A Multi-Strategy based Pre-Training Method for Cold-Start Recommendation

Bowen Hao, Hongzhi Yin, Jing Zhang, Cuiping Li, and Hong Chen

Abstract—Cold-start problem is a fundamental challenge for recommendation tasks. The recent self-supervised learning (SSL) on Graph Neural Networks (GNNs) model, PT-GNN, pre-trains the GNN model to reconstruct the cold-start embeddings and has shown great potential for cold-start recommendation. However, due to the over-smoothing problem, PT-GNN can only capture up to 3-order relation, which can not provide much useful auxiliary information to depict the target cold-start user or item. Besides, the embedding reconstruction task only considers the intra-correlations within the subgraph of users and items, while ignoring the inter-correlations across different subgraphs. To solve the above challenges, we propose a multi-strategy based pre-training method for cold-start recommendation (MPT), which extends PT-GNN from the perspective of model architecture and pretext tasks to improve the cold-start recommendation performance. Specifically, in terms of the model architecture, in addition to the short-range dependencies of users and items captured by the GNN encoder, we introduce a Transformer encoder to capture long-range dependencies. In terms of the pretext task, in addition to considering the intra-correlations of users and items by the embedding reconstruction task, we add embedding contrastive learning task to capture inter-correlations of users and items. We train the GNN and Transformer encoders on these pretext tasks under the meta-learning setting to simulate the real cold-start scenario, making the model easily and rapidly being adapted to new cold-start users and items. Experiments on three public recommendation datasets show the superiority of the proposed MPT model against the vanilla GNN models, the pre-training GNN model on user/item embedding inference and the recommendation task.

Index Terms—Recommender system, cold-start, pre-training, self-supervised learning

1 Introduction

Recommendation systems [17], [27] have been extensively deployed to alleviate information overload in various web services, such as social media, E-commerce websites and news portals. To predict the likelihood of a user adopting an item, collaborative filtering (CF) is the most widely adopted principle. The most common paradigm for CF, such as matrix factorization [36] and neural collaborative filtering [27], is to learn the user/item embeddings based on the user-item interactions. However, these models fail to learn high-quality embeddings for the cold-start users/items with sparse interactions.

To address the cold-start problem, traditional recommender systems incorporate the side information such as content features of users and items [25], [26] or external knowledge graphs (KGs) [4], [5] to compensate the low-quality embeddings caused by sparse interactions. However, the content features are not always available, and it is not easy to link the items to the entities in KGs due to the incompleteness and ambiguities of the entities. Another research line is to adopt the Graph Neural Networks (GNNs) to solve the problem, examples include PinSage [15], NGCF [16] and LightGCN [17]. The basic idea is to incorporate the high-order neighbors to refine the representations of the users/items. However, the GNN models for recommendation can not thoroughly solve the cold-start problem, as the embeddings of the cold-start users/items aren't explicitly optimized, and the cold-start neighbors have not been considered in these GNNs.

In our previous work, we propose PT-GNN [30], which pre-trains the GNN model to reconstruct the user/item embeddings under the meta-learning setting [8]. To further reduce the impact from the cold-start neighbors, PT-GNN incorporates a self-attention based meta aggregator to enhance the aggregation ability of each graph convolution step, and an adaptive neighbor sampler to select proper high-order neighbors according to the feedbacks from the GNN model.

However, PT-GNN still suffers from the following challenges: 1) In terms of the model architecture, PT-GNN suffers from lacking the ability to capture long-range dependencies. Existing researches such as Chen et al. [46] pointed out that GNN can only capture up to 3-hop neighbors due to the overfitting and over-smoothing problems [17]. However, in the cold-start scenario, both low-order neighbors ($L \leq 2$, where L is the convolution layer) and high-order neighbors ($L > 2$) are important to the target cold-start user/item. On one hand, the low-order neighbors are crucial for representing the target cold-start users/items, as the first-order neighbors directly reflect the user's preference or the item's target audience, and the second-order neighbors directly reflect the signals of the collaborative users/items. On the other hand, due to the extremely sparse interactions of the cold-start users/items, the first- and second-order neighbors are insufficient to represent a user or an item. Thus, for the target users/items, it is crucial to capture not only the short-range dependencies from their low-order neighbors, but also the long-range dependencies from their

- *B. Hao, J.Zhang, C. Li, and H. Chen are with the School of Information, Renmin University of China, Beijing 100872, China. E-mail: {jeremyhao, zhang-jing, licuiping, chong}@ruc.edu.cn.*
- *H. Yin is with the School of Information Technology & Electric Engineering, The University of Queensland, St Lucia, QLD 4072, Australia. E-mail: h.yin1@uq.edu.au.*

Manuscript received April 19, 2015; revised August 26, 2015.

	User/Item Embedding Reconstruction	User/Item Embedding Contrastive Learning
GNN Encoder	Intra-correlation ✓ Short-range Dependency	Inter-correlation ✗ Short-range Dependency
Transformer Encoder	Intra-correlation ✗ Long-range Dependency	Inter-correlation ✗ Long-range Dependency

Fig. 1: The improvement of MPT on the original PT-GNN. ✓ denotes the capacity of PT-GNN and ✗ denotes the missing capacity of PT-GNN.

high-order neighbors. Nevertheless, due to the limitation of the network structure, most of the existing GNN models can only capture the short-range dependencies, which inspires the first research question: *how to capture both short- and long-range dependencies of users/items?* 2) In terms of the pretext task, PT-GNN only considers the intra-correlations within the subgraph of the target user or item, but ignores the inter-correlations between the subgraphs of different users or items. More concretely, given a target cold-start user or item, PT-GNN samples its high-order neighbors to form a subgraph, and leverages the subgraph itself to reconstruct the cold-start embedding. Essentially, the embedding reconstruction task is a generative task, which focuses on exploring intra-correlations between nodes in a subgraph. On the contrary, some recent pre-training GNN models (e.g., GCC [20], GraphCL [54]) depending on the contrastive learning mechanism, can capture the similarities of nodes between different subgraphs, i.e., the inter-correlations. Thus, this inspires the second research question: *how to capture the inter-correlations of the target cold-start users/items in addition to the intra-correlations?*

Present work. We propose a Multi-strategy based Pre-Training method for cold-start recommendation (MPT), which extends PT-GNN from the perspective of model architecture and pretext tasks to improve the cold-start recommendation performance. The improvements over PT-GNN are shown in Fig. 1.

First, in terms of the model architecture, in addition to the original GNN encoder which captures the short-range dependencies from the user-item edges in the user-item graph, we introduce a Transformer encoder to capture the long-range dependencies from the user-item paths, which can be extracted by performing random walk [41] starting from the target user or item in the user-item graph. The multi-head self-attention mechanism in the Transformer attends nodes in different positions in a path [37], which can explicitly capture the long-range dependencies between users and items. Besides, we change the RL-based neighbor sampling strategy in the original GNN encoder into a simple yet effective dynamic sampling strategy, which can reduce the time complexity of the neighbor sampling process.

Second, in terms of the pretext task, in addition to the original embedding reconstruction task which considers the intra-correlations within the subgraph of the target user or item, we add embedding contrastive learning [50] task to capture the inter-correlations across the subgraphs or paths of different users or items. Specifically, we first augment a concerned subgraph or path by deleting or replacing the nodes in it. Then we treat the augmented subgraphs or paths of the concerned user or item as its positive counterparts, and those of other users or items as the negative counterparts. By contrastive learning upon the set of the positive and negative instances, we can pull the similar user or item embeddings together while pulling away dissimilar embeddings.

We train the GNN and Transformer encoders by the reconstruction and contrastive learning pretext tasks under the meta-learning setting [8] to simulate the cold-start scenario. Specifically, following PT-GNN, we first pick the users/items with sufficient interactions as the target users/items, and learn their ground-truth embeddings on the observed abundant interactions. Then for each target user/item, in order to simulate the cold-start scenario, we mask other neighbors and only maintain K first-order neighbors, based on which we form the user-item subgraph and use random walk [41] to generate the paths of the target user or item. Next we perform graph convolution multiple steps upon the user-item subgraph or self-attention mechanism upon the path to obtain the target user/item embeddings. Finally, we optimize the model parameters with the reconstruction and contrastive losses.

We adopt the pre-training & fine-tuning paradigm [20] to train MPT. During the pre-training stage, in order to capture the correlations of users and items from different views (i.e., the short- and long-range dependencies, the intra- and inter-correlations), we assign each pretext task an independent set of initialized embeddings, and train these tasks independently. During the fine-tuning state, we fuse the pre-trained embeddings from each pretext task and fine-tune the encoders by the downstream recommendation task. The contributions can be summarized as:

- We extend PT-GNN from the perspective of model architecture. In addition to the short-range dependencies of users and items captured by the GNN encoder, we add a Transformer encoder to capture long-range dependencies.
- We extend PT-GNN from the perspective of pretext tasks. In addition to considering the intra-correlations of users and items by the embedding reconstruction task, we add embedding contrastive learning task to capture inter-correlations of users and items.
- Experiments on both intrinsic embedding evaluation and extrinsic downstream tasks demonstrate the superiority of our proposed MPT model against the state-of-the-art GNN model and the original proposed PT-GNN.

2 PRELIMINARIES

In this section, we first define the problem and then introduce the original proposed PT-GNN .

2.1 Notation and Problem Definition

Bipartite Graph. We formalize the user-item interaction data for recommendation as a bipartite graph and denoted it as $\mathcal{G} = (\mathcal{U}, \mathcal{I}, \mathcal{E})$, where $\mathcal{U} = \{u_1, \cdots, u_{|\mathcal{U}|}\}$ is the set of users and $\mathcal{I} = \{i_1, \cdots, i_{|\mathcal{I}|}\}$ is the set of items. $\mathcal{U}$ and $\mathcal{I}$ comprise two types of the nodes in $\mathcal{G}$. $\mathcal{E} \subseteq \mathcal{U} \times \mathcal{I}$ denotes the set of edges that connect the users and items. We use $\mathcal{N}^l(u)$

to represent the l-order neighbors of user u. When ignoring the superscript, $\mathcal{N}(u)$ indicates the first-order neighbors of u. $\mathcal{N}^l(i)$ and $\mathcal{N}(i)$ are defined similarly for items.

User-Item Path. The user-item path is generated by the random walk strategy from the user-item interaction data, and has the same node distribution with the graph $\mathcal{G}$, as Perozzi et al. [41] proposed. Formally, there exists two types of paths: $\mathcal{P}$ = UIUI or $\mathcal{P}$ = IUIU, where U denotes the node type is user and I denotes the node type is item.

Problem Definition. Let $f : \mathcal{U} \cup \mathcal{I} \rightarrow \mathbf{R}^d$ be the encoding function that maps the users or items to d-dimension real-valued vectors. We denote a user u and an item i with initial embeddings $\mathbf{h}_u^0$ and $\mathbf{h}_i^0$, respectively. Given the bipartite graph $\mathcal{G}$ or the path $\mathcal{P}$, we aim to pre-train the encoding function f that is able to be applied to the downstream recommendation task to improve its performance. Note that for simplicity, in the following sections, we mainly take user embedding as an example to explain the proposed model, as item embedding can be explained in the same way.

2.2 A Brief Review of PT-GNN

The basic idea of PT-GNN is to leverage the vanilla GNN model as the encoder f to reconstruct the target cold-start embeddings to explicitly improve their embedding quality. To further reduce the impact from the cold-start neighbors, PT-GNN incorporates a self-attention based meta aggregator to enhance the aggregation ability of each graph convolution step, and an adaptive neighbor sampler to select proper high-order neighbors according to the feedbacks from the GNN model.

2.2.1 Basic Pre-training GNN Model

The basic pre-training GNN model reconstructs the cold-start embeddings under the meta-learning setting [8] to simulate the cold-start scenario. Specifically, for each target user u, we mask some neighbors and only maintains at most K^l neighbors (K is empirically selected as a small number, e.g., K=3) at the l-th layer. Based on which we perform graph convolution multiple steps to predict the target user embedding. Take GraphSAGE [29] as the backbone GNN model as an example, the graph convolution process at the l-th layer is:

$$\mathbf{h}_u^l = \sigma(\mathbf{W}^l \cdot (\mathbf{h}_u^{l-1} \,||\, \mathbf{h}_{\mathcal{N}(u)}^l)), \tag{1}$$

where $\mathbf{h}_u^l$ is the refined user embedding at the l-th convolution step, $\mathbf{h}_u^{l-1}$ is the previous user embedding ($\mathbf{h}_u^0$ is the initialized embedding). $\mathbf{h}_{\mathcal{N}(u)}^l$ is the averaged embedding of the neighbors, in which the neighbors are sampled by the random sampling strategy. σ is the sigmoid function, $\mathbf{W}^l$ is the parameter matrix, and $||$ is the concatenate operation.

We perform graph convolution L-1 steps, aggregate the refined embeddings of the first-order neighbors $\{\mathbf{h}_1^{L-1}, \cdots, \mathbf{h}_K^{L-1}\}$ to obtain the smoothed embedding $\mathbf{h}_{\mathcal{N}(u)}^L$, and transform it into the target embedding $\mathbf{h}_u^L$, i.e., $\mathbf{h}_{\mathcal{N}(u)}^L = \text{AGGREGATE}(\{\mathbf{h}_i^{L-1}, \forall i \in \mathcal{N}(u)\})$, $\mathbf{h}_u^L = \sigma(\mathbf{W}^L \cdot \mathbf{h}_{\mathcal{N}(u)}^L)$. Then we calculate the cosine similarity between the predicted embedding $\mathbf{h}_u^L$ and the ground-truth embedding $\mathbf{h}_u$ to optimize the parameters of the GNN model:

$$\Theta^* = \arg\max_{\Theta} \sum_u \cos(\mathbf{h}_u^L, \mathbf{h}_u), \tag{2}$$

where the ground-truth embedding $\mathbf{h}_u$ is learned by any recommender algorithms (e.g., NCF [27], LightGCN [17][1]), Θ is the model parameters. Similarly, we can obtain the item embedding $\mathbf{h}_i^L$ and optimize model parameters by (Eq. (2)).

2.2.2 Enhanced Pre-training Model: PT-GNN

The basic pre-training strategy has two problems. On one hand, it can not explicitly deal with the high-order cold-start neighbors during the graph convolution process. On the other hand, the GNN sampling strategies such as random sampling [29] or importance sampling [55] strategies may fail to sample high-order relevant cold-start neighbors due to their sparse interactions.

To solve the first challenge, we incorporate a meta aggregator to enhance the aggregation ability of each graph convolution step. Specifically, the meta aggregator uses self-attention mechanism [37] to encode the initial embeddings $\{\mathbf{h}_1^0, \cdots, \mathbf{h}_K^0\}$ of the K first-order neighbors for u as input, and outputs the meta embedding $\tilde{\mathbf{h}}_u$ for u. The process is:

$$\begin{aligned} \{\mathbf{h}_1, \cdots, \mathbf{h}_K\} &\leftarrow \text{SELF\_ATTENTION}(\{\mathbf{h}_1^0, \cdots, \mathbf{h}_K^0\}), \\ \tilde{\mathbf{h}}_u &= \text{AVERAGE}(\{\mathbf{h}_1, \cdots, \mathbf{h}_K\}). \end{aligned} \tag{3}$$

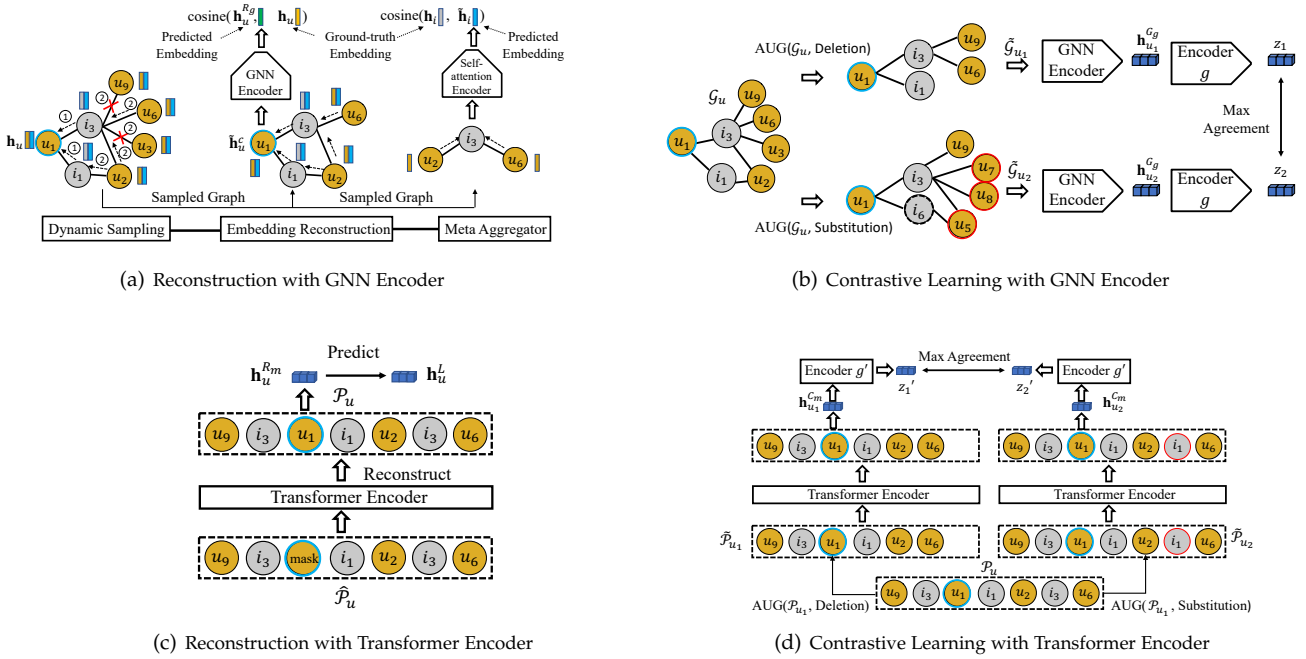
We use the same cosine similarity described in Eq. (2) to measure the similarity between the predicted meta embedding $\tilde{\mathbf{h}}_u$ and the ground truth embedding $\mathbf{h}_u$.

To solve the second challenge, we propose an adaptive neighbor sampler to select proper high-order neighbors according to the feedbacks from the GNN model. The adaptive neighbor sampler is formalized as a hierarchical Markov Sequential Decision Process, which sequentially samples from the low-order neighbors to the high-order neighbors and results in at most K^l neighbors in the l-th layer for each target user u. After sampling the neighbors of the former L layers, the GNN model accepts the sampled neighbors to produce the reward, which denotes whether the sampled neighbors are reasonable or not. Through maximizing the expected reward from the GNN model, the adaptive neighbor sampler can sample proper high-order neighbors. Once the meta aggregator and the adaptive neighbor sampler are trained, the meta embedding $\tilde{\mathbf{h}}_u$ and the averaged sampled neighbor embedding $\tilde{\mathbf{h}}_{\mathcal{N}(u)}^l$ at the l-th layer are added into each graph convolution step in Eq. (1):

$$\mathbf{h}_u^l = \sigma(\mathbf{W}^l \cdot (\tilde{\mathbf{h}}_u \,||\, \mathbf{h}_u^{l-1} \,||\, \tilde{\mathbf{h}}_{\mathcal{N}(u)}^l)). \tag{4}$$

For the target user u, Eq. (4) is repeated $L-1$ steps to obtain the embeddings $\{\mathbf{h}_1^{L-1}, \cdots, \mathbf{h}_K^{L-1}\}$ for its K first-order neighbors, then the predicted embedding $\mathbf{h}_u^L$ is obtained by averaging the first-order embeddings. Finally, Eq. (2) is

1. They are good enough to learn high-quality user/item embeddings from the abundant interactions, and we will discuss it in Section 5.3.1



(a) Reconstruction with GNN Encoder

(b) Contrastive Learning with GNN Encoder

(c) Reconstruction with Transformer Encoder

(d) Contrastive Learning with Transformer Encoder

Fig. 2: Overview of the proposed pretext tasks. These pretext tasks are trained independently.

used to optimize the parameters of the pre-training GNN model. The pre-training parameter set $\Theta = \{\Theta_{gnn}, \Theta_f, \Theta_s\}$, where Θ_{gnn} is the parameters of the vanilla GNN, Θ_f is the parameters of the meta aggregator and Θ_s is the parameters of the neighbor sampler. The item embedding $\mathbf{h}_i^L$ can be obtained in a similar way.

2.2.3 Downstream Recommendation Task

We fine-tune PT-GNN in the downstream recommendation task. Specifically, for each target user u and his neighbors $\{\mathcal{N}^1(u), \cdots, \mathcal{N}^L(u)\}$ of different order, we first use the pre-trained adaptive neighbor sampler to sample proper high-order neighbors $\{\hat{\mathcal{N}}^1(u), \hat{\mathcal{N}}^2(u) \cdots, \hat{\mathcal{N}}^L(u)\}$, and then use the pre-trained meta aggregator to produce the user embedding $\mathbf{h}_u^L$. The item embedding $\mathbf{h}_i^L$ can be obtained in the same way. Then we transform the embeddings to calculate the relevance score between a user and an item, i.e., $y(u,i) = \sigma(\mathbf{W} \cdot \mathbf{h}_u^L)^{\mathrm{T}} \sigma(\mathbf{W} \cdot \mathbf{h}_i^L)$ with parameters $\Theta_r = \{\mathbf{W}\}$. Finally, we adopt the BPR loss [17] to optimize Θ_r and fine-tune Θ:

$$\mathcal{L}_{BPR} = \sum_{(u,i)\in E,(u,j)\notin E} -\ln \sigma(y(u,i) - y(u,j)). \quad (5)$$

However, as shown in Fig. 1, due to the limitation of the GNN model, PT-GNN can only capture the short-range dependencies of users and items; besides, the embedding reconstruction task only focuses on exploring the intra-correlations within the subgraph of the target user or item. Therefore, it is necessary to fully capture both short- and long-range dependencies of users and items, and both intra- and inter-correlations of users or items.

3 The Proposed Model MPT

We propose a novel Multi-strategy based Pre-Training method (MPT), which extends PT-GNN from the perspective of model architecture and pretext tasks. Specifically, in terms of the model architecture, in addition to using the GNN encoder to capture the short-range dependencies of users and items, we introduce a Transformer encoder to capture long-range dependencies. In terms of the pretext task, in addition to considering the intra-correlations of users and items by the embedding reconstruction task, we add embedding contrastive learning task to capture inter-correlations of users and items. Hence, by combining each architecture and each pretext task, there are four different implementations of pretext tasks, as shown in Fig. 2. We detail each pretext task implementation in Section 3.1 - Section 3.4, and then present the overall model training process in Section 3.5. Finally, we analyze the time complexity of the pretext task implementation in Section 3.6.

3.1 Reconstruction with GNN Encoder

In this pretext task, we primarily follow the original PT-GNN model to reconstruct the user embedding, as proposed in Section 2.2. We modify PT-GNN in terms of its neighbor sampling strategy to reduce the time complexity.

The adaptive neighbor sampling strategy in PT-GNN suffers from the slow convergence issue, as it is essentially a Monto-Carlo based policy gradient strategy (REINFORCE) [32], [33], and has the complex action-state trajectories for the entire neighbor sampling process. To solve this problem, inspired by the dynamic sampling theory that uses the current model itself to sample instances [38], [39], we propose the dynamic sampling strategy, which samples from the low-order neighbors to the high-order neighbors according to the enhanced embeddings of the target user

and each neighbor. The enhanced embedding for the target user (or each neighbor) is obtained by concatenating the meta embedding produced by the meta aggregator in PT-GNN (Section 2.2.2) and the current trained embedding. The meta embedding enables considering the cold-start characteristics of the neighbors, and the current trained embedding enables dynamically selecting proper neighbors. Formally, at the l-th layer, the sampling process is:

$$\begin{aligned} a_j^u &= \cos\left(\tilde{\mathbf{h}}_u \,||\, \mathbf{h}_u^l, \tilde{\mathbf{h}}_j \,||\, \mathbf{h}_j^l\right), \\ \{\mathbf{h}_1^l, \cdots, \mathbf{h}_{K^l}^l\} &\leftarrow \text{S\_TOP}(a_1^u, \cdots, a_j^u, \cdots, a_{|\mathcal{N}^l(u)|}^u), \end{aligned} \tag{6}$$

where $||$ is the concatenate operation, a_j^u is the cosine similarity between the enhanced target user embedding $\tilde{\mathbf{h}}_u \,||\, \mathbf{h}_u^l$ and the enhanced j-th neighbor embedding $\tilde{\mathbf{h}}_j \,||\, \mathbf{h}_j^l$, $\tilde{\mathbf{h}}_u$ and $\tilde{\mathbf{h}}_j$ are the meta embeddings produced by the meta aggregator, $\mathbf{h}_u^l$ and $\mathbf{h}_j^l$ are the current trained user embedding, $\{\mathbf{h}_1^l, \cdots, \mathbf{h}_{K^l}^l\}$ is the top K^l relevant neighbors. S_TOP is the operation that selects top neighbor embeddings with top cosine similarity. Compared with the adaptive sampling strategy in PT-GNN, the proposed dynamic sampling strategy not only speeds up the model convergence, but also has competitive performance. As it does not need multiple sampling process, and considers the cold-start characteristics of the neighbors during the sampling process.

Once the neighbors are sampled, at the l-th layer, we use the meta embedding $\tilde{\mathbf{h}}_u$, the previous user embedding $\mathbf{h}_u^{l-1}$ and the averaged dynamically sampled neighbor embedding $\tilde{\mathbf{h}}_{\mathcal{N}(u)}^l$ to perform graph convolution, as shown in Eq. (4). Then we perform graph convolution L steps to obtain the predicted user embedding $\mathbf{h}_u^{R_g}$, and use Eq. (2) to optimize the model parameters. Fig. 2(a) shows this task, and the objective function is as follows:

$$\mathcal{L}_1 : = \arg\max_{\Theta_1} \sum_u \cos(\mathbf{h}_u^{R_g}, \mathbf{h}_u). \tag{7}$$

3.2 Contrastive Learning with GNN Encoder

In this pretext task, we propose to contrast the user embedding produced by the GNN encoder across subgraphs, as shown in Fig. 2(b). similar as PT-GNN, we also train this task under the meta-learning setting to simulate the cold-start scenario. Specifically, we also select users with abundant interactions, randomly select at most $K^l (1 \le l \le L)$ neighbors at layer l to form the subgraph $\mathcal{G}_u$. Then we perform contrastive learning (CL) upon $\mathcal{G}_u$ to learn the representations of users. In the following part, we first present the four components in the CL framework and then detail the key component, i.e., the data augmentation operation.

- An augmentation module $\text{AUG}(\cdot)$ augments the user subgraph $\mathcal{G}_u$ by deleting or replacing the users or items in it, which results in two random augmentations $\tilde{\mathcal{G}}_{u_1}$ = $\text{AUG}(\mathcal{G}_u, seed_1)$ and $\tilde{\mathcal{G}}_{u_2}$ = $\text{AUG}(\mathcal{G}_u, seed_2)$, where $seed_1$ and $seed_2$ are two random seeds. Following SimCLR [35], we evaluate individual or compositional data augmentation operations.
- A GNN encoder performs the graph convolution L steps using Eq. (4) to generate the representation of the target user for each augmented subgraph. i.e., $\mathbf{h}_{u_1}^{C_g} = \text{GNN}(\tilde{\mathcal{G}}_{u_1})$ and $\mathbf{h}_{u_2}^{C_g} = \text{GNN}(\tilde{\mathcal{G}}_{u_2})$.
- A neural network encoder g maps the encoded augmentations $\mathbf{h}_{u_1}^{C_g}$ and $\mathbf{h}_{u_2}^{C_g}$ into two vectors $z_1 = g(\mathbf{h}_{u_1}^{C_g})$, $z_2 = g(\mathbf{h}_{u_2}^{C_g})$. This operation is the same with SimCLR [35], as its observation that adding a nonlinear projection head can significantly improve the representation quality.
- A contrastive learning loss module maximizes the agreement between positive augmentation pair $(\tilde{s_1}, \tilde{s_2})$ in the set $\{\tilde{s}\}$. We construct the set $\{\tilde{s}\}$ by randomly augmenting twice for all users in a mini-batch s (assuming s is with size N), which gets a set $\tilde{s}$ with size $2N$. The two variants from the same original user form the positive pair, while all the other instances from the same mini-batch are regarded as negative samples for them. Then the contrastive loss for a positive pair is defined as:

$$l_c(m,n) = -\log \frac{\exp(\cos(z_m, z_n)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{k \ne m} \exp(\cos(z_m, z_n)/\tau)}, \tag{8}$$

where $\mathbb{1}_{k \ne m}$ is the indicator function to judge whether $k \ne m$, τ is a temperature parameter, and $\cos(z_m, z_n) = {z_m}^{\mathrm{T}} z_n / (||z_m||_2 ||z_n||_2)$ denotes the cosine similarity of two vector z_m and z_n. The overall contrastive loss $\mathcal{L}_2$ defined in a mini-batch is:

$$\mathcal{L}_2 := \min_{\Theta_2} \sum_{m=1}^{2N} \sum_{n=1}^{2N} \mathbb{1}_{m=n}\, l_c(m,n), \tag{9}$$

where $\mathbb{1}_{m=n}$ is a indicator function returns 1 when m and n is a positive pair, returns 0 otherwise, Θ_2 is the parameters of the CL framework.

The key method in CL is the data augmentation strategy. In recommender system, since each neighbor may play an essential role in expressing the user profile, it remains unknown whether data augmentation would benefit the representation learning and what kind of data augmentation could be useful. To answer these questions, we explore and test two basic augmentations, **deletion** and **substitution**. We believe there exist more potential augmentations, which we will leave for future exploration.

- **Deletion.** At each layer l, we random select a% users or items and delete them. If a parent user or item is deleted, its child users or items are all deleted.
- **Substitution.** At each layer l, we randomly select b% users or items. For each user u or item i in the selected list, we randomly replace u or i with one of its parent's interacted first-order neighbors. Note that we keep a and b as the same ratio and leave the tuning of different ratios in the future work.

Once the CL framework is trained, we leave out other parts and only maintain the parameters of the trained GNN encoder. When a new cold-start user u comes in, same as Section 3.1, the GNN encoder performs graph convolution L steps to obtain the enhanced embedding $\mathbf{h}_u^{C_g}$.

3.3 Reconstruction with Transformer Encoder

In this pretext task, we propose using Transformer encoder to reconstruct the embeddings of target users in the user-item path, as shown in Fig. 2(c). This pretext task is also

trained under the meta-learning setting. Specifically, similar to PT-GNN, we first choose abundant users and use any recommender algorithms to learn their ground-truth embeddings. Then for each user, we sample $K^l(1 \leq l \leq L)$ neighbors, based on which, we use random walk [41] to obtain the user-item path $\mathcal{P}$ = IUIU (or $\mathcal{P}$ = UIUI) with path length T. Finally, we mask the target user in the input path with special tokens "[mask]", and use Transformer encoder to predict the target user embedding.

Formally, given an original path $\mathcal{P}$=$[x_1, \cdots, x_T]$, where each node x_i in $\mathcal{P}$ represents the initial user embedding $\mathbf{h}_u^0$ or the initial item embedding $\mathbf{h}_i^0$. We first construct a corrupted path $\hat{\mathcal{P}}$ by replacing the target user token in $\mathcal{P}$ to a special symbol "[mask]" (suppose the target user is at the t-th position in $\hat{\mathcal{P}}$). Then we use Transformer encoder to map the input path $\hat{\mathcal{P}}$ into a sequence of hidden vectors $\mathrm{Tr}(\hat{\mathcal{P}}) = [\mathrm{Tr}(\hat{\mathcal{P}})_1, \cdots, \mathrm{Tr}(\hat{\mathcal{P}})_T]$. Finally, we fetch the t-th embedding in $\mathrm{Tr}(\hat{\mathcal{P}})$ to predict the ground-truth embedding, and use cosine similarity to optimize the parameters Θ_3 of the Transformer encoder. For simplicity, we use notation $\mathbf{h}_u^{R_p}$ to represent the t-th predicted embedding, i.e., $\mathrm{Tr}(\hat{\mathcal{P}})_t$ = $\mathbf{h}_u^{R_p}$. The objective function is:

$$\mathcal{L}_3 : \quad = \quad \arg\max_{\Theta_3} \sum_u \cos(\mathbf{h}_u^{R_p}, \mathbf{h}_u). \tag{10}$$

Once the Transformer encoder is trained, we can use it to predict the cold-start embedding. When a new cold-start user u with his interacted neighbors comes in, we first use random walk to generate the path set $\{\mathcal{P}_1, \cdots, \mathcal{P}_t, \cdots, \mathcal{P}_T\}$, where in the t-th path $\mathcal{P}_t$, u is in the t-th position. Then we replace u with the "[mask]" signal to generate the corrupted path $\hat{\mathcal{P}}_t$. Next we feed all the corrupted paths $\{\hat{\mathcal{P}}_1, \cdots, \hat{\mathcal{P}}_T\}$ into the Transformer encoder, obtain the predicted user embeddings $\{\mathbf{h}_{u_1}^{R_p}, \cdots, \mathbf{h}_{u_T}^{R_p}\}$. Finally, we average these predicted embeddings to obtain the final user embedding $\mathbf{h}_u^{R_p}$.

3.4 Contrastive Learning with Transformer Encoder

In this task, we propose to contrast the user embedding produced by the Transformer encoder across different paths, as shown in Fig. 2(d). We train this task under the meta-learning setting. Same as Section 3.3, we choose abundant users, sample $K^l(1 \leq l \leq L)$ order neighbors, and use random walk to obtain the path $\mathcal{P}_u$. Then we perform the CL framework to learn the cold-start user embedding:

- An augmentation module $\mathrm{AUG}(\cdot)$ augments the path $\mathcal{P}_u$=$[x_1, \cdots, x_T]$ by randomly deleting or replacing the users or items in it, i.e., $\tilde{\mathcal{P}}_{u_1} = \mathrm{AUG}(\mathcal{P}_u, seed_1)$ and $\tilde{\mathcal{P}}_{u_2} = \mathrm{AUG}(T_u, seed_2)$. Similar to Section 3.2, we evaluate individual or compositional data augmentation operations.
- A Transformer encoder accepts $\tilde{\mathcal{P}}_{u_1}$, $\tilde{\mathcal{P}}_{u_2}$ as input, and encodes the target user from two augmented paths into latent vectors, i.e., $\mathbf{h}_{u_1}^{C_p} = \mathrm{Tr}(\tilde{\mathcal{P}}_{u_1})$ and $\mathbf{h}_{u_2}^{C_p} = \mathrm{Tr}(\tilde{\mathcal{P}}_{u_2})$.
- A neural network encoder g' that maps the encoded augmentations $\mathbf{h}_{u_1}^{C_p}$ and $\mathbf{h}_{u_2}^{C_p}$ into two vectors $z_1' = g'(\mathbf{h}_{u_1}^{C_p})$, $z_2' = g'(\mathbf{h}_{u_2}^{C_p})$.
- A contrastive learning loss module maximizes the agreement between positive augmentation pair $(\tilde{s_1}, \tilde{s_2})$ in the set $\{\tilde{s}\}$. Same as Section 3.2, we also construct the set $\{\tilde{s}\}$ by randomly augmenting twice for all users in a mini-batch s to get a set $\tilde{s}$ with size $2N$, use the two variants from the same original user as positive pair, use all the other instances from the same mini-batch as negative samples, and use Eq. (8) as the contrastive loss $l_c(m, n)$ for a positive pair. Similar to Eq. (9), the overall contrastive loss $\mathcal{L}_4$ defined in a mini-batch is:

$$\mathcal{L}_4 := \min_{\Theta_4} \sum_{m=1}^{2N} \sum_{n=1}^{2N} \mathbb{1}_{m=n}\, l_c(m, n), \tag{11}$$

where Θ_4 is the parameters of the CL framework.

The data augmentation strategies are as follows:

- **Deletion.** For each path $\mathcal{P}_u$, we random select a% users and items and delete them.
- **Substitution.** For each path $\mathcal{P}_u$, we randomly select b% users and items. For each user u or item i in the selected list, we randomly replace u or i with one of its parent's interacted first-order neighbors.

Once the CL framework is trained, we leave out other parts and only maintain the parameters of the Transformer encoder. When a new cold-start user u with his interacted neighbors comes in, same as Section 3.3, we generate the path set $\{\mathcal{P}_1, \cdots, \mathcal{P}_T\}$, use Transformer encoder to obtain the encoded embeddings $\{\mathbf{h}_{u_1}^{C_p}, \cdots, \mathbf{h}_{u_T}^{C_p}\}$, and average these embeddings to obtain the final embedding $\mathbf{h}_u^{C_p}$.

3.5 Model Pre-training & Fine-tuning Process

We adopt the pre-training and fine-tuning paradigm [20] to train the GNN and Transformer encoders.

During the pre-training stage, we independently train each pretext task using the objective functions Eq. (7), Eq. (9), Eq. (10) and Eq. (11) to optimize the parameters $\{\Theta_1, \Theta_2, \Theta_3, \Theta_4\}$. We assign each pretext task an independent set of initialized user and item embeddings, and do not share embeddings for these pretext tasks. Therefore, we can train these pretext tasks in a fully parallel way.

During the fine-tuning process, we initialized the GNN and Transformer encoders with the trained parameters, and fine-tune them via the downstream recommendation task. Specifically, for each target cold-start user and his interacted neighbors $\{\mathcal{N}^1(u), \cdots, \mathcal{N}^L(u)\}$ of each order, we first use the trained GNN and Transformer encoders corresponding to each pretext task to generate the user embedding $\mathbf{h}_u^{R_g}$, $\mathbf{h}_u^{C_g}$, $\mathbf{h}_u^{R_p}$ and $\mathbf{h}_u^{C_p}$. Then we concatenate the generated embeddings and transform them into the final user embedding:

$$\mathbf{h}_u^* = \mathbf{W} \cdot (\mathbf{h}_u^{R_g} || \mathbf{h}_u^{C_g} || \mathbf{h}_u^{R_p} || \mathbf{h}_u^{C_p}), \tag{12}$$

where $\Theta_r = \{\mathbf{W}\}$ is the parameter matrix. We generate the final item embedding $\mathbf{h}_i^*$ in a similar way. Next we calculate the relevance score as the inner product of user and item final embeddings, i.e., $y(u, i) = \mathbf{h}_u^{*\mathrm{T}} \mathbf{h}_i^*$. Finally, we use BPR loss defined in Eq. (5) to optimize Θ_r and fine-tune $\{\Theta_1, \Theta_2, \Theta_3, \Theta_4\}$.

TABLE 1: Statistics of the Datasets.

Dataset	#Users	#Items	#Interactions	#Sparse Ratio
MovieLens-1M	6,040	3,706	1,000,209	4.47%
MOOCs	82,535	1,302	458,453	0.42%
Gowalla	29,858	40,981	1,027,370	0.08%

3.6 Discussions

As we can see, the GNN and Transformer encoders are the main components of the pretext tasks. For the GNN encoder, the time complexity is $O(T_{GNN} + T_S)$, where $O(T_{GNN})$ represents the time complexity of the layer-wise propagation of the GNN model, and $O(T_S)$ represents the time complexity of the sampling strategy. Since different GNN model has different time complexity $O(T_{GNN})$, we select classic GNN models and show their $O(T_{GNN})$. LightGCN [17]: $O(|\mathcal{R}^+| + T_S)$, NGCF [16]: $O(|\mathcal{R}^+|\ d^2 + T_S)$, GCMC [49]:$O(|\mathcal{R}^+|\ d^2 + T_S)$, GraphSAGE [29]: $O(|\mathcal{N}^+| + T_S)$, where $|\mathcal{R}^+|$ denotes the number of nonzero entries in the Laplacian matrix, $|\mathcal{N}^+|$ is the number of totally sampled instances and d is the embedding size. We present the time complexity of dynamic sampling strategy $O(T_S) = O(E * d * |\mathcal{N}^+|)$ and the adaptive sampling strategy $O(T_S) = O(E * M * d * |\mathcal{N}^+|)$, where E is the number of convergent epochs, M is the sampling times. Compared with adaptive sampling strategy, the dynamic strategy does not need multiple sampling time M, and has fewer convergent epochs E, thus has smaller time complexity. For the Transformer encoder, the time complexity is $O(T^2 * d)$, where T is the length of the user-item path.

4 Experiments

Following PT-GNN [30], we conduct intrinsic evaluation task to evaluate the quality of user/item embeddings, and extrinsic task to evaluate the cold-start recommendation performance. We answer the following research questions:

- **RQ1:** How does MPT perform embedding inference and cold-start recommendation compared with the state-of-the-art GNN and pre-training GNN models?
- **RQ2:** What are the benefits of performing pretext tasks in both intrinsic and extrinsic evaluation?
- **RQ3:** How do different settings influence the effectiveness of the proposed MPT model?

4.1 Experimental Settings

4.1.1 Datasets

We evaluate on three public datasets MovieLens-1M (Ml-1M)[2] [42], MOOCs[3] [43] and Gowalla[4] [45]. Table 1 illustrates the statistics of these datasets.

4.1.2 Comparison Methods

We select three types of baselines, including the state-of-the-art neural matrix factorization model, the GNN models and the self-supervised graph learning models.

2. https://grouplens.org/datasets/movielens/
3. http://moocdata.cn/data/course-recommendation
4. https://snap.stanford.edu/data/loc-gowalla.html

- **NCF [27]**: is a neural matrix factorization model which combines Multi-layer Perceptron and matrix factorization to learn the embeddings of users and items.
- **PinSage [15]**: employs the GraphSAGE [29] model, which samples neighbors randomly and aggregates them by the AVERAGE function.
- **GCMC [49]**: employs the standard GCN [28] model to learn the embeddings of users and items.
- **NGCF [16]**: primarily follows the neural passing based GNN model [56], but additionally adds second-order interactions during the message passing process.
- **LightGCN [17]**: discards the feature transformation and nonlinear activation functions in NGCF.
- **SGL [58]**: contrasts the node representation within the graphs from multiple views, where node dropout, edge dropout and random walk are adopted to generate these views. We find edge dropout has the best performance.
- **PT-GNN [30]**: takes the embedding reconstruction as the pretext task to explicitly improve the embedding quality.

For each vanilla GNN model (e.g., GCMC, NGCF), notation GNN^+ means we apply GNN in PT-GNN, and notation GNN^* means we apply GNN in MPT. Since SGL adopts multi-task learning paradigm for recommendation, for fair comparison, we compare it in the extrinsic recommendation task and use notation GNN-SGL to denote it.

4.1.3 Intrinsic and Extrinsic Settings.

We divide each dataset into the meta-training set D_T and meta-test set D_N. We train and evaluate the proposed MPT model in the intrinsic user/item embedding inference task on D_T. Once the model is trained, we fine-tune it in the extrinsic recommendation task and evaluate it on D_N. We select the users/items from each dataset with sufficient interactions as the target users/items in D_T, as the intrinsic evaluation needs the true embeddings of users/items inferred from the sufficient interactions. In the cold-start user scenario, we divide the users with the number of the direct interacted items more than n_i into D_T and leave the rest users into D_N. We set n_i as 25, 10 and 25 for the dataset Ml-1M, MOOCs and Gowalla, respectively. Similarly, in the cold-start item scenario, we divide the items with the number of the direct interacted users more than n_u into D_T and leave the rest items into D_N, where we set n_u as 15, 10 and 15 for Ml-1M, MOOCs and Gowalla, respectively. We set K as 3 and 8 in the intrinsic task, and 8 in the extrinsic task. By default, we set d as 32, the learning rate as 0.003, Kas 3, L as 4, T as 6, a as 0.2, b as 0.2 and τ as 0.2.

4.1.4 Intrinsic Evaluations: Embedding Inference

We conduct the intrinsic evaluation task, which aims to infer the embeddings of cold-start users and items by the proposed MPT model. Both the evaluations on user embedding inference and item embedding inference are performed.

Training and Test Settings. We perform intrinsic evaluation on the meta-training set D_T. Specifically, same as PT-GNN, we train NCF [27] to get the ground-truth embeddings for the target users/items in D_T. We also explore whether MPT is sensitive to the ground-truth embedding generated by other models such as LightGCN [17]. We randomly split D_T into the training set $Train_T$ and the test set $Test_T$ with

TABLE 2: Overall performance of user/item embedding inference with Spearman correlation. L=4.

Methods	Ml-1M (user)		MOOCs (user)		Gowalla(user)		Ml-1M (item)		MOOCs (item)		Gowalla(item)	
	3-shot	8-shot	3-shot	8-shot	3-shot	8-shot	3-shot	8-shot	3-shot	8-shot	3-shot	8-shot
NCF	-0.019	0.066	-0.098	-0.062	0.042	0.117	-0.128	-0.007	-0.036	0.027	-0.036	-0.018
PinSage	0.036	0.115	0.085	0.128	0.104	0.134	0.117	0.146	0.116	0.182	0.112	0.198
PinSage$^+$	0.318	0.378	0.302	0.338	0.326	0.384	0.468	0.490	0.316	0.336	0.336	0.353
PinSage*	**0.381**	**0.396**	**0.336**	**0.346**	**0.375**	**0.391**	**0.482**	**0.488**	**0.323**	**0.339**	**0.346**	**0.367**
GCMC	0.026	0.069	0.092	0.138	0.092	0.125	0.118	0.133	0.108	0.118	0.106	0.114
GCMC$^+$	0.201	0.251	0.501	0.531	0.432	0.468	0.451	0.467	0.214	0.231	0.301	0.334
GCMC*	**0.258**	**0.264**	**0.566**	**0.582**	**0.482**	**0.497**	**0.488**	**0.529**	**0.217**	**0.233**	**0.351**	**0.364**
NGCF	0.069	0.109	0.063	0.095	0.082	0.114	0.022	0.033	0.007	0.018	0.007	0.013
NGCF$^+$	0.182	0.236	0.083	0.146	0.104	0.149	0.168	0.193	0.099	0.121	0.153	0.182
NGCF*	**0.291**	**0.301**	**0.197**	**0.203**	**0.165**	**0.179**	**0.219**	**0.227**	**0.213**	**0.223**	**0.181**	**0.188**
LightGCN	0.099	0.138	0.060	0.068	0.162	0.184	0.201	0.262	0.181	0.232	0.213	0.245
LightGCN$^+$	0.272	0.288	0.292	0.309	0.229	0.234	0.382	0.403	0.334	0.353	0.386	0.403
LightGCN*	**0.283**	**0.294**	**0.304**	**0.315**	**0.391**	**0.401**	**0.398**	**0.409**	**0.338**	**0.383**	**0.536**	**0.548**

a ratio of 7:3. To mimic the cold-start users/items on $Test_T$, we randomly keep K neighbors for each user/item, which results in at most K^l neighbors ($1 \leq l \leq L$) for each target user/item. Thus $Test_T$ is changed into $Test'_T$.

We train NCF transductively on the merged dataset $Train_T$ and $Test'_T$. We train the vanilla GNN models by BPR loss [17] on $Train_T$. We train PT-GNN on $Train_T$, where we first perform Eq. (4) L-1 steps to obtain the refined first-order neighbors, and then average them to reconstruct the target embedding; finally we use Eq. (2) to measure the quality of the predicted embedding. We train MPT on $Train_T$, where we first perform four pretext tasks by Eq. (7), Eq. (9), Eq. (10) and Eq. (11), and then use Eq. (12) to fuse the generated embeddings; finally we use Eq. (2) to measure the quality of the fused embedding. Note that the embeddings in all the models are randomly initialized. Following [44], [57], we use Spearman correlation to measure the quality of learned embeddings, as Lazaridou et al. [57] pointed out Spearman correlation can measure the agreement between the ground-truth human annotations and the machine-generated ones.[5]

4.1.5 Extrinsic Evaluation: Recommendation

We apply the pre-training GNN model into the downstream recommendation task and evaluate its performance.

Training and Testing Settings. We consider the scenario of the cold-start users and use the meta-test set D_N to perform recommendation. For each user in D_N, we select top c% (c=0.2 by default) of his interacted items in chronological order into the training set $Train_N$, and leave the rest items into the test set $Test_N$. We pre-train our model on D_T and fine-tune it on $Train_N$ according to Section 3.5.

The vanilla GNN and the NCF models are trained by the BPR loss on D_T and $Train_N$. For each user in $Test_N$, we calculate his relevance score to each of the rest 1-c% items. We adopt Recall@$\mathcal{K}$ and NDCG@$\mathcal{K}$ as the metrics to evaluate the items ranked by the relevance scores. By default, we set $\mathcal{K}$ as 20 for Ml-1M, MOOCs and Gowalla.

5. We also use cosine similarity to measure the agreement between the ground-truth and predicted embeddings, and find the results have the same trend.

TABLE 3: Overall recommendation performance with sparse rate c=20%, L=4.

Methods	Ml-1M		MOOCs		Gowalla	
	Recall	NDCG	Recall	NDCG	Recall	NDCG
NCF	0.004	0.009	0.132	0.087	0.039	0.013
PinSage	0.006	0.012	0.085	0.066	0.003	0.011
PinSage-SGL	0.040	0.033	0.172	0.086	0.011	0.021
PinSage$^+$	0.047	**0.038**	0.150	0.089	0.008	0.019
PinSage*	**0.054**	0.036	**0.182**	**0.099**	**0.045**	**0.021**
GCMC	0.008	0.006	0.232	0.172	0.006	0.033
GCMC-SGL	0.038	0.037	0.239	0.177	0.033	0.035
GCMC$^+$	0.044	0.041	0.248	0.187	0.018	0.032
GCMC*	**0.071**	**0.076**	**0.272**	**0.212**	**0.065**	**0.043**
NGCF	0.052	0.058	0.208	0.171	0.009	0.013
NGCF-SGL	0.062	**0.079**	0.216	0.177	0.028	0.014
NGCF$^+$	0.064	0.071	0.217	0.181	0.037	0.014
NGCF*	**0.083**	0.077	**0.241**	**0.208**	**0.047**	**0.016**
LightGCN	0.058	0.064	0.217	0.169	0.011	0.023
LightGCN-SGL	0.066	0.077	0.275	0.178	0.034	0.028
LightGCN$^+$	0.078	0.071	0.308	0.184	0.049	0.031
LightGCN*	**0.094**	**0.101**	**0.346**	**0.223**	**0.051**	**0.036**

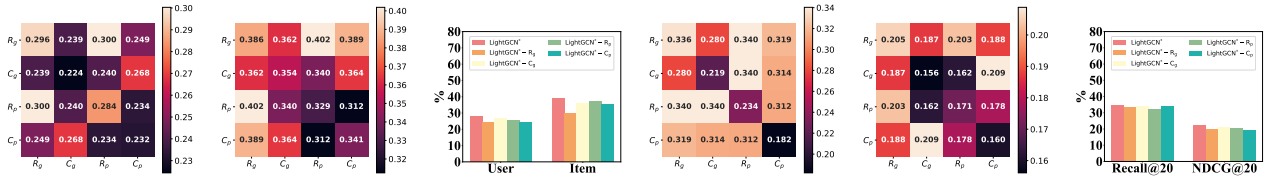
5 Experimental Results

5.1 Performance Comparison (RQ1)

5.1.1 Overall Performance Comparison

We report the overall performance of intrinsic embedding inference and extrinsic recommendation tasks in Table 2 and Table 3. We find that:

- MPT (denoted as GNN*) is better than NCF and the vanilla GNN model, which indicates the effectiveness of the pre-training strategy. Compared with the pre-training model PT-GNN (denoted as GNN$^+$) and SGL (denoted as GNN-SGL), MPT also performs better than them. This indicates the superiority of simultaneously considering the intra- and inter- correlations of nodes as well as the short- and long-range dependencies of nodes.
- In Table 2, when K decreases from 8 to 3, the performance of all the baseline methods drops by a large scale, while MPT drops a little. This indicates MPT is able to learn high-quality embeddings for users or items that have extremely sparse interactions.



(a) Spearman (user) % (b) Spearman (item) % (c) Spearman (d) Recall@20 (e) NDCG@20 (f) Recall & NDCG

Fig. 3: Intrinsic and extrinsic task evaluation under individual or composition of pretext tasks. K=3, L=4, T=6, c=20%. We evaluate the variant models on MOOCs (results on Ml-1M and Gowalla show the same trend which are omitted for space).

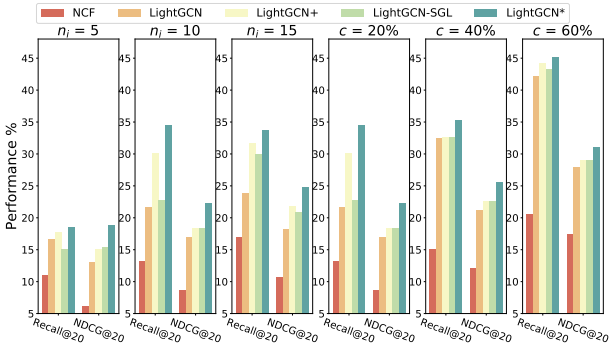
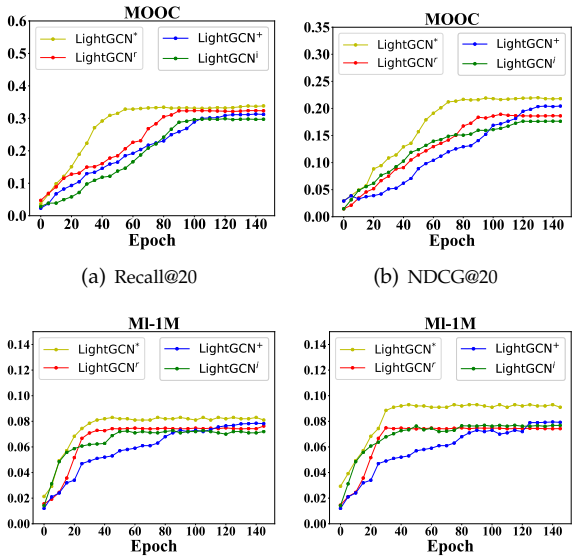


Fig. 4: Cold-start recommendation under different n_i and c.



(a) Recall@20 (b) NDCG@20 (c) Recall@20 (d) NDCG@20

Fig. 5: Comparison among different sampling strategies.

5.1.2 *Interacted Number and Sparse Rate Analysis*

It is still unclear how does MPT handle the cold-start users with different interacted items n_i and sparse rate c. To this end, we change the interaction number n_i in the range of $\{5, 10, 15\}$ and the sparse rate c in the range of $\{20\%, 40\%, 60\%\}$, select LightGCN as the backbone GNN model and report the cold-start recommendation performance on MOOCs dataset in Fig. 4. The smaller n_i and c are, the cold-start users in D_N have fewer interactions. We find that:

- MPT (denoted as LightGCN*) is consistently superior to all the other baselines, which justifies the superiority of MPT in handling cold-start recommendation with different n_i and c.
- When n_i decreases from 15 to 5, MPT has a larger improvement compared with other baselines, which verifies its capability to solve the cold-start users with extremely sparse interactions.
- When c decreases from 60% to 20%, MPT has a larger improvement compared with other baselines, which again verifies the superiority of MPT in handling cold-start users with different sparse rate.

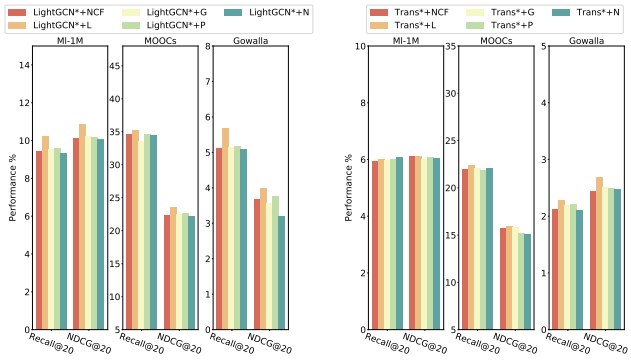
5.2 Ablation Study (RQ2)

5.2.1 *Impact of Pretext Tasks*

It is still not clear which part of the pretext tasks is responsible for the good performance in MPT. To answer this question, we apply LightGCN in MPT, perform individual or compositional pretext tasks, report the performance of intrinsic and extrinsic task in Fig. 3, where notation R_g, C_g, R_p and C_p denote the pretext task of reconstruction with GNN, contrastive learning with GNN, reconstruction with Transformer and contrastive learning with Transformer, respectively; notation LightGCN*-R_g means we only discard reconstruction task and use the other three pretext tasks. Other variant models are named in a similar way. Aligning Table 2 with Fig. 3, we find that: (1) Combining all the pretext tasks can benefit both embedding quality and recommendation performance. This indicates simultaneously consider intra- and inter-correlations of users and items can benefit cold-start representation learning and recommendation. (2) Performing contrastive learning with only GNN, Transformer encoder or their combinations performs not good in the intrinsic task, but has satisfactory performance in the recommendation task. The reason is that contrastive learning does not focus on predicting the target embedding, but can capture the inter-correlations of users or items, and thus can benefit the recommendation task.

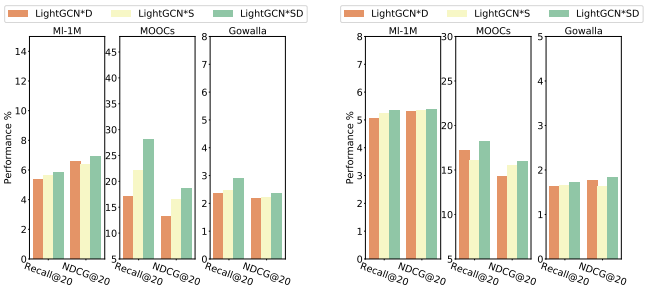
5.2.2 *Impact of the Sampling Strategy*

We study the effect of the proposed dynamic sampling strategy. For fair comparison, we compare four variants of the LightGCN model: LightGCN$^+$ that adaptively samples neighbors [30], LightGCN^i that samples neighbors according to the importance sampling strategy [55], LightGCN^r that randomly samples neighbors [29] and LightGCN* that dynamically samples neighbors. We report the recommen-



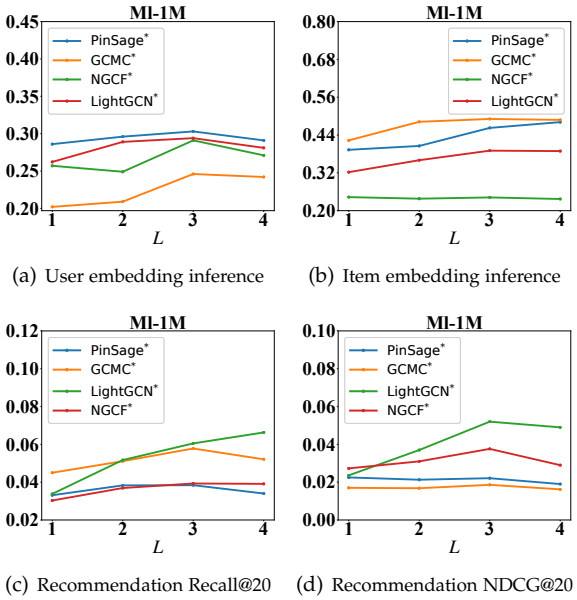
(a) LightGCN (b) Transformer Encoder

Fig. 6: Sensitive analysis of ground-truth embeddings.



(a) LightGCN (b) Transformer Encoder

Fig. 7: Recommendation performance under individual or compositional data augmentations. c= 20%, L=4.



(a) User embedding inference (b) Item embedding inference

(c) Recommendation Recall@20 (d) Recommendation NDCG@20

Fig. 8: Embedding inference and recommendation performance under different layer L.

dation performance on MOOCs and Ml-1M in Fig. 5. We find that: (1) Although the adaptive sampling strategy has competitive performance than the random and important sampling strategies, it has slow convergence speed due to the complex RL-based sampling process. (2) Compared with the other sampling strategies, the proposed dynamic sampling strategy not only has the best performance, but also has quick convergence speed.

5.3 Study of MPT (RQ3)

5.3.1 Effect of Ground-truth Embedding

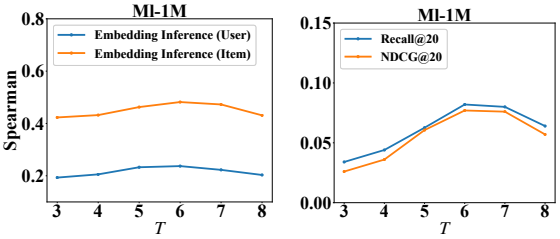
As mentioned in Section 2, when performing embedding reconstruction with GNN and Transformer encoders, we choose NCF to learn the ground-truth embeddings. However, one may consider whether MPT's performance is sensitive to the ground-truth embedding. To this end, we use the baseline GNN models to learn the ground-truth embeddings, and only perform embedding reconstruction task with LightGCN or Transformer. For NCF, we concatenate embeddings produced by both the MLP and GMF modules as ground-truth embeddings. While for PinSage, GCMC, NGCF and LightGCN, we combine the embeddings obtained at each layer to form the ground-truth matrix, i.e., $\mathbf{E} = \mathbf{E}^{(0)} + \cdots + \mathbf{E}^{(L)}$, where $\mathbf{E}^l \in \mathcal{R}^{(|\mathcal{U}|+|\mathcal{I}|)\times d}$ is the concatenated user-item embedding matrix at l-th convolution step. Fig. 6(a) and Fig. 6(b) shows the recommendation performance using LightGCN and Transformer encoder, respectively. Suffix +NCF, +L, +G, +P and +N denote that the ground-truth embeddings are obtained by NCF, LightGCN, GCMC, PinSage and NGCF, respectively. We find that: (1) All the models that equipped with different ground-truth embeddings achieve almost the same performance. This indicates our model is not sensitive to the ground-truth embeddings, as the NCF and vanilla GNN models are good enough to learn high-quality user or item embeddings from the abundant interactions. (2) Besides, when LightGCN is used to obtain the ground-truth embeddings (denoted as LightGCN*+L and Trans*+L), we can obtain marginal performance gain compared with other variant models.

5.3.2 Effect of Data Augmentation

To understand the effects of individual or compositional data augmentations in the contrastive learning (CL) task using GNNs or Transformer encoder, we investigate the performance of CL pretext tasks in MPT when applying augmentations individually or in pairs. We report the performance in Fig. 7. Notation LightGCN*D, LightGCN*S and LightGCN*SD mean we apply LightGCN into the CL task and use deletion, substitution and their combinations, respectively; Notation Trans*D, Trans*S and Trans*SD denote we use Transformer encoder into the CL task and use deletion, substitution and their combinations, respectively. We find that: (1) Composing augmentations can lead to better recommendation performance than performing individual data augmentation. (2) Aligning Fig. 7 and Table 3, we find combining substitution and deletion using GNNs has competitive performance than SGL. The reason is that, same as SGL, our proposed contrastive data augmentation can also change the graph structure, and thus inter-correlations of nodes can be captured.

5.3.3 Effect of Hyperparameters

We move on to study different designs of the layer depth L and user-item path length T in MPT. Due to the space



(a) User embedding inference (b) Item embedding inference

Fig. 9: Embedding inference and recommendation performance under different path length T.

limitation, we omit the results on MOOCs and Gowalla which have a similar trend to Ml-1M.

- We only perform embedding reconstruction with GNNs in MPT, and report the results in Fig. 8. We find that the performance first increases and then drops when increasing L from 1 to 4. The peak point is 3 at most cases. This indicates GNN can only capture short-range dependencies, which is consistent with LightGCN's [17] finding.
- We only perform embedding reconstruction with Transformer encoder in MPT, and report the results in Fig. 9. We find that the performance first improves when the path length T increases from 3 to 6, and then drops from 6 to 8. This indicates Transformer encoder is not good at capturing short-range dependencies of nodes, but can capture long-range dependencies.

6 Related Work

Pre-training GNNs. Recent advances on pre-training GNNs aim to empower GNNs to capture the correlations of nodes in an input graph, so that it can easily generalize to any downstream tasks with a few fine-tuning steps on the graphs. Examples include GPT-GNN [19], GCC [20], DGI [21], InfoGraph [22] and GraphCL [54]. More recently, some researchers explore pre-training GNNs on user-item graphs for recommendation. For example, PT-GNN [30] reconstructs embeddings under the meta-learning setting. SGL [51] contrasts node representation by node dropout, edge dropout and random walk data augmentation operations. PMGT [53] reconstructs graph structure and node feature using side information. GCN-P/COM-P [52] learns the representations of entities constructed from the side information. However, these methods suffer from ineffective long-range dependency problem, as the GNN model can only capture low-order correlations of nodes. Besides, the pretext tasks of these methods consider either intra-correlations [30], [52], [53] or inter-correlations [51] of nodes, rather than both. Further, the side information is not always available, making it difficult to construct the pretext tasks. To solve the above problems, we propose MPT, which considers both short- and long-range dependencies of nodes, and both intra- and inter-correlations of nodes.

Cold-start Recommendation. Cold-start issue is a fundamental challenge in recommender systems. On one hand, existing recommender systems incorporate the side information such as spatial information [25], [26], social trust path [1], [2], [3], [18] and knowledge graphs [4], [5] to enhance the representations of the cold-start users/items. However, the side information is not always available, making it intractable to improve the cold-start embedding's quality. On the other hand, researchers solve the cold-start issue by only mining the underlying patterns behind the user-item interactions. One kind of the method is meta-learning [6], [7], [8], [9], which consists of metric-based recommendation [10] and model-based recommendation [11], [12], [13], [14]. However, few of them capture the high-order interactions. Another kind of method is GNN-based recommendation, which incorporates signals of high-order neighbors to refine the user/item embedding. The representative models include PinSage [15], NGCF [16], LightGCN [17] and CAGR [18]. However, these GNN-based methods address the cold-start embeddings through optimizing the likelihood of a user adopting an item, which isn't a direct improvement of the embedding quality.

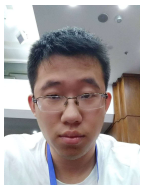
7 Conclusion

We propose a multi-strategy based pre-training method, MPT, which extends PT-GNN from the perspective of model architecture and pretext tasks to improve the cold-start recommendation performance. Specifically, in addition to the short-range dependencies of nodes captured by the GNN encoder, we add a transformer encoder to capture long-range dependencies. In addition to considering the intra-correlations of nodes by the embedding reconstruction task, we add embedding contrastive learning task to capture inter-correlations of nodes. Experiments on three datasets demonstrate the effectiveness of our proposed MPT model against the vanilla GNN and pre-training GNN models.

References

[1] Yin, H., Wang, Q., Zheng, K., Li, Z., Yang, J., & Zhou, X. Social influence-based group representation learning for group recommendation. In ICDE'19 (pp. 566-577).
[2] Wang, Q., Yin, H., Wang, H., Nguyen, Q. V. H., Huang, Z., & Cui, L. Enhancing collaborative filtering with generative augmentation. In SIGKDD'19 (pp. 548-556).
[3] Gharibshah, Z., Zhu, X., Hainline, A., & Conway, M. Deep learning for user interest and response prediction in online display advertising. Data Science and Engineering, 5(1), 12-26.
[4] Wang, H., Zhang, F., Zhao, M., Li, W., Xie, X., & Guo, M. Multi-task feature learning for knowledge graph enhanced recommendation. In WWW'19 (pp. 2000-2010).
[5] Wang, X., He, X., Cao, Y., Liu, M., & Chua, T. S. Kgat: Knowledge graph attention network for recommendation. In SIGKDD'19.
[6] Finn, C., Abbeel, P., & Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. ICML'17.
[7] Munkhdalai, T., & Yu, H. Meta networks. In ICML'17.
[8] Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., & Wierstra, D. Matching networks for one shot learning. In NeurIPS'16.
[9] Snell, J., Swersky, K., & Zemel, R. S. Prototypical networks for few-shot learning. In NeurIPS'17.
[10] Vartak, M., Thiagarajan, A., Miranda, C., Bratman, J., & Larochelle, H. A meta-learning perspective on cold-start recommendations for items. In NeurIPS'17.
[11] Du, Z., Wang, X., Yang, H., Zhou, J., & Tang, J. Sequential scenario-specific meta learner for online recommendation. In SIGKDD'19.
[12] Lee, H., Im, J., Jang, S., Cho, H., & Chung, S. MeLU: meta-learned user preference estimator for cold-start recommendation. In SIGKDD'19 (pp. 1073-1082).

[13] Lu, Y., Fang, Y., & Shi, C. Meta-learning on heterogeneous information networks for cold-start recommendation. In Proceedings of the 26th ACM SIGKDD'20 (pp. 1563-1573).
[14] Pan, F., Li, S., Ao, X., Tang, P., & He, Q. Warm up cold-start advertisements: Improving ctr predictions via learning to learn id embeddings. In SIGIR'19 (pp. 695-704).
[15] Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W. L., & Leskovec, J. Graph convolutional neural networks for web-scale recommender systems. In SIGKDD'18 (pp. 974-983).
[16] Wang, X., He, X., Wang, M., Feng, F., & Chua, T. S. Neural graph collaborative filtering. In SIGIR'19 (pp. 165-174).
[17] He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., & Wang, M. Lightgcn: Simplifying and powering graph convolution network for recommendation. In SIGIR'20 (pp. 639-648).
[18] Yin, H., Wang, Q., Zheng, K., Li, Z., & Zhou, X. Overcoming Data Sparsity in Group Recommendation. In TKDE'20.
[19] Hu, Z., Dong, Y., Wang, K., Chang, K. W., & Sun, Y. Gpt-gnn: Generative pre-training of graph neural networks. In SIGKDD'20.
[20] Qiu, J., Chen, Q., Dong, Y., Zhang, J., Yang, H., Ding, M., ... & Tang, J. Gcc: Graph contrastive coding for graph neural network pre-training. In M SIGKDD'20 (pp. 1150-1160).
[21] Velickovic, P., Fedus, W., Hamilton, W. L., Liò, P., Bengio, Y., & Hjelm, R. D. Deep Graph Infomax. In ICLR'19 (Poster).
[22] Sun, F. Y., Hoffmann, J., Verma, V., & Tang, J. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In ICLR'20.
[23] You, Y., Chen, T., Wang, Z., & Shen, Y. When does self-supervision help graph convolutional networks?. In ICML'20.
[24] Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., & Leskovec, J. (2019). Strategies for pre-training graph neural networks. In ICLR'20.
[25] Yin, H., Wang, W., Wang, H., Chen, L., & Zhou, X. Spatial-aware hierarchical collaborative deep learning for POI recommendation. In TKDE'17, 29(11), 2537-2551.
[26] Yin, H., Cui, B., Sun, Y., Hu, Z., & Chen, L. LCARS: A spatial item recommender system. ACM In TOIS'14, 32(3), 1-37.
[27] He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T. S. Neural collaborative filtering. In WWW'17 (pp. 173-182).
[28] Kipf, T. N., & Welling, M. Semi-supervised classification with graph convolutional networks. In ICLR'17.
[29] Hamilton, W. L., Ying, R., & Leskovec, J. Inductive representation learning on large graphs. In NeurIPS'17.
[30] Hao, B., Zhang, J., Yin, H., Li, C., & Chen, H. Pre-Training Graph Neural Networks for Cold-Start Users and Items Representation. In WSDM'21.
[31] Liu, X., Zhang, F., Hou, Z., Wang, Z., Mian, L., Zhang, J., & Tang, J. Self-supervised learning: Generative or contrastive. In TKDE'21.
[32] Sutton, R. S., McAllester, D. A., Singh, S. P., & Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In NeurIPS'99 (Vol. 99, pp. 1057-1063).
[33] Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning, 8(3-4), 229-256.
[34] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT'19
[35] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. A simple framework for contrastive learning of visual representations. ICML'20.
[36] Linden, G., Smith, B., & York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. IEEE Internet computing, 7(1), 76-80.
[37] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, A. N., ... & Polosukhin, I. Attention is all you need. In NeurIPS'17.
[38] Burges, C. J. (2010). From ranknet to lambdarank to lambdamart: An overview. Learning, 11(23-581), 81.
[39] Zhang, W., Chen, T., Wang, J., & Yu, Y. Optimizing top-n collaborative filtering via dynamic negative item sampling. In SIGIR'13.
[40] Sun, Y., Han, J., Yan, X., Yu, P. S., & Wu, T. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. In VLDB'11.
[41] Perozzi, B., Al-Rfou, R., & Skiena, S. Deepwalk: Online learning of social representations. In SIGKDD'14. (pp. 701-710).
[42] Harper, F. M., & Konstan, J. A. The movielens datasets: History and context. In TIIS'16. 5(4), 1-19.
[43] Zhang, J., Hao, B., Chen, B., Li, C., Chen, H., & Sun, J. Hierarchical reinforcement learning for course recommendation in MOOCs. In AAAI'19 (Vol. 33, No. 01, pp. 435-442).
[44] Hu, Z., Chen, T., Chang, K. W., & Sun, Y. Few-shot representation learning for out-of-vocabulary words. In ACL'19.
[45] Liang, D., Charlin, L., McInerney, J., & Blei, D. M. Modeling user exposure in recommendation. In WWW'16.
[46] Chen, T., & Wong, R. C. W.Handling information loss of graph neural networks for session-based recommendation. In SIGKDD'20.
[47] Wang, X., Liu, N., & Shi, C. Self-supervised Heterogeneous Graph Neural Network with Co-contrastive Learning. In SIGKDD'21.
[48] Wu, J., Wang, X., Feng, F., He, X., Chen, L., Lian, J., & Xie, X. Self-supervised Graph Learning for Recommendation. In SIGIR'21.
[49] Berg, R. V. D., Kipf, T. N., & Welling, M. Graph convolutional matrix completion. In SIGKDD'18.
[50] Wu, Z., Xiong, Y., Yu, S. X., & Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In CVPR'18.
[51] Wu, J., Wang, X., Feng, F., He, X., Chen, L., Lian, J., & Xie, X. Self-supervised Graph Learning for Recommendation. In SIGIR'21.
[52] Meng, Z., Liu, S., Macdonald, C., & Ounis, I. Graph Neural Pre-training for Enhancing Recommendations using Side Information. arXiv preprint arXiv:2107.03936.
[53] Liu, Y., Yang, S., Lei, C., Wang, G., Tang, H., Zhang, J., ... & Miao, C. Pre-training Graph Transformer with Multimodal Side Information for Recommendation. arXiv preprint arXiv:2010.12284.
[54] You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., & Shen, Y. Graph contrastive learning with augmentations. In NeurIPS'20.
[55] Chen, J., Ma, T., & Xiao, C. Fastgcn: fast learning with graph convolutional networks via importance sampling. In ICLR'18.
[56] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. Neural message passing for quantum chemistry. In PMLR'17.
[57] Lazaridou, A., Marelli, M., & Baroni, M. (2017). Multimodal word meaning induction from minimal exposure to natural text. Cognitive science, 41, 677-705.
[58] Wu, J., Wang, X., Feng, F., He, X., Chen, L., Lian, J., & Xie, X. (2021, July). Self-supervised graph learning for recommendation. In SIGIR'21 (pp. 726-735).



Bowen Hao is a PhD Candidate in the School of Information, Renmin University of China. His research interests include recommender system. He has published papers in AAAI, WSDM, ECML-PKDD and APWeb.

Hongzhi Yin received the PhD degree in computer science from Peking University, in 2014. He is a associate professor with the University of Queensland. His research interests include recommender system, user profiling, topic models, deep learning, social media mining, and location-based services.

Jing Zhang received the PhD degree from the Department of Computer Science and Technology, Tsinghua University. She is an associate professor in School of Information, Renmin University of China. Her research interests include knowledge graph constructing and mining.

Cuiping Li received the PhD degree from the Institute of Computing Technology, CAS. She is currently a professor of Renmin University of China. Her current research interests include database systems, social network analysis, and data mining.

Hong Chen received PhD degree from the Institute of Computing Technology, CAS. She is a professor of Renmin University of China. Her research interests include data privacy, big data management, and data analysis based on new hardwares. She is a distinguished member of the CCF.