Bayesian Heteroskedastic Choice Modeling on Non-identically Distributed Linkages

Liang Hu[1,2], Wei Cao[2], Jian Cao[1,*], Guandong Xu[2], Longbing Cao[2], Zhiping Gu[3]
[1]Shanghai Jiaotong University, [2]University of Technology Sydney, [3]Shanghai Electronic Information Institute
{lianghu, cao-jian}@sjtu.edu.cn, wei.cao@student.uts.edu.au, {guandong.xu, longbing.cao}@uts.edu.au, guzhiping@stiei.edu.cn

***Abstract*—Choice modeling (CM) aims to describe and predict choices according to attributes of subjects and options. If we presume each choice making as the formation of link between subjects and options, immediately CM can be bridged to link analysis and prediction (LAP) problem. However, such a mapping is often not trivial and straightforward. In LAP problems, the only available observations are links among objects but their attributes are often inaccessible. Therefore, we extend CM into a latent feature space to avoid the need of explicit attributes. Moreover, LAP is usually based on binary linkage assumption that models observed links as positive instances and unobserved links as negative instances. Instead, we use a weaker assumption that treats unobserved links as pseudo negative instances. Furthermore, most subjects or options may be quite heterogeneous due to the long-tail distribution, which is failed to capture by conventional LAP approaches. To address above challenges, we propose a Bayesian heteroskedastic choice model to represent the non-identically distributed linkages in the LAP problems. Finally, the empirical evaluation on real-world datasets proves the superiority of our approach.**

Keywords—link analysis and prediction, heteroskedastic choice model, non-IID Bayesian analysis, parallel Gibbs sampling

I. INTRODUCTION

Choice Modeling (CM) has proven to be effective for policy, labor, health, marketing, economics and psychology research over the decades [1]. The goal of CM is to model the decision process of a subject's choices among a set of options where the subjects refer to customers and options refer to products. As a result, CM can predict choices on the basis of the attributes of subjects and options. Link analysis and prediction (LAP) is a prominent topic in the data mining, for example, social network analysis studies linkages between people (on a unipartite graph) and collaborative filtering (CF) that studies linkages between users and their preferred items (on a bipartite graph).

CM studies decision procession to generate links between subjects and options, while LAP can also be considered a puzzle of modeling the factors of entities that lead to the choices of link formation. Therefore, it is possible to bridge CM to deal with the LAP problems but we need to remove some barrier between them. the only available data are links but the attributes of entities are often unavailable in real world, e.g., user attributes are often inaccessible in recommender systems due to privacy. Motivated by the prevalence of latent variable models [2, 3], we extend CM to model the attributes of subjects and options in a latent feature space. Moreover, real-world links between subjects and options are usually long-tail [4] distributed because different subjects may have their specialized choices. However, most current LAP methods assume independent and identically distributed (IID) linkages which may fail to capture the heterogeneity of choices between subjects and options. Inspired by heteroskedastic choice model [1, 5], we model choices, i.e. links in LAP, with non-identically linkage assumption so as to overcome above deficiency.

In this paper, we propose a *latent variable based Bayesian heteroskedastic choice model* (BHCM) where the term "*latent variable*" refers to three aspects: (1) *latent features* of subjects and options (2) *latent groups* of subjects and options (3) *latent utility* of each choice; and the term "*heteroskedastic*" points out the modeling of choices (linkages) under a non-IID assumption.

II. HETEROGENEITY OF LINKAGES

To get a deep insight into the motivation of BHCM, we first need to understand the nature of real-world data distribution, and the deficiency of current LAP methods under IID assumption.

A. Long-tail Distributed Linkages in Real World

It is known that most real-world data are often long-tail [4] distributed. In the LAP problem, we can often observe such a phenomenon: the minority of entities are associated with many links while the majority of entities are only associated with few links. From CM view, links correspond to choices and entities correspond to subjects and options. Formally, if a subject is associated with many choices, we define it as a ***core subject***, else it is defined as a ***trivial subject***. Similarly, if an option is chosen by many subjects, we define it as a ***core option***, else it is defined as a ***trivial option***.

Latent feature based approaches have become dominant in LAP [2, 6]. As illustrated in Fig. 1 (a), probabilistic matrix factorization (PMF) [7] is such a typical latent feature model which minimizes the negative log-joint-likelihood w.r.t. the normally distributed user feature vector $\boldsymbol{U}_i$ and item vector $\boldsymbol{V}_j$:

$$L = -\left[\sum_{ij} logP_{\sigma}\left(Y_{ij}\middle|\boldsymbol{U}_i, \boldsymbol{V}_j\right) + \sum_i logP_{\theta_U}(\boldsymbol{U}_i) + \sum_j logP_{\theta_V}(\boldsymbol{V}_j)\right]$$
$$\text{where } \boldsymbol{U}_i \overset{iid}{\sim} N(\boldsymbol{\theta_U}), \boldsymbol{V}_j \overset{iid}{\sim} N(\boldsymbol{\theta_U}) \text{ and } \boldsymbol{\theta_U} = \{\boldsymbol{\mu_U}, \sigma_{\boldsymbol{u}}\} \quad (1)$$

where $P\left(Y_{ij}\middle|\boldsymbol{U}_i, \boldsymbol{V}_j\right)$ acts as a loss function for fitting a rating Y_{ij} and $P_{\theta_U}(\boldsymbol{U}_i)$ serves as a Tikhonov regularizer $\lambda\|\boldsymbol{U}_i - \boldsymbol{\mu_U}\|^2$ where $\boldsymbol{\mu_U}$ is often assumed zero mean [7]. From Eq. (1), we can find that the $\boldsymbol{\mu_U}$ is heavily determined by core subjects because they account for the majority of data for estimates. The regularization term $\lambda\|\boldsymbol{U}_i - \boldsymbol{\mu_U}\|^2$ shrinks $\boldsymbol{U}_i$ towards $\boldsymbol{\mu_U}$. If a trivial subject has similar preferences to core subjects, such shrinkage is reasonable. However, if a trivial subject has heterogeneous preferences, such shrinkage may be undesirable. Since a trivial subject accounts for few data, the shrinkage

*Jian Cao is the corresponding author

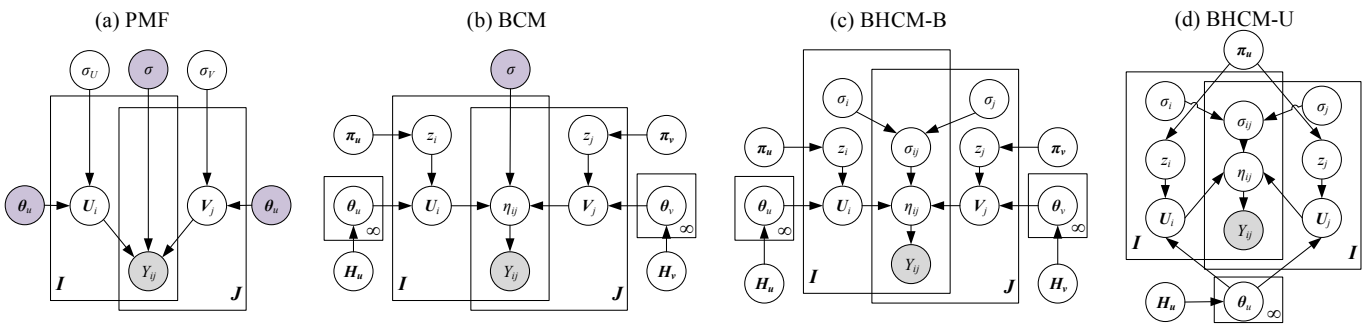

Fig. 1. The graphical representations of four models, where all hyperparameters are omitted for concision. (a) PMF models IID latent features and homoscedastic error (b) BCM models group-specific distributed lantent features and homoscedastic error; (c) BHCM-B models group-specific distributed lantent features and heteroskedastic error; (d) BHCM-U is a special case of BHCM-B with symmetric latent features (for undirected unipartite graph).

caused by regularization may overwhelm the estimates of the $\boldsymbol{U}_i$ by minimizing the loss (cf. Eq. (1)). It results in the failure to represent the heterogeneity of subjects. We address the above issue by estimating the $\boldsymbol{U}_i$ around a group mean $\boldsymbol{\mu}_g$ where all members are homogeneous in this group. Hence, the latent features $\boldsymbol{U}_i$ of subjects are drawn from their group-specific distributions instead of a global distribution. In fact, the degrees of heterogeneity in different datasets may be quite different, it is hard to manually specify the number of groups. Hence, we can employ Bayesian nonparametric prior to determine the number of groups adaptively. Similarly, it learns the latent features $\boldsymbol{V}_j$

B. Link Formation by Heteroskedastic Choice

In LAP problems, the binary linkage is usually assumed, i.e. observed links as positive instances and unobserved links as negative instances. For example, recommender systems often treated purchased items as positive instances with unpurchased as negative ones [6]. We argue that such a binary linkage assumption may turn out to be too strong, because unobserved links in many cases often are not truly negative instances, e.g., an author does not cite a paper because she is not aware of it rather than purposely omitting it. To address this issue, one may build LAP model under a weaker assumption that treats unobserved links as pseudo-negative instances instead of true. We call it unary linkage assumption since only observed links are treated as true instances.

Classic choice models were also built on binary linkages, i.e., choice/not-choice over each pair of subject and option so it needs to be revised to capacitate unary linkages. Intuitively, the true-positive choices can surely reflect the subjective decision whereas the decision on pseudo-negative choices is unsure, i.e., dislike or unawareness. Hence we model them via different priors [8]: informative priors, with small variances, are placed on true-positive choices while less informative priors, with larger variances, are placed on pseudo-negative choices. Intuitively, the choice made by a core subject is more informative than the choice made by a trivial subject, because the choice made by a core subject implies less randomness. Similarly, a choice made on a core option is less random than that made on a trivial option. Therefore, we place more informative priors on the choices associated with core subjects or options and less informative priors on the choices associated with trivial ones. The above analysis implies the non-IID nature of choices, i.e. linkages. Specially, we borrow the concept from heteroskedastic choice model [1, 5] to model the heteroskedastic errors over linkages.

III. Models

Discrete choice modeling often consists of two interrelated tasks: specification of the behavioral model and estimation of the parameters of that model [1]. Before presenting our BHCM, we first describe the preliminary about dichotomous choice.

A. Preliminary

In general, discrete choice models are often derived from random utility model (RUM) where the choice making is assumed to maximize utility [1]. On the basis of utility theory [9], let's consider a latent variable η_{ij} to model the utility of a choice by: $\eta_{ij} = \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_{ij}$, where $\boldsymbol{x}_{ij}$ is a feature vector consisting of the attributes of subject i and option j, $\boldsymbol{\beta}$ is a parameter vector to quantify the utility of each attribute of $\boldsymbol{x}_{ij}$, and ε_{ij} is the error term. Each observation $\mathbf{Y}_{ij}$ is related to the latent utility η_{ij} associated with a threshold parameter τ:

$$\mathbf{Y}_{ij} = \begin{cases} 1 & if\ \eta_{ij} > \tau \\ 0 & otherwise \end{cases}$$

That is, if the utility exceeds τ, the subject i chooses the option j. Typically, τ is set zero for binary data. Then, the probability of such a dichotomous choice can be given by:

$$\begin{aligned} \mathrm{P}\left(\mathbf{Y}_{ij} = 1 \middle| \boldsymbol{x}_{ij}\right) &= \mathrm{P}\left(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_{ij} > 0\right) \\ &= 1 - \mathrm{P}\left(\varepsilon_{ij} \le -\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}\right) = 1 - CDF\left(-\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}\right) \end{aligned} \quad (2)$$

where $CDF(\cdot)$ stands for some cumulative distribution function (CDF). For the binary case, a probit or logit function [10] is often chosen. In this paper, we choose the probit model, i.e., the CDF $\Phi(\cdot)$ of a normal distribution is used in Eq. (2), because it can provide a close-form inference for our model. Specially, we can write Eq. (2) as $\Phi\left(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}\right)$ due to the symmetry of normal distribution, i.e. $1 - \Phi\left(-\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}\right) = \Phi\left(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}\right)$.

The probit model assumes standard normally distributed error $\varepsilon_{ij} \sim N(0,1)$. That is, homoscedastic errors with constant variance $\sigma_{ij}^2 = 1$ are assumed over all choices. However, the parameter estimates will be biased and inconsistent if the errors are heteroskedastic [5]. Some researchers have proposed using a parametric model to avoid such biased estimation caused by the heteroskedasticity [11]. By modeling heteroskedastic error w.r.t. each choice, i.e. non-constant σ_{ij}^2, we can obtain Eq. (3) where σ_{ij} is often determined by some parametric function $f\left(\boldsymbol{\theta}_{ij}\right)$ [11] and $\boldsymbol{\theta}_{ij}$ may be related to subject or option. Given

$i \in \mathcal{I}$ to index subjects and $j \in \mathcal{J}$ to index options, the likelihood is given by Eq. (4).

$$\Phi\left(s_{ij}/\sigma_{ij}\right) = \mathrm{P}\left(\varepsilon_{ij}/\sigma_{ij} < \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}/\sigma_{ij}\right) \quad (3)$$

$$L(\boldsymbol{\beta}|\mathbf{Y}) = \prod_{i\in\mathcal{I},j\in\mathcal{J}} \Phi\left(s_{ij}/\sigma_{ij}\right)^{\mathbf{Y}_{ij}}\left[1-\Phi\left(s_{ij}/\sigma_{ij}\right)\right]^{1-\mathbf{Y}_{ij}} \quad (4)$$

B. Model Specifications

Given the above dichotomous choice model, it is possible to learn the parameter $\boldsymbol{\beta}$ by maximizing the likelihood (Eq. 4). However, explicit attributes $\boldsymbol{x}_{ij}$ are not always in the LAP problems. This can be handled by latent variable models through modeling subjects and options using latent features [2, 12]. Here, we denote $\boldsymbol{U}_i \in \mathbb{R}^d$ as the latent feature vector of subject i and $\boldsymbol{V}_j \in \mathbb{R}^d$ as the latent feature vector of option j. Then, we can immediately obtain the latent utility:

$$\eta_{ij} = \boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j + \varepsilon_{ij} \text{ where } \varepsilon_{ij} \overset{iid}{\sim} N(0,\sigma^2)$$

Moreover, in Section II, we argue that the latent features of each subject or option should be drawn from a group-specific distribution instead of a global one due to the heterogeneity. Therefore, we employ the Dirichlet Process (DP) [13] as a non-parametric prior to determine the number of groups adaptively and generate parameters for the corresponding group-specific distributions. As illustrated in Fig. 1 (b), the model BCM extends the dichotomous choice model with DP so as to generate group-specific latent features for both subjects and options.

BCM assumes the homoscedastic error over binary linkages. That is, the errors are IID standard normally distributed, i.e. $\varepsilon_{ij} \overset{iid}{\sim} N(0,1)$, to model all choices. However, homoscedastic error assumption is improper due to the long tail phenomenon. To tackle with this issue, we can model such linkages based on heteroskedastic choices, i.e. the utility of each choice is non-IID:

$$\eta_{ij} = \boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j + \varepsilon_{ij} \text{ where } \varepsilon_{ij} \sim N\left(0,\sigma_{ij}^2\right)$$

where σ_{ij}^2 varies with each choice instead of a constant. Theoretically, the larger variance σ_{ij}^2 means the more diffuse distribution, so it implies lower confidence level on making that choice. Therefore, we can model the error of positive choices with a small variance σ_{ij}^2 whereas a larger variance σ_{ij}^2 is used to model the error of pseudo-negative choices (unchosen data).

As presented in Section II, the uncertainty of each choice is related to both the choice maker and the option itself. For example, a core subject is more certain to make the choices or not-choices while a trivial subject tends to make the choice with more uncertainty. Similarly, the certainty of choices on the core and trivial options are also different. Therefore, we need to represent a heteroskedastic error for each choice, namely $\varepsilon_{ij} \sim N\left(0,\sigma_{ij}^2\right)$, where the variance σ_{ij}^2 is determined by both subject i and option j as depicted by the model BHCM-B in Fig. 1 (c). Specially, we can place different priors on the variance parameters [8] w.r.t. true positive choices and pseudo-negative choices. More detail is discussed in the following subsection.

Moreover, let's consider the LAP on an undirected unipartite graph where subjects and options are in an identical set with the symmetric links, so we should also enforce the symmetry of latent feature vectors. That is, the same latent feature vectors serve for both subject and options. Fig. 1 (d) shows such a symmetric model BHCM-U (applied to undirected unipartite graph), which is a variant of BHCM-B with symmetric features.

TABLE I. GENERATIVE PROCESS FOR BHCM-B

1. Stick-breaking construction:
   $\boldsymbol{\pi_u}|\alpha_u \sim GEM(\alpha_u) \qquad \boldsymbol{\pi_v}|\alpha_v \sim GEM(\alpha_v)$
2. For each subject i:
   a. Sample a group assignment: $\mathrm{z}_i \sim \boldsymbol{\pi}_u$;
   b. Sample a latent feature vector:
      $\boldsymbol{U}_i|\,\mathrm{z}_i, \{\theta_k\}_{k=1}^{\infty} \sim N(\theta_{zi})$ where $\theta_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k\}$
   c. Sample heteroskedastic variance parameters:
      i. Sample variance for positive choices:
         $\sigma_{i1}^2 \sim IG([a + N_{i1}s(\Phi_1) - N_{i1}]/2\,, b/2)$
      ii. Sample variance for negative choices:
         $\sigma_{i0}^2 \sim IG([a + N_{i0}s(\Phi_0) - N_{i0}]/2\,, b/2)$
3. For each option j:
   a. Sample a group assignment: $\mathrm{z}_j \sim \boldsymbol{\pi}_v$;
   b. Sample a latent feature vector:
      $\boldsymbol{V}_j\,|\mathrm{z}_j, \{\vartheta_k\}_{k=1}^{\infty} \sim N\left(\vartheta_{z_j}\right)$ where $\vartheta_k = \left\{\dot{\boldsymbol{\mu}}_k, \dot{\boldsymbol{\Lambda}}_k\right\}$
   c. Sample heteroskedastic variance parameters:
      i. Sample variance for positive choices:
         $\sigma_{j1}^2 \sim IG\left(\left[a + N_{j1}s(\Phi_1) - N_{j1}\right]/2\,, b/2\right)$
      ii. Sample variance for negative choices:
         $\sigma_{j0}^2 \sim IG\left(\left[a + N_{j0}s(\Phi_0) - N_{j0}\right]/2\,, b/2\right)$
4. For each subject-option pair $\langle i,j \rangle$:
   a. Sample latent utility ($\boldsymbol{V}_j$ is replaced by $\boldsymbol{U}_j$ for BHCM-U, $\sigma_{ij,1}^2$ and $\sigma_{ij,0}^2$ are set to 1 for BCM) :
      $$\eta_{ij} \sim \begin{cases} N\left(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j, \sigma_{ij,1}^2\right) & \delta_{ij} = 1 \\ N\left(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j, \sigma_{ij,0}^2\right) & \delta_{ij} = 0 \end{cases}$$
      where $\sigma_{ij,1}^2 = \left(\sigma_{i1}^2 + \sigma_{j1}^2\right)/2,\ \sigma_{ij,0}^2 = \left(\sigma_{i0}^2 + \sigma_{j0}^2\right)/2$
   b. Set link: $\mathbf{Y}_{ij} = \begin{cases} 1 & if\ \eta_{ij} > 0 \\ 0 & otherwise \end{cases}$

C. Bayesian Specificaion and Intepretaion

We can write down the generative process of the choices w.r.t. BHCM-B (BCM and BHCM-U are sub-models which can be generated similarly) in Table I where we introduce a set of binary variables δ_{ij} to indicate true-positive or pseudo-negative choice:

$$\delta_{ij} = \begin{cases} 1 & < i,j > \text{is a true positive choice} \\ 0 & < i,j > \text{is a pseduo negtive choice} \end{cases}$$

In Table I, $GEM(\alpha)$ stands for a stick-breaking process for DP [13]. $N_{i1} = |\boldsymbol{i1}|$ and $N_{i0} = |\boldsymbol{i0}|$ where $\boldsymbol{i1}$ stands for the positive choices and $\boldsymbol{i0}$ denotes the negative choices made by subject i. $\boldsymbol{j1}$ and $\boldsymbol{j0}$ are similarly defined w.r.t. option j. $s(\Phi)$ is a function with the parameters Φ. $IG(a,b)$ is an inverse-gamma distribution [14]. Due to conjugacy of the normal-gamma [14], we can easily obtain the posteriors of σ_{i1}^2 and σ_{i0}^2:

$$\sigma_{i1}^2|\boldsymbol{\eta_{i1}}, \boldsymbol{U}_i, \boldsymbol{V} \sim IG\left(\left[a + N_{i1}s(\Phi_1)\right]/2\,, b + \textstyle\sum_{j\in\boldsymbol{i1}} \varepsilon_{ij}^2/2\right)$$

$$\sigma_{i0}^2|\,\boldsymbol{\eta_{i0}}, \boldsymbol{U}_i, \boldsymbol{V} \sim IG\left(\left[a + N_{i0}s(\Phi_0)\right]/2\,, b + \textstyle\sum_{j\in\boldsymbol{i0}} \varepsilon_{ij}^2/2\right) \quad (5)$$

where $\varepsilon_{ij} = \eta_{ij} - \boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j$. The mode of $IG(\alpha,\beta)$ is $\beta/(\alpha+1)$, so if we set $a = -1$ and $b = 0$ in Eq. (5), then we can obtain a very simple form of the mode:

$$M(\sigma_{i1}^2) = \bar{\sigma}_{i1}^2/s(\beta_1,\omega_1,\gamma_1) \quad M(\sigma_{i0}^2) = \bar{\sigma}_{i0}^2/s(\beta_0,\omega_0,\gamma_0) \quad (6)$$

Each mode is a fraction where the numerator is the sample variance, i.e., $\bar{\sigma}_{i1}^2 = \sum_{j\in \boldsymbol{i1}} \varepsilon_{ij}^2/N_{i1}$ and $\bar{\sigma}_{i1}^2 = \sum_{j\in \boldsymbol{i0}} \varepsilon_{ij}^2/N_{i0}$, and the denominator is a function $s(\Phi)$. Here we define $s(\Phi)$ as a generalized logistic function [15]:

$$s(\Phi) = s(\beta,\omega,\gamma) = \frac{\omega-\gamma}{1+e^{-\beta(N_{i1}-1)}} + \gamma \quad (7)$$

where $\beta > 0$ controls the rate varying with N_{i1}. It is easy to see that Eq. (7) has the upper bound ω when N_{i1} is large and has a minimum value $(\omega+\gamma)/2$ when $N_{i1} = 1$. We can let $s(\Phi_1) > s(\Phi_0)$ by a larger ω_1 and a smaller ω_0 so as to differentiate the scale of $\bar{\sigma}_{i1}^2$ and $\bar{\sigma}_{i0}^2$. As a result, the mode $M(\sigma_{i1}^2)$ tends to be small and $M(\sigma_{i0}^2)$ tends to be large, cf. Eq. (6). Since the values of $\sigma_{i1}^2, \sigma_{i0}^2$ are more probably drawn around the modes, the utility η_{i1} of true-positive choices tend to associate with informative priors (i.e. a small σ_{i1}^2) while the utility η_{i0} of pseudo-negative choices tend to associate with less informative priors (i.e. large σ_{i1}^2). That is, it places a more informative prior on a core subject i's choices due to the larger N_{i1} while less informative prior on a trivial subject's choices. In the similar way, we can interpret priors on the variance parameters $\sigma_{j,1}^2, \sigma_{j,0}^2$ from the perspective of options so as to differentiate core options and trivial options.

As a result, the subject-oriented utility η_{ij}^s and the option-oriented utility η_{ij}^o of each choice (i,j) are respectively distributed as follows:

$$\eta_{ij}^s \sim \begin{cases} N(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j, \sigma_{i,1}^2) \\ N(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j, \sigma_{i,0}^2) \end{cases} \quad \eta_{ij}^o \sim \begin{cases} N(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j, \sigma_{j,1}^2) & \delta_{ij} = 1 \\ N(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j, \sigma_{j,0}^2) & \delta_{ij} = 0 \end{cases}$$

Further, we can create a joint view of subject and option to measure the utility η_{ij}. Here, we use a convex combination of η_{ij}^s and η_{ij}^o by the parameter a to represent the η_{ij}:

$$\eta_{ij} = a\eta_{ij}^s + (1-a)\eta_{ij}^o$$

According to the property of normal random variables, if we set $a = 0.5$, we can immediately obtain the distribution of utility η_{ij} as Step 4 of the generative process in Table I.

IV. Learning and Inference

So far, we have presented the detail of BHCMs to model the heterogeneities for linkages. In order to conduct the prediction task, we first need to design an efficient algorithm to learn the model parameters.

A. Model Parameter Learning

In fact, exact inference is obviously intractable for BHCMs. However, its structure nicely lends itself to approximate inference via Markov Chain Monte Carlo (MCMC). Specially, we design a Gibbs sampler to draw samples in parallel for each step by taking advantage of the factorial conditional distribution over the parameters. In Algorithm I, we give a brief sampling scheme for BHCM-B (the sub-model BCM and BHCM-U can be sampled similarly). Here we omit the detail of each sampling step due to the limited space, which may refer to [10, 13, 14].

Theoretically, the speed of this algorithm is linear with the number of CPUs if not considering the overhead of data communication. That is, if we can sample the parameters w.r.t. each choice on a separate CPU in parallel, then sampling the parameters for all choices can be finished in approximately the same time as the case with one choice, since each step can be executed in parallel.

ALOGRITHM I. PRALLEL GIBBS SAMPLING SCHEME FOR BHCM-B

- Draw group assignment $z_i|\Phi\backslash z_i$ for each subject i using DP
- Draw group assignment $\mathrm{z}_j|\Phi\backslash z_j$ for each option j using DP
- Draw latent features $\boldsymbol{U}_i|\Phi\backslash\boldsymbol{U}_i$ for each subject i in parallel
- Draw latent features $\boldsymbol{V}_j|\Phi\backslash\boldsymbol{V}_j$ for each option j in parallel
- Draw $\sigma_{i1}^2, \sigma_{i0}^2|\Phi\backslash\sigma_{i1}^2, \sigma_{i0}^2$ for each subject i in parallel
- Draw $\sigma_{j1}^2, \sigma_{j0}^2|\Phi\backslash\sigma_{j1}^2, \sigma_{j0}^2$ for each option j in parallel
- Draw utility $\eta_{ij}|\boldsymbol{U}_i, \boldsymbol{V}_j, \sigma_{ij}^2$ for each choice $\langle i,j\rangle$ in parallel
- Draw $\theta_k|\{\boldsymbol{U}_i\}_{i\in s(k)}$ for each subject group k in parallel
- Draw $\vartheta_k|\{\boldsymbol{V}_j\}_{j\in O(k)}$ for each option group k in parallel

B. Inference

One of the main tasks of LAP is to infer the likelihood of new interactions between entities. From the CM view, it is equivalent to ranking the predictive choices in terms of their utility. Higher utility means higher probability that a subject will make that choice, i.e. generate a link. Given a subject i, the predictive distribution of the utility over option j is given by:

$$\mathrm{P}(\eta_{ij}|\mathbf{Y}) \propto \int N(\eta_{ij}|\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j, \sigma_{ij,0}^2) dP(\boldsymbol{U}_i) dP(\boldsymbol{V}_j) dP(\sigma_{ij,0}^2)$$

In the MCMC method, the predictive expectation of η_{ij} can be retrieved through the Monte Carlo approximation from S samples. In practice, we use the expectation of $\eta_{ij}^{(s)}$ w.r.t. each sample to avoid unnecessary sampling noise. Therefore, we can estimate the utility $\hat{\eta}_{ij}$ using by:

$$\hat{\eta}_{ij} = \mathbb{E}(\eta_{ij}) \approx \frac{1}{S}\sum_{s=1}^{S} \mathbb{E}\left(N(\eta_{ij}^{(s)}|\boldsymbol{U}_i^{(s)\mathrm{T}}\boldsymbol{V}_j^{(s)}, \sigma_{ij,0}^{2(s)})\right) \propto \sum_{s=1}^{S} \boldsymbol{U}_i^{(s)\mathrm{T}}\boldsymbol{V}_j^{(s)} \quad (9)$$

Now, let $\boldsymbol{c}$ denote the set of candidate options for subject i. Then, we can sort the utility $\{\hat{\eta}_{ik}\}_{k\in\boldsymbol{c}}$ in a descending order to retrieve the rank over predictive choices

V. Related Work

LAP problems are originally studied on a unipartite graph with one set of entities, e.g. people, webpages. As studied in this paper, probabilistic models are often designed to represent the presence or absence of links. Mixed membership stochastic block models (MMSB) [16] study the membership of each object using the relational between each pair of nodes, which have been applied to the LAP on social networks and protein interaction networks. Latent feature based matrix factorization (MF) [3, 7] methods are dominant in the CF area. In fact, most MF models, including the PMF model are built on real-value data, e.g. ratings, so they are not suitable to model the binary/unary linkages. Exceptionally, the maximum margin MF (MMMF) [17], which aim to learn latent features for the maximum large-margin prediction, can perform binary classification on linkage data on a bipartite graph, but it cannot be applied to LAP with the constraint of symmetric features, i.e. LAP on an undirected unipartite graph. Recently, some other

MF methods, such as the latent feature log-linear (LFL) model [12] and the supervised MF (SMF) [6] have been proposed to deal with LAP problems on both bipartite and unipartite graphs. However, all these LAP methods are implicitly designed under the IID assumption and do not consider the heterogeneity of linkages as focused on in this paper.

To avoid modeling latent features for all users or item with a single distribution, DPMF [18] is proposed to model the latent features with group-specific distributions governed by Dirichlet process, which is similar to the BCM, but DPMF is mainly used to deal with real-value data, e.g. ratings, while BCM studies the utility to generate a link. Moreover, some approaches have been proposed to deal with the unary linkages. Weighted MF (WMF) [19, 20] extends traditional MF with weighted loss, where the loss on fitting positive instances are penalized with a large weight while the negative ones are penalized with a much smaller weight. Bayesian personalized ranking (BPR) learns the preference ordering over each pair of items [21]. In fact, such an idea can be viewed as paired preference analysis in the CM [22].

VI. Experiments

We conducted experiments on three real-world datasets to cover three representative LAP problems studied in this paper.

A. Comparative Methods

PMF, MMMF, LFL, SMF, MMSB and WMF are used as the state-of-the-art methods for comparison because they are applicable to our testing problems and their code is publicly available. Specially, PMF models the links as a matrix with real-value ratings, i.e. 1 for observed links and 0 otherwise. In the experiments, we initialize the hyper parameters and the dimensionality of features for each model following the settings in the original papers, and then tune them by cross validation.

B. Evaluation Metrics

In following experiments we use three commonly accepted metrics for LAP evaluation: (1) area under the ROC curve (AUC); (2) Precision; and (3) Recall.

- *AUC* measures the probability that the rank of positive instances is higher than the rank of negative ones, where $\boldsymbol{C}^+/\boldsymbol{C}^-$ denotes positive/negative instances in the testing set and $\delta(rk(i)<rk(k))$ returns 1 if $rk(i)<rk(k)$ and 0 otherwise:

$$AUC = \left[\sum_{i\in\boldsymbol{C}^+}\sum_{k\in\boldsymbol{C}^-}\delta\left(rk(i)<rk(k)\right)\right]/[|\boldsymbol{C}^+|\cdot|\boldsymbol{C}^-|]$$

- *Rec@K* measures recall of the top *K* retrieved items.
- *Pr@K* measures precision of the top *K* retrieved items.

C. Social Relationship Prediction

The NIPS coauthorship dataset has been used to evaluate quite a few LAP models [2, 6, 12]. Here, we randomly extracted 512 authors who coauthored at least 3 publications with others. Then, we can obtain a 512×512 symmetric binary matrix where the entries with value 1 indicate observed coauthorships. We used the leave-one-out strategy to construct the testing dataset, that is, we randomly hold out one observed coauthorship as the positive instance and nine authors without observed coauthorships as the negative instances for each author.

The Epinions dataset is provided by the consumer review site *Epinions.com* where members of the site can decide whether to "trust" each other. Hence, we can construct a directed who-trusts-whom network. In this experiment, we randomly extracted 1082 users to construct a directed graph represented by a 1082×1082 asymmetric binary matrix. For the testing dataset, we adopted the similar leave-five-out strategy (5 positive instances combining with 45 negative instances are held out) over the users who trust at least 10 other users, i.e., the testing users originally have at least 10 outlinks.

For the NIPS dataset, we adopted BHCM-U to model the latent features of authors due to the coauthorships being an undirected unipartite graph over authors. For the Epinions dataset, we adopted BHCM-B to respectively model the latent features of trusters and trustees since trust relation is directed. Moreover, we set $\{\omega_1=4, \gamma_1=0, \omega_0=1, \gamma_0=0.8\}$ for the generalized logistic function of Eq. (7), which was tested to produce good results.

TABLE II. The AUC of comparative methods

Model	NIPS	Epinions
PMF	NA	0.7769±0.157
MMMF	NA	0.7682±0.148
LFL	0.6203±0.269	0.6250±0.169
SMF	0.6379±0.253	0.6529±0.165
MMSB	0.6651±0.237	0.7335±0.160
BCM	**0.7089±0.205**	**0.8043±0.136**
BHCM	**0.7355±0.183**	**0.8196±0.129**

The average AUCs and standard deviations are reported in Table II. Thanks to heteroskedastic CM technique, BHCM produces a significant improvement over other comparative methods. The reason is that people always have different backgrounds and interests, which results in long-tail distributed choices; however, all the baseline methods adopt the IID assumptions to model both latent features and linkages, which fail to capture the heterogeneity among subjects and options. In comparison, BHCM models the latent features by group-specific distributions and the linkages by heteroskedastic distributions so it is more capable of capturing the underlying heterogeneity. Specially, it can be found that BHCM outperforms BCM, which reveals the fact that the unobserved links do not always mean true negative instances in the real-world scenarios. Hence, the unary-linkage assumption adopted by BHCM is more suitable for these datasets than the binary-linkages adapted by other methods. Furthermore, we can find that the standard deviation of BHCM is the smallest among all comparative models, which illustrates that BHCM can provide much better representation to capture the heterogeneity. So far, all these reasons result in BHCM having the best performance.

Fig. 2 reports the AUC values over the users grouped by different numbers of observed links on the Epinions dataset. We find that BHCM considerably outperforms other models even when users have few observed links. This result again proves the advantage of our non-IID LAP model with the unary-linkage assumption. The major reason is that most users are trivial users with relatively few observed links (as illustrated on the left of Fig. 2), the model parameters tend to be learned from the remainder of negative links under the binary-linkage assumption used by other IID LAP models. Moreover, some models may lead to over-regularization when learning the latent features for those trivial users with few data (see the analysis in Section II). In comparison, BHCM naturally overcomes these limitations with the non-IID linkage assumption, achieving better and more

stable performance than other comparative methods no matter how many observed links are associated with users.

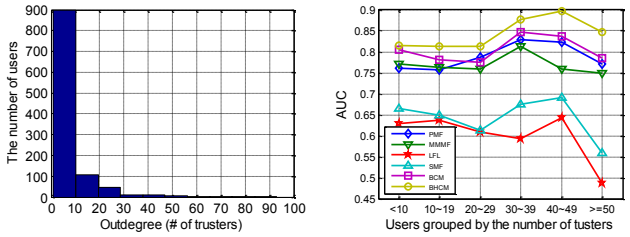

Fig. 2. Top: The long-tail linkage distribution of Epinions training set. Bottom: The results of AUC over users grouped by different numbers of trusters (links) for all comparative methods.

D. Item Recommendation

In a social networking site, it is only known what items users are interested in but there is often no data available to record what users dislike. Therefore, it is a typical unary-linkage based CF problem. In this experiment, we use the SNS data provided by KDD Cup 2012[1] where the items include users, groups, games, etc. We randomly sampled 2000 users and 1000 items, so we obtained a 2000×1000 matrix containing ones to indicate observed links. Then, we held out 20% of the observed links for each user as the ground truth for testing.

TABLE III. PR@5,10 AND REC@5,10 OF COMPARATIVE METHODS

Model	Pr@5	Pr@10	Rec@5	Rec@10
PMF	0.2074	0.1805	0.1669	0.2926
MMMF	0.2356	0.1987	0.1959	0.3051
LFL	0.1973	0.1760	0.1458	0.2723
WMF	0.2350	0.2075	0.2115	0.3272
BCM	**0.2437**	**0.2187**	**0.2450**	**0.3488**
BHCM	**0.2659**	**0.2334**	**0.2412**	**0.3577**

In this CF problem, the methods PMF, MMMF, LFL and BCM are built on binary-linkage assumption, i.e., the unchosen items are treated as true negative instances, whereas BHCM and WMF are constructed under unary-linkage assumption. In this experiment, we set $\{\beta_1^U=\beta_0^U=0.1, \beta_1^V=\beta_0^V=0.2\}$ for BHCM, and other parameters are set the same as previous experiments. As reported in Table III, BHCM achieves much better precision and recall than other binary-linkage assumption based LAP models. Specially, WMF outperforms PMF, because WMF models the linkages as a one-class CF problem, i.e. under the unary-linkage assumption. However, WMF underperforms BHCM, which may be attributed to three weak points: (1) WMF is a real-value based model; (2) The latent features of WMF are IID assumed; (3) WMF cannot adaptively find optimal weight parameters. Obviously, BHCM effectively addresses these weak points.

VII. CONCLUSION

In this paper, we present the model BHCM, which draws on the experience of choice modeling to deal with LAP problems. The core idea of BHCM is to model heterogeneity of linkages for LAP under a non-IID assumption. With such specifications, BHCM can elegantly model the heteroskedastic unary linkages which are ubiquitous in real world. The final experimental results manifest the sophistication of our models against other comparative state-of-the-art methods for different applications.

[1] http://www.kddcup2012.org/c/kddcup2012-track1

ACKNOWLEDGE

This work is partially supported by China NSF (Granted No. 61272438, 61472253), Research Funds of Shanghai Municipal Science and Technology Commission (Granted No. 14511107702, 12511502704).

REFERENCES

[1] K. Train, *Discrete choice methods with simulation*: Cambridge university press, 2003.

[2] K. T. Miller, T. L. Griffiths, and M. I. Jordan, "Nonparametric latent feature models for link prediction," in *Proceeding of Advances in Neural Information Processing Systems (NIPS)*, 2009, pp. 1276-1284.

[3] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer,* vol. 42, pp. 30-37, 2009.

[4] C. Anderson, *The long tail: Why the future of business is selling less of more*: Hachette Digital, Inc., 2006.

[5] L. Keele and D. K. Park, "Difficult choices: an evaluation of heterogeneous choice models," in *The 2004 Meeting of the American Political Science Association*, 2006, pp. 2-5.

[6] A. K. Menon and C. Elkan, "Link Prediction via Matrix Factorization," in *Machine Learning and Knowledge Discovery in Databases*. vol. 6912, D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, Eds., ed: Springer Berlin Heidelberg, 2011, pp. 437-452.

[7] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *Advances in neural information processing systems*, 2008, pp. 1257-1264.

[8] W. J. Browne, *Applying MCMC methods to multi-level models*: University of Bath, 1998.

[9] J. Von Neumann and O. Morgenstern, *Theory of games and economic behavior (commemorative edition)*: Princeton university press, 2007.

[10] J. H. Albert and S. Chib, "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association,* vol. 88, pp. 669-679, 1993/06/01 1993.

[11] R. M. Alvarez and J. Brehm, "American Ambivalence Towards Abortion Policy: Development of a Heteroskedastic Probit Model of Competing Values," *American Journal of Political Science,* vol. 39, pp. 1055-1082, 1995.

[12] A. K. Menon and C. Elkan, "A Log-Linear Model with Latent Features for Dyadic Prediction," in *Proceedings of the 2010 IEEE International Conference on Data Mining*, 2010, pp. 364-373.

[13] Y. W. Teh, "Dirichlet process," *Encyclopedia of machine learning,* pp. 280-287, 2010.

[14] K. P. Murphy, "Conjugate Bayesian analysis of the Gaussian distribution," *def,* vol. 1, p. 16, 2007.

[15] F. Richards, "A flexible growth function for empirical use," *Journal of experimental Botany,* vol. 10, p. 290, 1959.

[16] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, "Mixed Membership Stochastic Blockmodels," *J. Mach. Learn. Res.,* vol. 9, pp. 1981-2014, 2008.

[17] N. Srebro, J. D. M. Rennie, and T. Jaakkola, "Maximum-margin matrix factorization," in *Advances in neural information processing systems*, 2005, pp. 1329-1336.

[18] I. Porteous, A. U. Asuncion, and M. Welling, "Bayesian Matrix Factorization with Side Information and Dirichlet Process Mixtures," in *AAAI*, 2010.

[19] R. Pan, Y. Zhou, B. Cao, N. N. Liu, R. Lukose, M. Scholz, *et al.*, "One-class collaborative filtering," in *Eighth IEEE International Conference on Data Mining*, 2008, pp. 502-511.

[20] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," 2008, pp. 263-272.

[21] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, Montreal, Quebec, Canada, 2009, pp. 452-461.

[22] P. Courcoux and M. Semenou, "Preference data analysis using a paired comparison model," *Food Quality and Preference,* vol. 8, pp. 353-358, 9// 1997.