correlated with each other, we only kept the samples from every 5 iterations to collect independent samples. Moreover, we discarded the samples in the burn-in period (the first 20% of samples).

5.2 Topics Produced by URM and UCM

Given URM and UCM as Bayesian nonparametric models, both models are able to automatically determine the optimal number of latent topics based on the data. To estimate the posterior over the number of topics, during the



(a) Histogram of the number of tweet topics for URM

(b) Histogram of the number of retweet topics for URM

(c) Histogram of the number of tweet topics for UCM

(d) Histogram of the number of retweet topics for UCM

Figure 4: Histogram of the number of latent topics produced during the Gibbs sampling process

Gibbs sampling process we collected posterior samples after the Markov chain had converged. The plots in Figure 4 depict the histograms of the number of tweet/retweet topics produced by URM and UCM. From the histograms, it is seen that both models discovered 100 $\sim$ 120 topics from tweets/retweets. Since the uncovered latent topics reflect the effectiveness of URM and UCM, and provide insights about users' interest on Twitter, we will illustrate a sample of distilled latent topics later in this section.

A latent topic can be represented as a distribution over a fixed set of words in the vocabulary. For a tweet topic k, the posterior distribution of words can be calculated as:

$$\phi_{kw} = p(w|y = k) = \frac{e_{kw} + \tau_w}{\sum_{w=1}^{V}(e_{kw} + \tau_w)}, \tag{28}$$

where the counter e_{kw} gives the number of times word w is assigned to topic k across all tweets. Similarly, the posterior distribution of words for a retweet topic k can be computed as:

$$\sigma_{kx} = p(x|z = k) = \frac{g_{kx} + \epsilon_x}{\sum_{x=1}^{V}(g_{kx} + \epsilon_x)}, \tag{29}$$

where the counter g_{kx} gives the number of times word x is assigned to topic k across all retweets. Since the number of latent topics might vary during the Gibbs sampling process, we collected samples when the Markov chain had converged to a stationary distribution.

Table 1 shows a sample of latent topics produced by URM and UCM in some run of Gibbs sampling. Every topic is represented by the set of top five most probable words under this topic. Intuitively, it is clear that both models distilled meaningful topics from tweets and retweets. For example, the first row in this table, which lists words *music*, *album*, *band*, *play*, and *show*, indicates a music-related topic, and the topic given in the first row for UCM, which is represented by *god*, *jesus*, *lord*, *church*, and *his*, is clearly relevant to Christianity. Naturally, such anecdotal evidence is very hard to generalize. In the next section, we will present a quantitative measure to evaluate the quality of the distilled topics.

Table 1: A sample of latent topics produced by URM and UCM

Model	Topic	Top-5 words
URM	Tweet	music, album, band, play, show
		love, kids, mom, fun, baby
		real, estate, property, read, home
	Retweet	travel, hotel, flight, new, italy
		social, media, twitter, facebook, marketing
		book, read, amazon, writing, author
UCM	Tweet	god, jesus, lord, church, his
		video, music, live, album, show
		green, car, energy, hybrid, carbon
	Retweet	google, iphone, apple, ipad, app
		film, movie, avatar, tv, trailer
		bowl, super, nfl, football, sports

5.3 Topic Quality

We followed the *word intrusion* approach introduced in [10] to quantify the topic quality. Eight human experts participated in our word intrusion task. To evaluate the quality of a topic, the human experts were presented with six randomly order words, which consisted of the five words with the highest probability under the topic and a word from another topic from the same model. The human experts were then asked to find the word which was out of place or did not belong with the others, i.e., the *intruder*. In case of semantically coherent topic words, the intruder should be easily found. To further test the interaction between latent topics, the intruder was chosen from a set of words which had a low probability (out of the top 25 words) in the evaluated topic and a high probability (top 5 of the remaining words) in another topic. For each model, every human expert judged an average of 106 instances.

Let j_k^m denote the index of the intruder among the words generated from topic k distilled by model m. Further let i_{ks}^m denote the intruder selected by human expert s on the set of words generated from topic k distilled by model m, and let S denote the number of human experts ($S = 8$ in our experiments). According to [10], the model precision on topic k is defined by the fraction of human experts that agree with the model on the topic:

$$\text{MP}_k^m = \sum_{s=1}^{S} \mathbb{1}(i_{ks}^m = j_k^m)/S. \tag{30}$$

The precision of model m computes the average of MP_k^m over all K topics: $\text{MP}^m = \sum_{k=1}^{K} \text{MP}_k^m / K$.

We compared the results of URM and UCM with those of Hierarchical Dirichlet Processes (HDP), which is a different Bayesian nonparametric model. In HDP, the words of each user are generated from a unique probability measure, which is drawn from a DP. The probability measures for all users share the same base measure, which is a draw from another DP. More details of HDP can be found in Section 3. In our experiments, we built three independent HDPs as baselines based on different pieces of the data. One of the

Table 2: Comparison of model precisions

HDP-t	HDP-r	HDP-tr	URM	UCM
0.643	0.628	0.654	0.688	0.751

Table 3: Model precisions of URM and UCM over tweet/retweet topics

Topic	URM	UCM
Tweet topic	0.718	0.769
Retweet topic	0.640	0.731

HDPs, which we refer to as HDP-t, was run on the top of tweet text, while neglecting the information of the retweet structure. In other words, HDP-t considers the words in each users' tweets only to be generated from a user-specific probability measure. In contrast, another HDP, referred to as HDP-r, did the opposite by running on words in users' retweets without taking their tweets into account. The last HDP, which we refer to as HDP-tr, integrated information of tweets and retweets by aggregating the words from both tweets and retweets of each user, which were considered to be generated from a user-specific probability measure.

We computed overall model precisions for the three baselines HDP-t, HDP-r and HDP-tr, as well as our models URM and UCM. As shown in Table 2, HDP-tr performed better than both HDP-t and HDP-r, suggesting that integrating the content of tweets and retweets in a model produces higher-quality topics than separate modeling of tweets and retweets. Our models URM and UCM outperformed all the three baselines, which clearly demonstrates the capability of the proposed models to distill high-quality latent topics. Specifically, UCM gave a higher model precision than URM. To track the cause of the performance difference, we computed model precisions of URM and UCM over tweet topics and retweet topics separately. From Table 3, it is observed that UCM is superior to URM in terms of quality of both tweet topics and retweet topics. Moreover, UCM gave a much higher model precision than URM over retweet topics, which implies that it should be more appropriate to have one $\tilde{G}_r$ for each user than having one $\tilde{G}_j$ for each retweet, since the user-specific $\tilde{G}_r$ should have sufficient content from the user to characterize his or her retweet interest. The difference in modeling the retweet structure also improves the tweet topic quality of UCM over that of URM.

5.4 Predictive Power Analysis

As generative models, URM, UCM and HDP are all able to generate and predict unseen new data. We evaluated the predictive power and generalizability of these models using the standard *perplexity* metric [8]. The perplexity is monotonically decreasing in the likelihood of the held-out test data. Hence, a lower perplexity score indicates stronger predictive power. Formally, the perplexity is defined as:

$$perplexity(D_{\text{test}}) = \exp\left\{-\frac{\sum_{u\in D_{\text{test}}}\log p(\mathbf{w}_u)}{\sum_{u\in D_{\text{test}}}|\mathbf{w}_u|}\right\}, \quad (31)$$

where D_{test} denotes the test set of all Twitter users' words in tweets/retweets. To calculate the word perplexity, we held out 20% of the data D_{test} for test purposes and trained the models on the remaining 80%.

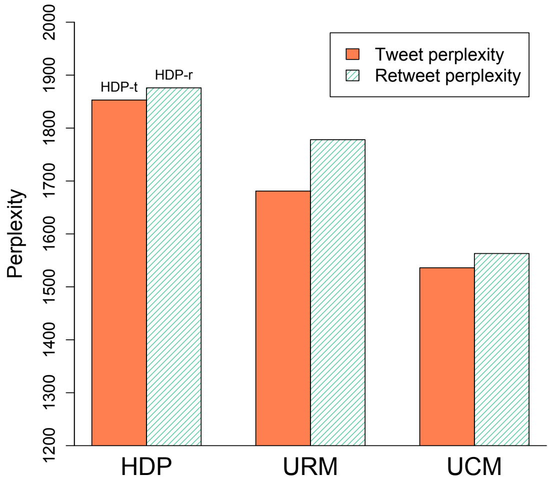


Figure 5: Comparison of word perplexity for HDP, URM and UCM (lower is better)

Figure 5 compares the word perplexity for HDP, URM and UCM. For our models URM and UCM, we calculated perplexity on the words in tweets as well as perplexity on the words in retweets. Since HDP-t and HDP-r applied to tweets and retweets, respectively, we calculated perplexity for HDP-t on the words in tweets and perplexity for HDP-r on the words in retweets. From this figure, we see that UCM gave the lowest perplexity on both tweets and retweets, confirming its strongest predictive power and the best generalizability. Although URM is inferior to UCM, the URM model outperformed the two HDP models in generating and predicting the words in both tweets and retweets. We also calculated overall perplexity on both tweets and retweets for HDP-tr, URM and UCM. As a result, UCM gave the lowest overall perplexity of 1540.6. The second-best URM had overall perplexity of 1723.1, which outperformed HDP-tr with overall perplexity of 1779.3. The experimental results are consistent with the results of the evaluation of topic quality. It validates the hypothesis that proper modeling of the retweet structure enhances the effectiveness of the model.

6. CONCLUSION

This paper presents two novel Bayesian nonparametric models, URM and UCM, for user behavior analysis on Twitter. The two models are able to leverage the signals from both tweet text and the retweet relationship. Furthermore, both models enable tight coupling of the analysis of text and the retweet network in the same Bayesian framework. As nonparametric models, URM and UCM can automatically figure out the optimal values of their parameters based on input data.

In particular, both URM and UCM have a probability measure G_u specific to each user, which characterizes his or her unique topical interest. Individual users' tweets are generated by drawing from the user-specific G_u. The difference between URM and UCM lies in modeling of the retweet structure. URM has one $\hat{G}_j$ for each retweet, while UCM has one $\hat{G}_r$ for each user. We conducted thorough experiments on real-world Twitter data to compare URM, UCM and the baselines. Experimental results show that both URM and UCM significantly outperform all the baselines in terms of the quality of distilled topics, model precision, and predictive power. We also demonstrate the further improvement of UCM over URM, due to UCM's more appropriate modeling of the retweet structure.

7. REFERENCES

[1] A. Ahmed and E. P. Xing. Dynamic non-parametric mixture models and the recurrent chinese restaurant process: with applications to evolutionary clustering. In *Proceedings of the SIAM International Conference on Data Mining, SDM 2008, April 24-26, 2008, Atlanta, Georgia, USA*, pages 219–230, 2008.

[2] C. E. Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 11 1974.

[3] Y. Artzi, P. Pantel, and M. Gamon. Predicting responses to microblog posts. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 602–606, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[4] E. Baralis, T. Cerquitelli, S. Chiusano, L. Grimaudo, and X. Xiao. Analysis of twitter data using a multiple-level clustering strategy. In A. Cuzzocrea and S. Maabout, editors, *MEDI*, volume 8216 of *Lecture Notes in Computer Science*, pages 13–24. Springer, 2013.

[5] B. Bi, Y. Tian, Y. Sismanis, A. Balmin, and J. Cho. Scalable topic-specific influence analysis on microblogs. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 513–522, New York, NY, USA, 2014. ACM.

[6] D. R. Bild, Y. Liu, R. P. Dick, Z. M. Mao, and D. S. Wallach. Aggregate characterization of user behavior in twitter and analysis of the retweet graph. *ACM Trans. Internet Technol.*, 15(1):4:1–4:24, Mar. 2015.

[7] D. Blackwell and J. B. MacQueen. Ferguson distributions via polya urn schemes. *The Annals of Statistics*, 1(2):353–355, 03 1973.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, Mar. 2003.

[9] D. Boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences*, HICSS '10, pages 1–10, Washington, DC, USA, 2010. IEEE Computer Society.

[10] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-graber, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 288–296. Curran Associates, Inc., 2009.

[11] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 199–208, New York, NY, USA, 2009. ACM.

[12] M. Cheong and V. Lee. Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base. In *Proceedings of the 2Nd ACM Workshop on Social Web Search and Mining*, SWSM '09, pages 1–8, New York, NY, USA, 2009. ACM.

[13] G. Comarela, M. Crovella, V. Almeida, and F. Benevenuto. Understanding factors that affect response rates in twitter. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, HT '12, pages 123–132, New York, NY, USA, 2012. ACM.

[14] Z. Dai, A. Sun, and X.-Y. Liu. Crest: Cluster-based representation enrichment for short text classification. In J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu, editors, *PAKDD (2)*, volume 7819 of *Lecture Notes in Computer Science*, pages 256–267. Springer, 2013.

[15] T. S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[16] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[17] T. H. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Trans. on Knowl. and Data Eng.*, 15(4):784–796, July 2003.

[18] L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in twitter. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 57–58, New York, NY, USA, 2011. ACM.

[19] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 137–146, New York, NY, USA, 2003. ACM.

[20] K. W. Lim, C. Chen, and W. Buntine. Twitter-Network topic model: A full bayesian treatment for social network and text modeling. In *NIPS2013 Topic Model workshop*, page 4, Australia, Dec 2013.

[21] S. A. Macskassy and M. Michelson. Why do people retweet? anti-homophily wins the day! In L. A. Adamic, R. A. Baeza-Yates, and S. Counts, editors, *ICWSM*. The AAAI Press, 2011.

[22] M. Mathioudakis and N. Koudas. Twittermonitor: Trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, pages 1155–1158, New York, NY, USA, 2010. ACM.

[23] R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of*

Computational and Graphical Statistics, 9(2):249–265, 2000.

[24] P. Orbanz and Y. W. Teh. Bayesian nonparametric models. In *Encyclopedia of Machine Learning*. Springer, 2010.

[25] I. Porteous. *Networks of Mixture Blocks for Non Parametric Bayesian Models with Applications*. PhD thesis, Long Beach, CA, USA, 2010. AAI3403449.

[26] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[27] K. D. Rosa, R. Shah, B. Lin, A. Gershman, and R. Frederking. Topical Clustering of Tweets. *Proceedings of the ACM SIGIR: SWSM*, 2011.

[28] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[29] Y. W. Teh and M. I. Jordan. Hierarchical Bayesian nonparametric models with applications. In N. Hjort, C. Holmes, P. Müller, and S. Walker, editors, *Bayesian Nonparametrics: Principles and Practice*. Cambridge University Press, 2010.

[30] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):pp. 1566–1581, 2006.

[31] M. J. Welch, U. Schonfeld, D. He, and J. Cho. Topical semantics of twitter links. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 327–336, New York, NY, USA, 2011. ACM.

[32] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 261–270, New York, NY, USA, 2010. ACM.

[33] J. Yang and S. Counts. Predicting the speed, scale, and range of information diffusion in Twitter. In *4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, May 2010.

[34] Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, and Z. Su. Understanding retweeting behaviors in social networks. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1633–1636, New York, NY, USA, 2010. ACM.

[35] T. R. Zaman, R. Herbrich, J. V. Gael, and D. Stern. Predicting information spreading in twitter. In *Computational Social Science and the Wisdom of Crowds Workshop (colocated with NIPS 2010)*, December 2010.

[36] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pages 338–349, Berlin, Heidelberg, 2011. Springer-Verlag.