

# Learning with Heterogeneous Side Information Fusion for Recommender Systems

Huan Zhao<sup>1</sup>, Quanming Yao<sup>1</sup>, Yangqiu Song<sup>1</sup>, James T. Kwok<sup>1</sup>, Dik Lun Lee<sup>1</sup>

*<sup>a</sup>Department of Computer Science and Engineering,  
Hong Kong University of Science and Technology,  
Clear Water Bay, Hong Kong.*

---

## Abstract

Recommender System (RS) is a hot area where artificial intelligence (AI) techniques can be effectively applied to improve performance. Since the well-known Netflix Challenge, collaborative filtering (CF) has become the most popular and effective recommendation method. Despite their success in CF, various AI techniques still have to face the data sparsity and cold start problems. Previous works tried to solve these two problems by utilizing auxiliary information, such as social connections among users and meta-data of items. However, they process different types of information separately, leading to information loss. In this work, we propose to utilize Heterogeneous Information Network (HIN), which is a natural and general representation of different types of data, to enhance CF-based recommending methods. HIN-based recommender systems face two problems: how to represent high-level semantics for recommendation and how to fuse the heterogeneous information to recommend. To address these problems, we propose to applying meta-graph to HIN-based RS and solve the information fusion problem with a “matrix factorization (MF) + factorization machine (FM)” framework. For the “MF” part, we obtain user-item similarity matrices from each meta-graph and adopt low-rank matrix approximation to get latent features for both users and items. For the “FM” part, we propose to apply FM with Group lasso (FMG) on the obtained features to simultaneously predict missing ratings and select useful meta-graphs. Experimental results on two large real-world datasets, i.e., Amazon and Yelp, show that our proposed approach is better than that of the state-of-the-art FM and other HIN-based recommending methods.

**Keywords:** Recommender Systems, Collaborative filtering, Heterogeneous Information Networks, Matrix Factorization, Factorization Machine

---



---

*Email addresses:* hzhaoaf@cse.ust.hk (Huan Zhao), qyaoaa@cse.ust.com (Quanming

## 1. Introduction

With the development of Internet technology, especially mobile Internet, in recent years, recommender systems (RSs) have become an indispensable tool in everyday life. Aiming at providing interesting items to users based on their preferences, RSs are widely used in many domains, e.g., product recommendation on Amazon, movie recommendation on Netflix and news recommendation on Facebook. One of the key components of RS is user modeling, which is to understand users' preferences based on their past behaviors on the Internet. Various artificial intelligence (AI) techniques have been applied to learn users' preferences automatically, including collaborative filtering [1, 2], content-based filtering [3], deep learning based methods [4], transfer learning based methods [5, 6, 7], and reinforcement learning based methods [8, 9]. CF has been the most popular recommending method in the last decade. It tries to predict users' preferences based on similar users. Despite the success of CF, it faces two problems: data sparsity and cold start. The first problem is due to the fact that users tend to interact with only a small number of items and the second is caused by new users or items without any behavior records. Both of these two problems impair the recommending performance. To address them, researchers tried to incorporate auxiliary information, or side information, to enhance CF. For example, social connections among users [10, 11] and reviews [12, 13] of items have been utilized to improve the recommending performance. However, the challenge is that various side information are processed independently, leading to information loss across different side information.

The problem becomes more severe on modern RSs, where rich heterogeneous side information can be captured. For example, on Amazon, products have categories and belong to brands, and users can write reviews on products. On Yelp, users can follow other users to form a social network, businesses have categories and locations, and users can write reviews on businesses. Consequently, real-world RSs need to consider rich semantics of different types of side information together rather than process them one by one. This rich heterogeneity requires the development of a mathematical representation to formulate it and a tool to compute over it.

Heterogeneous information networks (HINs) [14] have been proposed as a general data representation for different types of data, such as scholar network data [15], social network data [16], patient network data [17], and knowledge graph data [18]. In early works, HINs were used to handle entity search and similarity measure [15], where the query and result entities are assumed to have the same type (e.g., using *Person* to search *Person*). Later, it was extended

---

Yao), [yqsong@cse.ust.hk](mailto:yqsong@cse.ust.hk) (Yangqiu Song), [jamesk@cse.ust.hk](mailto:jamesk@cse.ust.hk) (James T. Kwok),  
[dlee@cse.ust.hk](mailto:dlee@cse.ust.hk) (Dik Lun Lee)

to handle heterogeneous entity recommendation problems (i.e., recommending *Items* to *Users*) [19, 20, 21]. To incorporate rich semantics, we first build a network schema of an HIN. Figure 1 shows an example HIN on Yelp, and Figure 2 shows a network schema defined over the entity types *User*, *Review*, *Word*, *Business*, etc. Then, the semantic relatedness constrained by the entity types can be defined by the similarities between two entities along meta-paths [15]. For CF methods, if we want to recommend businesses to users, we can build a simple meta-path  $Business \rightarrow User$  and learn from this meta-path to make generalizations. From HIN’s schema, we can define more complicated meta-paths like  $User \rightarrow Review \rightarrow Word \rightarrow Review \rightarrow Business$ , which defines a similarity to measure whether a user tends to like a business if his/her reviews are similar to those written by other users for the same business.

When applying meta-path based similarities to recommender systems, there are two major challenges. First, meta-path may not be the best way to characterize the rich semantics. Figure 1 shows a concrete example, where a meta-path  $User \rightarrow Review \rightarrow Word \rightarrow Review \rightarrow Business$  is used to capture users’ similarity since they both write reviews and mention the same aspect (seafood) about the same business. However, if we want to capture the semantic that  $U_1$  and  $U_2$  rate the same type of business (such as *Restaurant*), and at the same time, they mention the same aspect (such as *seafood*), the meta-path fails. Thus, we need a better way to capture such complicated semantics. Recently, Huang et al. [22] and Fang et al. [23] have proposed to use meta-graph (or meta-structure) to compute similarity between homogeneous type of entities (e.g., using *Person* to search *Person*) over HINs, which can capture more complex semantics that meta-path cannot. However, they did not explore meta-graph for entities of heterogeneous types. Thus, in this paper, we extend meta-graph to the recommendation problem. However, how to use the similarities between heterogeneous types of entities derived from HINs in recommendation is still unclear, which results in the second challenge.

Second, different meta-paths or meta-graphs result in similarities with different semantics. How to assemble them effectively is another challenge. Currently, there are two principled ways. Considering our goal is to achieve accurate predictions of the ratings users give to items, which can be formulated as a matrix completion problem of the user-item rating matrix. One way to predict the missing ratings based on HIN is to use meta-paths to generate many ad-hoc alternative similarities among users and items, and then learn a weighting mechanism to combine the similarities from different meta-paths explicitly to approximate the user-item rating matrix [21]. This approach does not consider implicit factors of each meta-path, and each alternative similarity matrix could be very sparse to contribute to the final ensemble. The other way is to first factorize each user-item similarity matrix to obtain user and item latent features,

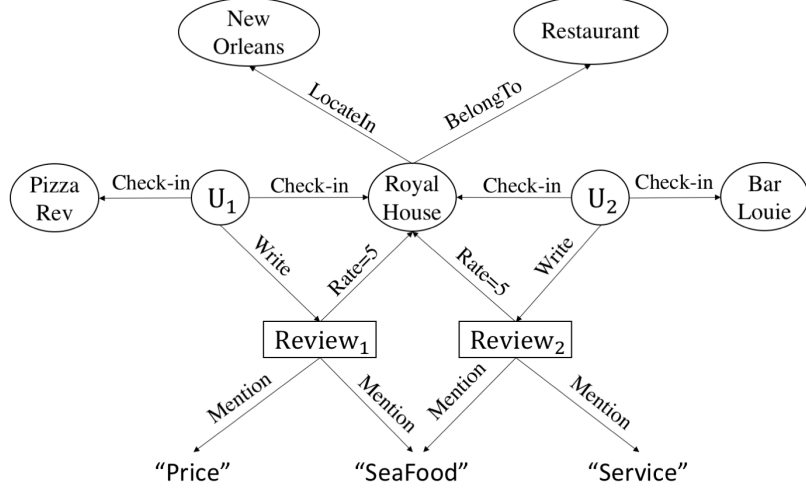


Figure 1: An example of HIN, which is built based on the web page for Royal House on Yelp.

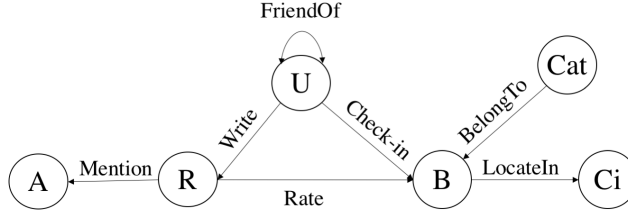


Figure 2: The Network Schema for the HIN in Figure 1. A: aspect extracted from reviews; R: reviews; U: users; B: business; Cat: category of item; Ci: city.

and then use all the latent features to recover a new user-item matrix [20]. This method resolves the sparsity problem of each similarity matrix. However, it does not fully make use of the latent features since when ensemble is performed, each meta-path cannot see others' variables but only the single value predicted by the others.

To address the above challenges, we propose a new principled way to make fully use of various side information on HIN. First, instead of using meta-path for heterogeneous recommendation [20, 21], we introduce the concept of meta-graph to the recommendation problem, which allows us to incorporate more complex semantics into our prediction problem. Second, instead of computing the recovered matrices directly, we use all of the latent features of all meta-graphs. Inspired by the famous work PCA+LDA used for face recognition [24], which first uses PCA (principle component analysis) to perform unsupervised dimensionality

reduction, and then applies LDA (linear discriminant analysis) to discover further reduced dimensions guided by supervision, we apply “matrix factorization (MF) + factorization machine (FM)” [25] to our recommendation problem. For each meta-graph, we first compute the user-item similarity matrix under its guidance, and then apply MF to it to obtain a set of user and item vectors, representing the latent features of users and items, respectively. Finally, with multiple sets of user and item latent features in hand, we use FM to assemble them to predict the missing ratings that users give to items.

Besides, to effectively select useful meta-graphs, we propose to use FM with Group lasso (FMG) to learn the parameters. To boost the performance on meta-graph selection, we further adopt a nonconvex variant of group lasso regularization. This leads to a nonconvex and nonsmooth optimization problem, which is difficult to solve. We propose two algorithms to efficiently solve the optimization problem; one is based on proximal gradient algorithm [26] and the other based on stochastic variance reduced gradient [27]. As a result, we can automatically determine for new incoming problems which meta-graphs should be used, and for each group of user and item features from a meta-graph, how they should be weighted. Experimental results on two large real-world datasets, Amazon and Yelp, show that our framework can successfully outperform other MF-based, FM-based, and existing HIN-based recommending methods. Our code is available at <https://github.com/HKUST-KnowComp/FMG>.

Preliminary results of this paper have been reported in [28]. In this full version, in addition to matrix factorization (MF), we also adopt nuclear norm regularization (NNR) to obtain latent features in Section 3.2.2. Furthermore, we adopt nonconvex regularization to boost meta-graph selection performance in Section 4.2 and design a new optimization algorithm, which is more efficient than the one used in [28], in Section 5.2. Finally, additional experiments are performed to support the above research in Section 6.4, 6.6 and 6.8.

### *Notation*

We denote vectors and matrices by lowercase and uppercase boldface letters, respectively. In this paper, a vector always denote row vector. For a vector  $\mathbf{x}$ ,  $\|\mathbf{x}\|_2 = (\sum_{i=1} |\mathbf{x}_i|^2)^{\frac{1}{2}}$  is its  $\ell_2$ -norm. For a matrix  $\mathbf{X}$ , its nuclear norm is  $\|\mathbf{X}\|_* = \sum_i \sigma_i(\mathbf{X})$ , where  $\sigma_i(\mathbf{X})$ 's are the singular values of  $\mathbf{X}$ ;  $\|\mathbf{X}\|_F = (\sum_{i,j} \mathbf{X}_{ij}^2)^{\frac{1}{2}}$  is its Frobenius norm and  $\|\mathbf{X}\|_1 = \sum_{i,j} |\mathbf{X}_{ij}|$  is its  $\ell_1$ -norm. For two matrices  $\mathbf{X}$  and  $\mathbf{Y}$ ,  $\langle \mathbf{X}, \mathbf{Y} \rangle = \sum_{i,j} \mathbf{X}_{ij} \mathbf{Y}_{ij}$ , and  $[\mathbf{X} \odot \mathbf{Y}]_{ij} = \mathbf{X}_{ij} \mathbf{Y}_{ij}$  denotes the element-wise multiplication. For a smooth function  $f$ ,  $\nabla f(\mathbf{x})$  is its gradient at  $\mathbf{x}$ .

## **2. “MF + FM” Framework**

The main contribution of this paper is the proposed “MF + FM” framework for HIN-based RS. By using HIN, we can incorporate various side information

into a unifying framework. In this section, we introduce how to recommend with the proposed “MF + FM” framework (Figure 3).

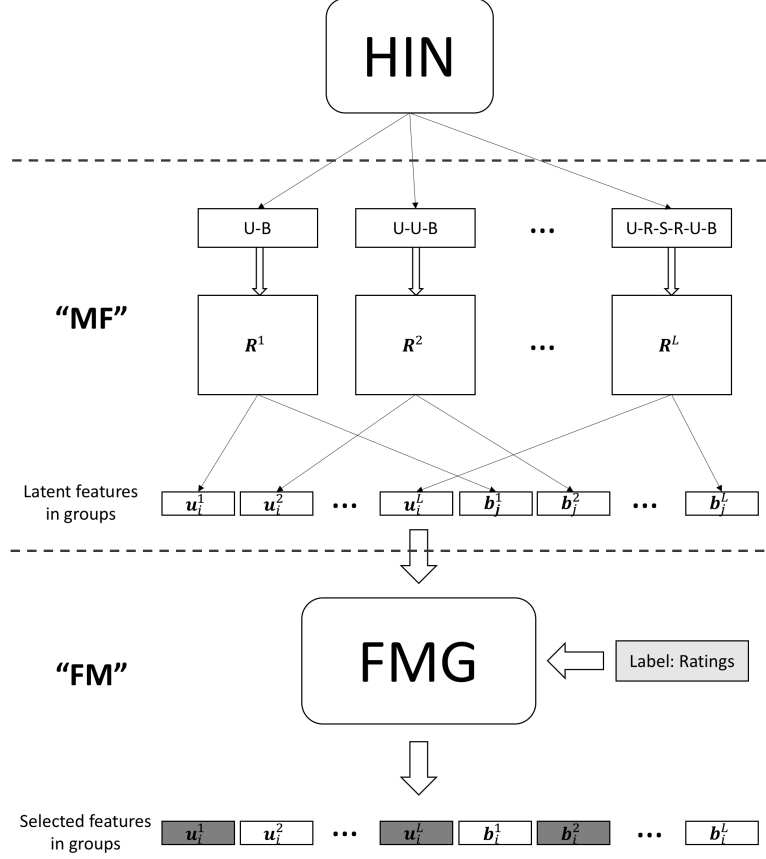


Figure 3: The proposed “MF + FM” Framework. The “MF” part: latent features are extracted from user-item similarity matrices, which is obtained from multiple meta-graphs based on an HIN (e.g., Figure 1). The “FM” part: latent features are concatenated and then fed into FMG model to predicted missing ratings. In the bottom, the features in grey are selected by FMG.

From Figure 3, we can see that the input of “MF” part is a HIN, e.g., the one in Figure 1, and the output is  $L$  groups of latent features of user and items, where  $L$  is the number of meta-graphs. The “MF” part, introduced in Section 3, generates latent features based on user-item similarity matrices using matrix factorization approaches.

The similarity matrices are computed based on multiple meta-graphs on HIN e.g., those in Figure 4. As existing methods only compute meta-path based similarities, we derive a new algorithm to compute the similarities between users and items under different meta-graphs.

Let  $\mathbf{R}^1, \mathbf{R}^2, \dots, \mathbf{R}^L$  be the  $L$  similarity matrices obtained. Since they tend to be very sparse, we use low-rank matrix approximation to factorize each similarity matrix into two low-dimensional matrices, representing the latent features of users and items, respectively.

The target of the “FM” part is to utilize these latent features to learn a better recommending model compared to previous HIN-based RSs. We propose to use FMG (See Section 4), which has two advantages over previous methods: 1) FM can capture non-linear interactions among features [25], which is more effective than linear ensemble model in the previous HIN-based RS [20]; 2) By introducing group lasso regularization, we can automatically select useful meta-graph based features, and thus determine which meta-graphs are better for new coming problems. Specifically, for a user-item pair, i.e., user  $u_i$  and item  $b_j$ , we first concatenate the latent features,  $\mathbf{u}_i^1, \mathbf{u}_i^2, \dots, \mathbf{u}_i^L$  and  $\mathbf{b}_i^1, \mathbf{b}_i^2, \dots, \mathbf{b}_i^L$ , from all the meta-graphs to create the feature vector, and the rating  $\mathbf{R}_{ij}$  is used as the label. We then train our FMG model with a special regularization method, which can select the useful features in groups, where each group corresponds to one meta-graph.

To efficiently solve the problem, in Section 5, we propose two algorithms, one is based on the proximal gradient algorithm [26] and the other based on the stochastic variance reduced gradient algorithm [27]. After training, FMG can select useful user and item latent features in groups, each of which corresponds to one meta-graph. The selected features are in grey in Figure 3.

### 3. Matrix Factorization (MF) for Feature Extraction

In this section, we elaborate the “MF” part for feature extraction. First, we compute the user-item similarity matrices in Section 3.1. Then, in Section 3.2, we obtain latent features based on these matrices using MF approaches.

#### 3.1. Meta-graph based Similarity Matrices Computation

According to [15, 22, 23], we first give the definitions of HIN, Network Schema for HIN, and Meta-graph, and then we introduce how to compute the meta-graph based similarities between users and items for recommendation.

**Definition 1** (Heterogeneous Information Network). A **heterogeneous information network** (HIN) is a graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  with an entity type mapping  $\phi: \mathcal{V} \rightarrow \mathcal{A}$  and a relation type mapping  $\psi: \mathcal{E} \rightarrow \mathcal{R}$ , where  $\mathcal{V}$  denotes the entity set,  $\mathcal{E}$  denotes the link set,  $\mathcal{A}$  denotes the entity type set, and  $\mathcal{R}$  denotes the relation type set, and the number of entity types  $|\mathcal{A}| > 1$  or the number of relation types  $|\mathcal{R}| > 1$ .

**Definition 2** (Network Schema). Given a HIN  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  with the entity type mapping  $\phi: \mathcal{V} \rightarrow \mathcal{A}$  and the relation type mapping  $\psi: \mathcal{E} \rightarrow \mathcal{R}$ , the **network schema** for network  $\mathcal{G}$ , denoted as  $\mathcal{T}_{\mathcal{G}} = (\mathcal{A}, \mathcal{R})$ , is a graph with nodes as entity types from  $\mathcal{A}$  and edges as relation types from  $\mathcal{R}$ .

In Figures 1 and 2, we give examples of a HIN and the network schema on the Yelp dataset, respectively. We can see that we have different types of nodes, e.g., *User*, *Review*, *Restaurant*, and different types of relations, e.g., *Write*, *Check-in*. The network schema defines the relations between node types, e.g., *User* Check-in *Restaurant*, *Restaurant* LocatedIn *City*. The definition of meta-graph is given below.

**Definition 3** (Meta-graph). A meta-graph  $\mathcal{M}$  is a directed acyclic graph (DAG) with a single source node  $n_s$  (i.e., with in-degree 0) and a single sink (target) node  $n_t$  (i.e., with out-degree 0), defined on an HIN  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ . Formally,  $\mathcal{M} = (\mathcal{V}_M, \mathcal{E}_M, \mathcal{A}_M, \mathcal{R}_M, n_s, n_t)$ , where  $\mathcal{V}_M \subseteq \mathcal{V}$  and  $\mathcal{E}_M \subseteq \mathcal{E}$  are constrained by  $\mathcal{A}_M \subseteq \mathcal{A}$  and  $\mathcal{R}_M \subseteq \mathcal{R}$ , respectively.

We show all the meta-graphs used in this paper on the Yelp dataset in Figure 4 and the Amazon dataset in Figure 5. We can see that they are DAGs with  $U$  (*User*) as the source node and  $B$  (*Business* for Yelp and *Product* for Amazon) as the target node. Here we use  $\mathcal{M}_3$  and  $\mathcal{M}_9$  in the Yelp dataset to illustrate the computation process for meta-graph based similarities.

Originally, commuting matrices [15] have been defined to compute the count-based similarity matrix for a meta-path. Suppose we have a meta-path  $\mathcal{P} = (\mathbb{A}_1, \mathbb{A}_2, \dots, \mathbb{A}_l)$ , where  $\mathbb{A}_i$ 's are node types in  $\mathcal{A}$ , and we can define a matrix  $\mathbf{W}_{\mathbb{A}_i \mathbb{A}_j}$  as the adjacency matrix between type  $\mathbb{A}_i$  and type  $\mathbb{A}_j$ . Then the commuting matrix for the path  $\mathcal{P}$  is defined by the multiplication of a sequence of adjacency matrices,

$$\mathbf{C}_P = \mathbf{W}_{\mathbb{A}_1, \mathbb{A}_2} \mathbf{W}_{\mathbb{A}_2, \mathbb{A}_3} \cdots \mathbf{W}_{\mathbb{A}_{l-1}, \mathbb{A}_l},$$

where  $\mathbf{C}_P(i, j)$ , the entry in the  $i$ -th row and  $j$ -th column, represents the number of path instances between object  $x_i \in \mathbb{A}_1$  and object  $x_j \in \mathbb{A}_l$  under meta-path  $\mathcal{P}$ . For example, for  $\mathcal{M}_3$  in Figure 4,  $\mathbf{C}_{M_3} = \mathbf{W}_{UB} \mathbf{W}_{UB}^\top \mathbf{W}_{UB}$ , where  $\mathbf{W}_{UB}$  is the adjacency matrix between type  $U$  and type  $B$ ,  $\mathbf{C}_{M_3}(i, j)$  represents the number of instances of  $\mathcal{M}_3$  between user  $u_i$  and item  $b_j$ . In this paper, for a meta-graph  $\mathcal{M}$ , we use the number of the instances of  $\mathcal{M}$  between a source object and target object as the similarity between them. In the remaining part of this paper, we use similarity matrix instead of commuting matrix for the sake of clarity.

From the above introduction, we can see that the meta-path based similarity matrix is easy to computed. However, for meta-graphs, the problem becomes more complicated. For example, consider  $\mathcal{M}_9$  in Figure 4, there are



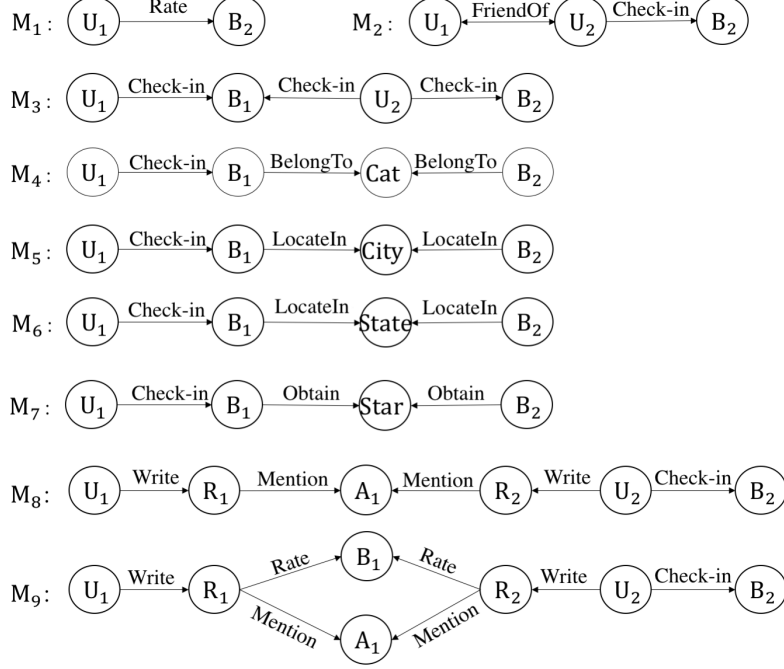


Figure 4: Meta-graphs used for the Yelp dataset (Star: the average stars a business obtained).

two ways to pass through the meta-graph, which are  $(U, R, A, R, U, B)$  and  $(U, R, B, R, U, B)$ . Note that  $R$  represents the entity type *Review* in HIN. In the path  $(U, R, A, R, U, B)$ ,  $(R, A, R)$  means if two reviews both mention the same  $A$  (*Aspect*), then they have some similarity. Similarly, in  $(U, R, B, R, U, B)$ ,  $(R, B, R)$  means if two reviews both rate the same  $B$  (*Business*), they have some similarity as well. We should define our logic of similarity when there are multiple ways for a flow to pass through the meta-graph from the source node to the target one. When there are two paths, we can allow a flow to pass through either path, or we constrain a flow to satisfy both of them. By analyzing the former strategy, we find that it is similar to simply split such meta-graph into multiple meta-paths and then adopt our later computation. Thus, we choose the latter, which requires one more matrix operation other than simple multiplication, i.e., element-wise product. Algorithm 1 depicts the algorithm for computing the count-based similarity for  $\mathcal{M}_9$  in Figure 4. After obtaining  $\mathbf{C}_{S_r}$ , it is easy to obtain the whole similarity matrix  $\mathbf{C}_{M_9}$  by the multiplication of a sequence of matrices. In practice, not limited to  $\mathcal{M}_9$  in Figure 4, the meta-graph defined in this paper can be computed by two operations (Hadamard product and multiplication) on the corresponding matrices.

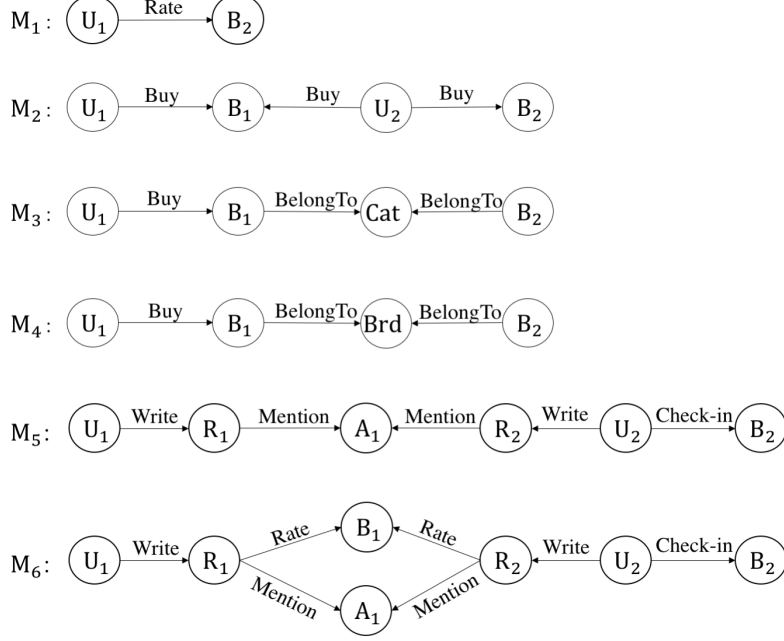


Figure 5: Meta-graphs used for the Amazon-200K dataset (Brd: brand of the item).

By computing the similarities between all users and items for the  $l$ -th meta-graph  $\mathcal{M}$ , we can obtain a user-item similarity matrix  $\mathbf{R}^l \in \mathbb{R}^{m \times n}$ , where  $\mathbf{R}_{ij}^l$  represents the similarity between user  $u_i$  and item  $b_j$  along the meta-graph, and  $m$  and  $n$  are the number of users and items, respectively. Note that  $\mathbf{R}_{ij}^l = \mathbf{C}_{M_l}(i, j)$ <sup>1</sup> if  $\mathbf{C}_{M_l}(i, j) > 0$  and 0 otherwise. By designing  $L$  meta-graphs, we can get  $L$  different user-item similarity matrices, denoted by  $\mathbf{R}^1, \dots, \mathbf{R}^L$ .

### 3.2. Meta-graph based Latent Feature Generation

In this section, we elaborate how to generate latent features of users and item from the obtained  $L$  user-item similarity matrices. Since the similarity matrices are usually very sparse, using them directly as features will lead to the high-dimensional learning problem, suffering from overfitting. Note that there exists similarity among users and items, motivated by the recent success of matrix completion for recommender systems [29, 2, 30], we propose to reduce the noise and deal with the sparsity of the similarity matrices by low-rank matrix approximation.

<sup>1</sup>To maintain consistency with the remaining sections, we change the notation  $\mathbf{C}$  into  $\mathbf{R}$ .

---

**Algorithm 1** Computing similarity matrix based on  $\mathcal{M}_9$ .

---

- 1: Compute  $\mathbf{C}_{P_1} : \mathbf{C}_{P_1} = \mathbf{W}_{RB} \mathbf{W}_{RB}^\top$ ;
  - 2: Compute  $\mathbf{C}_{P_2} : \mathbf{C}_{P_2} = \mathbf{W}_{RA} \mathbf{W}_{RA}^\top$ ;
  - 3: Compute  $\mathbf{C}_{S_r} : \mathbf{C}_{S_r} = \mathbf{C}_{P_1} \odot \mathbf{C}_{P_2}$ ;
  - 4: Compute  $\mathbf{C}_{M_9} : \mathbf{C}_{M_9} = \mathbf{W}_{UR} \mathbf{C}_{S_r} \mathbf{W}_{UR}^\top \mathbf{W}_{UB}$ .
- 

Specifically, the nonzero elements in a similarity matrix are treated as observations and the others are taken as missing ones, then we find a low-rank approximation to this matrix. Matrix factorization [2, 29] and nuclear norm regularization (NNR) [30, 31] are two popular approaches. We describe how latent features can be extracted using them in the sequel.

### 3.2.1. Matrix Factorization

Consider a user-item similarity matrix  $\mathbf{R} \in \mathcal{R}^{m \times n}$ . Let observed positions be indicated by 1's in  $\mathbf{\Omega} \in \{0, 1\}^{m \times n}$ , i.e.,  $[P_{\mathbf{\Omega}}(\mathbf{X})]_{ij} = \mathbf{X}_{ij}$  if  $\mathbf{\Omega}_{ij} = 1$  and 0 otherwise.  $\mathbf{R}$  is factorized as a product of  $\mathbf{U} \in \mathcal{R}^{m \times k}$  and  $\mathbf{V} \in \mathcal{R}^{n \times k}$  by solving the following optimization problem:

$$\min_{\mathbf{U}, \mathbf{B}} \frac{1}{2} \left\| P_{\mathbf{\Omega}}(\mathbf{UB}^\top - \mathbf{R}) \right\|_F^2 + \frac{\mu}{2} \left( \|\mathbf{U}\|_F^2 + \|\mathbf{B}\|_F^2 \right), \quad (1)$$

where  $k \ll \min(m, n)$  is the desired rank of  $\mathbf{R}$ , and  $\mu$  is the hyper-parameter controlling overfitting.

We adopt gradient descent based approach for optimizing (1), which has been popularly used in RS [2, 29]. After optimization, we take  $\mathbf{U}$  and  $\mathbf{B}$  as the latent features for users and items, respectively.

### 3.2.2. Nuclear Norm Regularization

Although MF can be simple, (1) is not a convex optimization problem, so there is no rigorous guarantee on the recovery performance. This motivates our adoption of the nuclear norm, which is defined as the sum of all singular values of a matrix. It is also the tightest convex envelope to the rank function. This leads to the following nuclear norm regularization (NNR) problem:

$$\min_{\mathbf{X}} \frac{1}{2} \left\| P_{\mathbf{\Omega}}(\mathbf{X} - \mathbf{R}) \right\|_F^2 + \mu \|\mathbf{X}\|_*. \quad (2)$$

where  $\mathbf{X}$  is the low-rank matrix to be recovered. Nice theoretical guarantee has been developed for (2), which shows that  $\mathbf{X}$  can be exactly recovered given sufficient observations [30]. These advantages make NNR popular for low-rank matrix approximation [30, 31].

Here, we also adopt (2) to generate latent features. We use the state-of-art AIS-Impute algorithm [32] for optimizing (2). It has fast  $O(1/T^2)$  convergence rate, where  $T$  is the number of iterations, with low per-iteration time complexity. In the iterations, a SVD decomposition of  $\mathbf{X} = \mathbf{P}\mathbf{\Sigma}\mathbf{Q}^\top$  is maintained ( $\mathbf{\Sigma}$  only contains the nonzero singular values). When the algorithm stops, we take  $\mathbf{U} = \mathbf{P}\mathbf{\Sigma}^{\frac{1}{2}}$  and  $\mathbf{B} = \mathbf{Q}\mathbf{\Sigma}^{\frac{1}{2}}$  as user and item latent features, respectively.

#### 4. Factorization Machine for Fusing Meta-graph based Features

In this section, we introduce our FM-based algorithm to fuse different groups of features generated by MF based on multiple meta-graphs. In Section 3.2, we obtain  $L$  groups of latent features of users and items, denoted as  $\mathbf{U}^1, \mathbf{B}^1, \dots, \mathbf{U}^L, \mathbf{B}^L$ , from  $L$  meta-graph based similarity matrices between users and items. For a sample  $\mathbf{x}^n$  in the observed ratings, i.e., a pair of user and item, denoted by  $\mathbf{u}_i$  and  $\mathbf{b}_j$ , we concatenate all of the corresponding user and item features from all of the  $L$  meta-graphs:

$$\mathbf{x}^n = [\underbrace{\mathbf{u}_i^1, \dots, \mathbf{u}_i^L}_{\sum_{l=1}^L F_l}, \underbrace{\mathbf{b}_j^1, \dots, \mathbf{b}_j^L}_{\sum_{l=1}^L F_l}] \in \mathbb{R}^d, \quad (3)$$

where  $d = 2 \sum_{l=1}^L F_l$ , and  $F_l$  is the rank of the factorization of the similarity matrix for the  $l$ -th meta-graph by (1) or (2), and  $\mathbf{u}_i^l$  and  $\mathbf{b}_j^l$ , respectively, represent user and item latent features generated from the  $l$ -th meta-graph.  $\mathbf{x}^n$  represents the feature vector of the  $n$ -th sample after concatenation. Then, each user and item can be represented by the  $\sum_{l=1}^L F_l$  latent features, respectively.

Given all of the features in (3), the predicted rating for the sample  $\mathbf{x}^n$  based on FM [25] is computed as follows.

$$\hat{y}^n(\mathbf{w}, \mathbf{V}) = b + \sum_{i=1}^d w_i x_i^n + \sum_{i=1}^d \sum_{j=i+1}^d \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i^n x_j^n, \quad (4)$$

where  $b$  is the global bias,  $\mathbf{w} \in \mathbb{R}^d$ , represents the first-order weights of the features, and  $\mathbf{V} = [\mathbf{v}_i] \in \mathbb{R}^{d \times K}$  represents the second-order weights for modeling the interactions among different features.  $\mathbf{v}_i$  is the  $i$ -th row of the matrix  $\mathbf{V}$ , which describes the  $i$ -th variable with  $K$  factors.  $\mathbf{x}_i^n$  is the  $i$ -th feature in  $\mathbf{x}^n$ . The parameters can be learned by minimizing the mean square loss:

$$\ell(\mathbf{w}, \mathbf{V}) = \frac{1}{N} \sum_{n=1}^N (y^n - \hat{y}^n(\mathbf{w}, \mathbf{V}))^2, \quad (5)$$

where  $y^n$  is an observed rating for the  $n$ -th sample.  $N$  is the number of all

observed ratings.

#### 4.1. Meta-graph Selection with Group Lasso

There are two problems when FM is applied to the meta-graph based latent features. The first problem is that it may bring noise when there are many meta-graphs, thus impairing the predicting capability of the model. Moreover, in practice, some meta-graphs can be useless since the strategies they represent may be useless. The second problem is the computational cost. All of the features are generated by MF, which means that the design matrix, i.e., features fed to FM, is dense. It increases the computational cost for learning the parameters of the model as well as that of online recommendation.

To alleviate the above two problems, we propose a novel regularization for FM, i.e., the group lasso regularization [33, 34], which is a feature selection method on a group of variables. Given the pre-defined non-overlapping  $G$  groups  $\{\mathcal{I}_1, \dots, \mathcal{I}_G\}$  on the parameter  $\mathbf{p}$ , the regularization is defined as follows.

$$\phi(\mathbf{p}) = \sum_{g=1}^G \eta_g \|\mathbf{p}_{\mathcal{I}_g}\|_2, \quad (6)$$

where  $\|\cdot\|_2$  is the  $\ell_2$ -norm. In our model, the groups correspond to the meta-graph based features. For example,  $\mathbf{U}^l$  and  $\mathbf{B}^l$  are the user and item latent features generated by the  $l$ -th meta-graph. For a pair of user  $i$  and item  $j$ , the latent features are  $\mathbf{u}_i^l$  and  $\mathbf{b}_j^l$ . There are two corresponding groups of variables in  $\mathbf{w}$  and  $\mathbf{V}$  according to (4). With  $L$  meta-graphs, the features of users or items from every single meta-graph can be put in a group. We have in total  $2L$  groups of variables in  $\mathbf{w}$  and  $\mathbf{V}$ , respectively.

For the first-order parameters  $\mathbf{w}$  in (4), which is a vector, group lasso is applied to the subset of variables in  $\mathbf{w}$ . Then we have:

$$\hat{\phi}(\mathbf{w}) = \sum_{l=1}^{2L} \hat{\eta}_l \|\mathbf{w}_l\|_2, \quad (7)$$

where  $\mathbf{w}_l \in \mathbb{R}^{F_l}$ , which models the weights for a group of user or item features from one meta-graph. For the second-order parameters  $\mathbf{V}$  in (4), we have the regularizer as follows.

$$\bar{\phi}(\mathbf{V}) = \sum_{l=1}^{2L} \bar{\eta}_l \|\mathbf{V}_l\|_F, \quad (8)$$

where  $\mathbf{V}_l \in \mathbb{R}^{F_l \times K}$ , the  $l$ -th block of  $\mathbf{V}$  corresponding to the  $l$ -th meta-graph based features in a sample, and  $\|\cdot\|_F$  is the Frobenius norm.

As a result, our model can simultaneously predict missing ratings, and automatically select useful meta-graphs.

#### 4.2. Nonconvex Regularization

While convex regularizers usually make the optimization easy, they often lead to biased estimation. For example, in sparse coding, the solution obtained by the  $\ell_1$ -regularizer is often not as sparse and accurate compared to [35]. Besides, in low-rank matrix learning, the estimated rank obtained with the nuclear norm regularizer is often very high [36]. To alleviate this problem, a number of nonconvex regularizers, which are variants of the convex  $\ell_1$ -norm, have been recently proposed. Empirically, these nonconvex regularizers usually outperform the convex ones. Motivated by the above observations, we propose to use nonconvex variant of (7) and (8) as follows.

$$\hat{\psi}(\mathbf{w}) = \sum_{l=1}^{2L} \hat{\eta}_l \kappa(\|\mathbf{w}_l\|_2), \quad \bar{\psi}(\mathbf{V}) = \sum_{l=1}^{2L} \bar{\eta}_l \kappa(\|\mathbf{V}_l\|_F), \quad (9)$$

where  $\kappa$  is a nonconvex penalty function. Here, we choose  $\kappa(|\alpha|) = \log(1 + |\alpha|)$  as the log-sum-penalty (LSP) [37], as it has been shown to give the best empirical performance on learning sparse vectors [38] and low-rank matrices [36].

#### 4.3. Comparison with Latent Feature Based Model

Yu et.al. studied recommendation based on HINs [20] and applied matrix factorization to generate latent features from different meta-paths and predict the rating by a weighted ensemble of dot product of user and item latent features from every single meta-path:  $\hat{r}(\mathbf{u}_i, \mathbf{b}_j) = \sum_{l=1}^L \theta_l \cdot \mathbf{u}_i^l (\mathbf{b}_j^l)^\top$ , where  $\hat{r}(\mathbf{u}_i, \mathbf{b}_j)$  is the predicted rating for user  $u_i$  and item  $b_j$ ,  $\mathbf{u}_i^l$  and  $\mathbf{b}_j^l$  are the latent features for  $u_i$  and item  $b_j$  from the  $l$ -th meta-path, respectively.  $L$  is the number of meta-paths used, and  $\theta_l$  is the weight for the  $l$ -th meta-path latent features. However, the predicting method is not adequate, as it fails to capture the interactions between inter-meta-path features, i.e., features across different meta-paths, and between the intra-meta-path features, i.e., features from the same meta-path. It may decrease the prediction performance for all of the features.

### 5. Model Optimization

Combining (5) and (9), we define our FM with Group lasso (FMG) model with the following objective function:

$$h(\mathbf{w}, \mathbf{V}) = \frac{1}{N} \sum_{n=1}^N (y^n - \hat{y}^n(\mathbf{w}, \mathbf{V}))^2 + \hat{\lambda} \hat{\psi}(\mathbf{w}) + \bar{\lambda} \bar{\psi}(\mathbf{V}). \quad (10)$$

Note that when  $\kappa(\alpha) = |\alpha|$  in (9), we get back (7) and (8). Thus, we directly use the nonconvex regularization in (10).

We can see that  $h$  is nonsmooth due to the use of  $\hat{\phi}^{\mathbf{w}}$  and  $\hat{\phi}^{\mathbf{V}}$ , and nonconvex due to the nonconvexity of loss  $\ell$  on  $\mathbf{w}$  and  $\mathbf{V}$ . To alleviate the difficulty on optimization, inspired by [38], we propose to reformulate (10) as follows.

$$\bar{h}(\mathbf{w}, \mathbf{V}) = \bar{\ell}(\mathbf{w}, \mathbf{V}) + \kappa_0 \hat{\lambda} \hat{\phi}(\mathbf{w}) + \kappa_0 \bar{\lambda} \bar{\phi}(\mathbf{V}), \quad (11)$$

where  $\bar{\ell}(\mathbf{w}, \mathbf{V}) = \ell(\mathbf{w}, \mathbf{V}) + g(\mathbf{w}, \mathbf{V})$ ,  $\kappa_0 = \lim_{\beta \rightarrow 0^+} \kappa'(|\beta|)$  and

$$g(\mathbf{w}, \mathbf{V}) = \hat{\lambda} [\hat{\psi}(\mathbf{w}) - \kappa_0 \hat{\phi}(\mathbf{w})] + \bar{\lambda} [\bar{\psi}(\mathbf{V}) - \kappa_0 \bar{\phi}(\mathbf{V})].$$

Note that  $\bar{h}$  is equivalent to  $h$ , based on Proposition 2.1 in [38]. A very important property for the augmented loss  $\bar{\ell}$  is that it is still smooth. As a result, while we are still optimizing a nonconvex regularized problem, we only need to deal with convex regularizers.

In the sequel, in Section 5.1, we show how the reformulated problem can be solved by the state-of-art proximal gradient algorithm [39]; moreover, such transformation enables us to design a more efficient optimization algorithm with convergence guarantee based on variance reduced methods [27]. Finally, the time complexity of the proposed algorithms is analyzed in Section 5.3.

### 5.1. Nonmonotonous Accelerated Proximal Gradient (nmAPG) Algorithm

To tackle the nonconvex nonsmooth objective function (11), we propose to use the PG algorithm [26]. Specifically, the state-of-the-art nonmonotonous accelerated proximal gradient (nmAPG) algorithm [39] is used. It targets at optimization problems of the form:

$$\min_{\mathbf{x}} F(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x}), \quad (12)$$

where  $f$  is a smooth (possibly nonconvex) loss function and  $g$  is a regularizer (can be nonsmooth and nonconvex). To guarantee the convergence of nmAPG, we also need  $\lim_{\|\mathbf{x}\|_2 \rightarrow \infty} F(\mathbf{x}) = \infty$ ,  $\inf_{\mathbf{x}} F(\mathbf{x}) > -\infty$  and there exists at least one solution to the proximal step, i.e.,  $\text{prox}_{\gamma g}(\mathbf{z}) = \arg \min_{\mathbf{x}} \frac{1}{2} \|\mathbf{x} - \mathbf{z}\|_2^2 + \gamma g(\mathbf{x})$ , where  $\gamma \geq 0$  is an arbitrary scalar [39].

The motivation of nmAPG comes from two facts. First, nonsmoothness comes from the proposed regularizers, which can be efficiently handled once the corresponding proximal steps have cheap closed-form solution. Second, the acceleration technique is useful for significantly speeding up first order optimization algorithms [38, 39, 40], and nmAPG is the state-of-art algorithm which can deal with general nonconvex problems with sound convergence

guarantee. The whole procedure is given in Algorithm 2. Note that while both  $\hat{\phi}$  and  $\bar{\phi}$  are nonsmooth in (11), they are imposed on  $\mathbf{w}$  and  $\mathbf{V}$  separately. Thus, for any  $\alpha, \beta \geq 0$ , we can also compute proximal operators independently for these two regularizers following [26]:

$$\text{prox}_{\alpha\hat{\phi}+\beta\bar{\phi}}(\mathbf{w}, \mathbf{V}) = \left( \text{prox}_{\alpha\hat{\phi}}(\mathbf{w}), \text{prox}_{\beta\bar{\phi}}(\mathbf{V}) \right). \quad (13)$$

These are performed in step 5 and 10 in Algorithm 2. The closed-form solution of those proximal operators can be obtained easily from Lemma 1 below. Thus, each proximal operator can be solved in one pass of all groups.

**Lemma 1** ([41]). *The closed-form solution of  $\mathbf{p}^* = \text{prox}_{\lambda\phi}(\mathbf{z})$  ( $\phi$  is defined in (6)) is given by  $\mathbf{p}_{\mathcal{I}_g}^* = \max\left(1 - \frac{\eta_g}{\|\mathbf{z}_{\mathcal{I}_g}\|_2}, 0\right) \mathbf{z}_{\mathcal{I}_g}$  for all  $g = 1, \dots, G$ .*

It is easy to verify that the above assumptions are satisfied by our objective  $h$  here. Thus, Algorithm 2 is guaranteed to produce a critical point for (11).

### 5.2. Stochastic Variance Reduced Gradient (SVRG) Algorithm

While nmAPG can be an efficient algorithm for (11), it is still a batch-gradient based method, which may not be efficient enough when the sample size is large. In this case, the stochastic gradient descent (SGD) [42] algorithm is preferred as it can incrementally update the learning parameters. However, the gradient in SGD is very noisy. To ensure the convergence of SGD, a decreasing step-size needs to be used, making it possibly even slower than batch-gradient methods.

Recently, the stochastic variance reduction gradient (SVRG) [27] algorithm has been developed. It avoids the diminishing step-size by introducing variance reduced techniques into gradient updates. As a result, it combines the best of both worlds, i.e., incremental update of the learning parameters while keeping non-diminishing step-size, to achieve significantly faster converging speed than SGD. Besides, it is also extended for problem in (12) with nonconvex objectives [43, 44]. This allows the loss function to be smooth (possibly nonconvex) but the regularizer still needs to be convex. Thus, instead of working on the original problem (10), we work on the transformed one in (11).

To use SVRG, we first define the augmented loss for the  $n$ -th sample as  $\bar{\ell}_n(\mathbf{w}, \mathbf{V}) = (y^n - \hat{y}^n(\mathbf{w}, \mathbf{V}))^2 + \frac{1}{N}g(\mathbf{w}, \mathbf{V})$ . The whole procedure is depicted in Algorithm 3. A full gradient is computed in step 4, a mini-batch  $\mathcal{B}$  of size  $m_b$  is constructed in step 6, and the variance reduced gradient is computed in step 7. Finally, the proximal steps can be separately executed based on (13) in step 8. As mentioned above, the nonconvex variant of SVRG [43, 44] cannot be directly applied on (10). Instead, we apply it on the transformed problem (11), where the regularizer becomes convex and the augmented loss is still smooth. Thus, Algorithm 3 is guaranteed to generate a critical point of (11).



---

**Algorithm 2** nmAPG [39] algorithm for (11).

---

```

1: Initiate  $\mathbf{w}_0, \mathbf{V}_0$  as Gaussian random matrices;
2:  $\bar{\mathbf{w}}_1 = \mathbf{w}_1 = \mathbf{w}_0, \bar{\mathbf{V}}_1 = \mathbf{V}_1 = \mathbf{V}_0, c_1 = \bar{h}(\mathbf{w}_1, \mathbf{V}_1); q_1 = 1, \delta = 10^{-3}, a_0 = 0, a_1 = 1$ , step-size  $\alpha$ ;
3: for  $t = 1, 2, \dots, T$  do
4:    $\mathbf{y}_t = \mathbf{w}_t + \frac{a_{t-1}}{a_t}(\bar{\mathbf{w}}_t - \mathbf{w}_t) + \frac{a_{t-1}-1}{a_t}(\mathbf{w}_t - \mathbf{w}_{t-1});$ 
    $\mathbf{Y}_t = \mathbf{V}_t + \frac{a_{t-1}}{a_t}(\bar{\mathbf{V}}_t - \mathbf{V}_t) + \frac{a_{t-1}-1}{a_t}(\mathbf{V}_t - \mathbf{V}_{t-1});$ 
5:    $\bar{\mathbf{w}}_{t+1} = \text{prox}_{\alpha\kappa_0\hat{\lambda}\hat{\phi}}(\mathbf{w}_t - \alpha\nabla_{\mathbf{w}}\bar{\ell}(\mathbf{w}_t, \mathbf{V}_t));$ 
    $\bar{\mathbf{V}}_{t+1} = \text{prox}_{\alpha\kappa_0\bar{\lambda}\bar{\phi}}(\mathbf{V}_t - \alpha\nabla_{\mathbf{V}}\bar{\ell}(\mathbf{w}_t, \mathbf{V}_t));$ 
6:    $\Delta_t = \|\bar{\mathbf{w}}_{t+1} - \mathbf{y}_t\|_2^2 + \|\bar{\mathbf{V}}_{t+1} - \mathbf{Y}_t\|_F^2$ 
7:   if  $\bar{h}(\bar{\mathbf{w}}_{t+1}, \bar{\mathbf{V}}_{t+1}) \leq c_t - \delta\Delta_t$ ; then
8:      $\mathbf{w}_{t+1} = \bar{\mathbf{w}}_{t+1}, \mathbf{V}_{t+1} = \bar{\mathbf{V}}_{t+1};$ 
9:   else
10:     $\hat{\mathbf{w}}_{t+1} = \text{prox}_{\alpha\kappa_0\hat{\lambda}\hat{\phi}}(\mathbf{w}_t - \alpha\nabla_{\mathbf{w}}\bar{\ell}(\mathbf{w}_t, \mathbf{V}_t));$ 
     $\hat{\mathbf{V}}_{t+1} = \text{prox}_{\alpha\kappa_0\bar{\lambda}\bar{\phi}}(\mathbf{V}_t - \alpha\nabla_{\mathbf{V}}\bar{\ell}(\mathbf{w}_t, \mathbf{V}_t));$ 
11:    if  $\bar{h}(\hat{\mathbf{w}}_{t+1}, \hat{\mathbf{V}}_{t+1}) < \bar{h}(\bar{\mathbf{w}}_{t+1}, \bar{\mathbf{V}}_{t+1})$  then
12:       $\mathbf{w}_{t+1} = \hat{\mathbf{w}}_{t+1}, \mathbf{V}_{t+1} = \hat{\mathbf{V}}_{t+1};$ 
13:    else
14:       $\mathbf{w}_{t+1} = \bar{\mathbf{w}}_{t+1}, \mathbf{V}_{t+1} = \bar{\mathbf{V}}_{t+1};$ 
15:    end if
16:  end if
17:   $a_{t+1} = \frac{1}{2}(\sqrt{4a_t^2 + 1} + 1);$ 
18:   $q_{t+1} = \eta q_t + 1, c_{t+1} = \frac{1}{q_{t+1}}(\eta q_t c_t + \bar{h}(\mathbf{w}_{t+1}, \mathbf{V}_{t+1}));$ 
19: end for
20: return  $\mathbf{w}_{T+1}, \mathbf{V}_{T+1}$ .

```

---

### 5.3. Complexity Analysis

For nmAPG in Algorithm 2, the main computation cost is incurred in performing the proximal steps (step 5 and 10) which cost  $O(NKd)$ ; then the evaluation of function value (step 7 and 11) costs  $O(NKd)$  time. Thus, per-iteration time complexity for Algorithm 2 is  $O(NKd)$ . For SVRG in Algorithm 3, the computation of the full gradient takes  $O(NKd)$  in step 5; then  $O(m_bBKd)$  is needed for steps 6-10 to perform mini-batch updates. Thus, one iteration in Algorithm 2 takes  $O((N + m_bB)Kd)$  time. Usually,  $m_bB$  shares the same order as  $N$  [27, 43, 44]. Thus, we set  $m_bB = N$  in our experiment. As a result, SVRG needs more time to perform one iteration than nmAPG. However, due to stochastic updates, SVRG empirically converges much faster as shown in Section 6.8.

---

**Algorithm 3** SVRG [43, 44] algorithm for (11).

---

```

1: Initiate  $\bar{\mathbf{w}}_0, \bar{\mathbf{V}}_0$  as Gaussian random matrices, mini-batch size  $m_b$ ;
2:  $\mathbf{w}_1^B = \bar{\mathbf{w}}_0, \mathbf{V}_1^B = \bar{\mathbf{V}}_0$  and step-size  $\alpha$ ;
3: for  $t = 1, 2, \dots, T$  do
4:    $\mathbf{w}_{t+1}^0 = \mathbf{w}_t^B, \mathbf{V}_{t+1}^0 = \mathbf{V}_t^B$ ;
5:    $\bar{\mathbf{g}}_{t+1}^{\mathbf{w}} = \nabla_{\mathbf{w}} \bar{\ell}(\bar{\mathbf{w}}_t, \bar{\mathbf{V}}_t), \bar{\mathbf{g}}_{t+1}^{\mathbf{V}} = \nabla_{\mathbf{V}} \bar{\ell}(\bar{\mathbf{w}}_t, \bar{\mathbf{V}}_t)$ ;
6:   for  $b = 0, 1, \dots, B - 1$  do
7:     Uniformly randomly sample a mini-batch  $\mathcal{B}$  of size  $m_b$ ;
8:      $\mathbf{m}_{\mathbf{w}}^b = \frac{1}{m_b} \sum_{i_b \in \mathcal{B}} (\nabla_{\mathbf{w}} \ell_{i_b}(\mathbf{w}_t^b, \mathbf{V}_t^b) - \nabla_{\mathbf{w}} \bar{\ell}_{i_b}(\bar{\mathbf{w}}_t, \bar{\mathbf{V}}_t)) + \bar{\mathbf{g}}_{t+1}^{\mathbf{w}},$ 
        $\mathbf{m}_{\mathbf{V}}^b = \frac{1}{m_b} \sum_{i_b \in \mathcal{B}} (\nabla_{\mathbf{V}} \ell_{i_b}(\mathbf{w}_t^b, \mathbf{V}_t^b) - \nabla_{\mathbf{V}} \bar{\ell}_{i_b}(\bar{\mathbf{w}}_t, \bar{\mathbf{V}}_t)) + \bar{\mathbf{g}}_{t+1}^{\mathbf{V}};$ 
9:      $\mathbf{w}_{t+1}^{b+1} = \text{prox}_{\alpha \kappa_0 \hat{\lambda} \hat{\phi}}(\mathbf{w}_{t+1}^b - \alpha \mathbf{m}_{\mathbf{w}}^b),$ 
        $\mathbf{V}_{t+1}^{b+1} = \text{prox}_{\alpha \kappa_0 \bar{\lambda} \bar{\phi}}(\mathbf{V}_{t+1}^b - \alpha \mathbf{m}_{\mathbf{V}}^b);$ 
10:   end for
11:    $\bar{\mathbf{w}}_{t+1} = \frac{1}{B} \sum_{b=1}^B \mathbf{w}_{t+1}^b, \bar{\mathbf{V}}_{t+1} = \frac{1}{B} \sum_{b=1}^B \mathbf{V}_{t+1}^b;$ 
12: end for
13: return  $\bar{\mathbf{w}}_{T+1}, \bar{\mathbf{V}}_{T+1}.$ 

```

---

## 6. Experiments

In this section, we conduct extensive experiments to demonstrate the effectiveness of our proposed framework. We first introduce the datasets, evaluation metric and experimental settings in Section 6.1. Then, in Section 6.2, we show the recommending performance of our proposed framework compared to several state-of-art recommending methods, including MF-based and HIN-based methods. Further, we analyze the influence of the parameter  $\lambda$  which controls the weight of convex regularization terms in Section 6.3, and then the influence of parameter  $\lambda$  for the nonconvex regularization term in Section 6.4. As a supplement, we show the performance of each single meta-graph in Section 6.5. In Section 6.6, we compare the performance between NNR and MF to extract the features. In Section 6.7, we show the influence of  $K$  of FMG. Finally, two proposed algorithms in Section 5 are compared in Section 6.8, and their scalability are demonstrated in Section 6.9.

### 6.1. Setup

To demonstrate the effectiveness of HIN for recommendation, we conduct experiments on four datasets with rich side information. The first dataset is Yelp, which is provided for the Yelp challenge.<sup>2</sup> Yelp is a website where a user can rate local businesses or post photos and reviews about them. The ratings fall

---

<sup>2</sup>[https://www.yelp.com/dataset\\_challenge](https://www.yelp.com/dataset_challenge)

in the range of 1 to 5, where higher ratings mean users like the businesses while lower rates mean users dislike businesses. Based on the information collected, the website can recommend businesses according to the users' preferences. The second dataset is Amazon Electronics,<sup>3</sup> which is provided in [45]. As we know, Amazon highly relies on RS to present interesting items to users who are surfing on the website. In [45] many domains of the Amazon dataset are provided, and we choose the electronics domain for our experiments. We extract subsets of entities from Yelp and Amazon to build the HIN, which includes diverse types and relations. The subsets of the two datasets both include around 200,000 ratings in the user-item rating matrices. Thus, we identify them as Yelp-200K and Amazon-200K, respectively. Besides, we also use the datasets provided in the CIKM paper [21], which we denote by CIKM-Yelp and CIKM-Douban. The statistics of our datasets are shown in Table 1. For the detailed information of CIKM-Yelp and CIKM-Douban, we refer the readers to [21]. Note that i) the number of types and relations in the first two datasets, i.e., Amazon-200K and Yelp-200K, in this paper are much more than those used in previous works [19, 20, 21]; ii) we give the density of the four datasets in Table 2. We can see that the densities of the rating matrices are much smaller than those previously used in [19, 20, 21].

Table 1: Statistics of the Yelp-200K and Amazon-200K datasets.

|        | Relations(A-B)    | Number<br>of A | Number<br>of B | Number<br>of (A-B) | Avg Degrees<br>of A/B |
|--------|-------------------|----------------|----------------|--------------------|-----------------------|
| Amazon | User-Review       | 59,297         | 183,807        | 183,807            | 3.1/1                 |
|        | Business-Category | 20,216         | 682            | 87,587             | 4.3/128.4             |
|        | Business-Brand    | 95,33          | 2,015          | 9,533              | 1/4.7                 |
|        | Review-Business   | 183,807        | 20,216         | 183,807            | 1/9.1                 |
|        | Review-Aspect     | 183,807        | 10             | 796,392            | 4.3/79,639.2          |
| Yelp   | User-Business     | 36,105         | 22,496         | 191,506            | 5.3/8.5               |
|        | User-Review       | 36,105         | 191,506        | 191,506            | 5.3/1                 |
|        | User-User         | 17,065         | 17,065         | 140,344            | 8.2/8.2               |
|        | Business-Category | 22,496         | 869            | 67,940             | 3/78.2                |
|        | Business-Star     | 22,496         | 9              | 22,496             | 1/2,499.6             |
|        | Business-State    | 22,496         | 18             | 22,496             | 1/1,249.8             |
|        | Business-City     | 22,496         | 215            | 22,496             | 1/104.6               |
|        | Review-Business   | 191,506        | 22,496         | 191,506            | 1/8.5                 |
|        | Review-Aspect     | 191,506        | 10             | 955,041            | 5/95,504.1            |

To evaluate the recommending performance, we adopt the root-mean-square-error (RMSE) as our metric, which is the most popular one for rating prediction

<sup>3</sup><http://jmcauley.ucsd.edu/data/amazon/>

Table 2: The density of rating matrices in the four datasets  $\text{Density} = \frac{\# \text{Ratings}}{\# \text{Users} \times \# \text{Items}}$ .

|         | Amazon-200K | Yelp-200K | CIKM-Yelp | CIKM-Douban |
|---------|-------------|-----------|-----------|-------------|
| Density | 0.015%      | 0.024%    | 0.086%    | 0.630%      |

in the literature [2, 10, 29]. It is defined as

$$\text{RMSE} = \sqrt{\frac{1}{|\mathcal{R}_{test}|} \sum_{y^n \in \mathcal{R}_{test}} (y^n - \hat{y}^n)^2},$$

where  $\mathcal{R}_{test}$  is the set of all the test samples,  $\hat{y}^n$  is the predicted rating for the  $n$ -th sample,  $y^n$  is the observed rating of the  $n$ -th sample in the test set. For RMSE, smaller value means better performance.

We compare the following models to our approaches.

- **RegSVD** [46]: The basic matrix factorization model with  $L_2$  regularization, which uses only the user-item rating matrix. We use the implementation in [47].
- **FMR** [25]: The factorization machine with only the user-item rating matrix. We adopt the method in Section 4.1.1 of [25] to model the rating prediction task. We use the code provided by the authors.<sup>4</sup>
- **HeteRec** [20]: It is based on meta-path based similarity between users and items. A weighted ensemble model is learned from the latent features of users and items generated by applying matrix factorization to the similarity matrices of different meta-paths. We implemented it based on [20].
- **SemRec** [21]: It is a meta-path based recommendation on weighted HIN, which is built by connecting users and items with the same ratings. Different models are learned from different meta-paths, and a weight ensemble method is used to predict the users' ratings. We use the code provided by the authors.<sup>5</sup>
- **FMG**: The proposed framework (Figure 3) with convex group lasso regularizer in (7) and (8) used with factorization machine.
- **FMG(LSP)**: Same as FMG, except that nonconvex group lasso regularizer in (9) is used.

<sup>4</sup><http://www.libfm.org/>

<sup>5</sup><https://github.com/zzqsmall/SemRec>

Note that it is reported in [21] that SemRec outperforms the method in [19], which uses meta-path based similarities as regularization terms in matrix factorization. Thus, we do not compare with [19] here.

The meta-graphs shown in Figure 4 and 5 are used in the experiments. To get the aspects (e.g.,  $A_1$  in Figure 4 and 5) from review texts, we use Gensim [48], a topic model software to extract topics, which are used as aspects. The number of topics is set to 10 empirically.

For the experimental settings, we randomly split 80% of the whole data set for training, 10% for validation and the remaining 10% for testing. The process is repeated five times and the average RMSE of the five rounds on test sets is reported. Our framework is implemented with Python 2.7, and all experiments run in a server (OS: CentOS release 6.9, CPU: Intel i7-3.4GHz, RAM: 32GB).

## 6.2. Recommendation Effectiveness

The results are shown in Table 3. Note that on CIKM-Yelp and CIKM-Douban datasets, we directly report the performance of SemRec from [21] as same amount of training data are used. Besides, the results of SemRec on Amazon-200K are not reported as the programs crashed due to large demand of memory. Firstly, we can see that our FMG model, including the convex and nonconvex ones, consistently outperforms all of the baselines on all four datasets. This demonstrates the effectiveness of the proposed framework shown in Figure 3. Note that the performance of FMG and FMG(LSP) are very close, but FMG(LSP) needs fewer features to achieve such performance, which verifies our motivation to use nonconvex regularization on selecting features. In the following two sections, we will compare in detail the two types of regularizers.

Table 3: Recommending performance in terms of RMSE. Percentages in the brackets are the reduction of RMSE comparing FMG with the corresponding approaches in the table header.

|          | Amazon-200K        | Yelp-200K          | CIKM-Yelp          | CIKM-Douban       |
|----------|--------------------|--------------------|--------------------|-------------------|
| RegSVD   | 2.9656<br>(-60.0%) | 2.5141<br>(-50.0%) | 1.5323<br>(-27.1%) | 0.7673<br>(-8.5%) |
| FMR      | 1.3462<br>(-11.0%) | 1.7637<br>(-28.7%) | 1.4342<br>(-11.0%) | 0.7524<br>(-6.7%) |
| HeteRec  | 2.5368<br>(-52.8%) | 2.3475<br>(-46.4%) | 1.4891<br>(-25.0%) | 0.7671<br>(-8.4%) |
| SemRec   | —<br>—             | 1.4603<br>(-13.8%) | 1.1559<br>(-3.4%)  | 0.7216<br>(-2.7%) |
| FMG      | <b>1.1953</b>      | <b>1.2583</b>      | <b>1.1167</b>      | <b>0.7023</b>     |
| FMG(LSP) | 1.1980             | 1.2593             | 1.1255             | 0.7035            |

Secondly, from Table 3, we can see that comparing to RegSVD and FMR,

which only use the rating matrix, SemRec and FMG, which use additional side information from meta-graphs, are significantly better. Especially, the sparser the rating matrix, the more benefit the additional information produces. For example, on Amazon-200K, FMG outperforms RegSVD by 60%, while for CIKM-Douban, the percentage of RMSE decrease is 8.5%. Note that the performance of HeteRec is worse than FMR, despite the fact that we have tried our best to tune the model. The reason is that, as we show in Section 4, using a weighting ensemble of dot product of latent features may lose information among the meta-graphs and fail to reduce noise caused by having too many meta-graphs.

When comparing the results of FMG and SemRec, we find that the performance gap between them are not that large, which means that SemRec is still a good method for rating prediction, especially when comparing SemRec to the other three baselines. The good performance of SemRec may be attributed to two reasons. First, incorporating rating values into HIN leads to a weighted HIN, which may better capture the meta-graph or meta-path based similarities. Second, the meta-graphs SemRec exploits are all of the style like  $U \rightarrow * \leftarrow U \rightarrow B$ , which has a good capability of predication. In Section 6.3, we will show that FMG can automatically select features constructed by meta-graphs like  $U \rightarrow * \leftarrow U \rightarrow B$  while removing those by meta-graphs like  $(U \rightarrow B \rightarrow * \leftarrow B)$ . In Section 6.5, we further study the prediction ability of each meta-graph, and also show that meta-graphs with style like  $U \rightarrow * \leftarrow U \rightarrow B$  are better than those like  $U \rightarrow B \rightarrow * \leftarrow B$ .

### 6.3. The Impact of Convex Regularizer

In this section, we study the impact of group lasso regularizer for our model. Specifically, we show the trend of RMSE of FMG by varying  $\lambda$  (with  $\hat{\lambda} = \bar{\lambda} = \lambda$  in (11)), which controls the weights of group lasso. For the sake of efficiency of parameter tuning, the experiments were conducted on Amazon-50K and Yelp-50K, where only 50,000 ratings are sampled and thus is a smaller version of Amazon-200K and Yelp-200K. The RMSE of Amazon-50K and Yelp-50K are shown in Figure 6(a) and (b), respectively. We can see that with  $\lambda$  increasing, RMSE decreases first and then increases, demonstrating that  $\lambda$  values that are too large or too small are not good for the performance of rating prediction. Specifically, on Amazon, the best performance is achieved when  $\lambda = 0.06$ , and on Yelp, the best is when  $\lambda = 0.05$ . Next, we give further analysis of these two parameters in terms of sparsity and the selected meta-graphs by group lasso.

#### 6.3.1. Sparsity of $\mathbf{w}$ , $\mathbf{V}$

We now study the sparsity of the learned parameters, i.e., the ratio of zeros in  $\mathbf{w}$ ,  $\mathbf{V}$  after learning. We define NNZ (number of non zeros) as  $\text{NNZ} = \frac{\text{nnz}}{w_n + v_n}$ , where  $\text{nnz}$  is the total number of nonzero elements in  $\mathbf{w}$  and  $\mathbf{V}$ , and  $w_n$  and  $v_n$

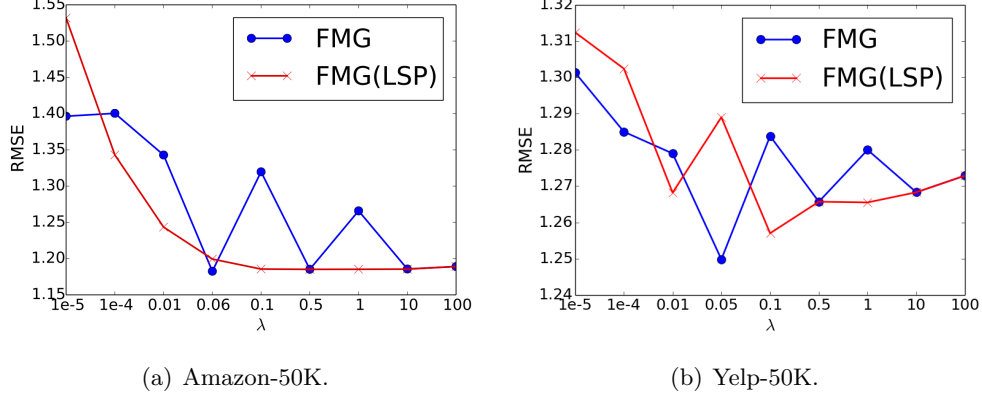


Figure 6: RMSE v.s various  $\lambda$  on the Amazon and Yelp datasets.

are the number of entries in  $\mathbf{w}$  and  $\mathbf{V}$ , respectively. The smaller NNZ is, the fewer nonzero elements in  $\mathbf{w}$  and  $\mathbf{V}$ , which means fewer meta-graph based features are left after training. The trend of NNZ with different  $\lambda$ 's is shown in Figure 7. We can see that with  $\lambda$  increasing, NNZ becomes smaller, which aligns with the effect of group lasso. Note that the trend is non-monotonous due to the nonconvexity of the objective.

### 6.3.2. The Selected Meta-graphs

In this section, we analyze the selected features by FMG. From Figure 6 (a) and (b), we can see that RMSE and sparsity are good when  $\lambda = 0.06$  on Amazon and  $\lambda = 0.05$  on Yelp. Thus, we want to show the selected meta-graphs in this configuration for user and item latent features. Recall that in Eq. (4), we introduce  $\mathbf{w}$  and  $\mathbf{V}$ , respectively, to capture the first-order weights for the features and second-order weights for interactions of the features. Thus, after training, the nonzero values in  $\mathbf{w}$  and  $\mathbf{V}$  represent the selected features, thus the selected meta-graphs. We list these selected meta-graphs corresponding to nonzero values in  $\mathbf{w}$  and  $\mathbf{V}$  from both the perspective of users and items in Table 4.

From Table 4, we can observe that the meta-graphs with style like  $U \rightarrow * \leftarrow U \rightarrow B$  are better than those like  $U \rightarrow B \rightarrow * \leftarrow B$ . Here, we use  $U \rightarrow * \leftarrow U \rightarrow B$  to represent meta-graphs like  $M_2, M_3, M_8, M_9$  in Figure 4 and  $M_2, M_5, M_6$  in Figure 5, and  $U \rightarrow B \rightarrow * \leftarrow B$  to represent meta-graphs like  $M_4, M_5, M_6, M_7$  in Figure 4 and  $M_3, M_4$  in Figure 5. For Yelp, we can see that meta-graphs like  $M_2, M_3, M_8, M_9$  tend to be selected while  $M_4 - M_7$  are removed, which means that on Yelp, recommendations by friends or similar users are better than those by similar items. Similar cases exist for Amazon, i.e.,  $M_3, M_4$  tend to be removed.

Table 4: The selected meta-graphs by FMG and FMG(LSP) on the Amazon and Yelp datasets. We show the selected latent features from the perspective of users and items on both first-order and second-order parameters.

|        |          | User-Part            |                     | Item-Part            |                 |
|--------|----------|----------------------|---------------------|----------------------|-----------------|
|        |          | first-order          | second-order        | first-order          | second-order    |
| Amazon | FMG      | $M_1-M_3, M_5$       | $M_1-M_6$           | $M_2, M_3, M_5, M_6$ | $M_2, M_5, M_6$ |
|        | FMG(LSP) | $M_1, M_5$           | -                   | $M_2, M_5$           | -               |
| Yelp   | FMG      | $M_1-M_4, M_6, M_8$  | $M_1-M_3, M_5, M_8$ | $M_1-M_5, M_8, M_9$  | $M_3, M_8$      |
|        | FMG(LSP) | $M_1, M_3, M_4, M_8$ | $M_2, M_3, M_8$     | $M_1-M_5, M_8$       | $M_8$           |

Besides, on both datasets, complex structures like  $M_9$  in Figure 4 and  $M_6$  in Figure 5 are determined to be important for item latent features, which demonstrates the importance of capturing these kinds of relations, which are ignored by previous meta-path based RSs [19, 20, 21].

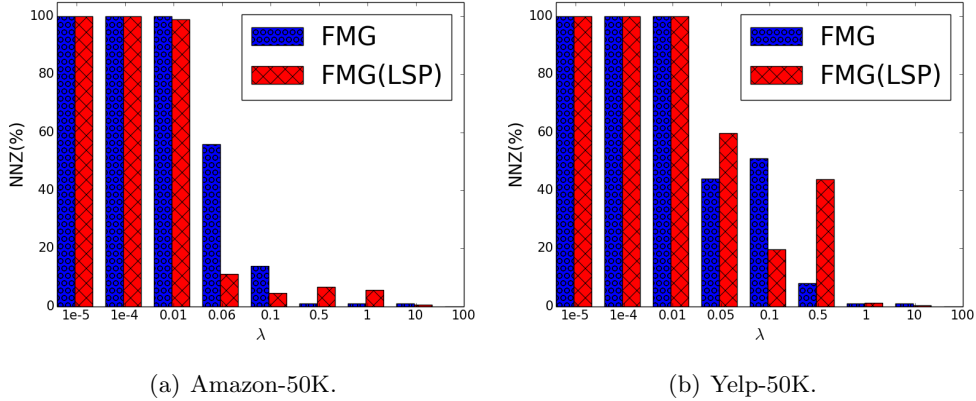


Figure 7: The trend of NNZ by varying  $\lambda$  on the Amazon and Yelp datasets.

#### 6.4. Impact of Nonconvex Regularizer

In this section, we study the performance of the nonconvex regularizer. To compare the results of the convex and nonconvex regularizers consistently, we also conduct experiments on Amazon-50K and Yelp-50K datasets.

The results are reported in the same manner as those in Section 6.3. The RMSEs of the nonconvex regularizer on Amazon-50K and Yelp-50K are shown in Figures 6(a) and (b), respectively. We observe that the trend of the nonconvex regularizer is similar to that of the convex regularizer. Specifically, on Amazon,



the best performance is achieved when  $\lambda = 0.5$ , and on Yelp, the best is when  $\lambda = 0.1$ . Note that the best performance with FMG(LSP) is a bit weaker than that of FMG on both datasets.

Similar to Section 6.3.1, we also use NNZ to show the performance of FMG(LSP) in Figure 7. We can see that with  $\lambda$  increasing, NNZ becomes smaller. Note that the trend is also non-monotonous due to the nonconvexity of the objective. Besides, NNZ of the parameters of FMG(LSP) is much smaller than that of FMG when the best performance on both Amazon and Yelp datasets is achieved. This is due to the effect of nonconvexity of LSP, which can induce larger sparsity of the parameters with a smaller loss of performance gain.

Next, we analyze the selected features by FMG(LSP). As done in FMG, we show the selected meta-graphs when the best performance is achieved in Figure 6, i.e.,  $\lambda = 0.5$  on Amazon and  $\lambda = 0.1$  on Yelp. The results of Amazon and Yelp are also shown in Table 4, and the observation is very similar to that of FMG, i.e., meta-graphs with style like  $U \rightarrow * \leftarrow U \rightarrow B$  are better than those like  $U \rightarrow B \rightarrow * \leftarrow B$ . For Yelp, we can see that meta-graphs like  $M_2, M_3, M_8$  tend to be selected while  $M_4 - M_7$  are removed. Similar cases exist on Amazon, i.e.,  $M_3, M_4$  tend to be removed.

An interesting discovery of this experiment is that for both datasets, complex structures are removed, i.e.,  $M_9$  on Yelp and  $M_6$  on Amazon. This is due to the fact that they may not be the most important features for the overall recommending performance. However, considering the performance of these two types of regularizers, the loss of performance gain of FMG(LSP) comparing to FMG may be caused by this difference. Thus it further demonstrates the importance of incorporating complex structures for recommendation.

### 6.5. Recommending Performance with Single Meta-Graph

In this section, we compare the performance of different meta-graphs separately. In the training process, we use only one meta-graph for user and item features and then predict with FMG and evaluate the results obtained by the corresponding meta-graph. Specifically, we run experiments to compare RMSE of all the meta-graphs in Figure 4 and 5. The RMSE of each meta-graph is shown in Figure 8. Note that we show for comparison the RMSE when all meta-graphs are used, which is denoted by  $M_{all}$ .

From Figure 8, we can see that on both Amazon and Yelp, the performance is the best when all meta-graph based user and item features are used, which demonstrates the usefulness of the semantics captured by the designed meta-graphs in Figure 4 and 5. Besides, we can see that on Yelp, the performance of  $M_4 - M_7$  is the worst, and on Amazon, the performance of  $M_3 - M_4$  is also among the worst three. Note that they are both meta-graphs with style like  $U \rightarrow B \rightarrow * \leftarrow B$ . Thus, it aligns with the observation in the above two sections

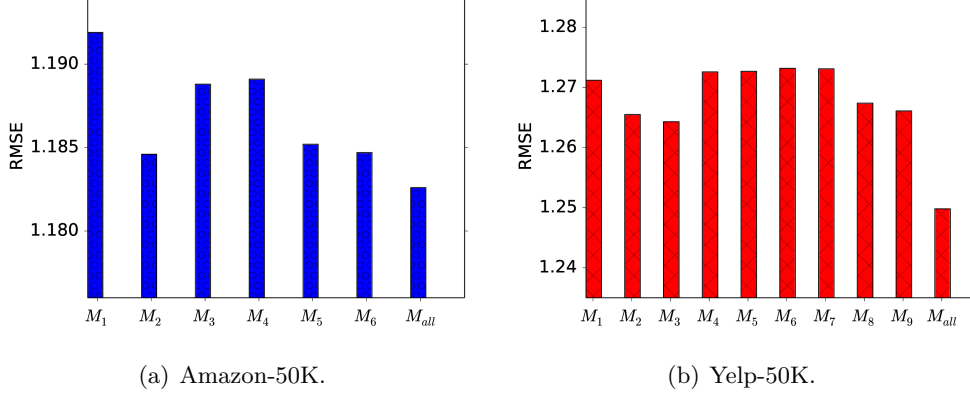


Figure 8: RMSE of single meta-graph on the Amazon and Yelp datasets.  $M_{all}$  is our model trained with all meta-graphs.

that meta-graphs with style like  $U \rightarrow * \leftarrow U \rightarrow B$  are better than those like  $U \rightarrow B \rightarrow * \leftarrow B$ .

Finally, for  $M_9$  on Yelp and  $M_6$  on Amazon, we can see that the performance of these two meta-graphs are among the best three, which demonstrates the usefulness of the complex semantics captured in  $M_9$  on Yelp and  $M_6$  on Amazon.

#### 6.6. Feature Extraction Methods

In this section, we compare the performance of NNR and MF, which are described in Section 3.1, for feature extraction. Note that, the parameter  $F$  of MF and  $\mu$  of NNR will lead to different number of latent features of each single similarity matrix. We show the performance with different  $d$ , i.e., total length of the input feature, in Figure 9. We can see that latent features from NNR have slightly better performance than MF, while the feature dimension resulting from NNR is much larger. Thus in practice, we use MF as the feature extraction method for the experiments in Section 6.2.

#### 6.7. Rank of Second-Order Weights

In this section, we show the performance trend by varying  $K$ , which is the rank of the second-order weights  $\mathbf{V}$  in the FMG model (see Section 4). For the sake of efficiency, we conduct extensive experiments using the smaller datasets, Amazon-50K and Yelp-50K. We use the MF-based latent features. We set  $K$  to values in the range of  $[2, 3, 5, 10, 20, 30, 40, 50, 100]$ , and the results are shown in Figure 10. We can see that the performance gets better with larger  $K$  on both datasets and becomes stable after  $K = 10$ . Thus, we fix  $K = 10$  for all other experiments.

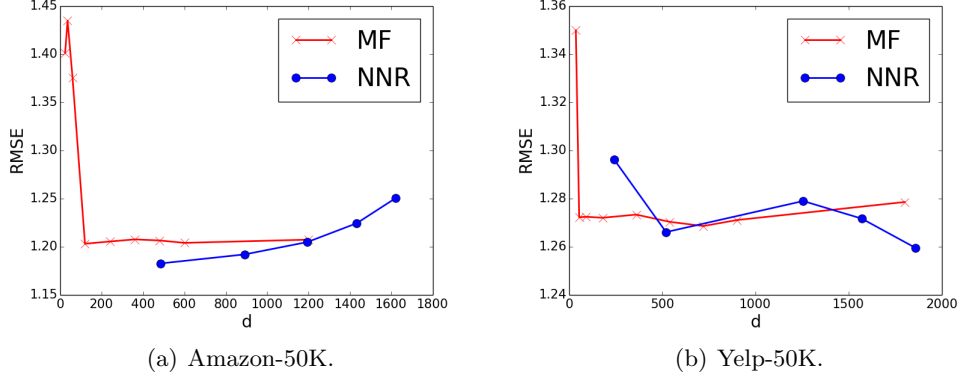


Figure 9: The performance of latent features obtained from MF and NNR.

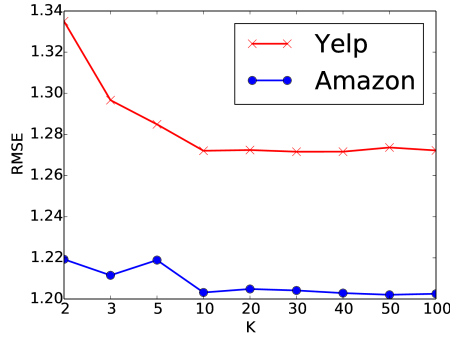


Figure 10: The trend of RMSE of FMG w.r.t.  $K$ .

### 6.8. Optimization Algorithm

In this section, we compare the SVRG and nmAPG algorithms proposed in Section 5. Besides, we also use SGD as a baseline as it is the most popular algorithm for models based on factorization machine [25, 49]. Again, we use the Amazon-50K and Yelp-50K datasets. As suggested in [27], we compare the efficiency of various algorithms based on testing RMSE w.r.t. the number of gradient computations divided by  $N$ .

The results are shown in Figure 11. As we can see, SGD is the slowest among all three algorithms and SVRG is the fastest. SGD can be faster than nmAPG at the beginning. However, due to the diminishing step-size, which is used to guarantee convergence of stochastic algorithms, SGD finally becomes the slowest. SVRG is also a stochastic gradient method, but it avoids the problem of diminishing step-size by using variance reduced technique, which results in even faster speed than nmAPG. Finally, as both SVRG and nmAPG are guaranteed

to produce a critical point of (10), they have the same empirical predicting performance.

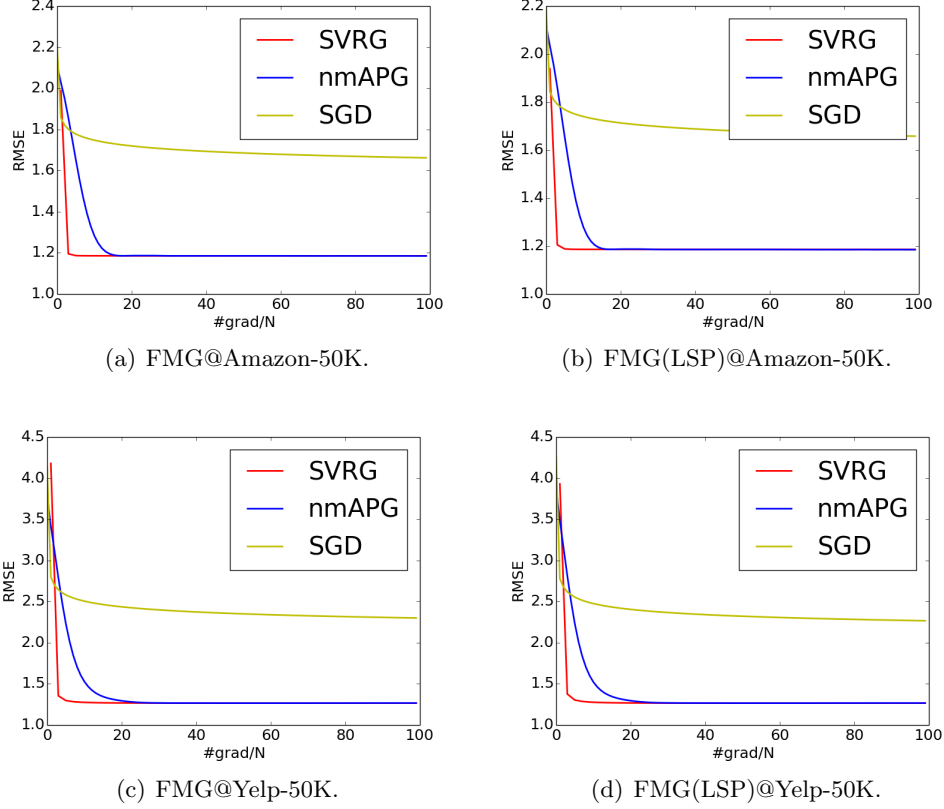


Figure 11: Comparison of various algorithms on the Amazon and Yelp datasets.

### 6.9. Scalability

In this section, we study the scalability of our framework. We extract a series of datasets of different scales from Amazon-200K and Yelp-200K datasets according to the number of observations in the user-item rating matrix. The specific values are  $[12.5K, 25K, 50K, 100K, 200K]$ .

The time cost on the Amazon and Yelp datasets are shown in Figure 12. For the sake of simplicity, we only show the results of FMG with SVRG and nmAPG algorithms. From Figure 12, the training time is almost linear to the number of observed ratings, which aligns with the analysis in Section 5.3 and demonstrates that our framework can be applied to large-scale datasets.

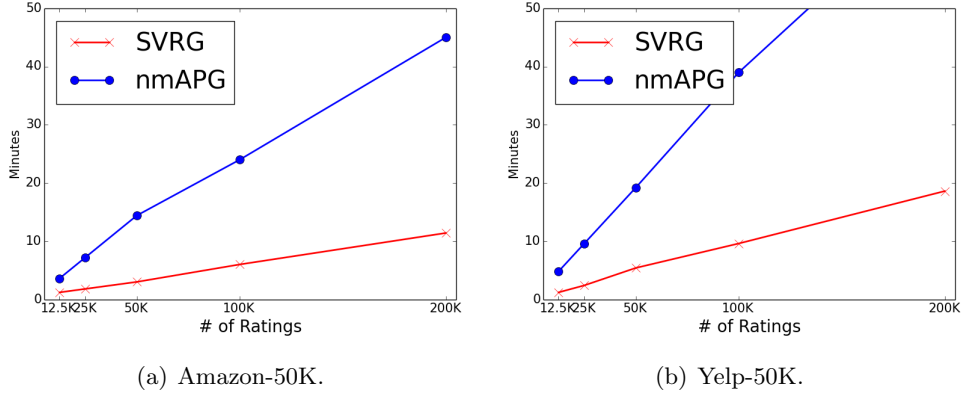


Figure 12: The training time of FMG with SVRG and nmAPG algorithms on the Amazon and Yelp datasets.

## 7. Related Work

In this section, we introduce existing works related to HIN, RS with side information, and FM.

### 7.1. Heterogeneous Information Networks (HINs)

HINs have been proposed as a general representation for many real-world graphs or networks [14, 15, 16, 17, 50]. Meta-path has been developed as a sequence of entity types defined by the HIN network schema. Based on a meta-path, several similarity measures, such as PathCount [15], PathSim [15], and PCRW [51] have been proposed. These measures have been shown to be useful for entity search and similarity measure in many real-world networks. After the development of meta-path, many data mining tasks have been enabled or enhanced, including recommendation [19, 20, 21], similarity search [15, 52], clustering [18, 53], classification [54, 55, 56, 57], link prediction [58, 59] and malware detection [60].

Recently, meta-graph (or meta-structure) has been proposed to define more complicated semantics in HIN [22, 23]. However, they still applied meta-graph to entity similarity problems where entities have the same type. In this paper, we extend meta-graph to the recommendation problem. The problem of recommendation requires us to approximate the large-scale user-item rating matrix. Thus, instead of computing each similarity efficiently online, we consider to compute the matrices offline, and design the best way to use the user-item matrices generated by different meta-graphs for the final prediction.

### 7.2. Recommendation with Heterogeneous Side Information

Modern recommender systems are able to capture rich side information such as social connections among users, meta-data and review text associated with items. Previous works have explored different methods to incorporate heterogeneous side information to enhance collaborative filtering based recommender systems. For example, Ma et al. [10] and Zhao et al. [11], respectively, incorporated social relations into low-rank and local low-rank matrix factorization to improve the recommending performance. In [12, 13], review texts are analyzed together with ratings in the rating prediction task. Ye et al. [61] proposed a probabilistic model to incorporate users’ preferences, social network and geographical information to enhance the point-of-interests recommendation. Zheng et al. [62] proposed to integrate users’ location data with history data to improve the performance of Point-of-Interest recommendation. These previous approaches have demonstrated the importance and effectiveness of heterogeneous information in improving recommendation accuracy. Pan et al. [6] proposed to use transfer learning to incorporate one type of auxiliary information to enhance collaborative filtering. Zhao et al. [7] proposed a novel transfer learning framework to utilize knowledge from other systems. However, most of these approaches dealt with different heterogeneous information disparately, hence losing important information that exists across them.

HIN-based recommendation has been proposed to avoid the disparate treatment of different types of information. Based on meta-path, several approaches have attempted to tackle the recommendation task based on HIN. In [19], meta-path based similarities are used as regularization terms in matrix factorization. In [20], multiple meta-paths are used to learn user and item latent features, which are then used to recover similarity matrices combined by a weighted mechanism. In [21], users’ ratings to items are used to build a weighted HIN, based on which meta-path based methods are used to measure the similarities of users for recommendation. The combination of different meta-paths are explicit, using the similarities instead of latent features. As discussed in the introduction, these approaches do not make full use of the meta-path based features, whereas our framework based on “MF + FM” aims to accomplish.

### 7.3. Factorization Machine (FM)

FM [25] is a popular and powerful recommendation framework, which can model non-linear interactions among features, e.g., the rating information, categories of items, texts, time. Many approaches and systems have been developed based on FM [49, 63]. Different from previous approaches which only consider explicit features, we generate latent features by low-rank approximation on similarity matrices generated from different meta-graphs. For FM using the

original explicit features, MF can be regarded as a step similar to PCA to perform dimensionality reduction to reduce the noise of the original features.

## 8. Conclusion

In this paper, we present a heterogeneous information network (HIN) based recommendation method. We introduce a principled way of fusing various side information on HIN. By using different meta-graphs derived from the HIN schema, we can capture complicated semantics between users and items. Then, we use matrix factorization and nuclear norm regularization to obtain the user and item latent features from each meta-path in an unsupervised way. After that, we use a group lasso regularized factorization machine to fuse different groups of semantic information extracted from different meta-graphs to predict the links. To solve the nonconvex nonsmooth optimization problem, we propose two algorithms, one is based on the proximal gradient algorithm and the other based on stochastic variance reduced gradient algorithm, to efficiently solve the optimization problem. Experimental results demonstrate the effectiveness of our approach.

In the future, we plan to explore automatic methods to generate meta-graphs instead of hand-crafting them as done in this paper. Thus, our framework can be quickly applied to new domains. Besides, our framework is a two-stage process, i.e., the “MF” part and “FM” part, where we did not use label information (ratings) when generating latent features from multiple meta-graphs. We plan to explore whether better latent features can be obtained if ratings are exploited in MF to generate latent features. To achieve this, a joint modeling of the two parts or an end-to-end deep learning model may be considered.

## 9. Acknowledgments

Huan Zhao and Dik Lun Lee are supported by the Research Grants Council HKSAR GRF (No. 615113). Quanming Yao and James T. Kwok are supported by the Research Grants Council HKSAR GRF (No. 614513). Yangqiu Song is supported by China 973 Fundamental R&D Program (No. 2014CB340304) and the Research Grants Council HKSAR GRF (No. 26206717).

## References

- [1] J. Herlocker, J. Konstan, A. Borchers, J. Riedl, An algorithmic framework for performing collaborative filtering, in: *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval*, 1999, pp. 230–237.
- [2] Y. Koren, Factorization meets the neighborhood: a multifaceted collaborative filtering model, in: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008, pp. 426–434.

- [3] M. Balabanović, Y. Shoham, Fab: content-based, collaborative recommendation, *Communications of the ACM* 40 (3) (1997) 66–72.
- [4] H. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, Wide & deep learning for recommender systems, in: *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, ACM, 2016, pp. 7–10.
- [5] B. Li, Q. Yang, X. Xue, Transfer learning for collaborative filtering via a rating-matrix generative model, in: *Proceedings of the 26th International Conference on Machine Learning*, 2009, pp. 617–624.
- [6] W. Pan, Q. Yang, Transfer learning in heterogeneous collaborative filtering domains, *Artificial intelligence* 197 (2013) 39–55.
- [7] L. Zhao, S. J. Pan, Q. Yang, A unified framework of active transfer learning for cross-system recommendation, *Artificial Intelligence* 245 (2017) 38–55.
- [8] N. Taghipour, A. Kardan, S. S. Ghidary, Usage-based web recommendations: A reinforcement learning approach, in: *Proceedings of the 2007 ACM conference on Recommender systems*, 2007, pp. 113–120.
- [9] H. Wang, Q. Wu, H. Wang, Factorization bandits for interactive recommendation., in: *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2017, pp. 2695–2702.
- [10] H. Ma, D. Zhou, C. Liu, M. R. Lyu, I. King, Recommender systems with social regularization, in: *Proceedings of the 4th International Conference on Web Search and Data Mining*, 2011, pp. 287–296.
- [11] H. Zhao, Q. Yao, J. Kwok, D. Lee, Collaborative filtering with social local models, in: *Proceedings of the 16th International Conference on Data Mining*, 2017, pp. 645–654.
- [12] J. McAuley, J. Leskovec, Hidden factors and hidden topics: Understanding rating dimensions with review text, in: *Proceedings of the 7th ACM Conference on Recommender Systems*, 2013, pp. 165–172.
- [13] G. Ling, M. R. Lyu, I. King, Ratings meet reviews, a combined approach to recommend, in: *Proceedings of the 8th ACM Conference on Recommender Systems*, 2014, pp. 105–112.
- [14] C. Shi, Y. Li, J. Zhang, Y. Sun, Y. Philip, A survey of heterogeneous information network analysis, *IEEE Transactions on Knowledge and Data Engineering* 29 (1) (2017) 17–37.
- [15] Y. Sun, J. Han, X. Yan, P. S. Yu, T. Wu, PathSim: Meta path-based top-k similarity search in heterogeneous information networks, in: *Proceedings of the VLDB Endowment*, 2011, pp. 992–1003.
- [16] X. Kong, J. Zhang, P. S. Yu, Inferring anchor links across multiple heterogeneous social networks, in: *Proceedings of the 22nd International Conference on Information and Knowledge Management*, 2013, pp. 179–188.
- [17] C. D. Joshua, Chapter 13: Mining electronic health records in the genomics era, *PLoS Computational Biology* 8 (12).
- [18] C. Wang, Y. Song, A. El-Kishky, D. Roth, M. Zhang, J. Han, Incorporating world knowledge to document clustering via heterogeneous information networks, in: *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1215–1224.
- [19] X. Yu, X. Ren, Q. Gu, Y. Sun, J. Han, Collaborative filtering with entity similarity regularization in heterogeneous information networks, *Tech. rep.*, University of Illinois at Urbana-Champaign (2013).
- [20] X. Yu, X. Ren, Y. Sun, Q. Gu, B. Sturt, U. Khandelwal, B. Norick, J. Han, Personalized entity recommendation: A heterogeneous information network approach, in: *Proceedings of the 7th International Conference on Web Search and Data Mining*, 2014, pp. 283–292.
- [21] C. Shi, Z. Zhang, P. Luo, P. S. Yu, Y. Yue, B. Wu, Semantic path based personalized recommendation on weighted heterogeneous information networks, in: *Proceedings of the 24th International Conference on Information and Knowledge Management*, 2015, pp. 453–



- [22] Z. Huang, Y. Zheng, R. Cheng, Y. Sun, N. Mamoulis, X. Li, Meta Structure: Computing relevance in large heterogeneous information networks, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1595–1604.
- [23] Y. Fang, W. Lin, W. Zheng, M. Wu, K. Chang, X. Li, Semantic proximity search on graphs with meta graph-based learning, in: Proceedings of the 32nd International Conference on Data Engineering, 2016, pp. 277–288.
- [24] P. Belhumeur, J. Hespanha, D. Kriegman, Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19 (7) (1997) 711–720.
- [25] S. Rendle, Factorization machines with libFM, *ACM Transactions on Intelligent Systems and Technology* 3 (3) (2012) 57:1–57:22.
- [26] N. Parikh, S. Boyd, Proximal algorithms, *Foundations and Trends in Optimization* 1 (3) (2014) 127–239.
- [27] L. Xiao, T. Zhang, A proximal stochastic gradient method with progressive variance reduction, *SIAM Journal on Optimization* 24 (4) (2014) 2057–2075.
- [28] H. Zhao, Q. Yao, J. Li, Y. Song, D. Lee, Meta-graph based recommendation fusion over heterogeneous information networks, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 635–644.
- [29] A. Mnih, R. Salakhutdinov, Probabilistic matrix factorization, in: *Advances in Neural Information Processing Systems*, 2007, pp. 1257–1264.
- [30] E. J. Candès, B. Recht, Exact matrix completion via convex optimization, *Foundations of Computational mathematics* 9 (6) (2009) 717.
- [31] E. Candès, X. Li, Y. Ma, J. Wright, Robust principal component analysis?, *Journal of the ACM* 58 (3) (2011) 11.
- [32] Q. Yao, J. Kwok, Accelerated inexact Soft-Impute for fast large-scale matrix completion, in: Proceedings of the 24th International Joint Conference on Artificial Intelligence, 2015, pp. 4002–4008.
- [33] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society: Series B* 68 (1) (2006) 49–67.
- [34] L. Jacob, G. Obozinski, J. Vert, Group lasso with overlap and graph lasso, in: Proceedings of the 26th International Conference on Machine Learning, 2009, pp. 433–440.
- [35] T. Zhang, Analysis of multi-stage convex relaxation for sparse regularization, *Journal of Machine Learning Research* 11 (2010) 1081–1107.
- [36] Q. Yao, J. Kwok, W. Zhong, Fast low-rank matrix learning with nonconvex regularization, in: Proceedings of the 15th International Conference on Data Mining, 2015, pp. 539–548.
- [37] E. Candès, M. Wakin, S. Boyd, Enhancing sparsity by reweighted  $\ell_1$  minimization, *Journal of Fourier Analysis and Applications* 14 (5-6) (2008) 877–905.
- [38] Q. Yao, J. Kwok, Efficient learning with a family of nonconvex regularizers by redistributing nonconvexity, in: Proceedings of the 33rd International Conference on Machine Learning, 2016, pp. 2645–2654.
- [39] H. Li, Z. Lin, Accelerated proximal gradient methods for nonconvex programming, in: *Advances in Neural Information Processing Systems*, 2015, pp. 379–387.
- [40] Q. Yao, J. Kwok, F. Gao, W. Chen, T.-Y. Liu, Efficient inexact proximal gradient algorithm for nonconvex problems, in: Proceedings of the 26th International Joint Conferences on Artificial Intelligence, 2017, pp. 3308–3314.
- [41] L. Yuan, J. Liu, J. Ye, Efficient methods for overlapping group lasso, in: *Advances in Neural Information Processing Systems*, 2011, pp. 352–360.
- [42] D. P. Bertsekas, *Nonlinear programming*, Athena Scientific, 1999.
- [43] S. Reddi, A. Hefny, S. Sra, B. Poczos, A. Smola, Stochastic variance reduction for nonconvex

- optimization, in: Proceedings of the 33rd International Conference on Machine Learning, 2016, pp. 314–323.
- [44] Z. Allen-Zhu, E. Hazan, Variance reduction for faster non-convex optimization, in: Proceedings of the 33rd International Conference on Machine Learning, 2016, pp. 699–707.
  - [45] R. He, J. McAuley, Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering, in: Proceedings of the 25th International Conference on World Wide Web, 2016, pp. 507–517.
  - [46] A. Paterek, Improving regularized singular value decomposition for collaborative filtering, Tech. rep., Institute of Informatics, Warsaw University (2007).
  - [47] G. Guo, J. Zhang, Z. Sun, N. Yorke-Smith, Librec: A Java library for recommender systems, Tech. rep., School of Information Systems, Singapore Management University (2015).
  - [48] R. Řehůřek, P. S., Software framework for topic modelling with large corpora (May 2010).
  - [49] L. Hong, A. S. Doumith, B. D. Davison, Co-factorization machines: Modeling user interests and predicting individual decisions in twitter, in: Proceedings of the 6th International Conference on Web Search and Data Mining, 2013, pp. 557–566.
  - [50] Y. Sun, J. Han, Mining heterogeneous information networks: A structural analysis approach, ACM SIGKDD Explorations Newsletter 14 (2) (2013) 20–28.
  - [51] N. Lao, W. Cohen, Relational retrieval using a combination of path-constrained random walks, Machine Learning 81 (1) (2010) 53–67.
  - [52] C. Shi, X. Kong, Y. Huang, Y. P. S., B. Wu, HeteSim: A general framework for relevance measure in heterogeneous networks, IEEE Transactions on Knowledge and Data Engineering 26 (10) (2014) 2479–2492.
  - [53] Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, X. Yu, PathSelClus: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks, ACM Transactions on Knowledge Discovery from Data 7 (3) (2013) 11:1–11:23.
  - [54] X. Kong, B. Cao, P. S. Yu, Multi-label classification by mining label and instance correlations from heterogeneous information networks, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013, pp. 614–622.
  - [55] C. Wang, Y. Song, H. Li, M. Zhang, J. Han, KnowSim: A document similarity measure on structured heterogeneous information networks, in: Proceedings of the 15th International Conference on Data Mining, 2015, pp. 1015–1020.
  - [56] C. Wang, Y. Song, H. Li, Y. Sun, Z. Zhang, J. Han, Distant meta-path similarities for text-based heterogeneous information networks, in: Proceedings of the 26th International Conference on Information and Knowledge Management, 2017, pp. 1629–1638.
  - [57] H. Jiang, Y. Song, C. Wang, M. Zhang, Y. Sun, Semi-supervised learning over heterogeneous information networks by ensemble of meta-graph guided random walks, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017, pp. 1944–1950.
  - [58] Y. Sun, J. Han, C. C. Aggarwal, N. V. Chawla, When will it happen?: Relationship prediction in heterogeneous information networks, in: Proceedings of the 5th International Conference on Web Search and Data Mining, 2012, pp. 663–672.
  - [59] J. Zhang, P. S. Yu, Z. Zhou, Meta-path based multi-network collective link prediction, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014, pp. 1286–1295.
  - [60] S. Hou, Y. Ye, Y. Song, M. Abdulhayoglu, HinDroid: An intelligent android malware detection system based on structured heterogeneous information network, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 1507–1515.
  - [61] M. Ye, P. Yin, W. C. Lee, D. L. Lee, Exploiting geographical influence for collaborative

- point-of-interest recommendation, in: Proceedings of the 34th International Conference on Research and Development in Information Retrieval, 2011, pp. 325–334.
- [62] V. W. Zheng, Y. Zheng, X. Xie, Q. Yang, Towards mobile intelligence: Learning from GPS history data for collaborative recommendation, *Artificial Intelligence* 184 (2012) 17–37.
- [63] S. Rendle, L. Schmidt-Thieme, Pairwise interaction tensor factorization for personalized tag recommendation, in: Proceedings of the 3rd International Conference on Web Search and Data Mining, 2010, pp. 81–90.