Improving Accuracy and Diversity in Matching of Recommendation with Diversified Preference Network

Ruobing Xie*, Qi Liu*, Shukai Liu, Ziwei Zhang, Peng Cui, *Senior Member, IEEE,* Bo Zhang, and Leyu Lin

Abstract—Real-world recommendation systems need to deal with millions of item candidates. Therefore, most practical large-scale recommendation systems usually contain two modules. The matching module aims to efficiently retrieve hundreds of high-quality items from large corpora, while the ranking module aims to generate specific ranks for these items. Recommendation diversity is an essential factor that strongly impacts user experience. There are lots of efforts that have explored recommendation diversity in ranking, while the matching module should take more responsibility for diversity. In this paper, we propose a novel Heterogeneous graph neural network framework for diversified recommendation (GraphDR) in matching to improve both recommendation accuracy and diversity. Specifically, GraphDR builds a huge heterogeneous preference network to record different types of user preferences, and conducts a field-level heterogeneous graph attention network for node aggregation. We conduct a neighbor-similarity based loss with a multi-channel matching to improve both accuracy and diversity. In experiments, we conduct extensive online and offline evaluations on a real-world recommendation system with various accuracy and diversity metrics and achieve significant improvements. GraphDR has been deployed on a well-known recommendation system named WeChat Top Stories, which affects millions of users. The source code will be released in https://github.com/lqfarmer/GraphDR.

Index Terms—recommender system, matching, heterogeneous graph, recommendation diversity, graph neural network.

1 Introduction

Recently, real-world personalized recommendation systems usually need to deal with hundreds of millions of items [1]. Therefore, it is challenging to conduct complicated end-to-end recommendation algorithms on the entire corpus, for even a linear time complexity w.r.t the corpus size is unacceptable [2]. To balance both effectiveness and efficiency in real-world scenarios, conventional recommendation systems usually consist of two modules, namely the matching module and the ranking module [3], [4]. The **matching** module, also regarded as the candidate generation in the Youtube DNN model [3], aims to retrieve a small subset of (usually hundreds of) items from the entire corpus efficiently. In contrast, the **ranking** module conducts sophisticated models on these retrieved items to get specific item ranks. Fig. 1 shows the classical two-step architecture. The matching module concentrates more on the diversity, efficiency and item coverage, while the ranking module focuses more on the accuracy of specific item ranks. This two-step architecture balances efficiency and effectiveness in practical recommendation systems.

Conventional recommendation models usually regard recommendation accuracy metrics such like Click-through-rate (CTR) as their central objectives, in which popular items clicked by users are more preferred. However, such objectives will lead to homogenization issues that reduce personalization and harm user experiences. To solve this issue, **recommendation diversity** is considered to evaluate the overall recommendation performances from another aspect [5]. It is measured in two classical ways, the individual diversity and the aggregate diversity [6], from different views. The **individual diversity** focuses on the local diversity in each recommended item list, which aims to balance user-item similarities and item-item dissimilarities [7]. In contrast, the **aggregate diversity** focuses on the global diversity in the overall recommendation, which is usually measured by the coverage of items that could be recommended by models in the entire corpus [8]. The significance of diversity has been widely verified to provide highly idiosyncratic items in recommendation [9]. Both diversities should be considered

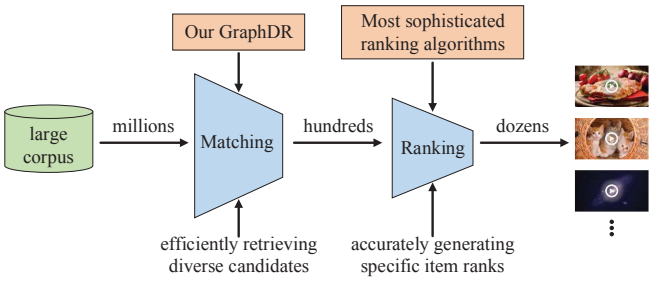


Fig. 1. An example of a real-world recommendation system. GraphDR focuses on the matching module, which aims to retrieve user-interested and diversified items efficiently. Note that the matching module cares whether good items are retrieved, not the specific item ranks.

- * *indicates equal contributions.*
- *Ruobing Xie (corresponding author), Qi Liu, Shukai Liu, Bo Zhang and Leyu Lin are with WeChat Search Application Department, Tencent, China, 100080.*
  E-mail: {ruobingxie,addisliu,shukailiu,nevinzhang,goshawklin}@tencent.com
- *Ziwei Zhang and Peng Cui are with the Department of Computer Science and Technology, Tsinghua University, Beijing, China, 100084.*
  E-mail: zw-zhang16@mails.tsinghua.edu.cn, cuip@tsinghua.edu.cn

in real-world recommendation systems.

There are lots of ranking models that have explored recommendation diversities with the help of dissimilarity factors [5], external taxonomy information [10], clustering [11] and graphic technologies [12]. However, most diversified recommendation models are specially designed for ranking, which are incredibly time-consuming to be used in matching with millions of items [12], while very few works systematically focus on the diversity in matching. In fact, matching should take more responsibility for diversity, since it cares more about the coverage of user-interested items rather than their specific item ranks. The recommendation diversity needs to be first guaranteed in the matching module. Otherwise, the homogenization of the item candidates generated by the matching module will inevitably lead to the lack of diversity in the final recommendation.

In this paper, we aim to improve both recommendation accuracy and diversity in the matching module, which is essential in real-world recommendation systems. We propose a novel **Heterogeneous graph neural network framework for diversified recommendation (GraphDR)**. Precisely, GraphDR mainly consists of three modules: (1) *Diversified preference network construction*, which aims to build a huge global heterogeneous network containing various interactions between different types of nodes including videos, tags, medias, users and words. These interactions between essential recommendation factors reflect user diverse preferences from a global view, which are the sources of diversity. (2) *Heterogeneous network representation learning (NRL)*, which learns node representations with a novel field-level heterogeneous graph attention network (FH-GAT). FH-GAT helps to better maintain and aggregate different types of interactions. We also innovatively conduct a neighbor-similarity based objective to encode users' diverse preferences into heterogeneous node representations. Different from CTR-oriented objectives that simply focus on click behaviors, the neighbor-similarity based objective highlights diversity by considering multiple factors of videos such as user watching habit, audience community, video content, video taxonomy, and content provider. (3) *Online multi-channel matching*, which generates a small subset of user-interested and diverse item candidates efficiently through multiple channels. The multi-channel strategy is conducted to further amplify the diversity in the recommended results. The diversity derives from the heterogeneous information well captured by the cooperation of all three modules in GraphDR.

In experiments, we conduct both offline and online evaluations on a real-world video recommendation system, which is widely used by hundreds of millions of users. We conduct extensive experiments to measure the recommendation accuracy and diversity with dozens of metrics. We also explore GraphDR with model analyses, ablation tests and case studies for better understanding. The main contributions are concluded as follows:

- We highlight and systematically explore the recommendation diversity issue in the matching module, which is essential in practical large-scale recommendation systems.
- We propose a novel GraphDR framework to jointly improve both recommendation accuracy and diversity in real-world matching. To the best of our knowledge, we are the first to introduce GNN on heterogeneous preference networks for diversified recommendation in matching.
- We propose a novel field-level heterogeneous GAT model to aggregate neighbors with different feature fields. We also innovatively conduct the neighbor-similarity based loss with online multi-channel matching to polish recommendation diversity.
- The offline and online evaluations demonstrate that GraphDR can improve both accuracy and diversity in practice. GraphDR is simple and effective, which has been deployed on a real-world recommendation system used by millions of users. It is also convenient to adopt GraphDR to other scenarios.

2 Related Works

In related works, we first give a brief introduction to the classical recommendation algorithms, and then introduce the efforts in recommendation diversity. We also include a discussion on graph neural networks in recommendation.

2.1 Recommendation Systems

Collaborative filtering (CF) is a representative method that recommends items with similar items or users [13]. Matrix factorization (MF) attempts to decompose user-item interaction matrix to get user and item representations [14]. FM [15] expands to model second-order feature interactions with latent vectors. However, most neural ranking models rely on user-item interactions for prediction. Hence, these complicated ranking models are hard to be directly used in matching, for they are extremely time-consuming when handling million-level items. With the thriving in deep learning, neural models like Deep Crossing [16], FNN [17], PNN [18], Wide&Deep [19], DCN [20] and DFN [21] are proposed to improve recommendation performances. DeepFM [22], AFM [23] and NFM [24] improve the original FM with DNN or attention. AutoInt [25] and BERT4Rec [26] also brings in self attention. Recently, AFN [27] and AutoFIS [28] are proposed to smartly model high-order feature interactions via logarithmic transformation or automatic feature selection. Most deep ranking models are challenging to be utilized in real-world matching modules, for they are extremely time-consuming dealing with millions of candidates.

In contrast, there are much fewer works specially designed for matching. Conventional systems usually use IR-based methods [29] or Collaborative filtering (CF) based methods [13] for fast retrieval. For neural models, embedding-based retrieval such as DSSM [30] is also widely deployed. Recently, Youtube [3] brings in deep models to learn user preference in matching. Moreover, TDM [2], JTM [31] and OTM [32] arrange items with tree structures to accelerate top-n item retrieval, which combine matching and ranking in a single model. ICAN [4] is specially designed for cold-start multi-channel matching. [33] also proposes an industrial embedding-based retrieval framework in Facebook search. However, these matching models mainly focus on CTR-oriented objectives. It is still challenging for these models to balance accuracy and diversity in real-world scenarios.

2.2 Diversified Recommendation

Merely using CTR-oriented objectives will make hot items hotter, which inevitably brings in serious homogenization issues that may degrade user experiences [34]. The significance of diversity has been verified by lots of efforts, since it could provide highly idiosyncratic items with less homogeneity for users in personalized recommendation [5], [9]. Recommendation diversity is mainly measured in individual diversity and aggregate diversity [6]. The **individual diversity** focuses on the local diversity in recommended lists. [5] and [10] focus on intra-list item dissimilarities. [34] proposes a novel item novelty, which measures the additional information from new items. Some works measure diversity with the varieties of taxonomy in item lists [10]. In contrast, the **aggregate diversity** measures the global diversity in overall systems. [8] measures this diversity with the coverage of recommended items. The higher item coverage indicates that the model could recommend more long-tail items, which implies a more diversified system from the global aspect. Both individual and aggregate diversities are essential metrics from different aspects. We evaluate both of them in Sec. 5.5 to measure the recommendation diversity.

There are some works that model diversity in ranking [35]. [5] bring dissimilarity factors to the loss functions to measure the individual diversity. External taxonomy information (e.g., tag, category and subtopic) [10], [36] and knowledge graph [37] are useful factors for diversity. Other technologies such as entropy regularizer [38], clustering [11], graph-based models [12], [39], and greedy map inference [7] have also been explored for diversified recommendation. Recently, diversified recommendation is armed with reinforcement learning [40] and adversarial learning [41]. Recommendation bandits [42], [43] are also well explored. However, most diversified models are specially designed for ranking, which are hard to be directly used in matching. To the best of our knowledge, we are the first to use GNN on the global heterogeneous interactions to improve both accuracy and diversity in the matching module.

2.3 Graph Neural Network (GNN)

Recently, GNN has been widely explored and verified in various fields. GCN [44] introduces convolution to graphs based on spectral graph theory. GraphSAGE [45] conducts an inductive representation learning on large graphs. Graph attention network (GAT) [46] brings in graph attention mechanism. HetGNN [47] and HAN [48] extend GNN to heterogeneous networks. In recommendation, [49], [50], [51], and [52] further use GNN on session, social and bundle recommendation. [53] explores multi-relational graphs. Heterogeneous graphs are also widely adopted to model different types of essential objects such as users, items, tags and providers in recommendation [54], [55]. Inspired by these models, we conduct a heterogeneous graph to model various types of feature interactions, and also use a heterogeneous GNN model for node aggregation.

3 METHODOLOGY

In this paper, we propose GraphDR to improve both accuracy and diversity in matching by considering user diverse preferences. In this section, we first show the overall framework of GraphDR (Sec. 3.1). Second, we introduce the construction of nodes and edges in the diversified preference network, which is the source of diversity in our model (Sec. 3.2). Next, we introduce the Diversity-aware network representation learning model FH-GAT used to generate node representations for all types of nodes (Sec. 3.3). Finally, we give a detailed discussion on the proposed Diversity-enhanced training objective (Sec. 3.4). We further introduce the online multi-channel matching module in Sec. 4.

3.1 Overall Architecture

The GraphDR framework mainly contains three modules as in Fig 2, including diversified preference network construction, network representation learning, and online multi-channel matching. In offline NRL, GraphDR first collects various informative interactions between heterogeneous nodes to build a huge global diversified preference network. Next, we propose a field-level HGAT model to learn node embeddings with the neighbor-similarity based objective. In online serving, the multi-channel matching retrieves hundreds of accurate and diverse item candidates efficiently with multiple channels. The offline NRL conducts time-consuming training to encode user diverse preferences into node embeddings, while the online serving efficiently uses these learned embeddings for fast and diversified multi-channel retrieval. All three modules contribute to the diversity (see Sec. 5.5 for more details).

3.2 Diversified Preference Network

The diversified preference network is the fundamental of diversity. We attempt to bring in heterogeneous interactions between essential objects in recommendation to describe user diverse preferences. Precisely, we focus on five different types of nodes including **video**, **tag**, **media**, **user** and **word**, which are essential factors that may impact users in recommendation. Each video has a title (containing words) and several tags annotated by editors. The video provider is viewed as the media. To alleviate the data sparsity and reduce computation costs, we cluster users into user groups as communities according to their basic profiles (i.e., the gender-age-location attribute triplets in this work), and consider these user groups as user nodes. We group users via user basic profiles for higher coverage.

We assume that the interactions between these five types of objects can reflect user diverse preferences. In GraphDR, we consider six types of edges to record these multi-aspect preferences as:

- **Video-video edge.** We generate the video-video edge between two video nodes if they have appeared adjacently in a user's video session. To reduce noises, we only use the **valid watching behaviors**, where videos have been watched for more than 70% of their total time lengths. Video-video edges record the sequential user watching habits in sessions.
- **Video-user edge.** Video-user edges are built if a video is validly watched by a user group at least 3 times in a week. This edge stores user-item interactions via the classical and essential historical behavior information, which implies the audience community of videos.

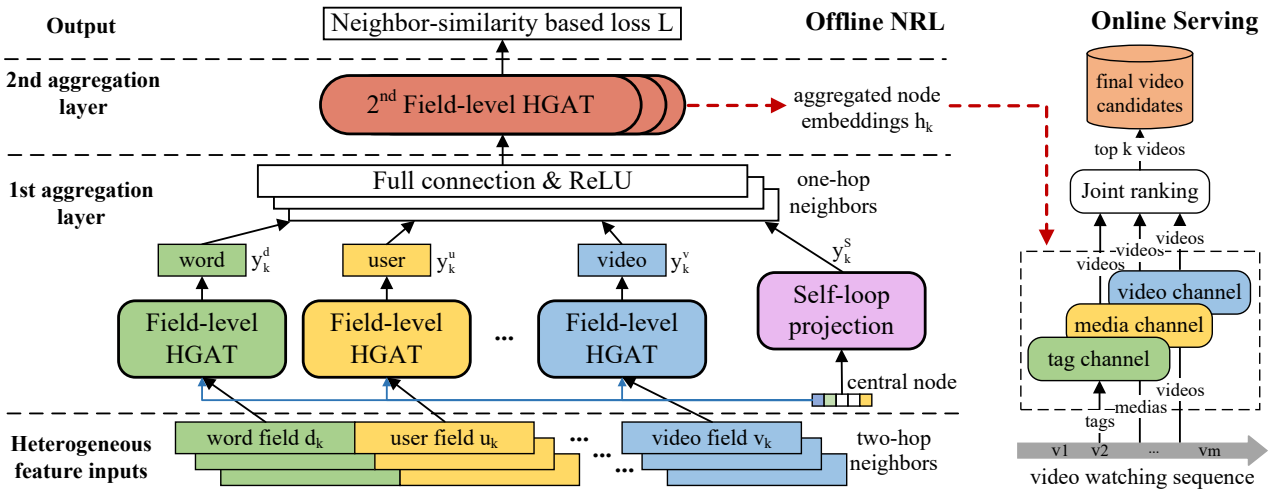


Fig. 2. The offline NRL and online serving parts of GraphDR for matching in recommendation. The left offline NRL part is the proposed FH-GAT model, which builds the aggregated node embeddings with heterogeneous GAT on the diversified preference network. The right online multi-channel matching part aims to retrieve hundreds of videos from large corpora efficiently. The recommendation diversity comes from the diversified preference network, FH-GAT trained with neighbor-similarity based loss, and the online multi-channel matching.

- **Video-tag edge.** Video-tag edge connects videos with their corresponding tags, which reflects the coarse-grained semantic preferences of taxonomy in videos.
- **Video-word edge.** Video-word edge links videos with their words in titles, reflecting the fine-grained semantic preferences of detailed word-level contents in videos.
- **Video-media edge.** Video-media edges are drawn between videos and their medias (i.e., video providers).
- **Tag-tag edge.** The tag-tag edges are built according to tag co-occurrence in a video, which highlights taxonomy relevance.

All edges are undirected since most co-occurrences between objects have no direction (we simply consider video-video edges as video co-occurrences in sessions for consistency and convenience). These heterogeneous edges bring in additional information of videos besides user-item click behaviors. They can reflect user diverse preferences in user watching habit, audience community, video content, taxonomy, and content provider. For instance, two related videos may be linked via the same user groups (video-user), video providers (video-media), tags (video-tag) or sessions (video-video), or even connected by multi-hop paths containing heterogeneous nodes. Hence, similar videos can be linked via multi-hop paths with heterogeneous nodes and edges from different aspects. They build up the potential reasons for recommendation under the proposed objective in Sec. 3.4, which are implicit, but diversified. We spend nearly 30 minutes on building this graph from billion-level instances in practice. The diversified preference network is the fundamental of diversity in GraphDR. It contains heterogeneous nodes and edges, which capture sufficient information to reflect user diversified preferences on different aspects of a video (e.g., tag, media, community, video relevance).

3.3 Diversity-aware Network Representation Learning

Network representation learning aims to encode user diverse preferences into node representations. Inspired by [48], [55], we propose a new **Field-level Heterogeneous Graph Attention Network (FH-GAT)**. Fig. 2 shows the 2-layer architecture of the FH-GAT model.

3.3.1 Heterogeneous Feature Layer

We first project all heterogeneous nodes into the same feature space. For the k-th node, its overall neighbor set N_k could be divided into five ***feature fields*** according to their types as $\{\bar{\boldsymbol{v}}_k, \bar{\boldsymbol{t}}_k, \bar{\boldsymbol{m}}_k, \bar{\boldsymbol{u}}_k, \bar{\boldsymbol{d}}_k\}$, where $\bar{\boldsymbol{v}}_k$, $\bar{\boldsymbol{t}}_k$, $\bar{\boldsymbol{m}}_k$, $\bar{\boldsymbol{u}}_k$ and $\bar{\boldsymbol{d}}_k$ indicate the one-hot representations of video, tag, media, user, word neighbors respectively. The node feature embeddings of the k-th node $\boldsymbol{h}_k$ is as follows:

$$\boldsymbol{f}_k = \text{concat}(\boldsymbol{v}_k, \boldsymbol{t}_k, \boldsymbol{m}_k, \boldsymbol{u}_k, \boldsymbol{d}_k), \tag{1}$$

in which $\boldsymbol{v}_k$ indicates the video-field feature embedding. In this work, we empirically set $\boldsymbol{v}_k = \boldsymbol{P}_v \bar{\boldsymbol{v}}_k$, where $\boldsymbol{P}_v \in \mathbb{R}^{d_v \times n_v}$ represents the lookup projection matrix generating $\boldsymbol{v}_k$ with the video neighbors. d_v is the dimension of $\boldsymbol{v}_k$ and n_v is the number of video nodes. For efficiency, the projection matrix is pre-defined as the indicator of top-frequent video neighbors and fixed during training. $\text{concat}(\cdot)$ is the concatenation operation. The tag, media, user and word field feature embeddings $\boldsymbol{t}_k$, $\boldsymbol{m}_k$, $\boldsymbol{u}_k$ and $\boldsymbol{d}_k$ are generated similarly as the video field feature embedding $\boldsymbol{v}_k$.

3.3.2 Field-level HGAT Layer

This layer takes the neighbor feature embeddings $\{\boldsymbol{f}_1, \cdots, \boldsymbol{f}_l\}$ of the k-th node as inputs. We set a weighting vector group $\{\boldsymbol{w}_k^v, \boldsymbol{w}_k^t, \boldsymbol{w}_k^m, \boldsymbol{w}_k^u, \boldsymbol{w}_k^d\}$ for each field, where $\boldsymbol{w}_k^v$ represents the k-th weighting vector of video. The output embedding $\boldsymbol{y}_k^v$ of the video field is defined as follows:

$$\boldsymbol{y}_k^v = \sum_{i=1}^{l} \alpha_{ki}^v \boldsymbol{v}_i, \quad \alpha_{ki}^v = \frac{\exp(\boldsymbol{w}_k^{v\top} \boldsymbol{v}_i)}{\sum_{j=1}^{n} \exp(\boldsymbol{w}_k^{v\top} \boldsymbol{v}_j)}, \tag{2}$$

where α_{ki}^v is the weight of the k-th node to its i-th neighbor in the video field. The construction of $\boldsymbol{y}_k^t$, $\boldsymbol{y}_k^m$, $\boldsymbol{y}_k^u$ and $\boldsymbol{y}_k^d$ are the same as $\boldsymbol{y}_k^v$. We concatenate these embeddings to form the final neighbor-based representation $\boldsymbol{y}_k^N$ as follows:

$$\boldsymbol{y}_k^N = \text{ReLU}(\boldsymbol{W}_n \cdot \text{concat}(\boldsymbol{y}_k^v, \boldsymbol{y}_k^t, \boldsymbol{y}_k^m, \boldsymbol{y}_k^u, \boldsymbol{y}_k^d)). \tag{3}$$

We further consider the self-loop projection as a supplement to highlight the central k-th node's information. We have:

$$\boldsymbol{y}_k^S = \mathrm{ReLU}(\boldsymbol{W}_s \cdot \boldsymbol{f}_k). \quad (4)$$

Next, we combine neighbor and self-loop based representations to get the 1st layer output $\boldsymbol{y}_k$, and use the 2nd FH-GAT layer to get the final aggregated representation $\boldsymbol{h}_k$ as:

$$\boldsymbol{h}_k = \mathrm{FH{-}GAT}(\boldsymbol{y}_k), \quad \boldsymbol{y}_k = \lambda_s \cdot \boldsymbol{y}_k^S + (1-\lambda_s) \cdot \boldsymbol{y}_k^N, \quad (5)$$

where λ_s is empirically set as 0.5.

FH-GAT aggregates heterogeneous neighbors separately in each feature field with different field-specific attention, which delicately encodes user diverse preferences related to specific fields to the final node representation. Other GNN models could also be easily adapted to our framework. Comparing with other heterogeneous GAT models like [48], FH-GAT is more like a multi-channel aggregation, which collects field-specific user preferences in categories from multi-hop neighbors for node aggregation. These aggregated node embeddings are regarded as the final representations for all types of nodes in both offline training and online matching.

3.4 Diversity-enhanced Training Objective

Conventional ranking models usually rely on supervised training with CTR-oriented objectives, which also brings in homogenization. In this work, instead of merely focusing on CTR, GraphDR aims to learn user diverse preferences from multi-aspect factors and improve both accuracy and diversity. Therefore, we conduct the **neighbor-similarity based loss** [55] instead of conventional CTR-oriented objectives to highlight diversity. Specifically, we regard nodes connected on the diversified preference network as neighbors, and assume that *all nodes' aggregated representations should be similar to their neighbors* on the diversified preference network regardless of their types. The neighbor-similarity based loss can be viewed as a simplified DeepWalk [56] loss with the path length set as 2 (too long paths may bring in more noises and computation costs), which is formalized as follows:

$$L = \sum_{h_k} \sum_{h_i \in N_k} \sum_{h_j \notin N_k} (\log(\sigma(\boldsymbol{h}_k^\top \boldsymbol{h}_j)) - \log(\sigma(\boldsymbol{h}_k^\top \boldsymbol{h}_i))). \quad (6)$$

$\boldsymbol{h}_k$ is the k-th aggregated node embedding given by FH-GAT, and N_k is the neighbor set of the k-th node. $\sigma(\cdot)$ indicates the sigmoid function. We utilize Adam [57] with negative sampling for training.

We summarize the motivation and advantages of the neighbor-similarity based loss as follows: (1) the neighbor-similarity based assumption projects all heterogeneous aggregated nodes into the same space, making videos to be more similar with their taxonomies, providers, audiences, and related videos in the same sessions than negative samples. If we only consider the video-user edges, this loss will degrade into the classical ranking loss as in MF methods [14]. The loss on other edges brings in user diverse preferences from different aspects. The similar ideas have also been verified in [55] and [58]. (2) Videos that a user may be interested in are very likely to be connected via (multi-step) paths in the diversified preference network. For example, the multi-step path *video:Apple event* $\leftrightarrow$ *tag:iPhone* $\leftrightarrow$ *tag:fast charge* $\leftrightarrow$ *video:new technique of charge* connects two related videos users may be interested. Through the neighbor-similarity based loss, related heterogeneous nodes linked by multi-hop paths in the diversified preference network will have similar representations. (3) GraphDR focuses on the matching module which values efficiency. Hence, the online multi-channel matching in Sec. 4 conducts an embedding-based retrieval to meet the requirement of efficiency, which ranks videos according to the similarities between different types of embeddings. The neighbor-similarity based loss perfectly matches the embedding-based retrieval for efficient, accurate and diverse matching.

Cooperating with the diversified preference network, the neighbor-similarity based loss can well balance both accuracy and diversity, since it calculates video similarities with multiple factors including user watching habit in session, audience community, video content, taxonomy, and content provider. Precisely, the click-based supervised information used in classical ranking models is collected by two global interactions in GraphDR: video-video edges (for sequential click information in session) and video-user edges (for community-aggregated user-item interactions). These two types of click-based interactions are still the dominating interactions (taking nearly 83% of all interactions in our dataset given in Table 1 to ensure the recommendation accuracy. In contrast, the other four interactions related to tags, medias and words mainly provide the generalization ability of node representations to ensure the recommendation diversity. Comparing with classical CTR-oriented losses that merely focus on clicks, GraphDR jointly considers user diverse preferences from multiple heterogeneous interactions, and thus could achieve better accuracy and diversity in matching.

4 Online Serving

We have deployed our GraphDR on the matching module of a widely-used video recommendation system in WeChat Top Stories, which has nearly billion-level daily views generated by million-level users. We introduce the details of the online multi-channel matching, system and efficiency.

4.1 Online Multi-channel Matching

Online multi-channel matching aims to retrieve hundreds of items from millions of candidates rapidly. GraphDR first builds the user representation with his/her valid watching behaviors $\{\hat{v}_1, \cdots, \hat{v}_m\}$ of videos. To improve the diversity, we conduct a multi-channel matching strategy as in Fig 2, which jointly retrieves video candidates from multiple aspects of representative tags, medias and videos in user historical behaviors.

In the video channel, each video in the valid watching behavior sequence retrieves top 100 videos with the cosine similarity between two aggregated video embeddings (pre-calculated and indexed for efficiency). The weighting score of the i-th video v_i in the video channel is formulated as:

$$score_i^v = \sum_{j=1}^{m} x_v(ij) \times complete_j \times time_j \times sim(v_i, \hat{v}_j). \quad (7)$$

$x_v(ij)$ equals 1 only if the i-th video v_i is in the top 100 nearest videos of the j-th video $\hat{v}_j$ in valid watching sequence, and otherwise equals 0. $complete_j$ is the watching time length percentage of $\hat{v}_j$, which measures the user's satisfaction of $\hat{v}_j$. $sim(v_i, \hat{v}_j)$ represents the cosine similarity calculated by the aggregated node embeddings of v_i and $\hat{v}_j$. We also use $time_j$ to highlight the short-term interests of users as follows:

$$time_j = \eta \cdot time_{j+1}, \quad time_m = 1, \tag{8}$$

in which $\eta = 0.95$ is a time decay factor.

In the tag and media channels, we first learn user preferences on tags and medias from user historical behaviors. For example, the i-th tag's preference score p_i^t is defined as:

$$p_i^t = \sum_{j=1}^{m} z_t(ij) \times complete_j \times time_j, \tag{9}$$

where $z_t(ij)$ equals 1 when the i-th tag belongs to $\hat{v}_j$, and otherwise equals 0. To reduce noises, we only select top 10 tags $\hat{t}_j$ ranked by p_i^t to form the user preferred tag set T_u. Next, each tag in T_u retrieves top 100 videos with the cosine similarities between tag and video aggregated embeddings. The weighting score of the i-th video in tag channel is calculated as:

$$score_i^t = \sum_{\hat{t}_j \in T_u} x_t(ij) \times \frac{p_j^t}{\sum_{\hat{t}_k \in T_u} p_k^t} \times sim(v_i, \hat{t}_j). \tag{10}$$

$x_t(ij)$ equals 1 if v_i is in the top 100 nearest videos of $\hat{t}_j$, and otherwise equals 0. $sim(v_i, \hat{t}_j)$ indicates the cosine similarity between v_i and $\hat{t}_j$. The weighting score of v_i in media channel $score_i^m$ is calculated similarly as $score_i^t$ of tag channel.

Finally, we combine all three multiple channels in the joint ranking to get the final video weighting scores as:

$$score_i = \lambda_v \cdot score_i^v + \lambda_t \cdot score_i^t + \lambda_m \cdot score_i^m. \tag{11}$$

We rank all videos with their final video weighting scores and select top-500 videos as the output of GraphDR. Note that GraphDR is deployed in matching, which is responsible for retrieving as many high-quality items as possible. A ranking module is then used to generate top 10 items for the final display to each user. All top hyper-parameters are empirically set according to the practical performances and system limitations. We do not use the user group embedding learned by FH-GAT for the online matching, since they are coarse-grained user community representations, and user historical behaviors are more informative for individuals. We also abandon the word channel in multi-channel matching, considering the ambiguity in words and its information redundancy with the tag channel.

4.2 Online System, Deployment and Efficiency

Online system and deployment. We deploy GraphDR on a well-known recommendation feed named WeChat Top Stories, which distributes million-level heterogeneous items for tens of millions of users. The online recommendation system mainly contains two modules including ranking and matching. The ranking module adopts sophisticated models such as AutoInt [25] and AFN [27] to model feature interactions between all types of features including users and items. Reinforcement learning is also used for list-wise rewards [59] and multiple objectives [60]. In contrast, the matching module aims to retrieve as many appropriate items as possible. Therefore, the matching module contains dozens of different types of matching strategies from various aspects. Our GraphDR and other matching methods work as one of the matching strategies in the matching module. All matching strategies compete with each other to generate item candidates that may be selected by the ranking module. The trade-off over accuracy and diversity can be controlled by (1) setting higher/lower loss weights for user-video edges in Eq. (6), and (2) setting higher/lower channel weights for the video channel (see Sec. 5.7). Sec. 5.6 gives the implementation details of our online evaluation.

Model efficiency. Online matching especially values efficiency to deal with million-level candidates. In GraphDR, all embedding similarities like $sim(v_i, \hat{v}_j)$ are pre-calculated and the top 100 nearest videos are indexed in offline, which enables fast retrieval. The online computation mainly locates in the scoring and sorting part in Sec. 4.1. Hence, its online computation complexity is $O(\log(100m))$ w.r.t the historical behavior length m, which is much superior to most deep ranking models that involve complicated user-item interactions in item scoring (whose computation costs are usually no less than $O(n)$ w.r.t the corpus size n). We spend nearly 11 hours on offline training, which is acceptable with billion-level instances. In online serving, we take nearly 10ms for each request, which meets the online time limitation.

5 Experiments

In experiments, we conduct extensive offline and online evaluations to verify that GraphDR can improve both accuracy and diversity. In this section, we attempt to answer the following five research questions: (**RQ1**): How does the proposed GraphDR model perform against different types of competitive models on recommendation accuracy in the matching module (see Sec. 5.4)? (**RQ2**): How does GraphDR perform against competitive baselines on diversity at the element level, list level and global level (see Sec. 5.5)? (**RQ3**): How does PAPERec perform in online system with various online accuracy and diversity related evaluation metrics (see Sec. 5.6)? (**RQ4**): How do different essential parameters affect GraphDR on recommendation accuracy and diversity (see Sec. 5.7)? (**RQ5**): Will node representations learned by GraphDR be successfully encoded with user diverse preferences (see Sec. 5.8)?

5.1 Datasets

Since there are few large-scale datasets for evaluating recommendation accuracy and diversity in matching, we build a novel dataset DivMat-2.1B extracted from WeChat Top Stories to simulate the real-world scenarios. We randomly select nearly 15 million users, and collect their 2.1 billion video watching instances along with video attributes such as tags, medias, titles and timestamps. Note that the dataset is collected before the GraphDR's online deployment, and there are dozens of matching strategies in online (Sec. 4.2).

Hence, there is no unfair bias in DivMat-2.1B. We split the dataset into a train set and a test set using the chronological order, where the test set contains $8,132,719$ valid watching behavior instances. Following Sec. 3.2, we build a huge diversified preference network via the train set, which has approximately 1.6 million heterogeneous nodes and 120 million edges. To prevent too many noises, we only consider top-frequent words, high-quality videos that have passed our anti-spam filters, and valid watching behaviors. The 15 million users are aggregated into 93 thousand user groups stated in Sec. 3.2 (users in the same user group have the same gender, age, location attributes). All data and attributes are collected after user approvals and data masking to protect user privacy. The detailed statistics of the dataset and the diversified network are in Table 1.

TABLE 1
Statistics of the DivMat-2.1B dataset.

video	user	tag	media	word	instance
1.2M	15M	103K	74K	150K	2.1B
#v-v	#v-t	#v-m	#v-w	#v-u	#t-t
97M	6.1M	1.2M	8.1M	2.3M	5.3M

5.2 Competitors

We implement several classical models as baselines, and categorize these competitors into four groups.

IR-based Methods. We implement three IR-based methods including Category-based, Tag-based and Media-based IR methods [29]. For Tag-based method, we build a tag-video inverted index, where videos for each tag are ranked by their popularity. The online matching retrieves videos with user preferred tags. Other IR-based methods are similar to Tag-based IR method.

CF-based Methods. We implement Item-CF [13] to retrieve similar videos with video co-occurrence. Moreover, we also implement BERT-CF, which uses semantic similarity to measure video similarity. Precisely, we calculate the semantic similarity of two videos with their title embeddings learned by BERT [61], and conduct CF to learn video embeddings for fast retrieval.

Homogeneous NRL Methods. We implement some typical NRL models on the homogeneous video network built with video sessions. The compared methods include DeepWalk [56] and GraphSAGE [45]. These learned video representations are then used for online embedding-based matching with the video channel.

Neural-based Methods. Youtube candidate generation model [3] is a classical deep model for matching. We further improve the original Youtube model with behavior-level attention [62] and neural FM [24] as Youtube+ATT+FM, which is a strong industrial baseline in practice. Moreover, we implement DSSM [30], which retrieves items according to the user-item similarities. We also implement AutoInt [25] to model feature interactions. These models are optimized under supervised learning with video behaviors.

Ablation Test Settings. We implement the heterogeneous versions of GraphSAGE [45] and GAT [46] to replace FH-GAT in the NRL module for ablation tests. We use GraphDR(GraphSAGE) and GraphDR(GAT) to represent these two settings respectively.

We conduct a nearest neighbor server for all embedding-based fast retrieval. We implement most representative industrial matching models including supervised-based methods (from semantic/behavior view) and retrieval-based (from taxonomy/media views) as baselines. Currently, there are few works that focus on diversity in matching. Note that we do not compare with complicated diversified recommendation models specially designed for ranking, due to their tremendous computation costs in matching [8]. We do not compare with TDM/JDM [2], [31] for the static tree-based retrieval verified on E-commerce is challenging to handle various aspects of diversities in video recommendation.

5.3 Experimental Settings

In GraphDR, the node feature embedding dimension is 900, where the video field's dimension d_v is 300 and others' are 150. The dimensions of two output embeddings in FH-GAT are 120. The numbers of neighbor sampling in the first and second layers are 30 and 20. In training, we randomly select 20 negative samples for each positive sample, and set batch size as 512. In online matching, we consider top 200 recent watched videos and retrieve top 500 candidates for ranking. The weighting scores λ_v, λ_t and λ_m are equally set to be 1. We conduct the grid search for parameter selection. For fair comparisons, other graph-based methods and ablation settings also use the same sampling strategy and embedding dimensions. All models follow the same settings in evaluation.

5.4 Recommendation Accuracy (RQ1)

We first evaluate all GraphDR models and baselines on recommendation accuracy in offline DivMat-2.1B dataset.

5.4.1 Evaluation Protocols

We focus on matching that aims to generate **hundreds of** item candidates. Differing from ranking, matching only cares **whether good items are retrieved**, not the specific item ranks. Therefore, we use hit rate (HIT@N) [26] as the evaluation metric for accuracy, where an instance is "hit" if the clicked item is ranked in top N. We do not use classical ranking metrics such as MAP and NDCG since matching does not care specific ranks. To simulate the real-world scenarios, we conduct HIT@N with larger N set as 100, 200, 300 and 500. Since we retrieve **top 500** items in the online recommendation system (see Sec. 4 for details), HIT@500 is considered to be the most essential accuracy metric.

5.4.2 Experimental Results

In Table 3 we can observe that:

(1) GraphDR(FH-GAT) significantly outperforms all baselines on HIT@500 with the significance level $\alpha = 0.01$. It indicates that GraphDR(FH-GAT) could retrieve accurate items in matching. Differing from conventional CTR-oriented models, GraphDR considers user diverse preferences related to video session, community, taxonomy, semantics and provider, which makes the matching results

TABLE 2
Results of different evaluation metrics on recommendation diversity.

Model	Element-level diversity			List-level diversity			Global-level diversity		
	tag	cate	media	tag	cate	media	coverage	long-tail	novelty
Category-based	17.64	1.00	13.15	206.98	4.26	98.76	0.0012	0.0836	0.0043
Tag-based	24.39	1.91	12.20	346.42	23.31	315.48	0.0270	0.1432	0.0343
Media-based	29.67	2.95	1.00	434.30	43.41	9.58	0.0309	0.1327	0.0543
BERT-CF	26.29	2.27	11.06	387.45	30.52	207.41	0.3829	0.2631	0.5734
Item-CF	31.86	3.66	11.47	499.42	55.42	234.31	0.1786	0.0000	0.3143
DeepWalk	30.64	3.24	13.23	476.76	52.53	246.33	0.1642	0.0000	0.3821
GraphSAGE	31.67	2.84	13.65	426.32	41.11	285.52	0.1806	0.0000	0.3532
DSSM	25.15	2.13	13.94	363.41	29.65	211.32	0.1688	0.0525	0.2843
AutoInt	26.31	2.41	13.21	372.31	32.12	242.31	0.1762	0.0612	0.2971
Youtube+ATT+FM	31.22	2.79	12.83	457.15	41.93	217.67	0.1532	0.0734	0.3523
GraphDR(GraphSAGE)	33.19	3.61	14.91	498.31	51.21	327.28	0.4892	0.2854	0.6742
GraphDR(GAT)	34.77	3.79	15.34	516.93	56.62	358.82	0.4934	0.3242	0.7032
GraphDR(FH-GAT)	**37.15**	**3.96**	**16.43**	**538.32**	**63.41**	**379.12**	**0.5132**	**0.3678**	**0.7352**

TABLE 3
Results of recommendation accuracy. HIT@500 is the most essential metric in our evaluation of recommendation accuracy, since we set N=500 in the matching module of our online system.

HIT@N	N=100	N=200	N=300	N=500
Category-based	0.0010	0.0018	0.0021	0.0031
Tag-based	0.0157	0.0207	0.0240	0.0287
Media-based	0.0235	0.0297	0.0337	0.0383
BERT-CF	0.0337	0.0469	0.0556	0.0669
Item-CF	0.0748	0.0904	0.1214	0.1459
DeepWalk	0.0799	0.0998	0.1130	0.1340
GraphSAGE	0.0932	0.1242	0.1568	0.1862
DSSM	0.1012	0.1326	0.1631	0.2031
AutoInt	0.1087	0.1488	0.1892	0.2401
Youtube+ATT+FM	**0.1392**	**0.1892**	0.2194	0.2549
GraphDR(GraphSAGE)	0.1013	0.1442	0.1818	0.2372
GraphDR(GAT)	0.1088	0.1674	0.2108	0.2731
GraphDR(FH-GAT)	0.1241	0.1885	**0.2384**	**0.3102**

more diversified. GraphDR is perfectly suitable for matching, since it concerns more about item coverage than their specific ranks.

(2) GraphDR(FH-GAT) performs comparable or slightly worse than Youtube+ATT+FM when N is small. It is intuitive since the neighbor-similarity based loss should balance accuracy and diversity, which inevitably harms ranking accuracy (not matching). In contrast, Youtube is a strong supervised baseline that benefits from its CTR-oriented objective. However, it suffers from overfitting and homogenization, and thus performs much worse than GraphDR(FH-GAT) when N grows bigger (which is the practical scenario). The diversity issue will be discussed in Sec. 5.5.

(3) Both IR-based methods and BERT-CF are not satisfactory. It indicates that the taxonomy and semantic similarities contribute less to accuracy compared to user behaviors. In contrast, Neural-based methods focus on CTR-oriented objectives and thus get better accuracies. However, they still perform worse than GraphDR, for they fail to consider heterogeneous interactions and thus lack coverage.

Ablation study. Among different GraphDR versions, we find that FH-GAT outperforms GAT and GraphSAGE. It confirms the power of field-specific aggregation in modeling user diverse preferences. Moreover, we further conduct an ablation test to verify that all different types of nodes are necessary for the diversified recommendation. For instance, the HIT@500 will drop to 29.31% if we wipe out all word nodes in DivMat-2.1B.

5.5 Recommendation Diversity (RQ2)

The matching module should be more responsible for the recommendation diversity. In this subsection, we evaluate all models on both individual diversity and aggregate diversity in matching with various evaluation metrics.

5.5.1 Evaluation Protocols

We conduct nine typical diversity metrics and group them into three classes, namely the element-level diversity, the list-level diversity and the global-level diversity. The former two diversities indicate the individual diversity, while the latter diversity measures the aggregate diversity [6] (more details are in Sec. 2.2). The **element-level diversity** focuses on the diversity in each element (or its representation), such as the tag, category, media in IR-based methods and their embeddings in other baselines. Precisely, we regard the average deduplicated tag/category/media numbers in top 20 videos retrieved by these elements as the element-level diversity. The **list-level diversity** measures the diversity at the recommended lists level (i.e., top 500 items), which is the most related metric to reflect the diversity in the matching module. It is time-consuming to measure the dissimilarity-based diversity metric [5] between all top 500 items for different matching methods on million-level instances. Hence, we use the average deduplicated tag/category/media numbers in the final recommended lists as the list-level diversity inspired by [10], [36]. For the **global-level diversity**, *coverage* indicates the percentage of items that could be recommended in the overall corpus [8]. *Long-tail* indicates the percentage of long-tail items in all results (videos that

TABLE 4
Online A/B test on recommendation accuracy and diversity in a real-world system.

Model	VV	VWT/c	VWT/v	PT	DIV	Tag diver	Cate diver
GraphSAGE	+3.08%	+6.20%	+1.48%	+4.66%	+2.43%	+6.24%	+10.27%
GraphDR(GraphSAGE)	+4.37%	+7.61%	+1.68%	+6.04%	+3.97%	+9.16%	+12.42%
GraphDR(GAT)	+5.30%	+9.36%	+2.49%	+6.07%	+8.00%	+12.81%	+15.57%
GraphDR(FH-GAT)	**+6.08%**	**+10.79%**	**+3.10%**	**+6.10%**	**+10.43%**	**+14.68%**	**+17.00%**

have not been watched for 15 days are empirically viewed as the long-tail videos). *Novelty* represents the percentage of unique new items (i.e., items that can only be recommended in this model) in the overall recommended item set, which is inspired by the item novelty in [34].

5.5.2 Experimental Results

Table 2 shows the results of various diversity metrics, from which we can know that:

(1) GraphDR(FH-GAT) achieves the best performances in all diversity metrics. The improvement derives from all three modules: (i) in diversified preference network, the heterogeneous interactions store user diverse preferences on taxonomy, semantics, community, video session and provider to link similar videos via multi-hop paths. (ii) In NRL, FH-GAT and its neighbor-similarity based loss successfully encode user diverse preferences into node representations. (iii) In online matching, the multi-channel strategy retrieves items from tag/media/video aspects, which also amplifies diversity. In addition, GraphDR(GraphSAGE) and GraphDR(GAT) generally outperform all baselines but still inferior to GraphDR(FH-GAT). It reconfirms the power of FH-GAT in diversity.

(2) The element-level and list-level diversities indirectly measure the individual diversity with diversities in tag, category and media. We assume that more tags/medias/categories in recommended lists indicate a more diversified recommendation. We find that behavior-based models like Youtube and GraphSAGE perform better than other baselines in individual diversities. Nevertheless, GraphDR has better results since it considers other types of interactions.

(3) The global-level diversity measures the aggregate diversity, where coverage, long-tail and novelty focus on different aspects. Behavior-based models only consider video watching behaviors, which are hard to handle long-tail and new items. In contrast, BERT-CF focuses on content similarity and achieves good aggregate diversity. Still, GraphDR considers user diverse preferences in various fields and achieves the best aggregate diversity. With the help of the neighbor-similarity based loss and the multi-channel matching, more related long-tail videos are retrieved via tag/media/word related similarities rather than user behaviors, which alleviates the popularity bias.

5.6 Online Evaluation (RQ3)

The offline evaluation has verified the improvements on accuracy and diversity in the matching module. We further conduct an online A/B test to evaluate GraphDR in real-world industrial-level scenarios.

5.6.1 Evaluation Protocols

We implement GraphDR on the matching module of WeChat Top Stories following Sec. 4. The original online matching model is an ensemble model containing multiple IR-based, CF-based and Neural-based methods in Sec. 5.2, which is widely used in practice. We regard GraphDR as an additional matching channel to the existing online ensemble model, with the ranking module unchanged. All videos retrieved by different matching channels will jointly compete with each other in the following ranking module. This online evaluation builds a real-world scenario for different matching algorithms to cooperate with the ranking module and compete with each other.

In online A/B test, we focus on the following seven representative metrics to evaluate accuracy and diversity: (1) video views per capita (VV), (2) video watching time per capita (VWT/c), (3) video watching time per video (VWT/v), (4) page turns per capita (PT), (5) deduplicated impressed videos per capita (DIV), (6) watched tag per capita (Tag diver), and (7) watched category per capita (Cate diver). The former five metrics mainly measure accuracy, while the latter two metrics measure diversity. We conduct the A/B test for 5 days with nearly 3.8 million users involved, and report the improvement percentages over the ensemble base model. The online evaluation can be viewed as an online ablation test for graph-based methods.

5.6.2 Experimental Results

Table 4 shows the results of online evaluation with multiple metrics, from which we find that:

(1) All GraphDR models outperform the ensemble base model, among which GraphDR(FH-GAT) achieves the best performances in accuracy and diversity with the significance level $\alpha = 0.01$. We have also passed the homogeneity test in online evaluation, which confirms that the system and traffic split are unbiased and the improvements are stable. It verifies the effectiveness of GraphDR in real-world scenarios. Moreover, the improvements from GraphSAGE to FH-GAT also imply the significances of FH-GAT.

(2) The significant improvements in the former five metrics reflect better accuracy. A better video view metric indicates that users are more willing to click videos, while a better video watching time indicates users are genuinely interested in their clicked videos. Moreover, the page turns and deduplicated impressed video metrics also reflect user experiences indirectly. Users will slide down and browse more videos if they are satisfied with the results.

(3) The average watched tags and categories measure the diversity. The better tag/category diversity derives from two factors: (a) more diverse videos impressed to users, and (b) better personalized results that can attract users to watch

more videos. These diverse items help us to explore users' potential interests and give surprising results, which could even contribute to the long-term performances. GraphDR achieves significant improvements on diversity, which also indicates that our model can bring in additional diversities compared to the diversified ensemble baseline.

5.7 Model Analyses (RQ4)

We conduct several analyses on different channels and user behavior sequence lengths to better understand GraphDR.

5.7.1 Analysis on Multi-channel Matching

In GraphDR, the online multi-channel matching module plays an important role in improving diversity. We evaluate the GraphDR(FH-GAT) on HIT@N and list-level diversity metrics with different channels individually. From Table 5 we find that: the video channel achieves better HIT@N results, since video embeddings are directly influenced by video watching behaviors. In contrast, the tag and media channels are more responsible for diversity. To balance accuracy and diversity, we combine all three channels for the online multi-channel matching in GraphDR.

TABLE 5
Results of different matching channels.

Channel	tag	media	video	joint
HIT@100	0.1027	0.0934	**0.1323**	0.1241
HIT@200	0.1571	0.1497	**0.1943**	0.1885
HIT@300	0.2143	0.2032	**0.2512**	0.2384
HIT@500	0.2787	0.2583	**0.3312**	0.3102
Tag diversity	**573.43**	543.31	468.42	538.32
Cate diversity	**71.31**	68.32	53.63	63.41
Media diversity	387.48	**401.58**	344.32	379.12

5.7.2 Analysis on Behavior Sequence Length

We also analyze the impacts of different behavior sequence lengths in online matching. In Table 6, as the behavior sequence length increases, HIT@N metrics achieve consistent improvements, while diversity metrics become slightly worse. It indicates that considering user long-term preferences can better understand users in recommendation. However, user long-term preferences are more stable, which inevitably harm the diversity. In GraphDR, we set the length as 200 since the improvements in accuracy are more significant than diversity.

TABLE 6
Results of different behavior sequence lengths.

Length	m=20	m=50	m=100	m=200
HIT@100	0.0791	0.0883	0.1072	**0.1241**
HIT@200	0.1237	0.1373	0.1653	**0.1885**
HIT@300	0.1742	0.1902	0.2114	**0.2384**
HIT@500	0.2393	0.2617	0.2763	**0.3102**
Tag diversity	**556.12**	552.22	547.43	538.32
Cate diversity	**69.52**	68.11	66.73	63.41
Media diversity	**395.45**	391.52	387.91	379.12

5.8 Case Study (RQ5)

In GraphDR, user diverse preferences are encoded in node embeddings. We give some tags and their nearest tags to explicitly display the diversity in Table 7. The interest in *Restaurant guide* may expand to specific food like *Foie gras* and their stories like *Food documentary*. The nearest tags of *El Nino phenomenon* reflect the interests in nature and science. Users like *iPhone 11 Pro Max* may also seek information on its hardware, software, and discount information. These nearest tags reflect both similarities in semantics and user preferences, since the node representations are learned under the neighbor-similarity based objective with a diversified preference graph containing various heterogeneous feature interactions. The similar phenomenon can be found in other nodes, showing that each heterogeneous representation could contain user diversified preferences.

TABLE 7
Examples of tags and their nearest tags.

Tag	Nearest tags
Restaurant guide	Roasted goose; Food documentary; Melaleuca cake; Foie gras; Hong Kong cuisine
El Nino phenomenon	Superluminal speed; Easter island; Darwin; Absolute zero; Parallel worlds theory
iPhone 11 Pro Max	iPhone SE; Fast charge; Mobile phone test; Voice assistant; iPhone discount

Table 8 shows the nearest tags of some typical user groups. According to the node embeddings and aggregated behaviors, young men users in our dataset are more interested in sports, while young women focus more on fashion. Differing from the youth, the elderly in Beijing concentrate on traditional Chinese art and culture. The geographic distance also leads to fine-grained differences in interested sports (e.g., golf V.S. soccer). The preference divergences in different communities verify the success of diversity modeling.

TABLE 8
Examples of user groups with nearest tags.

Sex	Age	City	Nearest tags
M	21	Beijing	Sports news; Entrepreneur; Comedy; Scientific anecdotes; Soccer
F	21	Beijing	Summer wear; Constellation; Product promotion; Diet food; Potted plant
M	59	Beijing	Calligraphy; Social documentary; Tai Chi; Exercise; Family
M	21	London	London Olympics; The Celtic; Scientists; Golf; 100 metres race

6 Conclusion and Future Work

In this work, we propose a simple and effective GraphDR framework to improve both accuracy and diversity in

matching. We propose a new diversified preference network to capture heterogeneous interactions between essential objects in recommendation. We also design a novel FH-GAT model with a neighbor-similarity based loss to encode user diverse preferences from heterogeneous interactions. In experiments, we conduct extensive offline and online evaluations, model analyses and case studies. The significant improvements verify the effectiveness and robustness of GraphDR in jointly improving accuracy and diversity.

In the future, we will explore more types of interactions and weighted edges in GraphDR. Moreover, we will enhance the multi-channel matching with more sophisticated models. Better graph neural networks and the combinations with existing diversity-aware methods in the ranking module are also worth being studied.

REFERENCES

[1] J. Wang, P. Huang, H. Zhao, Z. Zhang, B. Zhao, and D. L. Lee, "Billion-scale commodity embedding for e-commerce recommendation in alibaba," in *Proceedings of KDD*, 2018.

[2] H. Zhu, X. Li, P. Zhang, G. Li, J. He, H. Li, and K. Gai, "Learning tree-based deep model for recommender systems," in *Proceedings of KDD*, 2018.

[3] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for youtube recommendations," in *Proceedings of RecSys*, 2016.

[4] R. Xie, Z. Qiu, J. Rao, Y. Liu, B. Zhang, and L. Lin, "Internal and contextual attention network for cold-start multi-channel matching in recommendation," in *Proceedings of IJCAI*, 2020.

[5] K. Bradley and B. Smyth, "Improving recommendation diversity," in *Proceedings of AICS*, 2001.

[6] M. Kunaver and T. Požrl, "Diversity in recommender systems–a survey," *Knowledge-Based Systems*, 2017.

[7] L. Chen, G. Zhang, and E. Zhou, "Fast greedy map inference for determinantal point process to improve recommendation diversity," in *Proceedings of NIPS*, 2018.

[8] M. Ö. Karakaya and T. Aytekin, "Effective methods for increasing aggregate diversity in recommender systems," *knowledge and Information Systems*, 2018.

[9] L. Zhang, Q. Yan, J. Lu, Y. Chen, and Y. Liu, "Empirical research on the impact of personalized recommendation diversity," in *Proceedings of HICSS*, 2019.

[10] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen, "Improving recommendation lists through topic diversification," in *WWW*, 2005.

[11] T. Aytekin and M. Ö. Karakaya, "Clustering-based diversity improvement in top-n recommendation," *Journal of Intelligent Information Systems*, 2014.

[12] S. Nandanwar, A. Moroney, and M. N. Murty, "Fusing diversity in recommendations in heterogeneous information networks," in *Proceedings of WSDM*, 2018.

[13] B. M. Sarwar, G. Karypis, J. A. Konstan, J. Riedl *et al.*, "Item-based collaborative filtering recommendation algorithms." in *Proceedings of WWW*, 2001.

[14] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, 2009.

[15] S. Rendle, "Factorization machines," in *Proceedings of ICDM*, 2010.

[16] Y. Shan, T. R. Hoens, J. Jiao, H. Wang, D. Yu, and J. Mao, "Deep crossing: Web-scale modeling without manually crafted combinatorial features," in *Proceedings of KDD*, 2016.

[17] W. Zhang, T. Du, and J. Wang, "Deep learning over multi-field categorical data," in *European conference on information retrieval*, 2016.

[18] Y. Qu, H. Cai, K. Ren, W. Zhang, Y. Yu, Y. Wen, and J. Wang, "Product-based neural networks for user response prediction," in *Proceedings of ICDM*, 2016.

[19] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir *et al.*, "Wide & deep learning for recommender systems," in *Proceedings of the 1st workshop on deep learning for recommender systems*, 2016.

[20] R. Wang, B. Fu, G. Fu, and M. Wang, "Deep & cross network for ad click predictions," in *Proceedings of ADKDD*, 2017.

[21] R. Xie, C. Ling, Y. Wang, R. Wang, F. Xia, and L. Lin, "Deep feedback network for recommendation," in *Proceedings of IJCAI*, 2020.

[22] H. Guo, R. Tang, Y. Ye, Z. Li, and X. He, "Deepfm: a factorization-machine based neural network for ctr prediction," in *Proceedings of IJCAI*, 2017.

[23] J. Xiao, H. Ye, X. He, H. Zhang, F. Wu, and T.-S. Chua, "Attentional factorization machines: Learning the weight of feature interactions via attention networks," in *Proceedings of IJCAI*, 2017.

[24] X. He and T.-S. Chua, "Neural factorization machines for sparse predictive analytics," in *Proceedings of SIGIR*, 2017.

[25] W. Song, C. Shi, Z. Xiao, Z. Duan, Y. Xu, M. Zhang, and J. Tang, "Autoint: Automatic feature interaction learning via self-attentive neural networks," in *Proceedings of CIKM*, 2019.

[26] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang, "Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer," in *Proceedings of CIKM*, 2019.

[27] W. Cheng, Y. Shen, and L. Huang, "Adaptive factorization network: Learning adaptive-order feature interactions," in *Proceedings of AAAI*, 2020.

[28] B. Liu, C. Zhu, G. Li, W. Zhang, J. Lai, R. Tang, X. He, Z. Li, and Y. Yu, "Autofis: Automatic feature interaction selection in factorization models for click-through rate prediction," in *Proceedings of KDD*, 2020.

[29] M. K. Khribi, M. Jemni, and O. Nasraoui, "Automatic recommendations for e-learning personalization based on web usage mining techniques and information retrieval," in *Proceedings of ICALT*, 2008.

[30] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning deep structured semantic models for web search using clickthrough data," in *Proceedings of CIKM*, 2013.

[31] H. Zhu, D. Chang, Z. Xu, P. Zhang, X. Li, J. He, H. Li, J. Xu, and K. Gai, "Joint optimization of tree-based index and deep model for recommender systems," in *Proceedings of NIPS*, 2019.

[32] J. Zhuo, Z. Xu, W. Dai, H. Zhu, H. Li, J. Xu, and K. Gai, "Learning optimal tree models under beam search," in *Proceedings of ICML*, 2020.

[33] J.-T. Huang, A. Sharma, S. Sun, L. Xia, D. Zhang, P. Pronin, J. Padmanabhan, G. Ottaviano, and L. Yang, "Embedding-based retrieval in facebook search," in *Proceedings of KDD*, 2020.

[34] M. Zhang and N. Hurley, "Avoiding monotony: improving the diversity of recommendation lists," in *Proceedings of RecSys*, 2008.

[35] J. Hao, T. Zhao, J. Li, X. L. Dong, C. Faloutsos, Y. Sun, and W. Wang, "P-companion: A principled framework for diversified complementary product recommendation," in *Proceedings of CIKM*, 2020.

[36] L. Wu, Q. Liu, E. Chen, N. J. Yuan, G. Guo, and X. Xie, "Relevance meets coverage: A unified framework to generate diversified recommendations," *TIST*, 2016.

[37] L. Gan, D. Nurbakova, L. Laporte, and S. Calabretto, "Enhancing recommendation diversity using determinantal point processes on knowledge graphs," in *Proceedings of SIGIR*, 2020.

[38] L. Qin and X. Zhu, "Promoting diversity in recommendation by entropy regularizer," in *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[39] X. Zhu, A. Goldberg, J. Van Gael, and D. Andrzejewski, "Improving diversity in ranking using absorbing random walks," in *Proceedings of NAACL*, 2007.

[40] Y. Liu, Y. Zhang, Q. Wu, C. Miao, L. Cui, B. Zhao, Y. Zhao, and L. Guan, "Diversity-promoting deep reinforcement learning for interactive recommendation," *arXiv preprint arXiv:1903.07826*, 2019.

[41] Q. Wu, Y. Liu, C. Miao, B. Zhao, Y. Zhao, and L. Guan, "Pd-gan: Adversarial learning for personalized diversity-promoting recommendation," in *Proceedings of IJCAI*, 2019.

[42] S. Li, A. Karatzoglou, and C. Gentile, "Collaborative filtering bandits," in *Proceedings of SIGIR*, 2016.

[43] K. Mahadik, Q. Wu, S. Li, and A. Sabne, "Fast distributed bandits for online recommendation systems," in *Proceedings of ICS*, 2020.

[44] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proceedings of ICLR*, 2017.

[45] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proceedings of NIPS*, 2017.

[46] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *Proceedings of ICLR*, 2018.

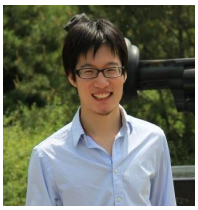
[47] C. Zhang, D. Song, C. Huang, A. Swami, and N. V. Chawla, "Heterogeneous graph neural network," in *Proceedings of KDD*, 2019.
[48] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P. S. Yu, "Heterogeneous graph attention network," in *Proceedings of WWW*, 2019.
[49] S. Wu, Y. Tang, Y. Zhu, L. Wang, X. Xie, and T. Tan, "Session-based recommendation with graph neural networks," in *Proceedings of AAAI*, 2019.
[50] W. Fan, Y. Ma, Q. Li, Y. He, E. Zhao, J. Tang, and D. Yin, "Graph neural networks for social recommendation," in *Proceedings of WWW*, 2019.
[51] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "Lightgcn: Simplifying and powering graph convolution network for recommendation," in *Proceedings of SIGIR*, 2020.
[52] J. Chang, C. Gao, X. He, D. Jin, and Y. Li, "Bundle recommendation with graph convolutional networks," in *Proceedings of SIGIR*, 2020.
[53] X. L. Dong, X. He, A. Kan, X. Li, Y. Liang, J. Ma, Y. E. Xu, C. Zhang, T. Zhao, G. Blanco Saldana *et al.*, "Autoknow: Self-driving knowledge collection for products of thousands of types," in *Proceedings of KDD*, 2020.
[54] Y. Lu, R. Xie, C. Shi, Y. Fang, W. Wang, X. Zhang, and L. Lin, "Social influence attentive neural network for friend-enhanced recommendation," in *Proceedings of ECML-PKDD*, 2020.
[55] Q. Liu, R. Xie, L. Chen, S. Liu, K. Tu, P. Cui, B. Zhang, and L. Lin, "Graph neural network for tag ranking in tag-enhanced video recommendation," in *Proceedings of CIKM*, 2020.
[56] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of KDD*, 2014.
[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of ICLR*, 2015.
[58] K. Zhou, H. Wang, W. X. Zhao, Y. Zhu, S. Wang, F. Zhang, Z. Wang, and J.-R. Wen, "S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization," in *Proceedings of CIKM*, 2020.
[59] R. Xie, S. Zhang, R. Wang, F. Xia, and L. Lin, "Hierarchical reinforcement learning for integrated recommendation," in *Proceedings of AAAI*, 2021.
[60] R. Xie, Y. Liu, S. Zhang, R. Wang, F. Xia, and L. Lin, "Personalized approximate pareto-efficient recommendation," in *Proceedings of WWW*, 2021.
[61] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL*, 2019.
[62] G. Zhou, X. Zhu, C. Song, Y. Fan, H. Zhu, X. Ma, Y. Yan, J. Jin, H. Li, and K. Gai, "Deep interest network for click-through rate prediction," in *Proceedings of KDD*, 2018.

Ruobing Xie is a senior researcher of WeChat, Tencent. He received his BEng degree in 2014 and his master degree in 2017 from the Department of Computer Science and Technology, Tsinghua University. His research interests include recommender system, knowledge graph and natural language processing. He has published over 30 papers in top-tier conferences and journals including KDD, WWW, SIGIR, TKDE, ACL, and AAAI.

Qi Liu received the master degree in Institute College of Information Science and Technology, University of Science and Technology of China, in 2016. He is currently a senior researcher of WeChat Recommendation Product Center, Tencent. His researcher interests include deep learning, machine translation, graph neural network, and its applications on recommendation system and NLP.

Shukai Liu is a researcher of the Recommendation Product Center, WeChat Search Application Department, Tencent. He got his BEng degree in 2012 from School of Computer Science, Xidian University, China, and got his Master Degree in 2015 from School of Computer Science and Technology, Beihang University, China. His research interests include recommendation system, knowledge graph and social network.

Ziwei Zhang received his B.S. from the Department of Physics, Tsinghua University, in 2016. He is currently pursuing a Ph.D. Degree in the Department of Computer Science and Technology at Tsinghua University. His research interests focus on network representation learning and machine learning on graph data. He has published several papers in prestigious conferences and journals, including KDD, AAAI, IJCAI, and TKDE.



Peng Cui is an Associate Professor with tenure in Tsinghua University. He got his PhD degree from Tsinghua University in 2010. His research interests include causally-regularized machine learning, network representation learning, and social dynamics modeling. He has published more than 100 papers in prestigious conferences and journals in data mining and multimedia. His recent research won the IEEE Multimedia Best Department Paper Award, SIGKDD 2016 Best Paper Finalist, ICDM 2015 Best Student Paper Award, SIGKDD 2014 Best Paper Finalist, IEEE ICME 2014 Best Paper Award, ACM MM12 Grand Challenge Multimodal Award, and MMM13 Best Paper Award. He is PC co-chair of CIKM2019 and MMM2020, SPC or area chair of WWW, ACM Multimedia, IJCAI, AAAI, etc., and Associate Editors of IEEE TKDE, IEEE TBD, ACM TIST, and ACM TOMM etc. He received ACM China Rising Star Award in 2015, and CCF-IEEE CS Young Scientist Award in 2018. He is now a Distinguished Member of ACM amd CCF, and a Senior Member of IEEE.

Bo Zhang is a senior researcher of WeChat Recommendation Product Center, Tencent. He got his BEng degree in 2009 from School of Computer Science, Xidian University, China, and got his Master Degree in 2012 from School of Computer Science and Technology, Zhejiang University, China. His research interests include recommender system and its applications.

Leyu Lin received the master degree in Institute of Computing Technology, Chinese Academy of Sciences, in 2008. He is currently the director of WeChat Recommendation Product Center, Tencent. His research interests include machine learning and its applications, such as search system, recommendation system and computational advertising.