Demographic Inference via Knowledge Transfer in Cross-Domain Recommender Systems

Abstract—**User demographics such as age and gender are very useful in recommender systems for applications such as personalization services and marketing, but may not always be available for individual users. Existing approaches can infer users' private demographics based on ratings, given labeled data from users who share demographics. However, such labeled information is not always available in many e-commerce services, particularly small online retailers and most media sites, for which no user registration is required. We introduce a novel probabilistic matrix factorization model for demographic transfer that enables knowledge transfer from the source domain, in which users' ratings and the corresponding demographics are available, to the target domain, in which we would like to infer unknown user demographics from ratings. Our proposed method is based on two observations: (1) Items from different but related domains may share the same latent factors such as genres and styles, and (2) Users who share similar demographics are likely to prefer similar genres across domains. This approach can align latent factors across domains that share neither common users nor common items, associating user demographics with latent factors in a unified framework. We also develop an iterative algorithm for model parameter estimation and theoretically show its convergence. Experiments on cross-domain datasets demonstrate that the proposed method consistently improves demographic classification accuracy over existing methods.**

Index Terms—**Demographic inference, Recommender systems, Matrix factorization**

I. Introduction

User demographics are important attributes for enriching online services that include personalization, marketing, and targeted advertisement. However, demographic information is not always available for online users, typically because either they decline to provide it [1], [2] or the online service is not designed to collect it. Instead, user interactions such as ratings, clicks, and purchases in recommender systems can sometimes provide sufficient information to infer user demographic attributes. For example, a Netflix user's preference for family-oriented and occasional children's movies may indicate that the user is a parent. Existing attempts [3], [4] suggest that it is possible to infer user genders based on ratings with as high as 80% accuracy given labeled data from users who share demographic information in recommender systems. However, such labeled user demographic information is not always available in many e-commerce services, particularly small online retailers (e.g., an outdoor shopping site) and most media sites (e.g., Yahoo News), in which no user registration is required.

It is therefore very useful to transfer knowledge from the source domain, in which users' ratings and the corresponding demographics are available, to the target domain, in which we would like to predict demographics from ratings. Initial work in this area includes de-anonymization of movie ratings datasets [5] by matching rating patterns between the source domain IMDB database and the target domain Netflix, in which the user identities are inferred. The success of the approach is based on the assumption that a subset of common items is rated by common users in both domains. Other approaches such as [6] require that different domains have some auxiliary information such as item content features for linking and grouping users or items. However, due to privacy concerns, the sharing of user and item information may be limited in practice.

Our task is to infer demographics from ratings in a target domain by transferring the knowledge from a source domain with both ratings and user demographics. Note that this task is totally different from traditional transfer learning via user modeling, since no demographic information in the target domain is available and no entities (e.g., users, items) can be linked across domains. Demographic transfer learning under this scenario is possible based on two observations. First, when two domains such as movies and books are related, different items may share the same latent factors such as genres and styles. For example, the "The Matrix" movie and the "Neuromancer" book by William Gibson both belong to the 'cyberpunk' science fiction genre. Second, users who share similar demographics are more likely to prefer similar genres across domains. For example, multiple studies (e.g., [7], [8]) have identified group-level differences in movie and book preferences between men and women (e.g., male preference for action-adventure and sports themes vs. female preference for relationship-based themes).

Inspired by these observations, we propose a probabilistic cross-domain matrix factorization model called *Transfer Matrix Factorization* (TMF), which can effectively transfer the knowledge of user demographics from the source domain to the target domain. Traditional matrix factorization approaches characterize both users and items with latent factors but fail in cross-domain demographic inference tasks since the latent factors across different domains are not aligned and there is no association between demographic labels and latent factors. Our proposed method is based on the joint matrix factorization of two user-item rating matrices from different domains with an important twist: it characterizes a user profile as an integration of both a group-level profile that captures the preference of users within the same demographic group, and a personal profile that captures the personal preference of each user. The group-level profile is further decomposed into the product of two components: the user membership of demographic groups and the association between demographic groups and latent factors. Since both the latent factors and the association between demographic groups and latent factors are shared across domains, the knowledge from the source domain can be used to improve the demographic inference in the target domain.

To summarize, the main contributions of our work are: (1) Our model explores effectively the correlation between demographics and ratings across different domains that share neither common users nor common items and infers the demographics precisely without giving any information in the target domain. (2) We develop an iterative algorithm for this optimization and theoretically show its convergence. (3) Extensive experiments using real-world datasets demonstrate that our model can achieve higher classification accuracy than existing methods, regardless of the amounts of labeled users, the sparsity of ratings, and the difference of demographic distributions in the source and target domains.

II. Related Work

Private user demographics have been inferred from various online activities such as friendship on Facebook [9], [10], linguistic features of tweets [11]–[14], reviews [15], and location check-ins [16]. Other studies have explored the association between user demographics and ratings in recommender systems [2]–[4], [17]. Initial attempts (e.g., [3]) have shown that it is possible to infer the genders of users in recommender systems based solely on their ratings with as high as 80% accuracy. Most of the existing work assumes that a small fraction of labeled data is available from users who are willing to provide their demographics. However, such information is limited due to privacy concerns.

To overcome this lack of sufficient labeled data, initial attempts have explored the association between demographics and online activities through cross-domain social computing platforms [18], [19]. Studies including [20] have demonstrated that it is possible to infer user demographics from search queries by transferring the knowledge from labeled Facebook data to unlabeled search engine data. The method maps both search queries and Facebook 'Liked' pages into the same text categories and transfers the association between demographics and text categories between two domains. A variety of online activities such as social connections, temporal access patterns, and geographic tags [18], [19] have been used to identify users across different social platforms such as Twitter and Flickr.

Accurate demographic inference in cross-domain recommender systems based only on ratings is challenging due to the lack of content information such as review text and social connections. A few attempts have explored user identities through cross-domain recommender systems, assuming that a subset of common items are rated by common users in both domains. For example, the work on Netflix de-anonymization [5] matches rating patterns between Netflix rating data and the public IMDB database to infer user identities. Traditional matrix factorization based methods [21]–[23] have been used to explain ratings through latent factors in a single domain. Those methods seek to map users and items in a low-dimensional space to capture intrinsic similarities. However, latent factors are not aligned across different domains.

Transfer learning has been used in cross-domain recommender systems to predict ratings. For example, collective matrix factorization (CMF) [24] can be applied in cross-domain recommendation assuming that entities such as users and items are shared across domains. A recent study [6] has integrated auxiliary content information, such as user and item features, to improve recommendation accuracy. Another group of work has improved rating prediction in domains where neither items nor users are shared [25]–[30]. Some representative methods such as [25], [26] are rating generative models based on the assumption that ratings are drawn from a shared cluster-level model. Our work focuses on a different perspective of recommender systems where we would like to infer private user traits from ratings. The idea of transferring group-level knowledge is also applied to cross-domain document categorization [31], [32]. Specifically, these types of approaches extend previous work [26], [27] with document class labels and transfer the association between word clusters and document classes based on nonnegative matrix factorization. In comparison with document modeling, our work models each user with both group-level preference related to demographics and individual preference, which is more suitable for recommender systems.

Our model is inspired by constrained probabilistic matrix factorization (CPMF) [33] and its extension [34], but our approach is different in the following ways. First, the CPMF models user preferences in a single domain with observed metadata such as demographics, but ours models user preferences across domains where the associations between demographics and user latent features are shared and the demographics of the target domain are unknown. Second, in CPMF the demographic indicator of a user is assumed to be an observed binary. However, we generalize this to be the latent probability of a user belonging to one of the demographic clusters, which follows a normal distribution.

III. Model

In this section, we introduce our TMF model for inferring demographics from rating matrices in cross-domain recommender systems. The key innovation of our model is that a user profile is characterized as both a demographic dependent profile that can be shared across domains and a personal profile that captures the user's personal preference in each domain. We also present an efficient algorithm to optimize the objective function, together with a rigorous convergence proof.

A. Basic Concept and Notation

Throughout this paper, we denote the real number set and nonnegative real number set as $\mathbb{R}$ and $\mathbb{R}_+$, respectively. The element at the *i*-th row and *j*-th column of a matrix M is denoted by $M_{(ij)}$. Frequently used notation is summarized in Table I.

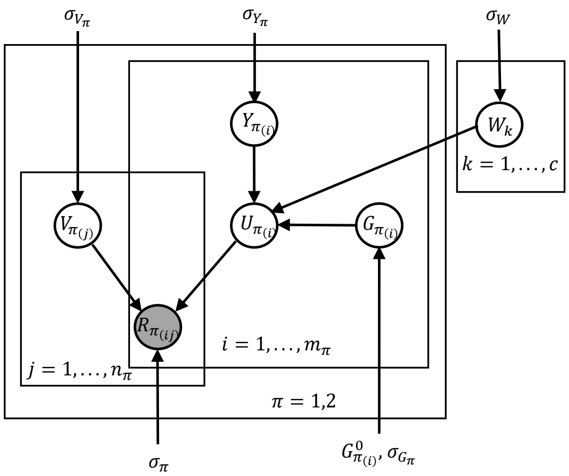


Fig. 1. The graphical model of Transfer Matrix Factorization (TMF).

TABLE I
FREQUENTLY USED NOTATION.

Notation	Description
π	domain indices, $\pi \in \{1, 2\}$
R_1, R_2	rating matrix in source, target domain
m_1, m_2	number of users in source, target domain
n_1, n_2	number of items in source, target domain
Y_1, Y_2	user personal feature matrix in source, target
V_1, V_2	item feature matrix in source, target
G_1, G_2	latent demographic matrix in source, target
G_1^0	true demographic indicator matrix in source
W	association matrix between demographics and latent features
σ_{Y_π}, σ_{V_π}	variance of Y_π, V_π
σ_{G_π}, σ_π	variance of G_π, R_π
σ_W	variance of W
α, β, γ, λ	regularization parameters
Γ, Λ	Lagrange multipliers

Given user ratings and demographic labels in the source domain, our goal is to predict the user demographic labels from ratings in the target domain. Note that domains share neither common users nor common items. Specifically, the source domain rating matrix is denoted by $R_1 \in \mathbb{R}^{m_1 \times n_1}$ with m_1 users rating on n_1 items, and the target rating matrix is denoted by $R_2 \in \mathbb{R}^{m_2 \times n_2}$ with m_2 users rating on n_2 items. Assume there are a total of c demographic categories in both domains, e.g., $c = 2$ for binary categories (e.g., married vs. not married) and $c > 2$ for multi-class categories. Let $G_1^0 \in \mathbb{R}_+^{m_1 \times c}$ represent the true demographic label indicator matrix in the source domain. The column of matrix G_1^0 indicates the class membership, that is $G^0_{1_{(ij)}} = 1$ if the *i*-th user is in the *j*-th demographic category and $G^0_{1_{(ij)}} = 0$ otherwise.

B. Transfer Matrix Factorization (TMF) Model

In traditional Matrix Factorization (MF), the rating matrix R is approximated with the product of two low-rank matrices: $U \in \mathbb{R}^{m \times k}$ that represents the latent user feature matrix, and $V \in \mathbb{R}^{k \times n}$ that represents the latent movie feature matrix. Each entry in R is approximated by the inner product of a row vector in U and a column vector in V: $R \approx UV$.

The key innovation of our model is to associate demographic information with latent user features as:

$$U_i = Y_i + \frac{\sum_{k=1}^c G_{ik} W_k}{\sum_{k=1}^c G_{ik}}, \tag{1}$$

where $Y \in \mathbb{R}^{m \times k}$ is the user personal feature matrix, $G \in \mathbb{R}^{m \times c}$ is the latent demographic matrix, and $W \in \mathbb{R}^{c \times k}$ is the association matrix between demographic categories and latent user features. In particular, we assume that the association matrix W can be shared in different but related domains. Informally, the row of the matrix W models the effect that a user with a specific demographic label has on the prior mean of the corresponding feature vector. Therefore, users with similar demographics will have feature vectors with similar prior distributions. The final feature vector of user i is obtained by adding offset Y_i to the mean of the prior distribution, which is important since ratings rely on not only demographics but also users' individual preferences, and usually the latter is much more crucial in predicting ratings. Without Y, the user feature matrix would only rely on demographics, meaning that, for example, users of the same gender will give the same rating score on one movie, and that is unrealistic. Inspired by the constrained probabilistic matrix factorization [33], our probabilistic graphic model is shown in Figure 1.

The likelihood of the observed ratings in each domain $\pi \in \{1, 2\}$ is as follows:

$$p(R_\pi | Y_\pi, V_\pi, W, G_\pi, \sigma_\pi^2) = \prod_{i=1}^{m_\pi} \prod_{j=1}^{n_\pi} [\mathcal{N}(R_{\pi_{(ij)}} \Big| \left(Y_{\pi_{(i\cdot)}} + \frac{\sum_{k=1}^c G_{\pi_{(ik)}} W_k}{\sum_{k=1}^c G_{\pi_{(ik)}}} \right) V_{\pi_{(\cdot j)}}, \sigma_\pi^2)]^{I_{\pi(ij)}}, \tag{2}$$

where $\mathcal{N}(x|\mu, \sigma^2)$ is the probability density of the normal distribution with mean μ and variance σ^2, and I_{ij} is the indicator function that is equal to 1 if user i rated movie j and equal to 0 otherwise. We also regularize all latent vectors by imposing Gaussian priors as follows: $\mathcal{N}(Y_{\pi_{(i\cdot)}} | \mathbf{0}, \sigma^2_{Y_\pi} \boldsymbol{I})$, $\mathcal{N}(V^T_{\pi_{(j\cdot)}} | \mathbf{0}, \sigma^2_{V_\pi} \boldsymbol{I})$, $\mathcal{N}(W_k | \mathbf{0}, \sigma^2_W \boldsymbol{I})$, $\mathcal{N}(G_{2_{(i\cdot)}} | \mathbf{0}, \sigma^2_{G_2} \boldsymbol{I})$, and $\mathcal{N}(G_{1_{(i\cdot)}} | G^0_{1_{(i\cdot)}}, \sigma^2_{G_1} \boldsymbol{I})$. Note that G_1^0 contains the true demographic label in the source domain and is the mean of the prior distribution of G_1.

We can obtain the *maximum a posteriori* (MAP) estimates of model parameters Y_π, V_π, G_π, and W with hyperparameters such as the prior variance σ_{Y_π} and the observation variance σ_π kept fixed by minimizing the following sum-of-squared-error objective function E:

$$E = \frac{1}{2} \sum_{\pi=1}^{2} \sum_{i=1}^{m_\pi} \sum_{j=1}^{n_\pi} I_{\pi_{(ij)}} [R_{\pi_{(ij)}} - \left(Y_{\pi_{(i\cdot)}} + \frac{\sum_{k=1}^c G_{\pi_{(ik)}} W_k}{\sum_{k=1}^c G_{\pi_{(ik)}}} \right) V_{\pi_{(\cdot j)}}]^2 + \sum_{\pi=1}^{2} \lambda_{Y_\pi} \sum_{i=1}^{m_\pi} \| Y_{\pi_{(i\cdot)}} \|^2 + \sum_{\pi=1}^{2} \lambda_{V_\pi} \sum_{j=1}^{n_\pi} \| V_{\pi_{(\cdot j)}} \|^2 + \lambda_W \sum_{k=1}^{c} \| W_k \|^2 + \alpha \sum_{i=1}^{m_1} \| G_{1_{(i\cdot)}} - G^0_{1_{(i\cdot)}} \|^2 + \gamma \sum_{i=1}^{m_2} \| G_{2_{(i\cdot)}} \|^2, \tag{3}$$

where regularization parameters $\lambda_{Y_\pi} = \sigma_\pi^2/2\sigma_{Y_\pi}^2$, $\lambda_{V_\pi} = \sigma_\pi^2/2\sigma_{V_\pi}^2$, $\lambda_W = \sum_{\pi=1}^{2} \sigma_\pi^2/2\sigma_W^2$, $\alpha = \sigma_1^2/2\sigma_{G_1}^2$, and $\gamma = \sigma_2^2/2\sigma_{G_2}^2$.

Furthermore, since matrices G_1 and G_2 indicate the probabilities that users belong to demographic classes, we revise the objective function by adding non-negative constraints to model parameters. To make it simpler, we minimize the loss function as follows:

$$\min_{Y_\pi, V_\pi, G_\pi, W} \|[R_1 - (Y_1 + G_1 W) V_1] \circ I_1\|^2 + \beta \|[R_2 - (Y_2 + G_2 W) V_2] \circ I_2\|^2$$
$$+ \sum_{\pi=1}^{2} \lambda_{Y_\pi} \|Y_\pi\|^2 + \sum_{\pi=1}^{2} \lambda_{V_\pi} \|V_\pi\|^2 + \lambda_W \|W\|^2 + \alpha \|G_1 - G_1^0\|^2 + \gamma \|G_2\|^2$$
$$s.t. \quad \sum_{j=1}^{c} G_{\pi_{(ij)}} = 1, G_\pi, V_\pi, Y_\pi, W \geq 0, \pi \in \{1,2\}, \tag{4}$$

where $\circ$ denotes element-wise product and β is the non-negative trade-off factor controlling the balance between the number of observations in the source and the target domains. Since G_1^0 contains the true demographic label information in the source domain, the regularization term enforces the similarity between G_1 and the prior G_1^0 in the source domain. We will then present an efficient algorithm to learn model parameters Y_π, V_π, G_π, and W to minimize the objective function in Eq. 4. The probability matrix G_2 in the target domain obtained through optimization will be used to predict user demographic labels. Specifically, the predicted demographic class label of the *i*-th user in the target domain is the index of the category with the largest probability. In addition, regularization parameters such as λ_{Y_π}, λ_{V_π}, λ_W, α, β, and γ provide a flexible way to regularization. To determine these parameters, we consider a set of reasonable parameter values for each of them, train the model for each setting, and choose the ones that perform best on the validation data.

C. Learning Algorithm

We now present the learning algorithm to find the optimal solution to our optimization problem in Eq. 4, which is achieved through the following theorem. As we use sparse matrices in the experiments, we ignore the indicator function matrix I_π in the following equations.

Theorem 1. *Updating Y_1, Y_2, V_1, V_2, G_1, G_2, W with eqs.* (5) *to* (9) *and normalizing G_1, G_2 to satisfy the equality constraints with eq.* (10) *in each iteration will monotonically decrease the objective function in eq.* (4) *until convergence.*

$$Y_\pi \leftarrow Y_\pi \circ \sqrt{\frac{[R_\pi V_\pi^T - G_\pi W V_\pi V_\pi^T]}{[Y_\pi (V_\pi V_\pi^T + \lambda)]}}, \tag{5}$$

$$V_\pi \leftarrow V_\pi \circ \sqrt{\frac{[(Y_\pi + G_\pi W)^T R_\pi]}{[((Y_\pi + G_\pi W)^T (Y_\pi + G_\pi W) + \lambda) V_\pi]}}, \tag{6}$$

$$G_1 \leftarrow G_1 \circ \sqrt{\frac{[(R_1 - Y_1 V_1) V_1^T W^T + \alpha G_1^0]}{[G_1 (W V_1 V_1^T W^T + \alpha)]}}, \tag{7}$$

$$G_2 \leftarrow G_2 \circ \sqrt{\frac{[(R_2 - Y_2 V_2) V_2^T W^T]}{[G_2 (W V_2 V_2^T W^T + \gamma)]}}, \tag{8}$$

$$W \leftarrow W \circ \frac{[\sqrt{\beta G_2^T (R_2 V_2^T - Y_2 V_2 V_2^T) + G_1^T (R_1 V_1^T - Y_1 V_1 V_1^T)}]}{[\sqrt{\beta G_2^T G_2 W V_2 V_2^T + G_1^T G_1 W V_1 V_1^T + \lambda W}]}, \tag{9}$$

$$G_{\pi_{(i\cdot)}} \leftarrow \frac{G_{\pi_{(i\cdot)}}}{\sum_{j=1}^{c} G_{\pi_{(ij)}}}, \tag{10}$$

where $\circ$ denotes element-wise product, $\frac{[\cdot]}{[\cdot]}$ denotes element-wise division, and $\sqrt{\cdot}$ denotes element-wise square root.

Algorithm 1: Transfer Matrix Factorization (TMF) for Cross-domain Recommender Systems

---

Input: Source domain rating matrix R_1 and true demographic label matrix G_1^0; target domain rating matrix R_2
Output: The demographic probability matrix in the target domain G_2
begin
  Initialize the matrix variables as Y_1, Y_2, V_1, V_2, W, G_1, G_2 and set parameters α, β, λ and γ. The initialization method will be detailed in the experimental section.
  for *iter* $\leftarrow$ *1* ***to*** *maxIter* **do**
    1. update Y_1, Y_2, V_1, V_2, W, G_1, G_2 by eqs. (5) to (9).
    2. normalize G_1, G_2 by eq. (10).
  end
  Output the matrix G_2 containing demographic labels.
end

The proof of Theorem 1 is given in the following theoretical analysis section and the learning algorithm for the model optimization is summarized in Algorithm 1.

D. Computational Complexity

We measure the computational complexity for eqs. (5) to (10) in a similar way as [28], [35]. The computational complexity for TMF in each iteration is of order $3m_1 n_1 k + cm_1 k + k^2 m_1 + k^2 n_1$ for eq. (5). In general, the latent dimension k and the number of categories c are much smaller than the size of rating matrices, that is, $k, c \ll min\{m, n\}$. Suppose $N = max\{m, n\}$, so the computational complexity is $O(N^2)$ in each iteration. Similarly, the computational complexity is $O(N^2)$ for eqs. (6) to (9) and is $O(N)$ for eq. (10) in each iteration. We assume this algorithm needs $maxIter$ iterations to converge. Therefore, multiplying these orders by $maxIter$ and then summating all the orders, we have the overall computational complexity as $O(maxIter \cdot N^2)$. Considering this is the worst case and the matrices are usually sparse in experiments, these matrix multiplications can be computed more efficiently in most cases on computers.

E. Theoretical Analysis

In this section, we will prove that the updating rules described in Theorem 1 will monotonically decrease the objective function in eq. (4) until convergence.

Proof. We first check the convergence of updating G_1 as described in eq. (7) and eq. (10) when Y_1,Y_2,V_1,V_2,W,G_2 are fixed. Following the standard theory of constrained optimization, we introduce Lagrangian multipliers and minimize the Lagrangian function based on optimizing function E in eq. (4):

$$\begin{aligned} L = E + Tr[\Gamma(G_1\boldsymbol{u}^T - \boldsymbol{v}_1^T)(G_1\boldsymbol{u}^T - \boldsymbol{v}_1^T)^T] \\ +Tr[\Lambda(G_2\boldsymbol{u}^T - \boldsymbol{v}_2^T)(G_2\boldsymbol{u}^T - \boldsymbol{v}_2^T)^T], \end{aligned} \tag{11}$$

where $\Gamma \in \mathbb{R}^{m_1\times m_1}$ and $\Lambda \in \mathbb{R}^{m_2\times m_2}$ are the Lagrange multipliers for the two constraints. All elements in vectors $\boldsymbol{u} \in \mathbb{R}^{1\times c}$, $\boldsymbol{v}_1 \in \mathbb{R}^{1\times m_1}$, and $\boldsymbol{v}_2 \in \mathbb{R}^{1\times m_2}$ are equal to one. Then we derive the updating rule for G_1 using the KKT complementarity condition for the constraints on G_1, that is:

$$\begin{aligned} \nabla_{G_1} L \circ G_1 = \{-2[R_1-(Y_1+G_1W)V_1]V_1^TW^T\} \circ G_1 \\ +[2\alpha(G_1-G_1^0)+2\Gamma(G_1\boldsymbol{u}^T-\boldsymbol{v}_1^T)\boldsymbol{u}] \circ G_1 = \mathbf{0}. \end{aligned} \tag{12}$$

Specifically, the role of Γ here is to drive the solutions to satisfy the constraint that the sum of the values in each row of G_1 is one. A similar normalization technique is presented by Zhuang et al. [31]. The effects of eq. (7) and eq. (10) are approximately equivalent to eq. (13) in terms of convergence, which is proven in Theorem 2. Similarly, the convergence of the updating rules for Y_1, Y_2, V_1, V_2, W, and G_2 can be proved according to Theorem 2 and the Multiplicative Update Rules [36]. Each update step in Algorithm 1 will not increase eq. (4) and the objective is lower bounded by zero, which guarantees the convergence. □

Theorem 2. *The Lagrangian function L in eq.* (11) *is monotonically decreasing (non-increasing) under the update rule eq.* (13).

$$G_1 \leftarrow G_1 \circ \sqrt{\frac{[(R_1-Y_1V_1)V_1^TW^T+\alpha G_1^0+\Gamma\boldsymbol{v}_1^T\boldsymbol{u}]}{[G_1(WV_1V_1^TW^T+\alpha+\Gamma\boldsymbol{u}^T\boldsymbol{u})]}}, \tag{13}$$

assuming the numerator and the denominator are both more than or equal to zero.

Proof. The key is to construct the auxiliary function of Lagrangian function L. According to Lemma 3, Lemma 4, and Proposition 5, we can obtain the Lagrangian function $L(G_1^{(0)}) = H(G_1^{(0)}, G_1^{(0)}) \geq H(G_1^{(1)}, G_1^{(0)}) \geq L(G_1^{(1)}) \cdots \geq L(G_1^{(maxIter)})$. Thus, L is monotonically decreasing other variables are fixed. □

Lemma 3. *[36] $Z(h,h')$ is an auxiliary function of $F(h)$ if the conditions $Z(h,h') \geq F(h)$ and $Z(h,h) = F(h)$ are satisfied. If Z is an auxiliary function for F, then F is non-increasing under the update:*

$$h^{(t+1)} = arg\,min_h Z(h,h'). \tag{14}$$

Lemma 4. *[37] For any matrices $A \in \mathbb{R}_+^{n\times n}$, $B \in \mathbb{R}_+^{k\times k}$, $S \in \mathbb{R}_+^{n\times k}$, and $S' \in \mathbb{R}_+^{n\times k}$, and A and B are symmetric, the following inequality holds:*

$$\sum_{i=1}^{n}\sum_{p=1}^{k}\frac{(AS'B)_{ip}S_{ip}^2}{S'_{ip}} \geq Tr(S^TASB). \tag{15}$$

Proposition 5. *Let $L(G_1)$ denote the sum of all terms in L that contain G_1. An auxiliary function for $L(G_1)$ is the following:*

$$\begin{aligned} H(G_1,G_1') = \sum_{ij}[G_1'(WV_1V_1^TW^T+\alpha+\Gamma\boldsymbol{u}^T\boldsymbol{u})]_{(ij)}\frac{(G_{1_{(ij)}})^2}{G'_{1_{(ij)}}} - \\ 2\sum_{ij}[(R_1-Y_1V_1)V_1^TW^T+\alpha G_1^0+\Gamma\boldsymbol{v}_1^T\boldsymbol{u}]_{(ij)}G'_{1_{(ij)}}(1+\log\frac{G_{1_{(ij)}}}{G'_{1_{(ij)}}}). \end{aligned} \tag{16}$$

Furthermore, it is a convex function with respect to G_1 and has a global minimum.

Proof. This can be proved similarly as in [37]. We have the Lagrangian function $L(G_1)$ based on the definition of trace and Frobenius norm of a matrix:

$$\begin{aligned} L(G_1) = Tr[-2(R_1-Y_1V_1)V_1^TW^TG_1^T - 2\alpha G_1^0G_1^T - \\ 2\Gamma\boldsymbol{v}_1^T\boldsymbol{u}G_1^T + G_1WV_1V_1^TW^TG_1^T + \alpha G_1G_1^T + \Gamma G_1\boldsymbol{u}^T\boldsymbol{u}G_1^T]. \end{aligned} \tag{17}$$

Giving the assumption in Theorem 2 that the numerator and the denominator are both more than or equal to zero, according to Lemma 4, we have

$$\begin{aligned} Tr(G_1WV_1V_1^TW^TG_1^T+\alpha G_1G_1^T+\Gamma G_1\boldsymbol{u}^T\boldsymbol{u}G_1^T) \leq \\ \sum_{ij}(G_1'WV_1V_1^TW^T+G_1'\alpha+G_1'\Gamma\boldsymbol{u}^T\boldsymbol{u})_{(ij)}\frac{(G_{1_{(ij)}})^2}{G'_{1_{(ij)}}}. \end{aligned} \tag{18}$$

Because $z \geq 1+\log(z), \forall z > 0$, let $z = \frac{G_{1_{(ij)}}}{G'_{1_{(ij)}}}$, we have

$$\begin{aligned} Tr[(R_1-Y_1V_1)V_1^TW^TG_1^T+\alpha G_1^0G_1^T+\Gamma\boldsymbol{v}_1^T\boldsymbol{u}G_1^T] \geq \\ \sum_{ij}[(R_1-Y_1V_1)V_1^TW^T+\alpha G_1^0+\Gamma\boldsymbol{v}_1^T\boldsymbol{u}]_{(ij)}G'_{1_{(ij)}}(1+\log\frac{G_{1_{(ij)}}}{G'_{1_{(ij)}}}). \end{aligned} \tag{19}$$

Summing over all the bounds of eqs. (18) and (19), we can obtain $H(G_1, G_1')$, which clearly satisfies: (1) $H(G_1,G_1') \geq L(G_1)$ and (2) $H(G_1,G_1) = L(G_1)$.

Then, fixing G_1', we minimize $H(G_1,G_1')$:

$$\begin{aligned} \frac{\partial H(G_1,G_1')}{\partial G_{1_{(ij)}}} = 2[G_1'(WV_1V_1^TW^T+\alpha+\Gamma\boldsymbol{u}^T\boldsymbol{u})]_{(ij)}\frac{G_{1_{(ij)}}}{G'_{1_{(ij)}}} \\ -2[(R_1-Y_1V_1)V_1^TW^T+\alpha G_1^0+\Gamma\boldsymbol{v}_1^T\boldsymbol{u}]_{(ij)}\frac{G'_{1_{(ij)}}}{G_{1_{(ij)}}}, \end{aligned} \tag{20}$$

and the Hessian matrix of $H(G_1,G_1')$ is

$$\begin{aligned} \frac{\partial^2 H(G_1,G_1')}{\partial G_{1_{(ij)}}G_{1_{(kl)}}} = 2\delta_{ik}\delta_{jl}\{[G_1'(WV_1V_1^TW^T+\alpha+\Gamma\boldsymbol{u}^T\boldsymbol{u})]_{(ij)}\frac{1}{G'_{1_{(ij)}}} \\ +[(R_1-Y_1V_1)V_1^TW^T+\alpha G_1^0+\Gamma\boldsymbol{v}_1^T\boldsymbol{u}]_{(ij)}\frac{G'_{1_{(ij)}}}{(G_{1_{(ij)}})^2}\}, \end{aligned} \tag{21}$$

which is a diagonal matrix with positive diagonal elements. Therefore, $H(G_1,G_1')$ is a convex function of G_1, and we can obtain the global minimum of $H(G_1,G_1')$ by setting $\partial H(G_1,G_1')/\partial G_{1_{(ij)}} = 0$.

Solving for G_1, the minimum is

$$G_{1_{(ij)}} = G'_{1_{(ij)}}\sqrt{\frac{[(R_1-Y_1V_1)V_1^TW^T+\alpha G_1^0+\Gamma\boldsymbol{v}_1^T\boldsymbol{u}]_{(ij)}}{[G_1'(WV_1V_1^TW^T+\alpha+\Gamma\boldsymbol{u}^T\boldsymbol{u})]_{(ij)}}}, \tag{22}$$

which is consistent with the updating formula eq. (13) derived from the KKT condition mentioned in Theorem 2. □

TABLE II
DATASET DESCRIPTIONS.

Data	User	Item	Rating	Gender(M/F)	Age(Y/O)
Movie	6040	3461	1M	71%/29%	NA
Flixster	6000	3500	2M	38%/62%	59%/41%
Book	6461	3680	0.2M	NA	20%/80%

IV. EXPERIMENTS

We evaluate the proposed TMF approach on the union of three real-world rating datasets **MovieLens (M)**, **Flixster (F)**, and **BookCrossing (B)** to demonstrate its effectiveness.

A. Datasets and Evaluation Criteria

Based on the availability of data, we use MovieLens to infer gender information in Flixster, use Flixster to infer gender information in MovieLens, and use Flixster to infer age information in BookCrossing. The details of each dataset are shown in Table II.

MovieLens[1]: The MovieLens dataset contains 3952 movies, 6040 users, and about 1 million ratings (scales 1-5). Each user has more than 20 ratings. We select 3461 movies with more than 3 ratings for the experiment. There are 999792 ratings and the density is 4.78%. The fraction of male users is 71.1%.

Flixster: The Flixster movie dataset collected by Jamali et al. [38]. We randomly select users with at least 200 ratings and movies with at least 100 ratings, which results in a subset of 2608105 ratings for 3500 movies by 6000 users. The rating density is 12.43% and the fraction of males is 38.3%. Note that the skew in gender distribution is the opposite of the one in MovieLens. In the task of age prediction, we follow an arbitrary convention of setting 25 years as the threshold between 'young' and 'old'. People who are below the threshold are labeled as 'young', and otherwise are labeled as 'old'. The fraction of users with an 'old' label is 41.0%.

BookCrossing[2]: For the BookCrossing dataset, for consistency across the evaluation datasets, we normalize the rating scales from 1 to 5 and select 6461 users and 3680 books with more than 15 and 20 ratings, respectively. There are 170134 ratings and the rating density is 0.72%. The fraction of users with an 'old' label is 79.6%.

To evaluate, we withheld the ground truth labels in the target domain and measured the classification accuracy using weighted-*precision*, *recall*, and *f-score*. In addition, we measured the *f-score* in each demographic category. Note that demographic labels in target domains are only used for evaluation, and not for training.

B. Baseline Methods and Parameter Settings

Our baselines include **MF-Logistic**, which uses matrix factorization [39] to decompose the rating matrices to latent vectors of size k in the source and target domains independently, and trains a logistic regressor to predict the user demographic labels given the low-rank user feature matrix in the source domain. Finally, the regressor is used to predict the user labels in the target domain. The drawback of this model is that the latent vectors in both domains may not be well aligned and in fact may represent quite different latent characteristics in source and target domains.

RMGM [26] is a rating matrix generative model based on the assumption that ratings are drawn from a shared cluster-level model. The core idea of this method is that each rating matrix R_i can be decomposed via tri-factor matrix factorization. The decomposition consists of a user membership matrix, an item membership matrix, and a core rating matrix that represents the mean item rating of the cluster and is shared across domains. The mixture generative model can be applied to demographic prediction with a simple modification in which the demographic label of the majorities in the source domain can be used as the predicted labels for users in the target domain. This approach transfers demographic knowledge in an unsupervised fashion.

MTrick is applied to cross-domain text document classification by Zhuang [31] based on tri-factor matrix factorization. The decomposition consists of document membership matrices, word membership matrices, and an association matrix between word clusters and document classes shared between source and target domains. The method can be adjusted to fit the cross-domain demographic classification. Specifically, our demographic probability matrix is analogous to the document membership matrix. One drawback is that the model assumes users in the same demographic cluster will give the same rating scores on the same item and ignores the individual preferences of each user. Regularization terms for MTrick were later added by Wang et al. [32] and Long et al. [28] to improve model generalization, a baseline we call **DKT** [32].

Our model **TMF** differs from others in that it characterizes a user profile as an integration of both a group-level profile that captures the preference of users within the same demographic group and a personal profile that captures users' personal preferences. In addition, we consider our TMF model with a full selection of regularization terms λ, γ, which we call **Reg-TMF**. **TMF** involves fewer regularization terms, i.e., no regularization parameter λ, γ for latent vectors Y_π, V_π, G_2 and W. We also explore regularization parameter sensitivities in Section IV-D. For most of the experiments, $\gamma = \lambda = 0.01$ is used for all latent variables. The trade-off parameters are $\alpha = 0.2$ and $\beta = 1$. We evaluate the objective function under different numbers of latent dimensions from 5 to 50 and choose the best latent dimension k. At the beginning, we randomly initialize Y_1, Y_2, V_1, V_2, W with non-negative values. We randomly initialize the demographic information matrices G_1 and G_2 with entries in the range of 0 to 1. The maximum number of iterations *maxIter* used in the optimization is 300.

C. Demographic Prediction Results

The results are reported in Table III, where F denotes the Flixster dataset, M denotes the MovieLens dataset and B denotes the BookCrossing dataset. In this section, we will first briefly summarize the experiment results and then discuss the

[1]https://grouplens.org/datasets/movielens/
[2]http://www2.informatik.uni-freiburg.de/~cziegler/BX/

TABLE III
AVERAGE PERFORMANCE AND PER-GROUP FSCORE FOR MULTIPLE METHODS INFERRING DIFFERENT TYPES OF DEMOGRAPHICS.

Data	Metrics	Methods					
		MF-Logistic	RMGM	MTrick	DKT	TMF	Reg-TMF
F to M infer gender	Precision	0.3750	0.4672	0.5659	0.5639	0.7194	**0.7195**
	Recall	0.4015	0.4860	0.5440	0.5418	0.7304	**0.7311**
	Fscore	0.3878	0.4764	0.5547	0.5526	0.7249	**0.7253**
	Female-Fscore	0.2835	0.3554	0.3405	0.3398	**0.4767**	0.4759
	Male-Fscore	0.4457	0.5460	0.6765	0.6744	0.8083	**0.8085**
M to F infer gender	Precision	0.4278	0.5098	0.6338	0.6337	**0.6980**	0.6978
	Recall	0.6214	0.5350	**0.7692**	0.7689	0.6963	0.6961
	Fscore	0.5068	0.5221	0.6950	0.6948	**0.6971**	0.6970
	Female-Fscore	0.3315	0.6224	0.7468	0.7467	**0.7519**	0.7517
	Male-Fscore	0.4999	0.3016	0.3389	0.3387	**0.6141**	0.6140
F to B infer age	Precision	0.2843	0.4728	0.4337	0.4331	**0.6282**	0.6262
	Recall	**0.7696**	0.5461	0.6173	0.6179	0.5921	0.5898
	Fscore	0.4152	0.5068	0.5095	0.5092	**0.6096**	0.6074
	Young-Fscore	0.3310	0.3271	**0.3504**	**0.3504**	0.2545	0.2502
	Old-Fscore	0.2306	0.5667	0.4980	0.4970	**0.7524**	0.7511

effects of the difference of demographic distributions in the source and target domains, the amounts of labeled users, and the sparsity of ratings on the classification accuracy.

First, comparing the performance of all six models, we can see that our model TMF and its variation Reg-TMF consistently outperform others in all three types of demographic prediction. According to Weinsberg's [3] work, gender prediction in a single domain has reached 80% while our method achieves up to 73% across certain domains. The MF-Logistic approach performs the worst since it does not align the two domains jointly. RMGM is better than MF-Logistic but the weighted F-score is just slightly above 0.5. The reason for the poor performance is that the demographic labels are not correlated with the generation of the ratings. MTrick and TMF perform much better than others because they both correlate demographic labels with rating generation in a supervised fashion. The crucial reason for TMF achieving the highest performance is that it characterizes a user profile with both an individual preference profile and a non-demographic user preference.

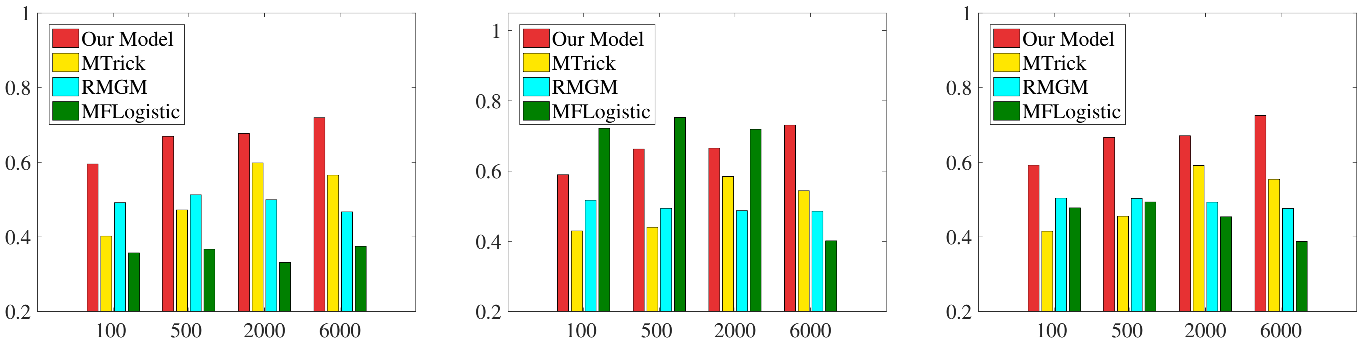
For all three types of demographic prediction, the distribution of the demographic labels in the source domain is totally opposite to that in the target domain, which makes the prediction tasks very challenging. In particular, the MovieLens dataset has a majority of males while the Flixster dataset has many more female users. In age prediction, the proportion of youth and old in the source domain is also opposite to that in the target domain. As shown in Table III, the MF-Logistic approach has extremely high recall but extremely low precision when we predict ages from Flixster to the BookCrossing dataset. This is because the method has a strong tendency to predict the majority class in the source domain for most of the users when the source domain is unbalanced. A good method should balance between recall and precision. Compared to other methods, our model TMF is more robust and consistently outperforms others regardless of the difference in demographic distributions in the source and target domains.

TABLE IV
PREDICTION PERFORMANCE OF DIFFERENT DEMOGRAPHIC DISTRIBUTIONS USING FLIXSTER TO INFER AGE IN BOOKCROSSING.

Age	Precision	Recall	Fscore	Source Youth(%)	Target Youth(%)
20	0.6290	0.6047	0.6166	38.03	7.41
25	0.6262	0.5898	0.6047	58.87	20.40
30	0.6002	0.6003	0.6003	72.75	39.00
35	0.5889	0.5929	0.5909	81.48	57.02

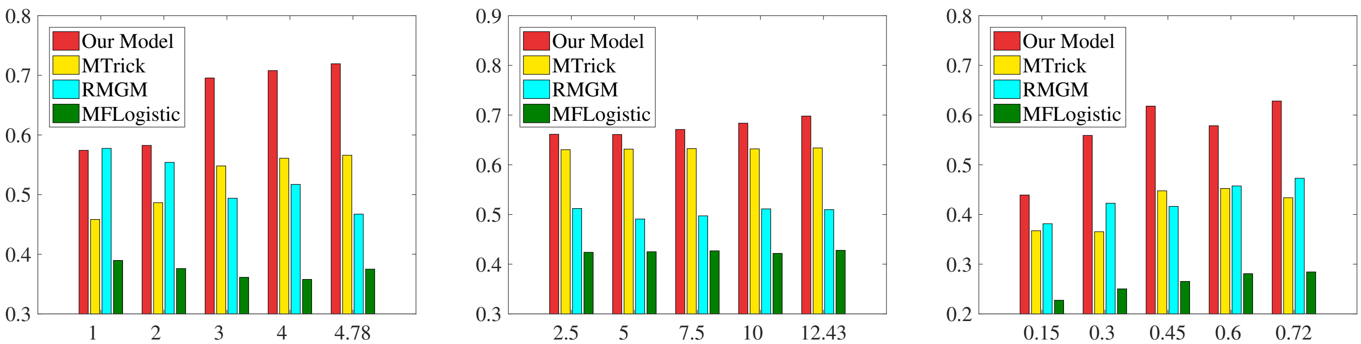
We can also see that the age prediction performance from Flixster to BookCrossing is weaker than the other two types of demographic prediction in Table III. One of the reasons is that BookCrossing ratings are extremely sparse with a rating density around 0.72%, compared to MovieLens (4.78%) and Flixster (12.43%). Demographic prediction tends to be much better with sufficient data in the target domain. We discuss the effects of the amounts of labeled users and the sparsity of ratings in the following paragraphs.

Effect of Demographic Distributions. To explore the effect of demographic distributions on the prediction performance in more detail, we carry out a series of experiments using Flixster, inferring age information in BookCrossing. We compare our model with all others except DKT since MTrick and DKT perform similarly. Specifically, we adjust the age threshold from 20 to 35 and divide the users into two different groups: youth and not youth. As the threshold changes, the distribution of age groups in the source and target domains will also change. From Table IV, we can see that the prediction performance of TMF is quite stable regardless of the change in demographic distribution. Performance is typically higher when the demographic distribution in the source domain is balanced. In particular, when the age threshold is 20 or 25, the proportion of youth in the source domain is more balanced and the prediction performance is better, but as the demographic



(a) Precision (b) Recall (c) F-score

Fig. 2. Prediction performance with respect to different amounts of labeled users in source domain using Flixster to infer gender in MovieLens.



(a) F to M infer Gender (b) M to F infer Gender (c) F to B infer Age

Fig. 3. Accuracy with respect to densities of testing data using different methods.

distribution in the source domain becomes more and more unbalanced, the prediction performance decreases somewhat. On the other hand, if the demographic distribution in the target domain is similar to that in the source domain, the performance is generally good. When the demographic distribution in the target domain is opposite to that in the source domain, the performance decreases. In Table IV, we can see that the ratio of youth vs. non-youth is 4:6 when the age threshold is 20, and the ratio is 6:4 when the threshold is 25. However, for the former case, the distributions for the source and target domains are more similar and thus performance is better.

Effect of Amounts of Labeled Users. We also evaluated our model's prediction performance with different amounts of labeled users in the source domain, inferring the MovieLens gender information using the Flixster dataset. We randomly selected 100, 500, 2000, and the full dataset (6000 users) from the Flixster dataset, and constructed several subsets as the source domain rating datasets, respectively.

Figure 2 demonstrates that the prediction performance of TMF steadily increases with increasing amounts of labeled users that we know in the source domain. Even if we only have a few labeled users, TMF still can reach high prediction performance. Figure 2 shows that only 500 labeled users can get more than 66% demographic prediction for precision, recall, and f-score. The other methods all perform weaker than ours as shown in Figure 2(b), with the exception that the recall of MF-Logistic is sometimes extremely high. The reason is that MF-Logistic tends to predict the majority label for most of the users in one group when the source domain is unbalanced. In such cases, the MF-Logistic's gambling policy can achieve either an extremely high recall or an extremely low recall, which is also shown in Table III.

Effect of Rating Sparsity. We also evaluated the sensitivity of our model TMF with respect to different levels of rating sparsity. Generally, prediction performance can be affected by the density of the rating matrices from either the source domain or the target domain dataset. In fact, we are more curious about the performance of the model when the rating density in the target domain varies: the rating data in the source domain are generally sufficient for constructing the shared association matrix between demographics and latent features. In addition to the original datasets, we subsampled datasets with an additional 4 levels of density in the target domain, for each type of demographic prediction task. Specifically, due to different density levels in the original data, for the target domain datasets, we randomly dropped out some ratings from the original matrix and obtained subsets with 1%, 2%, 3% and 4% rating ratios for Movielens, subsets with 2.5%, 5.0%, 7.5% and 10.0% rating ratios for Flixster, and subsets with 0.15%, 0.30%, 0.45% and 0.60% rating ratios for BookCrossing. In Figure 3(a), we used the 2.5% density Flixster dataset, from which we randomly selected 6000 users and 3500 movies with more than 60 and 12 ratings, respectively, as the source domain rating matrix.

The experiment results in Figure 3 show that for all models, the performance improves (prediction accuracy increases) as

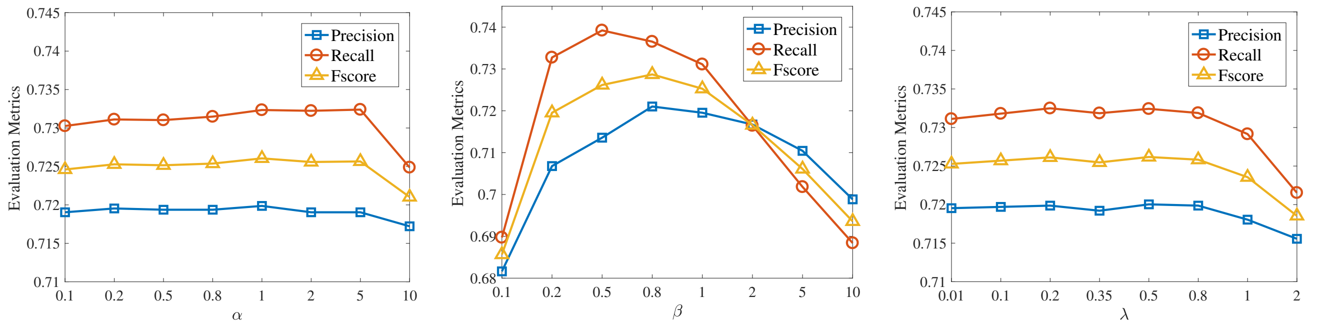


Fig. 4. Gender prediction performance with respect to different parameter settings of TMF using Flixster to infer gender information in MovieLens.

the density of the target domain increases. Moreover, the figures demonstrate that our model TMF is more sensitive to the density when the target domain data is not dense enough. However, when the first level is 2.5% in Figure 3(b), all the models do not improve much further with respect to the density. This shows TMF can reach good demographic prediction performance using less data, which also means higher efficiency in practice. Finally, it is obvious that the factor models (MTrick and TMF) perform much better than the mixture model (RMGM), which indicates the advantages of factor models that integrate demographic information from users' labels.

D. Parameter Effects

We used the gender inference experiment from Flixster to MovieLens dataset to investigate the effects of regularization parameters. The proposed TMF model has several regularization parameters including α, β, γ, and three types of λ in eq. (4). We present the parameter sensitivity results of this experiment, using Flixster to infer gender in MovieLens, in Figure 4. Parameter α specifies how strictly the demographic information should match the label. Bigger α enforces a closer match to the label when the users' demographics affect the ratings. We can see that the prediction performance is quite stable with respect to a wide range of α except for extreme values. Parameter β controls the balance between the number of observations in the source and target domains. Specifically, the rating loss should not be dominated by the domain with more observed ratings. We can see that good prediction performance depends on the balance between the rating densities of the two domains, implying that either a very large or small β is not appropriate. We have found the same phenomenon when using other types of datasets for inferring demographics. Parameter λ_W provides regularization for the latent matrix W. Figure 4 demonstrates that the prediction performance is also stable with respect to the regularization parameter, but in this case its sensitivity is slightly more than observed for α. In our experiments, the effects of the three types of λ and γ parameters are similar, so we chose λ_W for illustration. In sum, for α, λ and γ, our model's prediction performance is quite stable with respect to a wide range of parameter values.

V. Discussion

Our framework can be easily applied to inferring demographics beyond the binary age and gender labels used in these experiments, in order to solve more general transfer learning problems across domains, under the assumption that entities (e.g., users or items) with certain common characteristics tend to exhibit similar behavior in related domains. For example, instead of binary classification for each type of user trait prediction, our model could also be used for multi-class classification: to predict a multi-class user trait (e.g. education level), we can simply adopt one-vs-all classification while still keeping the current model formulation. In addition, we can predict different types of user traits jointly by grouping users at finer granularity (e.g., young/old female, young/old male, etc.) and set the cluster number appropriately (e.g., $c = 4$). Examples of other types of demographic information that could be analyzed in this way include political or religious view, race, education level, geo-location, and employment status. An appropriately validated version of our approach can be also used to infer personality traits for group-level study purposes when obtaining information through questionnaires (e.g, for big five personality traits [40]) in some domains may be difficult or impossible.

VI. Conclusions

We introduced a novel method called **Transfer Matrix Factorization (TMF)** to solve the problem of predicting user demographics using ratings in a target domain, through knowledge transfer from the source domain, in which users' ratings and the corresponding demographics are available. Our main contributions are: (1) Our model explores effectively the correlation between demographics and ratings across different domains that share neither common users nor common items. (2) We develop an iterative algorithm for this optimization and theoretically show its convergence. (3) Extensive experiments using real-world datasets demonstrate that our model can achieve higher classification accuracy, regardless of the amounts of labeled users, the sparsity of ratings, and the difference of demographic distributions in source and target domains. Our approach can be used as an analytical tool to assess the privacy impact for users of providing specific kinds of user information in one or more source domains, in the context of the existence of complementary data in a target domain. In future work, we would like to use our results to investigate effective strategies for operations such as obfuscating ratings that better protect user privacy. The potential success of various applications of the proposed method has positive impact on not only computer science but also psychology and social science.

References

[1] N. F. Awad and M. Krishnan, "The personalization privacy paradox: an empirical evaluation of information transparency and the willingness to be profiled online for personalization," *MIS Quarterly*, vol. 30, no. 1, pp. 13–28, 2006.

[2] J. Calandrino, A. Kilzer, A. Narayanan, E. Felten, and V. Shmatikov, ""you might also like:" privacy risks of collaborative filtering," in *Proc. of the IEEE Symposium on Security and Privacy*, 2011, pp. 231–246.

[3] U. Weinsberg, S. Bhagat, S. Ioannidis, and N. Taft, "Blurme: inferring and obfuscating user gender based on ratings," in *Proc. of the ACM Conference on Recommender Systems (RECSYS)*, 2012, pp. 195–202.

[4] S. Bhagat, U. Weinsberg, S. Ioannidis, and N. Taft, "Recommending with an agenda: active learning of private attributes using matrix factorization," in *Proc. of the ACM Conference on Recommender Systems (RECSYS)*, Oct. 2014, pp. 65–72.

[5] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. of the IEEE Symposium on Security and Privacy*, 2008, pp. 111–125.

[6] A. M. Elkahky, Y. Song, and X. He, "A multi-view deep learning approach for cross domain user modeling in recommendation systems," in *Proc. of the International World Wide Web Conference (WWW)*, 2015, pp. 278–288.

[7] M. B. Oliver, J. B. Weaver, III, and S. L. Sargent, "An examination of factors related to sex differences in enjoyment of sad films," *Journal of Broadcasting & Electronic Media*, vol. 44, no. 2, pp. 282–300, 2000.

[8] M. Thelwall, "Reader and author gender and genre in goodreads," *Journal of Librarianship and Information Science*, p. 0961000617709061, 2017.

[9] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proceedings of the National Academy of Sciences*, vol. 110, no. 15, pp. 5802–5805, 2013.

[10] E. Zheleva and L. Getoor, "To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles," in *Proc. of the International World Wide Web Conference (WWW)*, 2009, pp. 531–540.

[11] M. Pennacchiotti and A. Popescu, "Democrats, republicans and starbucks afficionados: user classification in twitter," in *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2011, pp. 430–438.

[12] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in twitter," in *Proc. of the 2nd international workshop on Search and mining user-generated contents*, 2010, pp. 37–44.

[13] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang, "Towards social user profiling: unified and discriminative influence model for inferring home locations," in *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2012, pp. 1023–1031.

[14] J. Zhang, X. Hu, Y. Zhang, and H. Liu, "Your age is no secret: Inferring microbloggers' ages via content and interaction analysis," in *Tenth International AAAI Conference on Web and Social Media*, 2016.

[15] J. Otterbacher, "Inferring gender of movie reviewers: exploiting writing style, content and metadata," in *Proc. of the ACM International Conference on Information and Knowledge Management (CIKM)*, 2010, pp. 369–378.

[16] Y. Zhong, N. Yuan, W. Zhong, F. Zhang, and X. Xie, "You are where you go: inferring demographic attributes from location check-ins," in *Proc. of the ACM International Conference on Web Search and Data Mining (WSDM)*, 2015, pp. 295–304.

[17] P. Wang, J. Guo, Y. Lan, J. Xu, and X. Cheng, "Your cart tells you: Inferring demographic attributes from purchase data," in *Proc. of the ACM International Conference on Web Search and Data Mining (WSDM)*, 2016, pp. 173–182.

[18] X. Mu, F. Zhu, E. Lim, J. Xiao, J. Wang, and Z. Zhou, "User identity linkage by latent user space modelling," in *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 1775–1784.

[19] M. Korayem and D. Crandall, "De-anonymizing users across heterogeneous social computing platforms," in *Proc. of the Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[20] B. Bi, M. Shokouhi, M. Kosinski, and T. Graepel, "Inferring the demographics of search users: social data meets search queries," in *Proc. of the International World Wide Web Conference (WWW)*, 2013, pp. 131–140.

[21] Y. Koren, "Factor in the neighbors: scalable and accurate collaborative filtering," *ACM Transactions on Knowledge Discovery from Data*, vol. 4, no. 1, pp. 1–24, 2010.

[22] R. Salakhutdinov and A. Mnih, "Bayesian probabilistic matrix factorization using markov chain monte carlo," in *Proc. of the International Conference on Machine Learning (ICML)*, 2008.

[23] B. Lakshminarayanan, G. Bouchard, and C. Archambeau, "Robust bayesian matrix factorisation," in *Proc. of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

[24] A. P. Singh and G. J. Gordon, "Relational learning via collective matrix factorization," in *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2008, pp. 650–658.

[25] T. Iwata and K. Takeuchi, "Cross-domain recommendation without shared users or items by sharing latent vector distributions," in *Proc. of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015, pp. 379–387.

[26] B. Li, Q. Yang, and X. Xue, "Transfer learning for collaborative filtering via a rating-matrix generative model," in *Proc. of the International Conference on Machine Learning (ICML)*, 2009, pp. 617–624.

[27] ——, "Can movies and books collaborate?: cross-domain collaborative filtering for sparsity reduction," in *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2009, pp. 2052–2057.

[28] M. Long, J. Wang, G. Ding, D. Shen, and Q. Yang, "Transfer learning with graph co-regularization," in *Proc. of the AAAI Conference on Artificial Intelligence*, 2012.

[29] W. Pan, N. N. Liu, E. W. Xiang, and Q. Yang, "Transfer learning to predict missing ratings via heterogeneous user feedbacks," in *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 22, no. 3, 2011, p. 2318.

[30] L. Zhao, S. J. Pan, E. W. Xiang, E. Zhong, Z. Lu, and Q. Yang, "Active transfer learning for cross-system recommendation." in *Proc. of the AAAI Conference on Artificial Intelligence*, 2013.

[31] F. Zhuang, P. Luo, H. Xiong, Z. Shi, Q. He, and Y. Xiong, "Exploiting associations between word clusters and document classes for cross-domain text categorization," in *Proc. of the SIAM International Conference on Data Mining (SDM)*, 2010, pp. 13–24.

[32] H. Wang, H. Huang, F. Nie, and C. Ding, "Cross-language web page classification via dual knowledge transfer using nonnegative matrix tri-factorization," in *Proc. of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2011, pp. 933–942.

[33] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *Proc. of the Annual Conference on Neural Information Processing Systems (NIPS)*, 2008.

[34] C. Severinski, "Augmenting probabilistic matrix factorization models for rare users," Ph.D. dissertation, University of Toronto (Canada), 2016.

[35] C. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 32, no. 1, pp. 45–55, 2010.

[36] D. Lee and H. Seung, "Algorithms for non-negative matrix factorization," in *Proc. of the Annual Conference on Neural Information Processing Systems (NIPS)*, 2001.

[37] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering," in *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2006, pp. 126–135.

[38] M. Jamali and M. Ester, "A matrix factorization technique with trust propagation for recommendation in social networks," in *Proc. of the ACM Conference on Recommender Systems (RECSYS)*, 2010, pp. 135–142.

[39] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[40] P. T. Costa and R. R. MacCrae, *Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO-FFI): professional manual*. Psychological Assessment Resources, Incorporated, 1992.