BayDNN: Friend Recommendation with Bayesian Personalized Ranking Deep Neural Network

Daizong Ding¹, Mi Zhang¹, Shao-Yuan Li², Jie Tang³, Xiaotie Chen¹, Zhi-Hua Zhou²
¹Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University, China
{17110240010.mi_zhang,13300180119}@fudan.edu.cn
²National Key Lab for Novel Software Technology, Nanjing University, China
{lisy,zhouzh}@lamda.nju.edu.cn
³Department of Computer Science and Technology, Tsinghua University, China
jietang@tsinghua.edu.cn

ABSTRACT
Friend recommendation is a critical task in social networks. In this paper, we propose a Bayesian Personalized Ranking Deep Neural Network (BayDNN) model for friend recommendation in social networks. BayDNN first extracts latent structural patterns from the input network data and then uses the Bayesian ranking to make friend recommendations. With BayDNN we achieve significant performance improvement on two public datasets: Epinions and Slashdot. For example, on Epinions dataset, BayDNN significantly outperforms the state-of-the-art algorithms, with a 5% improvement on NDCG over the best baseline.

The advantages of the proposed BayDNN mainly come from a novel Bayesian personalized ranking (BPR) idea, which precisely captures the user’s personal bias on the extracted deep features, and its underlying convolutional neural network (CNN), which offers a mechanism to extract latent deep structural feature representations of the complicated network data. To get good parameter estimation for the neural network, we present a fine-tuned pre-training strategy for the proposed BayDNN model based on Poisson and Bernoulli probabilistic models.

CCS CONCEPTS
- Information systems → Social advertising;

KEYWORDS
Bayesian Personalized Ranking Deep Neural Network; Probabilistic Model; Pre-training Strategy

1 INTRODUCTION
Recently, online social networks (OSN) such as Facebook, Twitter and Weibo have been growing exponentially. People from all over the world get connected to each other online, making friends with those sharing similar interests [1]. As the cornerstone of social networks, friendship and its formation has attracted tremendous attention from both academia and industry [20, 27]. For example, Granovetter categorized friendship into strong and weak, and found that novel information flows to individuals through weak ties rather than strong ties [9]. In online social networks, one phenomenon is that the leading reason for users to create new friendships is due to recommendation [28]. Therefore, friend recommendation becomes an essential task of social networks.

In this paper, our goal is to design a high-accuracy method for friend recommendation in online social networks. Different from previous works [11, 44] that rely on specific context information, we aim to design a method that is general enough to be applied to different social networks. We propose a deep neural network for friend recommendation using only network structure information.

Previous work on friend recommendation. From the algorithmic perspective, existing methodologies for the friend recommendation problem roughly fall into three categories: classification, fitting, and ranking [10, 22]. The classification method treats friendship between users as a binary classification problem, and trains a classifier to predict the likelihood of a friendship to be created between users based on pre-defined features. The fitting method transforms the friendship between users into a real-valued rating matrix, and utilizes collaborative filtering approaches such as matrix factorization to predict the probabilities of unknown friendships. Taking into account the common imbalance issue in most OSN [13], i.e., the amount of friendship is usually much less than that of non-friendships, which would lead the classification or fitting methods biased towards non-friendship, the ranking method has been proposed. Regarding friend recommendation as a learning to rank task — i.e., for each user, by ranking the probability of friendship for the users connected to it larger than that of users not connected to it — the ranking method has been proved to be effective combating the imbalance issue. Among them, the Bayesian personalized ranking (BPR) model [18, 30, 34] defined Bayesian pairwise ranking relation between users, i.e., the friendship probability of observed friends should be larger than the unobserved users pairs, was validated to achieve quite good performance.

Previous work on deep neural networks (DNN). Recently, deep neural networks have achieved pervasively excellent performance in various fields including image classification and phrase representations [4, 17]. Compared to traditional methods such as kernel machines [42] or matrix factorization [16] with shallow architecture or limited levels of feature extraction, deep neural networks like CNN with properly designed structure and careful parameter
We review existing methodologies for friend recommendation. However, till now few deep learning study has been conducted to address the task of friend recommendation.

To achieve high-accuracy recommendations, one straightforward idea is to apply deep neural network (DNN) to friend recommendation. However, the advantage of DNN is to learn a better feature representation, but it fails to consider the essential imbalance issue in friend recommendation. While a few efforts have been made using BPR (Bayesian Personalized Ranking) or DNN (Deep Neural Network) [41] for friend recommendation, no work has been done combining the strength of these two ideas.

Our solution and contributions. In this paper, we propose a deep Bayesian Personalized Ranking Deep Neural Network (BayDNN) to combine the advantage of BPR (Bayesian Personalized Ranking) and DNN (Deep Neural Network). The proposed BayDNN seamlessly combines DNN and Bayesian ranking, and integrates CNN at the stage of feature representation learning, which not only reduces the number of parameters compared to traditional full connect neural network, but also improves the effectiveness on feature mapping.

The proposed BayDNN model has several unique advantages compared with the previous methods. First, the model uses convolutional neural network (CNN) to extract latent deep structural feature representations of the complicated network data. Second, it uses a novel Bayesian personalized ranking learning algorithm to better capture users’ personal bias based on the extracted deep features. Third, to avoid poor parameter estimation, we design a fine-tuned pre-training strategy for the proposed BayDNN model based on Poisson and Bernoulli probabilistic models. Moreover, to enhance the BPR, we group the friendship between users into three sets, positive, negative, and unknown. And then, during the learning process, we enforce the positive friendship to be ranked before the unknown/negative ones, and the unknown friendship to be ranked before the negative ones.

We conduct experiments on multiple real datasets. The results show that BayDNN not only extracts better feature presentation, which is different from classical or neighbor based methods, but also performs well on large-scale datasets in friend recommending in OSN. In terms of accuracy performance it outperforms the best baseline and achieves 5% improvement on NDCG. In terms of time cost for the model training, with the proposed pre-training strategy the number of convergence epochs is largely reduced by 92.5%.

Organization. The rest of the paper is organized as following. We briefly review related work in Section 2. In Section 3 we propose our deep Bayesian Personalized Ranking Deep Neural Network (BayDNN) model and specify our pre-training strategies for BayDNN. In Section 4 we design several experiments to validate our model and conclude in Section 5.

2 RELATED WORK
We review existing methodologies for friend recommendation, highlighting the most relevant ones and explaining how our work distinguishes from them. Then, we briefly summarize recent development on deep neural networks.

Friend Recommendation. At the high level, methodologies dealing with friend recommendation can be grouped into three categories [22]: classification, fitting and ranking. The classification methods are based on extracted features, e.g., features between two nodes like path-based metric Katz [14] or neighbor-based metric Adamic/Adar [23]. By regarding the friendship status between two users as a binary classification task, they train classifiers such as SVM [8], logistic regression [29], and factor graph [40] to predict the friendship values between users. The fitting methods represent the friendship between users by real values and try to approximate the values for observed friendship as close as they can, e.g., [12, 25] used matrix factorization to predict the value of unknown friendship. Considering that the value of friendship probability is non-negative, Non-negative Matrix Factorization (NMF) was introduced by [19] constraining that two factorized matrix are non-negative, which showed better performance in non-negative data [3] compared to MF. However the classification and fitting methods fail to deal with the severe data imbalance issue due to the sparseness of OSN data, i.e., the amount of observed friendship is much smaller than the unobserved ones [26]. This leads to the models biased towards making low friendship probability[31]. Hence ranking methods that regard this problem as a learning to rank task are proposed. For each user, by predicting the likelihood of friendship between friends larger than non-friends, ranking methods are proved to be effective combating the imbalance issue [18, 30, 34]. Widely used in item recommendation tasks, one popular Bayesian personalized ranking (BPR) [30, 34] model was integrated with matrix factorization [18], called Bayesian Personalized Ranking Matrix Factorization (BPRMF). They defined a Bayesian pairwise ranking relation, i.e., observed data to be rated before unobserved data by user-item link-age probability, which is calculated by matrix factorization. We are inspired by this idea and apply the BPR idea to friendship network, by combining it with a DNN model rather than a shallow MF model.

Deep Neural Network. Recently, deep neural networks (DNN) have archived significant success in many machine learning and data mining tasks, e.g., image labeling and voice recognizing [38]. Bengio et al. [2] pointed out that the shallow architecture of kernel machines [42] and matrix factorization [16] restricts their learning performance; meanwhile DNN achieves superior performance with properly designed structure and carefully chosen parameters. Especially in non-linear feature mapping, convolutional neural network outperforms most models in image [5], text [36] and other fields. In existing literature, few studied have been conducted to apply DNN methods to the friend recommendation task. In fact, there is no straightforward way to do this. One reason is that there exists no corresponding pairwise ranking model for neural networks. Another reason is that DNN usually needs proper pre-training strategy to avoid poor parameter initialization, e.g., use pre-trained deep neural network like VGG-Net to initialize parameters of model [24, 32]. However, it is unclear which pre-training strategy should be used for DNN in friend recommendation. Wang et al. [41] tried to use a four-layer neural network with a simple pre-training strategy called Pairwise Input Neural Network (P2NN) to do link prediction, by training and predicting the probability of a link point-wisely. However, it did not achieve significant performance improvement.
3 THE PROPOSED FRAMEWORK

In this section, we first introduce the pair-wise ranking task of the friend recommendation problem. Then we propose the deep Bayesian Personalized Ranking Deep Neural Network (BayDNN) model and describe the pre-training strategies.

3.1 Preliminaries

A social network can be denoted as a graph \( G = \langle V, E \rangle \), where \( V \) is a set of \( |V| = N \) users, and \( E \in \{0, 1\}^{N \times N} \) represents the undirected friendship structure between users. Notation \( e_{ui} = E(u, i) = 1 \) denotes that users \( u \in V \) and \( i \in V \) are friends, while \( e_{ui} = E(u, i) = 0 \) denotes that the friendship between \( u \) and \( i \) is unobserved. In practice, \( E \) is usually very sparse for two reasons: Dunbars number [6], which suggests that the number of one person’s stable social relationships lies between 100 and 230, which is much smaller compared to the whole network; and missing links, which means that the observed relationships in an online social network usually only represent a limited part of one’s real social circles [26].

We define a friendship probability matrix \( R \in [0, 1]^{N \times N} \), where \( r_{ui} = R(u, i) = P(e_{ui} = 1) \) denotes the probability that a friendship between user \( u \) and \( i \) is formed. In this way, for each user \( u \), all the other users can be grouped into two disjoint sets—i.e., a set with positive friends \( P_u = \{i | e_{ui} = 1\} \) and the rest with unobserved friends \( U_u = \{i | e_{ui} = 0\} \). For the task of friend recommendation, our idea is to construct a ranking model that is able to rank the positive friends before the unobserved friends. Inspired by [37], we further divide the unobserved friends \( U_u \) into two sets, the unknown friends \( M_u \) and the high probably negative friends \( N_u \), i.e., \( U_u = M_u \cup N_u \). In this way we construct the set of negative friends \( N_u \) by selecting the users that can not reach \( u \) through 6 people (i.e., within 6 hops in the social network). It has been concluded that people can on average connect to another through 6 people [21], which means the probability for those people being friends is quite low. Based on this division, e.g., for any other users \( i \) and \( j \), if \( i \in P_u \) and \( j \in U_u \), or \( i \in M_u \) and \( j \in N_u \), the probability of friendship \( r_{ui} \) should be greater than \( r_{uj} \). To describe this relation we define a Partial Relation \( i >_u j \).

**Definition 1.** \( i >_u j \): For a user \( u \), and two other users \( i \) and \( j \) in the social network, \( i \) has a higher probability of friendship with \( u \) than \( j \), i.e., \( r_{ui} > r_{uj} \), we say that \( i \) and \( j \) have a Partial Relation on \( u \).

Note that with this definition, partial relation on \( u \) has property of transitivity, i.e., if \( i >_u j \) and \( j >_u k \), we have \( i >_u k \).

Based on \( P_u \), \( U_u \), and \( N_u \), in the social network there are two kinds of partial relation on \( u \). First is the partial relation between \( P_u \) and \( U_u \), i.e., \( \{i >_u j \ | \ i \in P_u, j \in U_u\} \). Then within \( U_u \) (\( U_u = M_u \cup N_u \)), there is another partial relation between \( M_u \) and \( N_u \), i.e., \( \{i >_u j \ | \ i \in M_u, j \in N_u\} \). The set of all the partial relations for \( u \) is depicted as \( R_u = \{(i, j) | i >_u j\} \).

In this way we formalize friend recommendation as a ranking problem. Compared to other options such as classification or rating problems that suffer from the imbalance problem, the ranking formulation can more or less avoid it [18, 30, 34]. Our model aims to maximize the ranking likelihood probability as follows:

\[
\max \prod_{u \in V} \prod_{(i, j) \in R_u} P(i >_u j)
\]

3.2 Bayesian Personalized Ranking Deep Neural Network (BayDNN)

We propose a novel strategy that embed BPR learning framework in DNN to produce ranking result. Fig. 1(b) illustrates the architecture of the proposed Bayesian Personalized Ranking Deep Neural Network (BayDNN).

For user \( u \) and \((i, j)\) BayDNN uses their relationship information \( E_u, E_i, E_j = E(u, i), E(i, j), E(j, i) \in \{0, 1\}^N \) as input. Specifically, for feature mapping from input to hidden layer \( H \), we design a Convolutional Feature Mapping (CFM) structure. The details of Fig. 1(b)’s CFM part is given in Fig. 2. The output layer aims to produce
the friendship probability $r_{ui} = P(r_{ui} = 1)$ and $r_{uj} = P(r_{uj} = 1)$ and Bayesian personalized ranking (BPR) layer predict the probability of this partial relation, i.e., $P(i > u j)$.

Convolutional Feature Mapping (CFM). We use convolutional neural network (CNN) to extract latent structural patterns of the input data. Specifically, in our problem we want to extract the non-linear structural features from the relationship vector $E_u$ for each user $u$. The convolution neural network (CNN) for feature mapping has been widely used in image classification, pattern recognition and so on, and has been proved to be very efficient and accurate for feature mapping [5]. We use one-dimensional CNN to be the convolution Feature Mapping layers in our model. Such a method has been also applied to synthesize lexical $n$-gram information in previous work such as [33] and has shown good performance. The design of the convolution feature mapping layer between the input $E_u$ and feature $H_u$ is given by Alg. 1 and depicted in Fig. 2.

We define a loop in CFM as a manipulation of one convolution layer and another pooling layer in order, where the convolution layer uses the output from last loop’s pooling layer and pass its calculation result to the following pooling layer, as described in Alg. 1. Here we initialize the CFM by regarding input $E_u \in \{0, 1\}^N$ for every user $u$ as pooling result from loop 0, and pass it on to loop 1. After $M$ loops we add another convolution layer upon CFM to reduce vectors to real value and concatenate them as extracted feature $H_u$. Specifically, for convolution layer in loop $l$, we have:

$$C^l_j = \tanh(K^l \cdot (s^{l-1}_j)^T + \text{bias}^l)$$

(2)

Here $(s^{l-1}_j)^T$ denotes the transpose of the vector $s^{l-1}_j$, which is the embedding of the element $j$ for all the output from loop $(l-1)$’s pooling layer $S^{l-1}$. In our model, we select an element $j$ and its nearby elements to form $s_j$, as shown in Fig. 2. $K^l$ and $\text{bias}^l$ are the convolution kernel and bias for loop $l$ respectively. And the $\tanh$ function outputs the result of convoluted feature $C^l_j$ for element $j$ in loop $l$. The most common choices of activation functions are the following: sigmoid function $\sigma(x)$, hyperbolic tangent function $\tanh(x)$, and rectified linear function ReLU($x$). Considering that we want to learn non-linear and non-negative features of the relationship information, we choose $\tanh$ function for the activation function.

The results from convolution layer are then passed to the next pooling layer, which aggregates the convoluted feature for further reduction. We define the following pooling operation in loop $l$:

$$S^l_j = \text{pool}^l(c^l_j)$$

(3)

where $c^l_j$ is the embedding of element $j$ for all the output from convolution layer $C^l_j$ in the same loop. Here $\text{pool}^l$ means the pooling function for loop $l$. In the first loop it is max function and in the following pooling layers they are $average$ function. We use this technique to avoid the sparsity of relationship information vector in first loop.

The pooling layer $S^M$ outputs many short vectors after $M$ loops, and the output differs with varied size of datasets. Then CFM will add a final convolution layer on it and convert these vectors to real values. We concatenate them to form a new vector as the extracted feature of CFM, defined as $F_u$ for user $u$, where:

$$F_u = \sigma(k^{M+1} \ast (S^M)^T + \text{bias}^{F^{M+1}})$$

(4)

Here considering the input of CFM are non-negative, we use $\sigma(x)$ as activation function to output non-negative features.

Algorithm 1 CFM for Relationship Vector

Input: Link vector $E_u$ for user $u$

Output: Extracted feature $F_u$ for user $u$

1: function CFM_FORWARD($E_u$)
2:  Let $S_0 = E_u$, $l = 1$
3:  for $l \le M$, $l = l + 1$
4:    Calculate convolution result $C^l$ from $S^{l-1}$
5:    Calculate pooling result $S^l$ from $C^l$
6:  Calculate convolution result $C^{M+1}$ from $S^M$
7:  return $F_u = C^{M+1}$
8: end function

Bayesian Personalized Ranking (BPR) Learning. The rest part of BayDNN is based on Feature Layer $H$, and we adopt the Bayesian Personalized Ranking technique to produce friend recommendation via ranking learning. The model is trained to maximize the likelihood defined in Eq. 1, hence the loss function is:

$$\mathcal{L}(\theta_c, \theta_r) = -\sum_u \sum_{(i,j) \in R_u} P(i > u j) + \lambda_1\|\theta_c\|^2 + \lambda_2\|\theta_r\|^2$$

(5)

where $\theta_c = \{K, \text{bias}F\}$, the parameters for kernels and biases in CNN layers, and $\theta_r = \{w, \text{biasH}, b\}$ are the parameters in the rest part of BayDNN. The rest structure of BayDNN and inference method is defined as below.

Hidden layer: Regarding $F_u$, $F_i$, $F_j$ as features of user $u$, $i$ and $j$ the hidden layer embedding their vectors into $H_u$, $H_i$ and $H_j$ for further calculation, e.g., for user $u$ the hidden layer produces $H_u$ by

$$H_u = \text{ReLU}(wF_u + \text{biasH})$$

(6)

Here we use $\text{ReLU}$ function instead of non-linear function for the consideration of convergence speed, which is simply defined as max(0, $x$). Parameters $w$ and $\text{biasH}$ are the weights and bias for units respectively.

Output Layer: Using $H_u$, $H_i$, $H_j$ as input, the following layer produces the friendship probability of $r_{ui}$ and $r_{uj}$, e.g.,
We design different pre-training strategies for Poisson distribution: $E$ a probabilistic-based solution by assuming that matrix $E$ follows:

However in the output layer, we allow each user $u$ to have its own bias parameter $\theta_u$ (Eq.7) to capture the user's personal preference in making friends, which has been shown to be an important factor in recommendation problems. We validate the value of modeling users' personal bias in our experiments.

The probability of a partial relation between $r_{ui}$ and $r_{uj}$ is:

$$P(i > u) = \frac{r_{ui} - r_{uj}}{2} + 0.5$$

**Inference:** To infer the parameters $\theta_c = \{K, biasF\}$ and $\theta_r = \{w, biasH, b\}$ for our BayDNN model through Eq.5, we employ back propagation (BP) algorithm by using stochastic gradient descent (SGD). However, since the objective function of Eq.1 is non-convex, proper initialization of parameter is critical to the performance of SGD-based method. In the following, we propose our pre-training strategies to learn proper initial parameters $\theta_c$ and $\theta_r$.

**Prediction:** Based on the output probability, the prediction for a friend is easy, e.g., for user $u$ and $i$ by forwarding on the network and fetch output layer $r_{ui}$ as their probability being friends. Then for different users around $u$ collect their probability and make a ranking list, based on the list the system could recommend potential friends to user $u$.

### 3.3 Pre-training Strategies

We design different pre-training strategies for $\theta_c$ in the CFM layers and $\theta_r$ in the rest layers, respectively, as depicted in Fig. 1(b). Firstly, to obtain $\theta_c$ in the convolution layers in Fig. 2, inspired by the idea of autoencoder, we expect the output of pre-training the same as input after specific mathematical operation, i.e., $E = F^T F$, and propose a Poisson probability model to derive the parameter value. Next, the pre-training is performed for parameters $\theta_r$ in ranking layers. We view each $e_{ui}$ as one observation sampled from its corresponding Bernoulli probability distribution $r_{ui}$, and transform the problem into a Bernoulli probability model.

**Pre-training for $\theta_c$.** Based on the two considerations that, 1)the output $F_u$ of the convolution parts can be regarded as the latent representation for each user $u$, and 2)the elements of $H$ are non-negative as the output of the sigmoid function, we can regard matrix $F$ as the non-negative latent representation matrix for users with linkage structure $E$. We use an approach proposed by [19] to derive the latent non-negative matrix $F$ based on $E$ by minimizing the following:

$$\min KL \left( E \| F^T F \right)$$

and [3] also proved that Eq. 9 is equivalent to maximize the likelihood of $E$, which is defined as:

$$L(F) = \log p(E | F) = \log \prod_u \mathcal{P}(e_{ui} | F^T_u F_i)$$

$$= \sum_u \left( e_{ui} \log(F^T_u F_i) - F^T_u F_i - \log \Gamma(e_{ui} + 1) \right)$$

Substituting $H$ defined by Eqs. 4 and 11 induces to:

$$L(\theta_c) = \sum_{u,i} \left( e_{ui} \log \sum_k [f(E_u)]_k \cdot [f(E_i)]_k - \sum_k [f(E_u)]_k \cdot [f(E_i)]_k - \log \Gamma(e_{ui} + 1) \right) - \lambda \| \theta_c \|^2$$

Here by regarding CFM operations as a non-linear function with input $E_u$ and parameters $\theta_c$, in Eq. 12, $f(E_u)$ represents the CFM’s output vector given $E_u$. Since it’s hard to infer $\theta_c$ from Eq.12, we propose to optimize Eq.12 by relaxing it to two sub-problems. The first sub-problem regards $f(E_u)$ simply as a variable rather than a function during the optimization, and aims to seek the optimal vectors $F^*_u$ for every user $u$ in Eq.12. Then given $F^*$ estimated in the first sub-problem, the second sub-problem tries to estimate $\theta_c$ so that function $f(E_u)$ given $E_u$ will best approximate $F^*_u$. These two sub-problems are iteratively solved in a relaxed EM-style. In the E step, the value of $f(E_u)$ is optimized given the parameters $\theta_c$; in the M step, the parameters $\theta_c$ are optimized given $F^*$. Specifically, the EM style algorithm is as following:

**E Step:** Regarding $F_u = f(E_u)$ as a variable, Eq.12 simplifies to:

$$\max_F \sum_{u,i} \left( e_{ui} \log \sum_k F_{uk} F_{ik} - F_{uk} F_{ik} - \log \Gamma(e_{ui} + 1) \right)$$

For which [3] proposed an iterative optimization solution, in each iteration, the elements of $F$ are updated using :

$$F^*_{uk} = \frac{\sum_e e_{uk} x_{ek}}{\sum_e x_{ek} F_{ek}} \cdot F^{k*} = \frac{\sum_e e_{uk} x_{ek}}{\sum_e x_{ek}} \cdot F^{k}$$

**M Step:** Given the estimated $F^*$ from E step, we update the parameters $\theta_c = \{v, biasF\}$ by solving a least-square problem:

$$\min_{v, biasH} \sum_u \| F^*_u - f(E_u) \|^2 + \lambda \| \theta_c \|^2$$

We use SGD to estimate $\theta_c$. In each iteration of SGD, each $\theta_c$ is updated as following:

$$\Delta \theta_c = \eta \cdot \left( [F^*_u - f(E_u)] \cdot \frac{\partial f}{\partial \theta_c} - \lambda \cdot \theta_c \right)$$

where $\eta$ is the learning rate and $\lambda$ is the regularizing parameter.

Note that previous work [41] has a similar two-step pre-training design for a different application. However, for the first sub-problem, it directly obtains the value for $F^*$ with a greedy algorithm. Then the second step approximate the fixed value $F^*$ by adjusting other parameters. This may work well in neural network with simple
Pre-training Framework for BayDNN

Initialize: Random initial $\theta_u, \theta_r$ with Gaussian

Require: Hyper-parameters like learning rate $\lambda$, $M$

1: procedure PRE-TRAINING($E_u$)
2: repeat
3: Calculate $F_u$ by CFM_FORWARD($E_u$) for every $u$
4: for all $u$, $i$,
5: Calculate optimized $F_{ui}^*$ with $F_{ui}$ (Eq. 14)
6: Calculate square error $\varepsilon = \|F^* - f(E)\|^2$ (Eq. 15)
7: Back propagation through CFM with error $\varepsilon$
8: until Convergence
9: Calculate $F_u$ by CFM_FORWARD($E_u$) for every $u$
10: repeat
11: for all $u$, $i$, $j$
12: Calculate $r_{ui}, r_{uj}$ from $F_u, E_i, E_j$ (Eq. 2, 3, 4, 6, 7)
13: Back propagation through BayDNN
14: until Convergence
15: for Epoch in AdaDelta
16: Random select a partial relation $u$, $i$, $j$
17: Calculate $\Delta \theta_{ui}$ using $r_{ui}, r_{uj}$ from $F_u, E_i, E_j$ (Eq. 15)
18: Update BayDNN with AdaDelta
19: end procedure

structure, however, in deep convolution neural network the problem may rise that the fixed value $F_u^*$ is not within the range of $f(E_u)$.

Pre-training for $\theta_r$. Using the embedded feature $H_u$ by the CFM part as input, and $r_{ui}, r_{uj}$ as the output, we pre-train the rest structure to estimate parameters $\theta_r = \{w, biasH, b\}$ for user $u$. Regarding the relationship between users $E_u$ as the observation sampled from the Bernoulli probability distribution with parameter $r_{ui}$:

$$p(e_{ui} | r_{ui}) = e^{r_{ui} \cdot (1 - r_{ui})^{1 - e_{ui}}},$$

Its corresponding log-likelihood is defined as:

$$L = \sum_{u,i,j} \left( e_{ui} \log r_{ui} + (1 - e_{ui}) \log (1 - r_{ui}) \right)$$

To estimate $\theta_r$, similarly we denote the non-linear function $g(F_u, F_i)$ given embedded feature $F$ through parameters $\theta_r$, which equals to output layer of our model as described in Eq. 7. To maximize the log-likelihood of Bernoulli distribution, for user $u$ we sample positive users $i$ where $e_{ui} = 1$ from set $\mathcal{P}_u$ and negative users $j$ where $e_{uj} = 0$ from set $\mathcal{N}_u$, where both sets are previously defined in Section 3.1. In addition we just sample $j$ from $\mathcal{N}_u$ instead of $\mathcal{U}_u$ because our purpose of pre-training is to accelerate the training of deep neural network. However if we sample $j$ from $\mathcal{U}_u$ like others traditional methods the time efficiency will be low. Our log-likelihood function is defined as:

$$L(\theta_r) = \sum_u \left( \sum_{i \in \mathcal{P}_u} \log g(F_u, F_i) + \sum_{j \in \mathcal{N}_u} \log (1 - g(F_u, F_j)) \right) - \lambda \|\theta_r\|^2$$

We use SGD to conduct the optimization. In each iteration of SGD, $\theta_r$ are updated as following:

$$\Delta \theta_r = \eta \cdot \left( r_{ui} - g(F_u, F_i) \right) \cdot \frac{\partial g}{\partial \theta_r} - \lambda \cdot \theta_r$$
Comparison Methods. We compare our BayDNN model with several state-of-the-art methods for friend recommendation, including two popular graph feature based methods, three matrix factorization based methods and one artificial neural network based method.

**Katz** [14]: It is a shortest path distance based metric for evaluating the similarity between two nodes in graph. Although Katz is a variant of shortest path distance, it generally works better for link prediction because this metric extract the more detailed path based features for the network. Based on these features between two users a proper trained classifier will judge whether there exists a link between them, as described by [29] logistic regression performed better in this situation than other classifiers like SVM. Hence we adopted the logistic regression based on metric Katz between every two users.

**Adamic/Adar** [1]: This is a metric for quantifying the similarity between two nodes in a web pages network based on common neighbors. Regarding common neighbors as features, [23] used this metric to do link prediction in online social networks. Experiments showed that this algorithm not only performed well on link prediction, but also it could extracted precise topology information of the graph. Here we also adopted the logistic regression as Katz.

**Matrix Factorization (MF)** [25]: Matrix Factorization is a latent factor model, widely used for rating prediction [16, 35]. Menon and Elkan et al. [25] proposed to use MF for link prediction problem by regarding positive links as 1 and the others as 0, which outperformed most baselines.

**Non-negative Matrix Factorization (NMF)** [19]: It is a non-negative version of matrix factorization, introduced by [19] with constraining that two factorized matrix are non-negative.

**Pairwise Input Neural Network (PINN)** [41]: It is a four-layer artificial neural network with simple pre-training strategy for link prediction, which regarded two nodes’ link vector as input and output the probability that there exists a link between them.

**Bayesian Personalized Ranking Matrix Factorization (BPRMF)** [37]: It is a matrix factorization based bayesian probabilistic model with typical ranking strategy that aims at maximizing the probability for all the partial relationship in dataset [34], which has been recognized as an effective ranking method for friend recommendation.

For our BayDNN method, we use Gaussian random variables to initialize the weight parameters in each layer and validate our model with dropout [39]. The batch size for BayDNN varies from 20 to 50 depending on the size of datasets, specifically, for Epinions we choose 40 as batch size and 30 for Slashdot. The experiments are conducted on a machine with one GPU (NVIDIA GTX-1080) and one CPU (INTEL Xeon E5-2620).

### 4.2 Recommendation Performances

In Table 1, the AUC and NDCG@5 results of the methods on the two datasets are demonstrated. In Fig. 3, the precision-recall curves and NDCG@n (with n ranging 3 to 10) results are plotted. From the results, we can see that our BayDNN outperforms greatly over the baselines on all the data sets and evaluation measures. On Epinions and Slashdot, it outperforms the best baseline BPRMFN respectively by 4.6% and 4.8% on NDCG@5, and much higher than the other baseline models.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Epinions</th>
<th>Slashdot</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>NDCG@5</td>
<td>AUC</td>
</tr>
<tr>
<td>Katz</td>
<td>0.751</td>
<td>0.554</td>
</tr>
<tr>
<td>Adamic/Adar</td>
<td>0.858</td>
<td>0.680</td>
</tr>
<tr>
<td>MF</td>
<td>0.888</td>
<td>0.838</td>
</tr>
<tr>
<td>NMF</td>
<td>0.934</td>
<td>0.897</td>
</tr>
<tr>
<td>BPRMF</td>
<td>0.948</td>
<td>0.928</td>
</tr>
<tr>
<td>PINN</td>
<td>0.843</td>
<td>0.746</td>
</tr>
<tr>
<td>BayDNN</td>
<td><strong>0.974</strong></td>
<td><strong>0.974</strong></td>
</tr>
</tbody>
</table>

Figure 3: The Precision-Recall curves and NDCG@n results of different methods on the two datasets.

For the baselines, comparing the two graph based methods Katz and Adamic/Adar, Adamic/Adar performs better than Katz as it extracts more precise features by paying attention to features of the neighbors, compared to Katz which only utilizes the features of the path between two users. Whereas comparing the two graph based methods with the other methods which learn feature representations, We see that overall the feature learning methods—MF, NMF, BPRMF, PINN and BayDNN clearly outperform the graph methods. Matrix factorization based algorithms show a big advantage over the graph-based algorithms that only uses the topology features. Taking into account the non-negative property of friendship, NMF outperforms MF consistently over all data sets, which verifies the value of non-negative.
of considering the specific property of tasks when designing the learning model.

What’s worth noting is that the deep neural network model PINN underperforms most algorithms, which confirms our previous intuition that straightforwardly applying DNN to friend recommendation without carefully designed network structure and pre-training strategy is far from sufficient.

Our BayDNN model shows a remarkable advantage over these baseline algorithms by combining the ranking idea of BPR and a finely designed DNN with our pre-training strategy. On Epinions dataset, in terms of NDCG@5, it achieves a 29% improvement over Adamic/Adar, 8% over NMF and 5% over BPRMF. With n increasing, the gaps become smaller. AUC also stated the same result.

### 4.3 Discussions

In this section, we discuss the effects of several different design parts in our BayDNN model from the following aspects: 1) How does the CFM (Convolutional Feature Mapping) part affect the learning? 2) How does the user bias affect the learning? 3) How does the pre-training affect the learning? 4) We further investigate the effect of negative friendship on learning by changing the way negative friendship are generated. To conduct the above three types of study, we compare the proposed BayDNN with several different variants of BayDNN.

In order to further study the effects of our deep convolutional neural network, we firstly replace the CFM part of BayDNN with a full connection neural network, to construct a method that combines BPR and a common neural network, i.e., BayNN, and compare our BayDNN method with it. The feature extraction part of BayNN works similarly as PINN.

To study the effect of user bias modeling, we compare BayDNN with BayDNN_{b}, which shares the same network structure with BayDNN except that, a uniform bias b for all the users is used in BayDNN_{b} rather than a unique b_{u} for each user for in BayDNN.

To study the effect of pre-training on learning, we compare BayDNN with two other models BayDNN (raw) and BayDNN (SAE) with different pre-training strategies. While BayDNN (raw) does not use any pre-training strategies, BayDNN (SAE) uses the stacked auto-encoder pre-training strategy proposed by [7]. It applies autoencoding on each layer of the neural network.

In BayDNN, the negative friendship are recognized as those user pairs who can’t reach each other within 6 hops. To study the effect of negative friendship on learning, we first compare BayDNN with BayDNN_{u} which uses no negative friendship during learning. We also change the way the negative friendship defined by varying the number of hops in \{0, 4, 6, 8\} to select negative friends.

The overall comparison results of the above 4 variants and BayDNN are shown in Table 2 and Fig. 4. From Table 2 and Fig. 4, our proposed BayDNN is optimal in all the three aspects. In the following, we discuss in more details from the respective aspects.

**How does the deep structure affect the learning?** We firstly study the effect of deep structure on BayDNN comparing to PINN with the same BPR learning strategy. Though with the help of BPR BayNN improves a bit than PINN (1.5% in AUC), it is still much worse than BayDNN (12% in AUC). We infer that the CFM part influence a lot because our convolution neural network extract better latent feature on link information rather than shallow architecture under the same learning strategy.

**How does the user bias affect the learning?** We then look into the effect of b_{u}, i.e., the personalized bias in Eq.7, by comparing BayDNN with BayDNN_{u}. Results in Table 2 show that our BayDNN with individual bias b_{u} for each user outperforms BayDNN_{u} with one single bias for all users 1% on average. Here we study the effects of the bias parameters by investigating the output value of the output layer of BayDNN and BayDNN_{u}. Fig. 5 depicts how the user bias influence the prediction value r_{ui} that will later be used for ranking.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Results</th>
<th>Epinions AUC</th>
<th>Epinions NDCG@5</th>
<th>Slashdot AUC</th>
<th>Slashdot NDCG@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>BayNN</td>
<td>0.855</td>
<td>0.765</td>
<td>0.852</td>
<td>0.658</td>
<td></td>
</tr>
<tr>
<td>BayDNN_{b}</td>
<td>0.964</td>
<td>0.957</td>
<td>0.959</td>
<td>0.945</td>
<td></td>
</tr>
<tr>
<td>BayDNN (raw)</td>
<td>0.821</td>
<td>0.788</td>
<td>0.814</td>
<td>0.767</td>
<td></td>
</tr>
<tr>
<td>BayDNN (SAE)</td>
<td>0.912</td>
<td>0.869</td>
<td>0.897</td>
<td>0.845</td>
<td></td>
</tr>
<tr>
<td>BayDNN_{u}</td>
<td>0.963</td>
<td>0.957</td>
<td>0.956</td>
<td>0.946</td>
<td></td>
</tr>
<tr>
<td>BayDNN</td>
<td>0.974</td>
<td>0.974</td>
<td>0.965</td>
<td>0.957</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: The AUC and NDCG@5 results of self comparison of BayDNN.
Figure 5: $r_{ui}$ distribution with a uniform bias, i.e., $r_{ui} = \sigma(O_u^T O_i + b)$, against $r_{ui}$ distribution with individual bias for each user, i.e., $r_{ui} = \sigma(O_u^T O_i + b_u + b_i)$, on Epinions dataset.

in the model on the Epinions dataset. From the figure we can see that the setting of personalized bias makes the values more distributed rather than within the range of a single sigmoid function, which brings in more flexibility and better performance.

Figure 6: The iterative AUC results during SGD optimization for BayDNN with different pre-training strategies on the Epinions dataset.

How does pre-training affect learning? We further validate our pre-training strategy by comparing the results with BayDNN (raw) and BayDNN (SAE). Table 2 and Fig. 4 show that without any pre-training strategy, the performance of BayDNN (raw) drops significantly. With a stacked auto-encoding pre-training BayDNN (SAE) improves largely over BayDNN (raw). In terms of NDCG@5, our strategy achieves a 15% improvement over BayDNN (raw) and 6% over BayDNN (SAE).

Moreover, we study the effects of pre-training strategies by investigating their step-wise performance during the SGD optimization. Fig. 6 depicts the iterative AUC results of BayDNN, BayDNN (raw) and BayDNN (SAE) during the SGD optimization process on the Epinions dataset. We can see that the proper pre-training strategy of BayDNN gives the model a very advantaged initialized space and the model converges very quickly in about 15 epochs, whereas BayDNN (raw) and BayDNN (SAE) start from a very low point and take a long process to converge, i.e., about 200 epochs. To this respect we can say that with our pre-training design, the time cost of the model is largely reduced by 92.5%.

Figure 7: Results changed by $k$ in sampling for $N$.

How does the negative friendship affect learning? The result comparing BayDNN with BayDNN-$U$ (i.e., no $N$) is given in Table 2 and Fig. 4 again, which depicts that it brings in 1% improvement in terms of NDCG. Here we further investigate how the introduction of negative friendship $N$ works and how it changes with different $k$ options, i.e., how many unreachable hops should it be in $N$ selection.

Moreover, let $k$ be the unreachable hop number we set, Fig. 7 depicts the precision results in the two datasets, respectively, with $k = 0$ (randomly select), $k = 4, 6, 8$ (unreachable by 4, 6, 8 hops). The results are consistent with the theory proposed by [21], i.e., as the number increases $k = 6$ reaches the best value of performance. The AUC and NDCG performance value with $k = 6$ are also given in Table 2. An interesting finding is that random selection ($k = 0$) is worse than no selection, which indicates that random selection brings in noise data (examples that do not actually belong to $N$) that is big enough to harm the performance.

Moreover, we see how it impact on the best performed baseline method, BPRMF. The results show that by introducing $N$ it works better than random selection and no $N$ selection, too. The figures show that BayDNN outperforms BPRNN in both situation (with or without $N$).

5 CONCLUSIONS

In this paper, we study the problem of friend recommendation and propose a novel BayDNN model by combining Bayesian Personalized Ranking and Deep Neural Networks. In BayDNN, we use an one-dimensional convolutional neural network (CNN) to extract latent deep structural feature representations from the input network data, and then use a Bayesian personalized ranking learning to captures users’ preference based on the extracted deep features.
BayDNN shows clearly superior performance to several state-of-the-art methods for friend recommendations on two different public datasets. To avoid poor parameter estimation, we also present a fine-tuned pre-training strategy for BayDNN based on Poisson and Bernoulli probabilistic models, respectively for different layers.

As the general idea of using deep neural networks for friend recommendation represents an interesting research direction, there are many potential future directions of this work. First, it is interesting to study incrementally learning the deep neural networks so that we can involve online user feedback into the learning process. Second, another potential is to infer the fined category (e.g., family, friend, and colleague) of social relationships [40]. Last, it would be interesting to connect the study with social theories to further understand how the topological structure of online social networks is formed.

ACKNOWLEDGMENTS

The authors would also like to thank the anonymous referees for their valuable comments and helpful suggestions. The work is supported by the National Natural Science Foundation of China under Grant No.: 61333014.

REFERENCES

[27] Yafeng Ren, Yue Zhang, Meishan Zhang, and Donghong Ji. 2016. Context-Sensitive Twitter Sentiment Classification Using Neural Network. AAAI.