Hierarchical Dirichlet Processes with Social Influence

Jin Qian, Yeyun Gong, Qi Zhang, Xuanjing Huang

Fudan University, Shanghai, China
{12110240030, yygong12, qz, xjhuang}@fudan.edu.cn

Abstract. The hierarchical Dirichlet process model has been successfully used for extracting the topical or semantic content of documents and other kinds of sparse count data. Along with the growth of social media, there have been simultaneous increases in the amounts of textual information and social structural information. To incorporate the information contained in these structures, in this paper, we propose a novel non-parametric model, social hierarchical Dirichlet process (sHDP), to solve the problem. We assume that the topic distributions of documents are similar to each other if their authors have relations in social networks. The proposed method is extended from the hierarchical Dirichlet process model. We evaluate the utility of our method by applying it to three data sets: papers from NIPS proceedings, a subset of articles from Cora, and microblogs with social network. Experimental results demonstrate that the proposed method can achieve better performance than state-of-the-art methods in all three data sets.

1 Introduction

Probabilistic topic models have demonstrated their effectiveness in analyzing sparse high-dimensional count data, including recent innovations such as probabilistic latent semantic analysis (PLSA) [8], latent Dirichlet allocation (LDA) [4], hierarchical Dirichlet processes (HDP) [17], and so on. Because of the ability of nonparametric Bayesian methods to handle an unbounded number of topics, among existing techniques, these nonparametric Bayesian methods have received more and more attention and have found broad applications, such as retrieval [6], image processing [15], topic detection and traction [18, 11], and so on.

With the dramatic increase in Web 2.0 applications, the textual information and social structural information have simultaneously grown. For example, from conference proceedings or online journal articles, we can obtain not only the content of articles, but also co-authorship networks of authors and citation networks. In Twitter-like services, except for the microblogs published by the users, the following and retweet relations also evolve social networks among users. In addition to these examples, we can easily find many other data collections with network structures attached, including emails, blogs, and forums. However, most of the current nonparametric Bayesian models usually take only the textual information into consideration. Hence, much more attention should be given to the development of methods to take advantage of network structures to advance the effectiveness in analyzing sparse count data.

Several works have studied the problem from different aspects. Mei et al. [13] proposed the use of a discrete regularization framework to extend the PLSA and LDA to achieve the task. Topic-link LDA model [10] tries to simultaneously perform topic modeling and community detection in a framework extended from LDA. Jie et al. [16] transferred the social influence problem into a topical factor graph model and proposed a topical affinity propagation on the factor graph to identify the topic-specific social influence. A relational topic model [5] was developed based on LDA and incorporated the links between documents as binary random variables. However, due to the unbounded number of topics, nonparametric Bayesian methods cannot be directly incorporated into these frameworks. There are a number of extensions of nonparametric Bayesian methods from different aspects, including temporal information [14, 7, 19], shared characteristics [9], and time or space distance dependent Chinese restaurant process (ddCRP) [3].

The ddCRP clusters data in a biased way: each data point is more likely to be clustered with other data that are near it in an external sense. For example we can use ddCRP to model the topics distribution of the documents. In the ddCRP, the topics distribution of document is sampled from the connected documents depend on the distance. However, our work reconstructed the topics distribution integrated other documents depend on the social structural information of authors.

In this paper, we propose a novel non-parametric model, social hierarchical Dirichlet process (sHDP), to take both textual and the social structural information of authors into consideration for modeling topics distribution. The topics distribution of document in sHDP is reconstructed based on the social structural information of authors. The work is motivated by the observation that if two authors have close relation in social networks, they may have similar interests and would talk about similar topics. Hence, the topic distributions of documents posted by them should have a high chance to be similar. Based on the assumption, we extend HDP model by incorporating the influence of social structure. Different from HDP, which models the dependence among groups through sharing the same set of discrete parameters, the proposed method sHDP is built on top of the social Chinese restaurant franchise (sCRF) process, which has the feature that mixture weights associated with parameters are different and influenced by social structure for different groups. To demonstrate the effectiveness of the proposed method, we use three data sets to evaluate the proposed method and compare the results with those of state-of-the-art methods. The experimental results demonstrate that the proposed method can achieve significantly better performance than previous methods with or without taking the structure information into consideration.

To summarize, the contributions of this paper are:

- We propose a novel non-parametric model, which involves both textual and structural information.
- We detail the method through the social Chinese restaurant franchise process and describe a Gibbs sampling algorithm for posterior inference.
- Experimental results show that the proposed method achieves better performance in three different kinds of data sets.

2 Social Hierarchical Dirichlet Processes

In this work, we aim to model the topic distributions for the data sets consist of both text documents (S) and associated network structure ($\mathfrak{G}$). Text documents can be web pages, microblogs, papers, and so on. The network structure can be social network, linking graph, co-author/citation graph, and so on. Let d_i and d_j to represent the ith and jth document in S respectively. $f_{d_i d_j}$ denotes the social influence between the two documents and can be calculated based on the network structure $\mathfrak{G}$. It measures the degree of how these two documents have the same topic.

2.1 The Preliminaries

A Dirichlet process is a random process, that is a probability distribution over distributions, parameterized by a scaling parameter γ and a base probability measure H. We denote it by $G_0 \sim DP(H, \gamma)$.

A perspective on the Dirichlet process is provided by the *Pólya urn scheme* [2]. A sequence of variables $\theta_1, \theta_2, ...$ are independent and identically distributed according to G_0. The Pólya urn representation of θ results from integrating out G_0 is as follows:

$$\begin{aligned} &\theta_i | \theta_1, ..., \theta_{i-1}, \gamma, H \sim \\ &\qquad \sum_{l=1}^{i-1} \frac{i}{i-1+\gamma} \delta_{\theta_l} + \frac{\gamma}{i-1+\gamma} H. \end{aligned} \tag{1}$$

Let $\phi_1, ..., \phi_K$ be the distinct values taken on by $\theta_1, ..., \theta_{i-1}$, and m_k be the number of values $\theta_{i'} = \phi_k$ for $1 \leq i' < i$. Then, the Eq.(1) can be re-expressed as:

$$\begin{aligned} &\theta_i | \theta_1, ..., \theta_{i-1}, \gamma, H \sim \\ &\qquad \sum_{k=1}^{K} \frac{m_k}{i-1+\gamma} \delta_{\theta_{\phi_k}} + \frac{\gamma}{i-1+\gamma} H. \end{aligned}$$

The Pólya urn scheme is closely related to the Chinese restaurant process (CR-P) [1]. In this metaphor, take θ_i to be a customer entering a restaurant with infinitely many tables, each serving a unique dish ϕ_k. Each arriving customer chooses a table, in proportion to how many customers are currently sitting at that table. With some positive probability proportional to γ, the customer starts a new, previously unoccupied table.

For each value θ_i, let z_i be an indicator random variable that picks out the unique value ϕ_k, such that $\theta_i = \phi_{z_i}$. We can get:

$$\begin{aligned} &z_i | z_1, ..., z_{i-1}, \gamma, H \sim \\ &\sum_{k=1}^{K} \frac{m_k}{i-1+\gamma} \delta(z_i, k) + \frac{\gamma}{i-1+\gamma} \delta(z_i, k^{new}), \end{aligned}$$

where K is the number of unique value. m_k is the number of indicator random variables taking the value k. k^{new} is a previously unseen value.

A second perspective on the Dirichlet process is stick-breaking construction. The stick-breaking construction considers a probability mass function $\{\beta_k\}_{k=1}^{\infty}$ on a countably infinite set, where the discrete probabilities are defined as follows:

$$\pi_k | \gamma \sim Beta(1, \gamma) \quad \beta_k = \pi_k \prod_{l=1}^{k-1} (1 - \pi_l), \tag{2}$$

A random draw $G_0 \sim DP(\gamma, H)$ can be expressed as:

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} \quad \theta_k | H \sim H, \quad k = 1, 2... \tag{3}$$

Dirichlet process can be used to model for a group data, while in many domains there are several groups of data produced by related, but distinct, generative processes. For this data, Teh et al. [17] proposed a hierarchical Dirichlet process (HDP) to link the group-specific Dirichlet processes. A HDP is a distribution over a set of random probability measures over probability space $(\Theta, \mathcal{B})$. Assume we have D groups of data and the dth group is denoted as $\{w_{dn}\}_{n=1,...,N_d}$. It defines a set of random probability measures $(G_d)_{d \in D}$. For the D data sets different group-specific G_d are drawn from $DP(\alpha_{d0}, G_0)$, in which G_0 is drawn from another DP with concentration parameter γ and base probability measure H. For each of these groups, w_{dn} is drawn from the model $w_{dn} \sim F(\theta_{dn})$ with parameters $\theta_{dn} \sim G_d$. Putting everything together, the generative model for HDP is represented as:

$$\begin{aligned} G_0 | \gamma, H &\sim DP(\gamma, H) \\ G_d &\sim DP(\alpha_{d0}, G_0) \\ \theta_{dn} &\sim G_d \\ w_{dn} &\sim F(\theta_{dn}), \end{aligned}$$

where $d \in D$ and $n = 1, ..., N_d$. In the hierarchical structure, different observations w_{dn} and $w_{dn'}$ in the same group share the same parameters θ^* based on the probability measure G_d. Moreover, since all G_d are composed of the same set of atoms $\{\theta_k^*\}_{k=1}^{\infty}$, the observations across different groups share parameters as a consequence of the discrete form of G_0. The clusters in each group d, assumed by the set $\{\theta_{dn}\}_{n=1,...,N_d}$, are inferred via the posterior density function on the parameters, with the likelihood function selecting the set of discrete parameters $\{\theta_k^*\}_{k=1}^{\infty}$ most consistent with the data $\{w_{dn}\}_{n=1,...,N_d}$. Meanwhile, clusters (and, hence, associated cluster parameters $\{\theta_k^*\}_{k=1}^{\infty}$ are shared across multiple data sets, as appropriate.

Since in the HDP different groups share the same parameters θ^*, the social relationship between groups has not been consider. The purpose of this paper is to extend the HDP to incorporate the structural information.

2.2 The Proposed Method

The proposed social hierarchical Dirichlet process extends from HDP, where documents share the same global measure G_0. In sHDP, each document has its own specific high-level random measure G_0^d. It is distributed as a Dirichlet process with concentration parameter γ and base probability measure H.

$$G_0^d | \gamma, H \sim DP(\gamma, H), \tag{4}$$

$G_0^d(d \in D)$ are tied together by the social influence. The random measures G_d are conditionally independent given G_0^d, with distributions given by a Dirichlet process

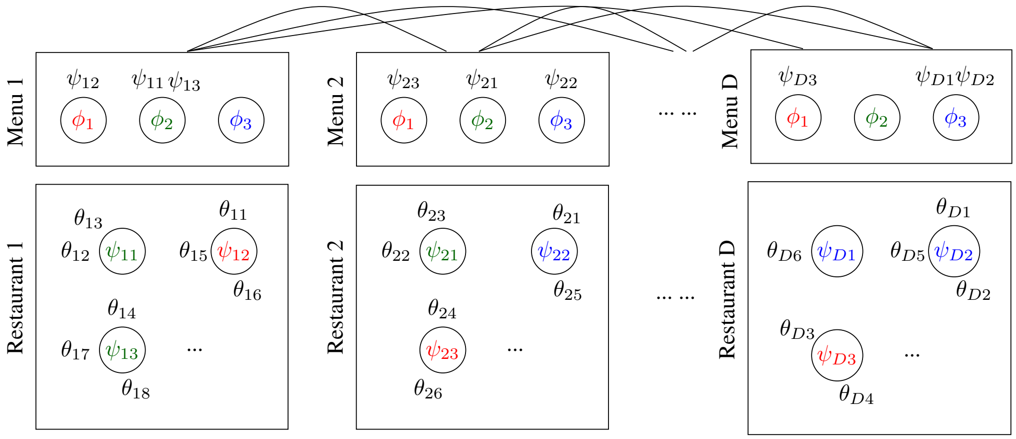

Menu 1
ψ_{12} ψ_{11} ψ_{13}
ϕ_1 ϕ_2 ϕ_3
Menu 2
ψ_{23} ψ_{21} ψ_{22}
ϕ_1 ϕ_2 ϕ_3
... ...
Menu D
ψ_{D3} $\psi_{D1}\psi_{D2}$
ϕ_1 ϕ_2 ϕ_3
Restaurant 1
θ_{13} θ_{11}
θ_{12} ψ_{11} θ_{15} ψ_{12}
θ_{16}
θ_{14}
θ_{17} ψ_{13} ...
θ_{18}
Restaurant 2
θ_{23} θ_{21}
θ_{22} ψ_{21} ψ_{22}
θ_{25}
θ_{24}
ψ_{23} ...
θ_{26}
... ...
Restaurant D
θ_{D1}
θ_{D6} ψ_{D1} θ_{D5} ψ_{D2}
θ_{D2}
θ_{D3}
ψ_{D3} ...
θ_{D4}

Fig. 1: A depiction of the social Chinese restaurant franchise (sCRF) process. Each restaurant is represented by a rectangle and has its own menu. The menus between different restaurants will influence each other. Customer θ_{dn} in restaurant d is seated at a table. Tables are represented by circles. A dish on a table is served from its menu ϕ_k. ψ_{dt} is an indicator to index items on the menu for a specific table.

with base probability measure G_0^d.

$$G_d|\alpha, G_0^d \sim DP(\alpha, G_0^d), \tag{5}$$

$$\begin{aligned} &G_0^d|\psi_{1:k}, H, \gamma \sim \\ &DP(\alpha + m_{d.}^*, \sum_k \frac{m_{dk}^*}{m_{d.}^* + \alpha}\delta(\psi_k) + \frac{\alpha}{m_{d.}^* + \alpha}H). \end{aligned} \tag{6}$$

Integrating out the random measure G_d and G_0^d through the Chinese restaurant process, we can get a social Chinese restaurant franchise process (sCRF). In this metaphor (see Figure 1), we have a restaurant franchise and each restaurant has a menu for itself. For each table of each restaurant, the first customer sits on will order a dish, and all customers sit on the same table will share the dish. In different restaurants and different tables, the dish can be same, which is controlled by the social influence. In this setup, the restaurants correspond to documents, and the customers correspond to the parameters θ. Let $\phi_1, \phi_2, ..., \phi_K$ denote K random variables distributed according to H. We use the variable ψ_{dt} to represent the dish served at table t in the restaurant d. Each θ_{dn} is related to one ψ_{dt}, we use t_{dn} as the index of them. Each ψ_{dt} is related to one ϕ_k, we use k_{dt} as the index. We use n_{dtk} to represent the number of customers in restaurant d at table t with dish k. And we use $n_{d.k}$ to represent the number of customers in restaurant d on the tables which serve the dish k. The notation m_{dk}^*, which equals to $\sum_{q \in D_d} f_{dq} m_{qk}$, represents the influenced number of dish k for restaurant d. D_d is the restaurant set which are connected with the restaurant d. m_{qk} denotes the number of tables in restaurant q with dish k. f_{dq} is the social influence between document d and q, we can calculate it based on the network structure. $m^*_{d.}$ represents the influenced number of tables for restaurant d. In summary, we can get the conditional distribution of θ_{dn}:

$$\theta_{dn}|\theta_{d1}, ..., \theta_{d,n-1}, \alpha, G_0^d \sim \\ \sum_{t=1}^{T_d} \frac{n_{dtk_{dt}}}{n-1+\alpha}\delta_{\psi_{dt}} + \frac{\alpha}{n-1+\alpha}G_0^d,$$

where T_d is a count of tables in restaurant d. And we can obtain the conditional distribution of ψ_{dt}:

$$\psi_{dt}|\psi_{11}, \psi_{12}, ..., \psi_{21}, ..., \psi_{dt-1}, \gamma, H \sim \\ \sum_{k=1}^{K} \frac{m^*_{dk}}{m^*_{d.}+\gamma}\delta_{\phi_k} + \frac{\gamma}{m^*_{d.}+\gamma}H.$$

The generative process of sCRF is shown in algorithm 1.

Algorithm 1 The generation process of sCRF

```
for each restaurant $d \in D$ do
    for each customer $\theta$ in restaurant $d$ do
        Choose table $t \propto \frac{n_{dtk_{dt}}}{n-1+\alpha}$
        Choose a new table $t^{new} \propto \frac{\alpha}{n-1+\alpha}$
        if Choose a new table $t^{new}$ then
            Sample a new dish for this table.
            Choose an existing dish $k \propto \frac{m^*_{dk}}{m^*_{d.}+\gamma}$
            Choose a new dish $k^{new} \propto \frac{\gamma}{m^*_{d.}+\gamma}$
        end if
    end for
end for
```

From algorithm 1, we can get the generation process of document d. Each word w_{dn} in document d is drawn from $F(\theta_{dn})$, the parameter θ_{dn} can select a cluster(table) t with probability $\frac{n_{dtk_{dt}}}{n-1+\alpha}$. Also θ_{dn} has probability $\frac{\alpha}{n-1+\alpha}$ to choose a new cluster(table) t^{new}, then it can choose a new topic k^{new} with probability $\frac{\gamma}{m^*_{d.}+\gamma}$ and increment K or choose an existing topic k with probability $\frac{m^*_{dk}}{m^*_{d.}+\gamma}$. In social hierarchical Dirichlet process, the topic distributions of different documents connected with social influence.

2.3 Inference

We use Gibbs sampling method to obtain samples of hidden variable assignment. In this model, we need to sample table $\mathbf{t}$ for each customer and dish $\mathbf{k}$ for each table.

The sampling probability of table t_{dn} is as follows:

$$p(t_{dn} = t|\mathbf{t}_{\neg dn}, \mathbf{k}) \propto \begin{cases} n_{dt}^{\neg dn} f_{\neg w_{dn}}^{k_{dt}}(w_{dn}), & t \text{ is an existing table} \\ \alpha p(w_{dn}|\mathbf{t}_{\neg dn}, t_{dn} = t^{new}, \mathbf{k}), & t \text{ is a new table} \end{cases},$$

where $n_{dt}^{\neg dn}$ is a count of customers at table t in restaurant d; $\neg dn$ denotes the counter calculated without considering the customer n in restaurant d; $f_{\neg w_{dn}}^{k_{dt}}(w_{dn})$ is the likelihood of generating w_{dn} for existing table t, which can be calculated by:

$$f_{\neg w_{dn}}^{k_{dt}}(w_{dn}) = \frac{\int f(w_{dn}|\phi_k) \prod_{d'n' \neq dn, z_{d'n'} = k} f(w_{d'n'}|\phi_k) h(\phi_k) d(\phi_k)}{\int \prod_{d'n' \neq dn, z_{d'n'} = k} f(w_{d'n'}|\phi_k) h(\phi_k) d(\phi_k)},$$

where $k = k_{dt}$ is the dish served at table t in restaurant d. And $p(w_{dn}|\mathbf{t}_{\neg dn}, t_{dn} = t^{new}, \mathbf{k})$ is the conditional distribution of w_{dn} for $t_{dn} = t^{new}$, which can be calculated by integrating out the possible values of $k_{dt^{new}}$ as follows:

$$p(w_{dn}|\mathbf{t}_{\neg dn}, t_{dn} = t^{new}, \mathbf{k}) = \sum_{k=1}^{K} \frac{m_{dk}^*}{m_{d.}^* + \gamma} f_{\neg w_{dn}}^{k}(w_{dn}) \quad + \frac{\gamma}{m_{d.}^* + \gamma} f_{\neg w_{dn}}^{k^{new}}(w_{dn}),$$

where m_{dk}^* is the influenced number of tables which assigned to dish k for restaurant d. $m_{d.}^*$ is the total influenced number of tables for restaurant d; $f_{\neg w_{dn}}^{k^{new}}(w_{dn}) = \int f(w_{dn}|\phi)h(\phi)d\phi$ is the prior density of w_{dn}; the prior probability that the new table t^{new} served a new dish k^{new} is proportional to γ.

If the sampled value of t_{dn} is equal to t^{new}, we can obtain a sample of $k_{dt^{new}}$ by sampling from:

$$p(k_{dt^{new}} = k|\mathbf{t}, \mathbf{k}_{\neg dt}) \propto \begin{cases} m_{dk}^* f_{\neg w_{dn}}^{k}(w_{dn}), & k \text{ is an existing dish} \\ \gamma f_{\neg w_{dn}}^{k^{new}}(w_{dn}), & k \text{ is a new dish} \end{cases}.$$

The probability that t_{dn} takes on a particular previously used value t is proportional to n_{dt}. So if some table t becomes unoccupied after updating t_{dn}, the probability that this table can be reoccupied will be zero. As a result, we may delete the corresponding k_{dt} from the data structure. If as a result of deleting k_{dt}, some dish k becomes unused for any table, we delete this kind of dish as well.

The sampling probability of dish k_{dt} for the table t is as follows:

$$p(k_{dt} = k|\mathbf{t}, \mathbf{k}_{\neg dt}) \propto \begin{cases} m_{dk}^{*\neg dt} f_{\neg \mathbf{w}_{dt}}^{k}(\mathbf{w}_{dt}), & k \text{ is an existing dish} \\ \gamma f_{\neg \mathbf{w}_{dt}}^{k^{new}}(\mathbf{w}_{dt}), & k = k^{new} \text{ is a new dish} \end{cases},$$

Table 1: Statistics of the three data sets

Data Set	Doc.	Words	Vocabulary Size	Links	Author
NIPS	5,179	1,607,205	11,890	15,404	13,784
CORA	9,842	358,824	2,620	78,721	21,101
SINA	468,177	1,358,010	11,596	12,329	1,318

where $\mathbf{w}_{dt}$ is the customers at table t in restaurant d. When we change k_{dt}, actually, the dish for all the customers at this table have changed, $f^{k}_{\neg w_{dt}}(\mathbf{w}_{dt})$ is the likelihood for the customers on this table.

2.4 Social Influence

As described in the previous section, network structures can be transferred into social influence and incorporated into sHDP. In this work, we propose a simple method to compute the influence between documents based on structural information. Each document has some specific authors. Firstly, we calculate the number of links between two authors. Then, based on the number of links between authors, we can inference the affect parameter f_{dq} between document d and document q as follows:

$$f_{dq} = \exp\left(\sum_{a_d \in A_d} \sum_{a_q \in A_q} \eta_{a_d,a_q} \frac{N_{a_d,a_q}}{0.5N_{a_d} + 0.5N_{a_q}}\right),$$

where A_d is the set of authors of document d, A_q is the set of authors of document q; $\eta_{a_d,a_q} = \frac{1}{Z}\frac{1}{I_{a_d} * I_{a_q}}$, Z is the normalization term; I_{a_d} and I_{a_q} are the rank number of author a_d in the document d and author a_q in the document q respectively; N_{a_d,a_q} is the count of links between author a_d and author a_q; N_{a_d} is the total number of links connect to author a_d; N_{a_q} is the number of links connect to author a_q.

Except the influence from the neighbours, we assume that it may also be influenced from other documents. This ensures sharing of global topic. Hence, we calculate the influence number of topic k for document d by the following equation:

$$m^{*}_{dk} = \sum_{q_o \in D^o} \lambda m_{q_o,k} + \sum_{q_r \in D^d} f_{d,q_r} m_{q_r,k},$$

where D^o is the document set which is not neighbours of document d; λ is the influence parameter from document set D^o.

3 Experiments

3.1 Data sets and Settings

To examine the effectiveness of the proposed sHDP model, we constructed three datasets: the papers from NIPS proceedings, articles published in CORA, and microblogs from

Sina Weibo[1]. The **NIPS** dataset contains full papers from NIPS proceedings between 1987 and 2013 and was obtained by crawling the documents and authors from the official NIPS website. The dataset contains 5,179 papers and 13,784 authors. The **CORA** dataset, which was constructed by McCallum et al. [12], contains abstracts from the CORA [2] computer science research paper archive. The title and abstract of each research paper is treated as the content. In total, there are 9,842 papers and 21,101 authors represented in the CORA dataset.

The **SINA** dataset contains microblogs crawled in the following manner. Firstly, we randomly selected 10 users as the central users. Then we collected the 1-*ego* network for all the central users based on their "following" relationships. All of the microblogs posted by these users from Jul. 1, 2013 to Sep. 30, 2013 were collected to construct the dataset. Using these steps, we gathered 468,177 microblogs belonging to 1,318 users in total. Each individual microblog posted by a user is treated as a separate document. Since there are no spaces between words in Chinese sentences, we used Stanford Word Segmenter[3] to split each microblog into a sequence of words.

The NIPS and CORA social networks are constructed based on co-author relationships, however the SINA dataset uses the *following* relationships between users for this purpose. For all of the documents, we removed the words whose frequency is more than 2000 or less than 50. For the NIPS and CORA datasets, we randomly selected 1,000 documents for testing and used the others as training data. For the SINA dataset, we randomly selected 100 documents for testing and 1,218 documents for training. The detailed statistics of the three datasets are given in Table 1.

For comparison with the proposed method, we also evaluated LDA, HDP, and the relational topic model (RTM) [5] using the same three datasets. LDA and HDP have been widely used for modeling topics. In this project, we evaluated them on the constructed datasets. However, as we mentioned in the previous sections, they do not take structural information into consideration. To compare the proposed method with the methods incorporating structural information, we evaluated RTM[4], which extends LDA and incorporates link information in the constructed datasets. To quantitatively evaluate the proposed method against the baselines, we used perplexity as the evaluation metric.

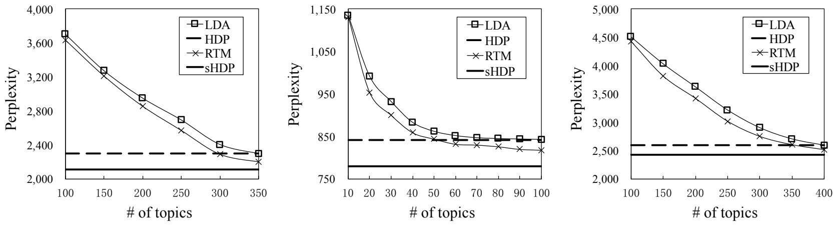
We ran sHDP and HDP with 1000 iterations of Gibbs sampling. Both of them use a symmetric Dirichlet distribution with parameters of 0.5 as the prior of base measure H over topic distributions. In sHDP and HDP, the concentration parameters are were given vague gamma priors, $\gamma \sim Gamma(5, 0.1)$ and $\alpha_0 \sim Gamma(5, 1)$. Posterior samples are were obtained with the Chinese restaurant franchise sampling scheme. The distributions over topics in LDA and RTM are were assumed to be symmetric Dirichlet with parameters $\alpha = 50.0/L$, with L being the number of topics, $\beta = 0.1$. γ is not used in LDA and RTM. These parameters are determined by 5-folds cross-validation on training data.

[1] Sina Weibo is one of the most popular websites providing microblogging services in China. http://www.weibo.com

[2] http://www.cora.justresearch.com

[3] http://nlp.stanford.edu/software/segmenter.shtml

[4] The toolkit was downloaded from the website of the authors. https://www.cs.princeton.edu/ blei/topicmodeling.html



(a) NIPS (b) CORA (c) SINA

Fig. 2: Perplexity results on the NIPS, CORA and SINA for LDA, HDP, RTM, and sHDP.

3.2 Evaluation results

Table 2 shows the comparison of the proposed method, sHDP, with the state-of-the-art methods on the three evaluation datasets. From the results, we see that sHDP achieves much better performance than the other methods in all three datasets. sHDP achieved percentage decreases in perplexity over HDP of 8.8%, 7.8%, and 6.6% when run on the NIPS, CORA, and SINA datasets, respectively. The relative improvements achieved by sHDP on the NIPS and CORA datasets are better than the improvements observed on the SINA dataset. One of the main reasons for this performance difference may be the method of constructing the social network since co-authors are more likely to have similar interests and may publish papers with similar topics. From the table, we also observe that RTM achieves better performance than HDP and LDA on all three datasets. This demonstrates that the availability of structural information can create a performance advantage for the task of topic modeling.

Table 2: Perplexity of different methods in all three data sets.

Methods	**NIPS**	**CORA**	**SINA**
LDA	2297	843	2595
HDP	2298	841	2591
RTM	2203	818	2519
sHDP	**2111**	**780**	**2431**

Consistent with the results reported by previous research, HDP achieves performance similar to that of LDA for all three datasets. Since the number of topics is one of the most sensitive hyperparameters for LDA and RTM, we evaluated it for all three datasets and show the results in Figure 2. From the results, we observe that the values which are used to achieve the best performance are different for different datasets. These results also demonstrate the advantages of non-parametric methods. This advantage is one of the key reasons why we tried to extend HDP to incorporate social influence in this project.

Table 3: An illustration of five topics for CORA data set. Five words with the highest conditional probability for each topic are given. We use the oval box to highlight the inappropriate words in the topics extracted by HDP.

	No.	Words
HDP	1	learning neural network training $\boxed{\text{method}}$
	2	tree graph $\boxed{\text{minimum}}$ $\boxed{\text{test}}$ $\boxed{\text{cost}}$
	3	language program object type $\boxed{\text{implementation}}$
	4	distribution $\boxed{\text{model}}$ probability $\boxed{\text{test}}$ random
	5	debug $\boxed{\text{single}}$ process program proof
sHDP	1	neural network learning training hidden
	2	graph tree node path edge
	3	code program parallel language java
	4	distribution markov probability bayesian statistical
	5	debug experiment error program analysis

Table 3 shows the topics extracted from the CORA dataset using HDP and sHDP. For each topic, we list the top five words with the highest probabilities. From the table, we observe that although HDP extracts reasonable topics, all five topics extracted by it have their limitations. For example, the word "method", which is a commonly used word in the computer science domain, is in the top list of Topic 1. Comparing this result with the topics extracted by sHDP, we see that the five topics identified by sHDP are much better. Topic 1 covers papers about neural networks, Topic 2 is related to the domain of the graph, Topic 3 corresponds to programming, Topic 4 includes statistics, and Topic 5 is closely related to experimental procedures.

4 Conclusions

In this paper we introduced a novel social hierarchical Dirichlet process model, sHDP, to incorporate structural information for modeling topics. In sHDP, the social network structure will transfer into the levels of influence between documents. We detailed sHD-P through the social Chinese restaurant franchise process and described a Gibbs sampling algorithm for posterior inference. For evaluating the proposed method, we constructed three datasets. Experimental results demonstrated that structural information can significantly benefit topical modeling, and the proposed method achieved better performance than the state-of-the-art methods for all three datasets.

References

1. Aldous, D.J.: Exchangeability and related topics. Springer (1985)
2. Blackwell, D., MacQueen, J.B.: Ferguson distributions via pólya urn schemes. The annals of statistics pp. 353–355 (1973)

3. Blei, D.M., Frazier, P.I.: Distance dependent chinese restaurant processes. The Journal of Machine Learning Research 12, 2461–2488 (2011)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research 3, 993–1022 (2003)
5. Chang, J., Blei, D.M.: Relational topic models for document networks. In: International Conference on Artificial Intelligence and Statistics. pp. 81–88 (2009)
6. Cowans, P.J.: Information retrieval using hierarchical dirichlet processes. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 564–565. ACM (2004)
7. Fox, E.B., Sudderth, E.B., Jordan, M.I., Willsky, A.S.: An hdp-hmm for systems with state persistence. In: Proceedings of the 25th International Conference on Machine Learning. pp. 312–319. ICML '08, ACM, New York, NY, USA (2008), `http://doi.acm.org/10.1145/1390156.1390196`
8. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. pp. 50–57. ACM (1999)
9. Kim, S., Smyth, P.: Hierarchical dirichlet processes with random effects. In: NIPS. pp. 697–704 (2006)
10. Liu, Y., Niculescu-Mizil, A., Gryc, W.: Topic-link lda: Joint models of topic and author community. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 665–672. ICML '09, ACM, New York, NY, USA (2009), `http://doi.acm.org/10.1145/1553374.1553460`
11. McCallum, A., Corrada-Emmanuel, A., Wang, X.: Topic and role discovery in social networks. Computer Science Department Faculty Publication Series p. 3 (2005)
12. McCallum, A., Nigam, K., Rennie, J., Seymore, K.: Automating the construction of internet portals with machine learning. Information Retrieval Journal 3, 127–163 (2000), www.research.whizbang.com/data
13. Mei, Q., Cai, D., Zhang, D., Zhai, C.: Topic modeling with network regularization. In: Proceedings of the 17th international conference on World Wide Web. pp. 101–110. ACM (2008)
14. Ren, L., Dunson, D.B., Carin, L.: The dynamic hierarchical dirichlet process. In: Proceedings of the 25th International Conference on Machine Learning. pp. 824–831. ICML '08, ACM, New York, NY, USA (2008), `http://doi.acm.org/10.1145/1390156.1390260`
15. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering objects and their location in images. In: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on. vol. 1, pp. 370–377. IEEE (2005)
16. Tang, J., Sun, J., Wang, C., Yang, Z.: Social influence analysis in large-scale networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 807–816. KDD '09, ACM, New York, NY, USA (2009), `http://doi.acm.org/10.1145/1557019.1557108`
17. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical dirichlet processes. Journal of the american statistical association 101(476) (2006)
18. Wang, X., McCallum, A.: Topics over time: a non-markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 424–433. ACM (2006)
19. Zhang, J., Song, Y., Zhang, C., Liu, S.: Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1079–1088. KDD '10, ACM, New York, NY, USA (2010), `http://doi.acm.org/10.1145/1835804.1835940`