Disentangled Multi-interest Representation Learning for Sequential Recommendation

Yingpeng Du
dyp1993@pku.edu.cn
Nanyang Technological University
Singapore

Ziyan Wang
wang1753@e.ntu.edu.sg
Nanyang Technological University
Singapore

Zhu Sun*
sunzhuntu@gmail.com
Singapore University of Technology and Design
Singapore

Yining Ma
yiningma@u.nus.edu
Nanyang Technological University
Singapore

Hongzhi Liu
liuhz@pku.edu.cn
Peking University
Beijing, China

Jie Zhang
zhangj@ntu.edu.sg
Nanyang Technological University
Singapore

ABSTRACT

Recently, much effort has been devoted to modeling users' multi-interests (aka multi-faceted preferences) based on their behaviors, aiming to accurately capture users' complex preferences. Existing methods attempt to model each interest of users through a distinct representation, but these multi-interest representations easily collapse into similar ones due to a lack of effective guidance. In this paper, we propose a generic multi-interest method for sequential recommendation, achieving disentangled representation learning of diverse interests technically and theoretically. To alleviate the collapse issue of multi-interests, we propose to conduct item partition guided by their likelihood of being co-purchased in a global view. It can encourage items in each group to focus on a discriminated interest, thus achieving effective disentangled learning of multi-interests. Specifically, we first prove the theoretical connection between item partition and spectral clustering, demonstrating its effectiveness in alleviating item-level and facet-level collapse issues that hinder existing disentangled methods. To efficiently optimize this problem, we then propose a Markov Random Field (MRF)-based method that samples small-scale sub-graphs from two separate MRFs, thus it can be approximated with a cross-entropy loss and optimized through contrastive learning. Finally, we perform multi-task learning to seamlessly align item partition learning with multi-interest modeling for more accurate recommendation. Experiments on three real-world datasets show that our method significantly outperforms state-of-the-art methods and can flexibly integrate with existing multi-interest models as a plugin to enhance their performances.

*Corresponding Author

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671800

CCS CONCEPTS

• **Information systems** → **Collaborative filtering**; **Recommender systems**.

KEYWORDS

Recommender Systems; Multi-Interests; Disentangled Learning; Item Partition.

ACM Reference Format:

Yingpeng Du, Ziyan Wang, Zhu Sun, Yining Ma, Hongzhi Liu, and Jie Zhang. 2024. Disentangled Multi-interest Representation Learning for Sequential Recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3637528.3671800

1 INTRODUCTION

Sequential recommendation, a key part of Recommender Systems (RSs), focuses on modeling users' dynamic preferences on items based on their sequential behaviors [39]. In the past few years, various techniques have been adopted for users' preference learning, such as Markov Chains [29], Recurrent Neural Networks (RNN) [10], attention mechanism [18], etc. However, these methods only model the user's preference with a single representation, which ignores the discrimination of different interests of users. Figure 1 illustrates the user's sequential purchase behaviors on an online shopping website, which indicates that the user usually shows multi-interests (aka multi-faceted preferences) on different kinds of items. Therefore, multi-interest methods [2, 3, 19, 46, 47] gain their popularity in RSs, which aims to capture users' different interests behind their behaviors and model each interest by a representation.

According to the information utilization for multi-interest modeling, existing methods can be categorized into two classes. One class of existing methods exploits auxiliary knowledge such as item categories to guide multi-interest learning [3, 22], but the high-quality and relevant auxiliary information may not always be accessible. The other class adopts capsule networks or the attention mechanism to purely derive from users' behaviors, implicitly allocating representations of user-engaged items into different interest facets [2, 46]. However, these multi-interest representations easily collapse into similar ones, which hardly capture the user's distinct interests in real-world applications.

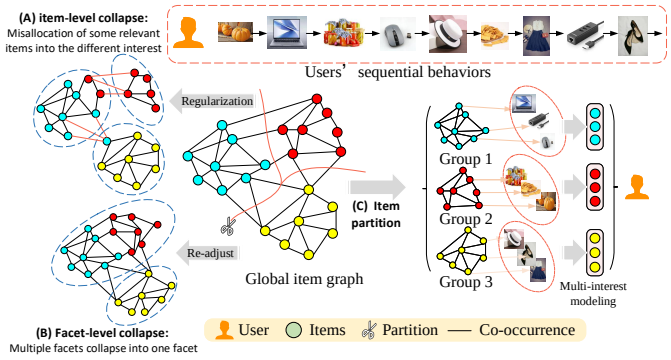

Figure 1: An illustrative example of users' multi-interest modeling with (A) item-level collapse; (B) Facet-level collapse; and (C) Item Partition. The global item graph measures the co-occurrence patterns of items.

To this end, the majority of existing methods propose to alleviate the collapse issues by regularization strategies, e.g., minimizing the correlations of interests [42], penalizing the sparsity of the item-to-interest routing matrix [47], and limiting the norm of interest representations [26, 36]. Nevertheless, these strategies only suggest "what should not be done" (e.g., avoiding the similarity of different interests) for multi-interest learning, which does not directly guarantee that "what should be done" (i.e., allocating relevant items into the same interest of users) in a proactive manner. As a result, they may suffer from the **item-level collapse of multi-interest** (Figure 1. A) due to the partial misallocation of items. Other methods [27, 51] attempt to proactively re-adjust the item-to-interest alignment to learn distinct interest representations. Specifically, they first identify representative items exhibiting similarity to a specific interest, and then pull the interest closer to these items while keeping away from others. Nevertheless, these methods may lead to **facet-level collapse of multi-interest** (Figure 1. B), e.g., multiple facets almost collapse into one facet, if two interests share the overlapped representative items in the training phase. To resolve these issues, we aim to partition relevant items into similar groups while ensuring each group is discriminated from the others as in Figure 1 (C). Therefore, items in each group exert a concentrated influence on the discriminated interest and show less impact on others through the attention and routing mechanism, contributing to better disentangled multi-interest representation learning.

Although straightforwardness of the idea, there remain challenges in the item partition problem: first, how to formulate it for alleviating both item-level and facet-level collapse issues without auxiliary information; second, how to conduct efficient partition learning to align with multi-interest recommendation tasks.

To address the first challenge, we propose to perform the item partition based on the global user-item interactions. Specifically, we first partition items into similar groups if they are likely to be co-purchased by users [16], because these items may complement each other to meet users' specific requirements, e.g., laptop, mouse, and adapter. Therefore, the item partition helps to model the discriminated interests of users, contributing to disentangled representation learning of multi-interests. Then, we formulate the item partition problem theoretically equivalent to spectral clustering on a global item graph that measures the co-occurrence patterns of items, which consists of two goals to seamlessly alleviate the item- and facet-level collapse issues. In particular, we prove the first goal is cutting off the item edges between different groups with the lowest price. Consequently, any misallocation of items (i.e., item-level collapse) leads to a higher price, so different groups will be discriminated from each other. The second goal is to maximize an entropy term to balance the item number for all groups, therefore any collapsing of two discriminated groups (i.e., facet-level collapse) will decrease the entropy.

To address the second challenge, we propose a Markov Random Field (MRF)-based optimization method to approximate the item partition problem with small-scale sub-graphs for efficiency, as it is costly to conduct spectral clustering directly [37, 43]. Specifically, we first define an item-confidence MRF and an item-partition MRF based on global co-occurrence patterns of items and item partition, respectively. Then, we sample sub-graphs from these two MRFs to reformulate the item partition problem as a cross-entropy loss inspired by [34], which can be solved through contrastive learning between these sub-graphs efficiently. To further align the item partition learning with multi-interest modeling, we propose conducting multi-task learning with the shared representations of items' partitions and embeddings, facilitating the item partition to help disentangle learning of users' multi-interests.

In summary, we propose a generic Disentangled-based Multi-Interest Representation method (named DisMIR) for sequential recommendation, alleviating both item-level and facet-level collapse of multi-interests without auxiliary information. The proposed item partition encourages items in each group to focus on a discriminated interest, thus achieving effective disentangled learning of multi-interests. We also establish its connection with spectral clustering theoretically and provide an efficient way for its optimization practically. Experiments on three real-world datasets show that the proposed method DisMIR not only consistently outperforms state-of-the-art methods, but also can flexibly integrate with existing multi-interest models as a plugin to enhance their performances.

2 RELATED WORK

2.1 Sequential Recommendation

Early work on sequential recommendation mainly employed the Markov chains to model the sequential patterns of users' behaviors [13, 29]. With the recent advances in deep learning and information technology, massive methods attempt to use different models and information sources for sequential recommendation. One class of existing methods explores and improves deep model architecture to enhance the model generalization capability, e.g., RNN [10], Memory Networks[8, 17], attentional mechanism [18, 31, 49], and graph-based models [4, 25, 48]. The other class of existing methods explores auxiliary information such as items' attributes to enhance the representations of users' preferences [1, 17, 50]. For example, Bai et al. [1] employ a hierarchical architecture to integrate attribute information of items. However, these methods rely on a singular representation to model user preferences, which overlooks the discrimination of different interests of users and is insufficient to capture users' multi-faceted preferences.

2.2 Multi-interest Recommendation

Recently, multi-interest learning-based methods have demonstrated significant promise in enhancing recommendation performance. According to the information utilization for users' multi-interest modeling, existing methods can be categorized into two classes: auxiliary knowledge-based multi-interest methods and interaction-based multi-interest methods.

The former methods attempt to guide multi-interest learning with the help of auxiliary knowledge such as users' profiles [3], items' categories [11, 20], multi-type behaviors [6, 21, 28], knowledge graphs [22, 44, 45], users' groups [23], and multi-modal information [40, 41]. For example, Liu et al. [22] propose to align implicit disentangled representations that are learned from user-item interactions with the explicit disentangled representations based on a knowledge graph. Chen et al. [6] propose a curriculum disentangled recommendation model to learn disentangled representations from users' multi-feedback (e.g., click, unclick, and dislike). However, these methods rely on high-quality auxiliary information that may not always be available. The latter methods adopt the attention mechanism [24, 26, 46] or capsule network [2, 19, 35, 47] to purely derive from users' engaged items without auxiliary information. For example, Li et al. [19] propose to capture users' various interests using dynamic routing through Capsule Network [30]. Tian et al. [35] propose a multi-interest learning method with a combination of sequential capsule network and graph convolutional aggregation, aiming to capture multi-interest and multi-level preferences of users. Although these methods have demonstrated significant promise in enhancing recommendation performance, multi-interest facets in these methods may easily collapse into similar ones and degrade into a single-interest [12], making it ineffective in capturing users' multi-interests. In addition, these methods provide limited theoretical analysis about alleviating the collapse of users' multi-interest representations.

2.3 Disentangled Representation Learning for Multi-interest Recommendation

To alleviate the collapse issue, many existing methods aim to achieve disentangled representation learning for users' multi-interests recently. The majority of existing methods propose to alleviate the collapse issues by additional regularization, e.g., minimizing the correlations of interests [7, 42], penalizing the sparsity of the item-to-interest routing matrix [47], and limiting the norm of interest representations [26, 36]. For example, Wang et al. [42] propose to minimize the correlation distance [32, 33] of multi-faceted representation of users and items, which encourages the factor-aware representations to be independent. Xie et al. [47] introduce the variance regularizer of the item-to-interest routing matrix, and penalize it for alleviating routing collapse in capsule networks.

$$\mathcal{L}_{\text{reg}} = ||diag(\boldsymbol{D})||^2, \quad \boldsymbol{D} = (\boldsymbol{B} - \bar{\boldsymbol{B}})^\top (\boldsymbol{B} - \bar{\boldsymbol{B}}), \tag{1}$$

where $\bar{\boldsymbol{B}}$ denotes the mean of the routing weights $\boldsymbol{B}$ along the first axis, and $diag(\cdot)$ represents the aggregation of diagonal elements for a matrix. However, these passive regularization strategies are not essential for ensuring effective item-to-interest alignment, i.e., allocating relevant items into the same interest of users, which may suffer from the item-level collapse issue. The other methods propose to re-adjust item-to-interest alignment based on the representations of multi-interests [27, 51]. For instance, Zhang et al. [51] introduce a backward-flow mechanism to identify representative items exhibiting similarity to a specific interest, and then pulled the interest closer to these items while keeping away from others,

$$\mathcal{L}_{\text{reg}} = -\sum_{\boldsymbol{m}} \frac{\exp(\boldsymbol{m}, \boldsymbol{e}_i)}{\exp(\boldsymbol{m}, \boldsymbol{e}_i) + \sum_j \exp(\boldsymbol{m}, \boldsymbol{e}_j)}, \tag{2}$$

where $\boldsymbol{m}$, $\boldsymbol{e}_i$, and $\boldsymbol{e}_j$ denote the interest, representative item i, and another item j, respectively. However, the user's multi-interests may suffer from the facet-level collapse issue in these methods, because some interests can coincidentally share the overlapped representative items in the training phase.

3 PRELIMINARY

3.1 Problem Formulation

Let $\mathcal{U}$ and $\mathcal{I}$ represent the sets of M users and N items, respectively. Each user $u \in \mathcal{U}$ has engaged a series of items $\mathcal{E}_u = [e_t | t = 1, \cdots, |\mathcal{E}_u|]$ in the chronological order, where $e_t \in \mathcal{I}$ denotes the t-th item that the user engaged. In this paper, we aim to effectively provide a recommendation list for the user, which contains items that he/she is likely to engage with in the near future. The main notations used throughout the paper are summarized in Table 6.

3.2 Capsule Networks for Multi-Interests

Among multi-interest methods, capsule networks [30] have gained popularity for multi-interest modeling due to their effectiveness [2, 19, 35, 47]. Specifically, the capsule network can generate users' F-interest representations derived from their recent κ sequential behaviors $\mathcal{E}'_u = [e_{|\mathcal{E}_u|-\kappa+1}, \cdots, e_{|\mathcal{E}_u|}]$, where F denotes the pre-defined number of users' multi-interests. Specifically, the f-th interest capsule $\boldsymbol{m}_f \in \mathbb{R}^d$ can be calculated as follows:

$$\boldsymbol{m}_f = \sum_{i=1}^{\kappa} b_{if} \cdot \boldsymbol{W} \cdot \boldsymbol{e}_i, \quad f = 1, \cdots, F,$$

where $\boldsymbol{e}_i \in \mathbb{R}^d$ denotes embeddings of the i-th items in the user's sequential behaviors $\mathcal{E}'_u$, and d denotes the dimension of latent space. $\boldsymbol{W} \in \mathbb{R}^{d \times d}$ is the transformation matrix, and b_{if} denotes the probability (the routing weight) of item e_i belonging to the f-th interest capsule. The routing weight b_{if} is calculated by the softmax operation of the routing logits c_{if} that measures the similarity between the item embedding and squashed vector $\boldsymbol{v_f}$, i.e.,

$$b_{if} = \frac{exp(c_{if})}{\sum_{f=1}^{F} exp(c_{if})}, \quad c_{if} = \boldsymbol{e}_i^\top \cdot \boldsymbol{W} \cdot \boldsymbol{v_f}, \quad \boldsymbol{v_f} = \frac{||\boldsymbol{m}_f||^2}{||\boldsymbol{m}_f||^2 + 1} \frac{\boldsymbol{m}_f}{||\boldsymbol{m}_f||},$$

where the values of routing logits c_{if} are initialized to zeros with the repeated routing process.

4 THE PROPOSED THEORY AND METHOD

The overall architecture of the proposed method is shown in Figure 2. First, we propose to formulate the item partition problem and prove its relationship to the spectral cluster (Section 4.1). Then, to efficiently optimize this problem, we propose to approximate it with a cross-entropy loss based on MRFs, which can be optimized by contrastive learning (Section 4.2). Finally, we conduct multi-interest learning to align item partition learning with multi-interest learning for more accurate recommendation (Section 4.3).

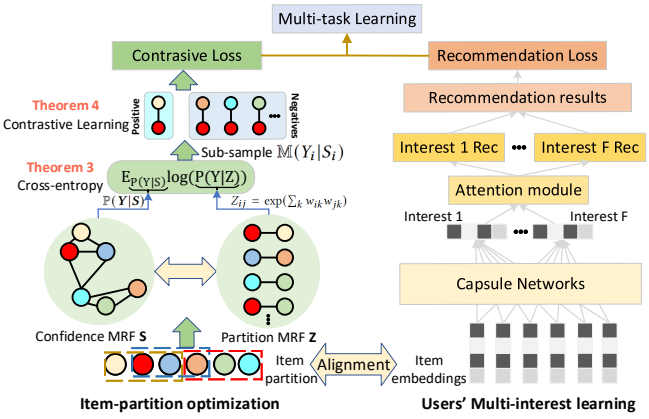

Figure 2: The architecture of the disentangled multi-interest representation learning framework.

4.1 Item Partition Formulation

To guide users' disentangled multi-interest learning, we propose to perform the item partition based on global user-item interactions. Specifically, we organize the user-item interactions into an item-confidence graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The nodes of $\mathcal{G}$ are all items, and their edges measure the confidence of items inspired by frequency patterns of users' behaviors [16], i.e., $S_{ij} = \frac{co(i,j)}{\sum_{j=1}^{N} co(i,j)}$, where $co(i, j)$ denotes the number of users who interact with the item i and the item j simultaneously, and we have $\sum_{j=1}^{N} S_{ij} = 1$. To comprehensively model the item partition in a generic framework, we propose two kinds of partition scenarios, namely Non-overlapped Partition and Overlapped Partition.

4.1.1 Non-overlapped partition. Non-overlapped partition handles the scenario where each item only belongs to one specific group. To effectively partition items into multiple groups (e.g., K), we propose to partition items based on the item-confidence graph, which contains global information on users' behavioral patterns:

$$\text{OPT} = \max \sum_{i \neq j} \sum_{k=1}^{K} w_{ik} \cdot w_{jk} \cdot S_{ij} - \sum_i \log \sum_j \sum_{k=1}^{K} w_{ik} \cdot w_{jk}, \quad (3)$$

$$\sum_{k=1}^{K} w_{ik} = 1, \quad w_{ik} \in \{0, 1\},$$

where partition weight $w_{ik} = 1$ means that the item i is partitioned into group k otherwise $w_{ik} = 0$. $\sum_{k=1}^{K} w_{ik} \cdot w_{jk}$ checks whether the pair of items i and j are partitioned into the same group. The first term aims to partition items with high confidence S_{ij} into the same group as much as possible. The second term punishes trivial solutions, e.g., partitioning all items into the same group. To explain how the item partition problem helps to alleviate both item-level and facet-level collapse issues, we illustrate its relationship to spectral clustering [38] as follows[1].

Theorem 1. ***(Non-overlapped spectral clustering)*** *The item partition problem in Equation (3) is equivalent to spectral clustering based on the item-confidence graph,*

$$OPT \propto -\sum_{k=1}^{K} Cut(\mathcal{A}_k, \bar{\mathcal{A}}_k) + N \cdot entropy(|\mathcal{A}_k|/N),$$

where $\{\mathcal{A}_1, \cdots, \mathcal{A}_K\}$ denote the item partition groups which satisfy $\cup_k \mathcal{A}_k = \mathcal{I}$ and $\mathcal{A}_{k_1} \cap \mathcal{A}_{k_2} = \emptyset$. $\bar{\mathcal{A}}_k = \mathcal{I} - \mathcal{A}_k$ denote the complement set of $\bar{\mathcal{A}}_k$. $Cut(A_{k_1}, A_{k_2})$ calculates the price (total weights) of cutting off all edges between $\mathcal{A}_{k_1}$ and nodes $\mathcal{A}_{k_2}$.

For the item-level collapse issue, any misallocation of items leads to a higher price. Specifically, if the item i is more relevant to items in group $\mathcal{A}_{k_1}$ than group $\mathcal{A}_{k_2}$, i.e., $Cut(i, \mathcal{A}_{k_1}) > Cut(i, \mathcal{A}_{k_2})$, we have

$$Cut(\mathcal{A}_{k_1} + i, \mathcal{A}_{k_2} - i) = Cut(\mathcal{A}_{k_1}, \mathcal{A}_{k_2}) - Cut(\mathcal{A}_{k_1}, i) + Cut(\mathcal{A}_{k_2}, i)$$
$$< Cut(\mathcal{A}_{k_1}, \mathcal{A}_{k_2}) + Cut(\mathcal{A}_{k_1}, i) - Cut(\mathcal{A}_{k_2}, i) = Cut(\mathcal{A}_{k_1} - i, \mathcal{A}_{k_2} + i),$$

which indicates that misallocating the item i in group $\mathcal{A}_{k_2}$ will introduce a higher price. As the first goal is cutting off the item edges between different groups with the lowest prices, $\sum_{k=1}^{K} Cut(\mathcal{A}_k, \bar{\mathcal{A}}_k) = 2 \cdot \sum_{k_1=1}^{K} \sum_{k_2=k_1+1}^{K} Cut(\mathcal{A}_{k_1}, \mathcal{A}_{k_2})$, it can alleviate the item-level collapse issue by forcing items from different groups to discriminate from each other while allocating relevant items into the same group. **For the facet-level collapse issue**, the second goal is to maximize the entropy term of items' numbers among different groups, balancing the item number for all groups. Specifically, suppose we have two groups with N_1 and M_1 items, where $N_1 \approx M_1$. The first goal is to cut off the item edges between different groups with the lowest price, however, it may generate an imbalance solution, i.e., $N_2 << M_2$, and $N_2 + M_2 = N_1 + M_1$, with fewer edges as $N_2 * M_2 < N_1 * M_1$. It may lead to facet-level collapse because the facet with N_2 items nearly disappears. Therefore, balancing the item number of discriminated groups can prevent imbalance solutions and alleviate the facet-level collapse issue. Therefore, any collapsing of two discriminated groups will decrease the entropy.

4.1.2 Overlapped partition. In real-world scenarios, an item usually belongs to several groups at the same time [22], e.g., pumpkins are not only limited to the kitchen but also used as decorations during Halloween. To this end, we propose to model the overlapped partition weight w_{ik} by relaxing it into a real value $\mathbb{R}$ as follows.

$$\text{OPT} = \max \sum_{i \neq j} \sum_{k=1}^{K} w_{ik} \cdot w_{jk} \cdot S_{ij} - \sum_i \log \sum_j \exp(\sum_{k=1}^{K} w_{ik} \cdot w_{jk}), \quad (4)$$

where $w_{ik} \in \mathbb{R}$ and we add the $\exp(\cdot)$ to ensure the positive value in regularization term. In this case, the partition goal is equivalent to overlapped spectral clustering [37, 43] as follows.

Theorem 2. ***(Overlapped spectral clustering)*** *Maximizing Equation (4) with $w_{ik} \in \mathbb{R}$ is equivalent to conducting overlapped spectral clustering based on the item-confidence graph, i.e.,*

$$OPT = -tr(\boldsymbol{W}^\top \boldsymbol{L} \boldsymbol{W}) + \sum_i \log \frac{\exp(\sum_k w_{i,k} \cdot w_{i,k})}{\sum_j \exp(\sum_k w_{i,k} \cdot w_{j,k})},$$

where $\boldsymbol{L} = \boldsymbol{I} - \boldsymbol{S}$ denotes the Laplacian matrix of matrix $\boldsymbol{S}$. It is equivalent to doing spectral clustering with a repulsion regularization.

In summary, the goal of item partition is to allocate items in a spectral clustering manner, which helps to partition relevant items into similar groups while ensuring each group is discriminated from the others. With the supervision of item partitions, items of each group can exert a concentrated influence on the specific interest and show less impact on others with the attention or routing mechanism, contributing to better disentangled multi-interest learning.

[1]Due to space limitation, we provide the proof of all Theorems and Lemmas in the Appendix

4.2 MRF-based Optimization

In the real-world application of RSs, the number of candidate items is usually very large, so it is costly to conduct spectral clustering directly [37, 43]. To tackle this challenge, we propose an MRF-based optimization method with sub-graph sampling for efficiency purposes inspired by [34], as comparing the whole graphs is costly. Specifically, we first define an item-confidence MRF and an item-partition MRF, and then introduce a probability distribution for sub-graph sampling of them. Second, we propose to approximate the item partition problem with cross-entropy loss between sub-graphs sampled from these two MRFs. Finally, we optimize the cross-entropy loss in a contrastive learning manner.

4.2.1 Induced probability distributions on MRF. To guide the item partition with global frequency patterns of users' behaviors, we propose to define an item-confidence MRF and an item-partition MRF in a symmetric way. Specifically, the item-confidence MRF models the frequency patterns of users' behaviors, whose nodes are all items and edges are defined to measure the confidence between nodes S_{ij}. The item-partition MRF aims to model the item partition, whose nodes are all items and edges are defined to measure the possibility of two items belonging to the same group, i.e., $Z_{ij} = \exp(\sum_{k=1}^{K} w_{ik} w_{jk})$. To reduce the complexity of large-scale MRFs, we introduce an induced probability distribution $\mathbb{P}(\cdot|\boldsymbol{R})$ for sub-graph sampling inspired by [34], where nodes in sampled subgraphs have only one out-going edge.

Definition 1. *Given an MRF with n variables with their relations $\boldsymbol{R} \in \mathbb{R}_{+}^{n \times n}$, we define a probability distribution $\mathbb{P}(\boldsymbol{Y}|\boldsymbol{R})$ on the directed unweighted subgraph $\boldsymbol{Y} \in \{0,1\}^{n \times n}$ with only one out-going edge for each node, i.e.,*

$$\mathbb{P}(\boldsymbol{Y}|\boldsymbol{R}) = \frac{V_R(\boldsymbol{Y}) \cdot \Omega(\boldsymbol{Y})}{\sum_{Y \in \mathcal{Y}} V_R(\boldsymbol{Y}) \cdot \Omega(\boldsymbol{Y})},$$

where $\mathcal{Y} = \{\boldsymbol{Y} \in \{0,1\}^{n \times n} | Y_{i,i} = 0\}$ denote all possible graphs with n nodes. $\Omega(\boldsymbol{Y}) \triangleq \prod_i \mathbb{I}(\sum_j Y_{ij} = 1)$ checks if each node has exactly one out-going edge. $V_R(\boldsymbol{Y}) \triangleq \prod_{Y_{ij}=1} R_{ij}$ combines the scores of each edge in the subgraph $\boldsymbol{Y}$.

4.2.2 Approximation. To estimate the proposed item partition problem, we first propose two lemmas to reformulate its two terms in Equation (4) based on MRFs.

Lemma 4.1. *The first term in the item partition problem can be formulated as the expectation of $\mathbb{P}(\cdot|\boldsymbol{S})$ about $\boldsymbol{Z}$:*

$$\sum_{i \neq j} \sum_{k=1}^{K} w_{ik} \cdot w_{jk} \cdot S_{ij} = \mathbb{E}_{\mathbb{P}(\boldsymbol{Y}|\boldsymbol{S})} \log \left(\Omega(\boldsymbol{Y}) \cdot \prod\nolimits_{Y_{ij}=1} Z_{ij} \right),$$

where $Z_{ij} = \exp(\sum_{k=1}^{K} w_{ik} w_{jk})$.

Lemma 4.2. *Maximizing the second terms in the item partition problem is equivalent to maximizing the aggregation of graphs whose nodes have exactly one out-going edge:*

$$\sum\nolimits_i \log \sum\nolimits_j \exp(\sum\nolimits_{k=1}^{K} w_{ik} \cdot w_{jk}) = -\log \left(\sum\nolimits_{Y \in \mathcal{Y}} \Omega(\boldsymbol{Y}) \cdot \prod\nolimits_{Y_{ij}=1} Z_{ij} \right).$$

Based on these two lemmas, we can approximate the proposed partition problem as a cross-entropy loss that bridges the item-confidence MRF and the item-partition MRF, verifying that frequency patterns of users' behaviors are important guidance for the item partition learning.

Theorem 3. ***(Approximation)*** *Maximizing the proposed partition problem is equivalent to optimizing the cross-entropy loss between the two MRFs:*

$$\max OPT \Leftrightarrow \max \mathbb{E}_{\mathbb{P}(\boldsymbol{Y}|\boldsymbol{S})} \log(\mathbb{P}(\boldsymbol{Y}|\boldsymbol{Z})),$$

where $\boldsymbol{S}$ and $\boldsymbol{Z}$ describe the relations in the item-confidence MRF and the item-partition MRF, respectively.

4.2.3 Optimization. To further reduce the complexity of cross-entropy loss between these two MRFs, we divide it into node-level comparisons as their out-going edges are independent to each other [34]. Specifically, suppose $\boldsymbol{Y}_i' \sim \mathbb{M}(R_{i_1}/\sum_j R_{i_j}, \cdots, R_{i_n}/\sum_j R_{i_j})$ (abbreviate as $\mathbb{M}(\boldsymbol{Y}_i'|\boldsymbol{R}_i)$) is a one-hot encoding sample from multinomial distribution $\mathbb{M}(\cdot)$, we have $\mathbb{P}(\boldsymbol{Y} = (\tilde{\boldsymbol{Y}}_1, \cdots, \tilde{\boldsymbol{Y}}_n)|\boldsymbol{R}) = \prod_i \mathbb{M}(\boldsymbol{Y}_i' = \tilde{\boldsymbol{Y}}_i|\boldsymbol{R}_i)$ as $\boldsymbol{Y}_1', \cdots, \boldsymbol{Y}_n'$ are independent of each other, where $\tilde{\boldsymbol{Y}}_i$ denotes the i-th row of $\tilde{\boldsymbol{Y}}$. Consequently, item partition optimization can be achieved by contrastive learning as follows.

Theorem 4. ***Contrastive learning.*** *The proposed partition goals in Equation (3) and Equation (4) can be solved in a contrastive learning manner, i.e.,*

$$\begin{aligned} \mathcal{L}_{Partition} &= \mathbb{E}_{\mathbb{P}(\boldsymbol{Y}|\boldsymbol{S})} \log(\mathbb{P}(\boldsymbol{Y}|\boldsymbol{Z})) \\ &= \sum_i \mathbb{E}_{\mathbb{M}(\boldsymbol{Y_i}|\boldsymbol{S_i})} \log \frac{\exp(\sum_k w_{ik} w_{jk})}{\sum_{j'} \exp(\sum_k w_{ik} w_{j'k})}, \end{aligned}$$

where $\boldsymbol{Y_i}$ is a one-hot encoding vector with the j-th element $\boldsymbol{Y}_{ij} = 1$.

In summary, this subsection provides an efficient way to optimize the proposed partition problem. Specifically, for each node i in the item-confidence graph, we can sample one node j from its neighbor according to $\boldsymbol{Y_i} \sim \mathbb{M}(\cdot|\boldsymbol{S_i})$ as the positive node and others as the negative nodes for the contrastive learning. Intuitively, as the items connected by an edge with higher item confidence are more likely to be sampled in the subgraph, they are more likely to be partitioned into the same group. Therefore, such clustering learning can make items belonging to different/same groups show distinguishable/similar representations, which helps the capsule network more precisely allocate items into multi-interest facets. As an item usually belongs to several groups in real-world scenarios [22], we take the real-value partition weights to model the overlapped item partition. For efficient optimization, we sample N_v (e.g., 100) items to approximate the rest of the items as the negative nodes. Therefore, the complexity of updating all nodes is around $O(N \cdot K \cdot N_v)$, which is more efficient than spectral clustering with computation complexity $O(N^3)$ [43] and $O(N^2 \log N)$ [37] because of $N >> N_v$ and $N >> K$. In addition, the proposed method provides a generic way to flexibly integrate with existing multi-interest models based on multi-task learning, which is challenging for spectral clustering and the Gaussian mixture model.

4.3 Multi-task Representation Learning

To bridge item partition and multi-interest modeling, we propose conducting multi-task learning of item partition and multi-interest recommendation tasks, because the appropriate item partition can help to disentangle users' multi-interest learning to alleviate interest collapse issues. In the following, we first introduce multi-interest modeling for recommendation, and then align item partition with the multi-interest recommendation tasks with multi-task learning.

Table 1: Statistics of the experimental datasets

Dataset	#Users	#Items	#Interactions
Gowalla	165,506	174,605	2,061,264
Retail Rocket	33,708	81,635	356,840
Amazon Books	603,668	367,982	8,898,041

To capture users' multi-interests, we adopt the capsule networks to derive from the user's behaviors $\mathcal{E}'_u$,

$$\boldsymbol{m}_{u1}, \cdots, \boldsymbol{m}_{uF} = Capsule([\boldsymbol{e}_1, \cdots, \boldsymbol{e}_K])$$

where $[\boldsymbol{e}_1, \cdots, \boldsymbol{e}_K]$ denotes the item embeddings among the user's behaviors $\mathcal{E}'_u$, and $\boldsymbol{m}_{uf}$ denotes his/her f-th interest. Then, we adopt the attention mechanism to capture the user's dynamic preferences on different interests,

$$a_{u,f} = \exp(\boldsymbol{m}_{uf}^\top \cdot \bar{\boldsymbol{e}}_u) / \sum_{f'=1}^{F} \exp(\boldsymbol{m}_{uf'}^\top \cdot \bar{\boldsymbol{e}}_u),$$

where $\bar{\boldsymbol{e}}_{\boldsymbol{u}} = Average([\boldsymbol{e}_1, \cdots, \boldsymbol{e}_K])$ measures the user's general preference by aggregating his/her historical behaviors. To predict items that the user may interact with, we fuse the user's preferences on items varying with different interests,

$$score(u, i) = \sum_{f=1}^{F} a_{u,f} \cdot \hat{r}_{u,f,i}, \quad \hat{r}_{u,f,i} = \boldsymbol{m}_{uf}^\top \cdot \boldsymbol{e}_{\{i\}}, \tag{5}$$

where $\hat{r}_{u,f,i}$ measures the score of item i w.r.t. user u's f-th interest, and $\boldsymbol{e}_{\{i\}}$ denotes the embedding of the item i. For the recommendation task, we adopt the pairwise loss to define the recommendation objective function as follows,

$$\mathcal{L}_{Rec} = \sum\nolimits_{u,i,j} \log \sigma \left(score(u, i) - score(u, j) \right),$$

where the train set $\mathcal{D} = \{(u, i, j) | u \in \mathcal{U}, i, j \in \mathcal{I}\}$ means that user u engaged item i instead of item j. $\sigma(\cdot)$ denotes the Sigmoid function. We select the item j with the highest score among negative item candidates as the hard negative instance inspired by [47].

To align the item partition with the multi-interest recommendation, we first connect the item's partition weights with its embeddings based on the shared representations, i.e., $\boldsymbol{e}_i = [w_{i1}, \cdots, w_{iK}]$ and $d = K$. Then, we propose to conduct multi-task learning with consideration of both item partition task and multi-interest recommendation task as follows:

$$\mathcal{L} = \mathcal{L}_{Rec} + \lambda \cdot \mathcal{L}_{Partition}, \tag{6}$$

where λ denotes the trade-off coefficient of these two tasks.

5 EXPERIMENT

In this section, we aim to validate the effectiveness of the proposed method DisMIR. Specifically, we conduct extensive experiments to study the following research questions:

RQ1: Whether the proposed DisMIR outperforms state-of-the-art multi-interest recommendation methods?
RQ2: Whether the proposed DisMIR benefits from the item partition for multi-interests disentangled learning?
RQ3: Whether the proposed DisMIR can enhance existing multi-interest models for reccomendation?
RQ4: How do hyper-parameters influence the performance of the proposed DisMIR?

5.1 Experimental Setup

5.1.1 Datasets. We adopt three public datasets, including Gowalla[2], RetailRocket[3], and Amazon Books[4]. The Gowalla is a checking-in dataset derived from a social network website named gowalla.com. RetailRocket is an online shopping dataset collected from the biggest E-Commerce website (named Taobao) in China. The Amazon Books dataset is the largest subset in the Amazon series datasets, which is collected from the Amazon platform with users' reviews on various types of books. All these datasets contain the timestamps or orders of user behaviors. Following prior studies [2, 47], we filter out users and items that have less than 5 records, then we treat check-in, view, and review behaviors as implicit feedback for these three datasets, respectively[5]. The characteristics of these datasets are summarized in Table 1.

5.1.2 Evaluation Methodology and Metrics. To make the proposed DisMIR comparable with existing multi-interest recommendation methods, we set the evaluation methodology and metrics in our experiments following the prior study[6] [2, 47]. Based on the timestamps of interactions, we chronologically split the interaction records into training, validation, and test sets by the proportion of 8:1:1 for each user. For evaluation, we test the last 20% items in the user's sequence based on the first 80% of users' behavior sequences for their preference inference. To evaluate the performance, we adopt three widely used evaluation metrics for top-n recommendation: Recall (R), Hit Rate (HR), and Normalized Discounted Cumulative Gain (ND), where n was set as 20/50 empirically. Experimental results are recorded as the average of five runs.

5.1.3 Baselines. We take the following state-of-the-art methods as the baselines, mainly including two single-interest models and six multi-interest models for comparison. -**DNN [9]** averages the embeddings of different features, and then feeds their concatenation into a deep neural network for recommendation. -**GRU4Rec [15]** utilizes the GRU module to model users' sequential behaviors for recommendation. -**MIND [19]** adopts the capsule network [30] module to capture users' various interests with a dynamic routing strategy. -**ComiRec [2]** adopts the capsule network with a self-attention dynamic routing method to capture users' different interests for diversity-accuracy goals. -**Re4 [51]** utilizes the backward flow as regularization for users' multi-interest preference learning, contributing to alleviating routing collapse of interest facets. -**UMI [3]** proposes to model users' multi-interest preferences based on users' profiles with a hard negative sampling strategy. We follow the same implementation of hard negative sampling strategy as in [47]. -**PIMIRec [5]** is a state-of-the-art multi-interest framework with consideration of both time information and interactivity among items for sequential recommendation. -**REMI [47]** is a state-of-the-art multi-interest framework with an importance-negative-sampling strategy and routing variation regularization. As MarcridVAE [26] fails to capture the dynamic multi-interest of users for sequential recommendation and suffers from GPU memory costs, we don't include it in baseline methods.

[2] https://snap.stanford.edu/data/
[3] 4https://www.kaggle.com/retailrocket/ecommerce-dataset
[4] http://jmcauley.ucsd.edu/data/amazon/links.html
[5] https://github.com/RUCAIBox/RecSysDatasets
[6] https://github.com/Tokkiu/REMI

Table 2: Performance of different methods. * indicates statistically significant improvement of the proposed method to baseline models on t-test (p < 0.05). 'Improve' indicates the relative improvements of DisMIR over the strongest baseline.

Dataset	Metric	DNN	GRU4Rec	MIND	ComiRec	Re4	UMI	PIMIRec	REMI	DisMIR	improve.
Gowalla	R@20	0.0864	0.0900	0.0923	0.0623	0.0843	0.0961	0.1193	0.1300	**0.1384***	6.48%
	HR@20	0.3211	0.3359	0.3122	0.2281	0.3104	0.3314	0.3843	0.4021	**0.4312***	7.25%
	ND@20	0.1384	0.1433	0.1336	0.0955	0.1287	0.1391	0.1603	0.1715	**0.1855***	8.15%
	R@50	0.1388	0.1458	0.1521	0.1181	0.1396	0.1642	0.1951	0.2109	**0.2169***	2.83%
	HR@50	0.4390	0.4577	0.4482	0.3778	0.4224	0.4719	0.5207	0.5498	**0.5599***	1.83%
	ND@50	0.1434	0.1494	0.1450	0.1199	0.1410	0.1505	0.1660	0.1792	**0.1849***	3.16%
Retail Rocket	R@20	0.1050	0.0827	0.1415	0.1035	0.1397	0.1519	0.1828	0.2129	**0.2385***	12.04%
	HR@20	0.1711	0.1376	0.2195	0.1602	0.2103	0.2364	0.2764	0.3183	**0.3524***	10.72%
	ND@20	0.0641	0.0517	0.0804	0.0609	0.0785	0.0875	0.1025	0.1198	**0.1330***	11.04%
	R@50	0.1608	0.1371	0.2148	0.1666	0.2194	0.2423	0.2811	0.3160	**0.3447***	9.08%
	HR@50	0.2518	0.2132	0.3183	0.2501	0.3174	0.3574	0.3969	0.4515	**0.4815***	6.64%
	ND@50	0.0701	0.0593	0.0880	0.0684	0.0884	0.0974	0.1093	0.1281	**0.1360***	6.20%
Amazon Books	R@20	0.0467	0.0441	0.0433	0.0539	0.0597	0.0690	0.0682	0.0839	**0.0887***	5.81%
	HR@20	0.1043	0.1004	0.0907	0.1108	0.1240	0.1423	0.1411	0.1674	**0.1805***	7.86%
	ND@20	0.0391	0.0378	0.0340	0.0406	0.0476	0.0527	0.0526	0.0629	**0.0678***	7.77%
	R@50	0.0722	0.0706	0.0677	0.0848	0.0690	0.1053	0.1056	0.1207	**0.1366***	13.21%
	HR@50	0.1607	0.1553	0.1379	0.1716	0.1975	0.2059	0.2062	0.2326	**0.2637***	13.38%
	ND@50	0.0457	0.0443	0.0390	0.0481	0.0576	0.0587	0.0583	0.0667	**0.0758***	13.63%

5.1.4 Implementation Details. For a fair comparison, all methods are optimized by the Adam optimizer with the same latent space dimension (i.e., 64) and learning rate (i.e., 1×10^{-3}). For hard negative sampling, we select the hard negative item among 10 item candidates except 50 item candidates for Amazon Books dataset due to its large scale. For the capsule network in the proposed DisMIR, we set its transformation matrix as the identity matrix inspired by the LightGCN [14] and adopt a one-time routing process for efficiency. For the hyper-parameters in the proposed DisMIR, we set trade-off coefficients $\lambda = [0.01, 0.1, 1.0]$ and interest numbers $F = [4, 8, 6]$ for Gowalla Retail, Rocket, and Amazon Books, respectively. Detailed analysis of the hyper-parameter can be found in Section 5.5. For baseline models, we set their parameters as authors' implementation if they exist, otherwise we tune them to their best.

5.2 Model Comparison (RQ1)

Table 2 shows the performance of different methods including single-interest models and multi-interest models. To make the table notable, we bold the best results and underline the best baseline results for each dataset with one specific evaluation metric. Firstly, the proposed method DisMIR significantly outperforms all baseline methods in every case, which shows the effectiveness of the proposed method. DisMIR demonstrates notable enhancements, with average improvements of 4.95%, 9.29%, and 10.28% when compared to the top-performing baseline models on the Gowalla, Retail Rocket, and Amazon Books, respectively. Second, several multi-interest methods such as MIND and ComiRec even underperform single-interest methods such as GRU4Rec and DNN in some cases. This may be attributed to their shortcoming in capturing users' dynamic preferences on different interest facets and the unsuitable training scheme. Finally, REMI achieves the best performance among the baseline methods, which mainly benefits from the negative sampling strategy and routing variation regularization. Re4 can help to improve the performance of its basic version ComiRec with the re-adjustment of the item-to-interest alignment. These also verify the importance of alleviating interest collapse issues.

5.3 Ablation Study (RQ2)

5.3.1 Comparison among the variants of the proposed method. To examine the efficacy of different components of DisMIR, we compare a number of its variants.

- **MIR**: It removes the item partition task for disentangled multi-interest representation learning, i.e., $\lambda = 0$.
- **MIR-RV**: It replaces the item partition task with the Routing Variation regularization (Equation 1) proposed in [47], which penalizes the variance regularizer of the item-to-interest routing matrix to alleviate routing collapse in capsule networks.
- **MIR-BF**: It replaces the item partition task with the Backward Flow re-adjustment strategy (Equation 2) proposed in [51], which aims to learn representative embeddings of different interests.
- **w/o DPI**: It removes users' Dynamic Preference modeling on different interests over time, assuming equal importance for different interests in Equation (5).

Table 3 shows the performance of the proposed method and its variant, i.e., MIR, MIR-RV, MIR-BF, DisMIR, and w/o DPI.

- On the one hand, DisMIR with the item partition consistently outperforms MIR without the item partition. On the other hand, the variant methods with regularization (i.e., MIR-RV) and re-adjustment (i.e., MIR-RV) outperform the variant MIR in most cases. This illustrates the necessity of alleviating the collapse issue for multi-interest preference learning.
- DisMIR consistently outperforms its variant approaches, namely MIR-RV and MIR-BF, across all cases. It indicates that the proposed item partition is more effective than existing strategies for multi-interest disentanglement, which can alleviate both item-level and facet-level collapse issues for users' multi-interests learning.

Table 3: Performance of variant DisMIR for ablation studies.

Dataset	Method	R@20	HR@20	ND@20	R@50	HR@50	ND@50
Gowalla	MIR	0.1332	0.4205	0.1778	0.2132	0.5524	0.1804
	MIR-RV	0.1162	0.3790	0.1473	0.1963	0.5247	0.1615
	MIR-BF	0.1298	0.4151	0.1712	0.2126	0.5570	0.1820
	w/o DPI	0.1303	0.4030	0.1708	0.2125	0.5552	0.1800
	DisMIR	**0.1384**	**0.4312**	**0.1855**	**0.2169**	**0.5599**	**0.1849**
Retail Rocket	MIR	0.2235	0.3355	0.1253	0.3267	0.4574	0.1277
	MIR-RV	0.2144	0.3213	0.1154	0.3303	0.4640	0.1261
	MIR-BF	0.2326	0.3450	0.1300	0.3370	0.4732	0.1335
	w/o DPI	0.2152	0.3204	0.1201	0.3198	0.4506	0.1266
	DisMIR	**0.2385**	**0.3524**	**0.1330**	**0.3447**	**0.4815**	**0.1360**
Amazon Books	MIR	0.0817	0.1691	0.0631	0.1309	0.2547	0.0728
	MIR-RV	0.0844	0.1739	0.0640	0.1318	0.2567	0.0718
	MIR-BF	0.0849	0.1733	0.0636	0.1343	0.2602	0.0731
	w/o DPI	0.0853	0.1689	0.0638	0.1223	0.2342	0.0669
	DisMIR	**0.0887**	**0.1805**	**0.0678**	**0.1366**	**0.2637**	**0.0758**

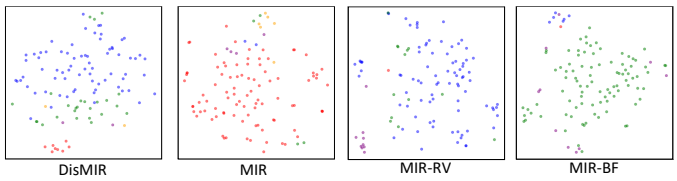

Figure 3: Items' representation visualization in latent space.

- DisMIR outperforms its variant w/o DPI in all cases, which indicates the necessity of capturing users' dynamic preferences on different interests.

5.3.2 A case study for the distribution of user interests. Moreover, we further investigate in how the proposed DisMIR contributes to the items' representation distribution in the latent space. To this end, we visualize the item representations learned by comparing DisMIR, MIR, MIR-RV, and MIR-BF. Specifically, we first randomly sample a user on the Gowalla dataset, and then adopt the t-SNE transformation to map embeddings of his/her engaged items into a 2-dimension plane as shown in Figure 3, where items (points) with different colors mean they are partitioned into different interests. Firstly, the DisMIR exhibits more noticeable clusters that reflect the user's different interests among these methods, indicating that the item partition can help with effective disentangled learning of multi-interests. Secondly, the variant MIR-RV misallocates the green and blue points with ineffective item-to-interest alignment, which can be attributed to the item-level collapse issue of the regularization strategy. Thirdly, the variant MIR-BF assigns most items into the same group (i.e., green one) compared to others, which can be attributed to the facet-level collapse issue of the re-adjustment strategy. Finally, the variant MIR shows both item-level and facet-level collapse issues, indicating the necessity of disentangled multi-interest learning.

5.3.3 A statistical analysis for the distribution of user interests. To delve deeper into the distribution of user interests, we carry out a statistical analysis by comparing our proposed method DisMIR with existing disentangled strategies [47, 51]. Firstly, to assess the item misallocation issue in users' interests (i.e., item-level collapse), we measure the average similarity of items that are from different interest facets for a user, i.e.,

$$Similarity = \frac{2}{F \cdot (F-1)} \sum_{f=1}^{F} \sum_{f'=f+1}^{F} sim(G_f, G_{f'})$$

$$sim(G_f, G_{f'}) = \sum_{i \in G_f, j \in G_{f'}} \frac{< e_i, e_j >}{|G_f||G_{f'}|}$$

where G_f denotes items that are partitioned into interest f for the user, and e_i denotes the representation of the item i. Typically, a high similarity between different interests usually means a more severe issue of item misallocation (i.e., item-level collapse) in multi-interest modeling. Second, to measure item balance in users' interests, we compute the minimal (MIN), maximal (MAX), and standard deviation (STD) of item numbers that are allocated different interest facets for a user. Typically, a high STD and large gap between MIN and MAX usually mean an imbalance issue (i.e., facet-level collapse) in multi-interest modeling.

Table 4: statistical analysis for the distribution of user interests among different methods.

Metric	DisMIR	MIR	MIR-RV	MIR-BF
Similarity	72.1	99.1	85.5	57.4
MIN/MAX	12.7/40.1	12.1/ 42.7	12.2/42.4	8.9/50.7
STD	10.8	12.0	11.8	16.8

We randomly sample 100 users and report their average values, statistical results among different methods (DisMIR, MIR, MIR-RV [47], MIR-BF [51]) are shown in Table 4. Firstly, MIR-BF exhibits the most pronounced item imbalance issue among the methods, as evidenced by the highest STD and the largest gap between MIN and MAX, confirming that the Backward Flow re-adjustment strategy [51] is prone to facet-level collapse. Although MIR-BF shows the lowest similarity between different interests, it is mainly caused by a facet-level collapse issue where only one interest is dominant (e.g., MIR-BF as shown in Figure 3), which is essentially a poor multi-interest distribution. Secondly, MIR-RV with Routing Variation regularization strategy [47] shows the high similarity between different interests, mixing up items from different interests, which indicates a severe issue of item misallocation (i.e., item-level collapse) in multi-interest modeling. Thirdly, the MIR without disentangled learning strategies shows both high STD and similarity, suffering from item-level and facet-level collapse issues. Finally, our DisMIR method achieves balanced item allocation (evidenced by a low STD) and distinguishable interest allocation (evidenced by low similarity between different interests), proving its effectiveness in addressing both item- and facet-level collapse issues.

5.4 Enhancement with Existing Multi-interest Recommendation Methods. (RQ3)

As the item partition task is orthogonal to existing multi-interest methods and frameworks, it is interesting to explore whether this strategy can be applied to enhance their performance. As such, we instantiate our DisMIR in three representative and state-of-the-art multi-interest models (i.e., MIND [30], ComiRec [2], and

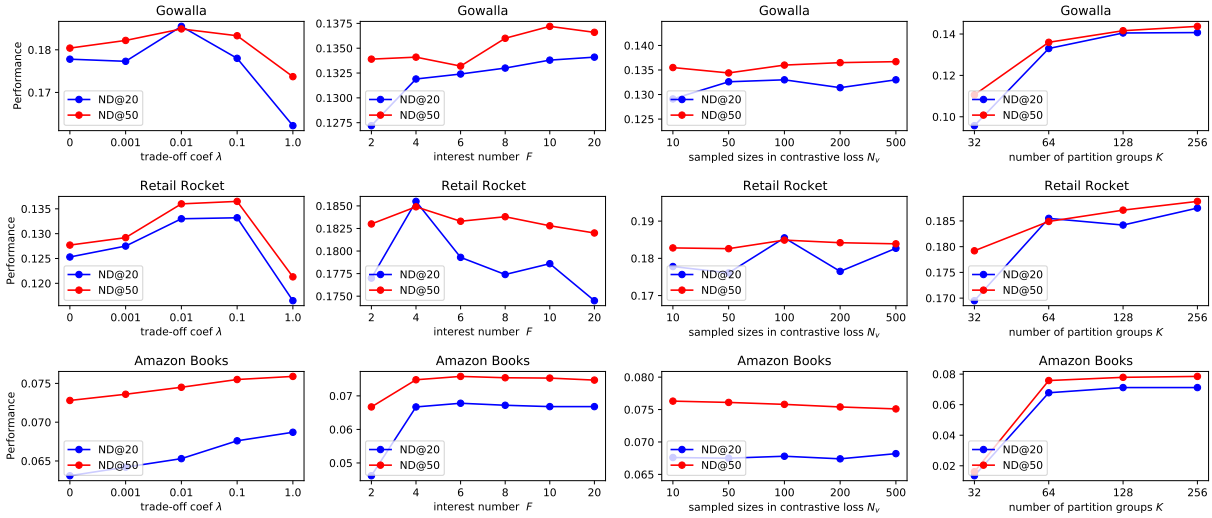

Figure 4: Performance of the proposed method varying with different trade-off coefficients and interest numbers.

Table 5: Performance of existing multi-interest recommendation methods with the item partition task.

Methods	Backbone	MIND		ComiRec		REMI	
	Enhance	✓	✗	✓	✗	✓	✗
Gowalla	R@20	**0.0923**	0.0881	**0.0670**	0.0623	**0.1303**	0.1300
	HR@20	**0.3122**	0.3122	**0.2403**	0.2281	**0.4030**	0.4021
	ND@20	0.1336	**0.1353**	**0.1021**	0.0955	0.1708	**0.1715**
Retail Rocket	R@20	**0.1415**	0.1386	**0.1118**	0.1035	**0.2152**	0.2129
	HR@20	**0.2195**	0.2124	**0.1703**	0.1602	**0.3204**	0.3183
	ND@20	**0.0804**	0.0764	**0.0641**	0.0609	**0.1201**	0.1198
Amazon Books	R@20	**0.0433**	0.0423	**0.0545**	0.0539	**0.0853**	0.0839
	HR@20	**0.0907**	0.0888	**0.1128**	0.1108	**0.1689**	0.1674
	ND@20	**0.0340**	0.0332	**0.0420**	0.0406	**0.0638**	0.0629

REMI [47]). The results are presented in Table 5. Firstly, models without regularization (i.e., MIND and ComiRec) benefit from the item partition task in most cases, which reflects the effectiveness of the proposed DisMIR in alleviating the interest collapse issue. Secondly, the REMI with the routing regularization still benefits from the item partition task, which indicates that the item partition task is more effective for multi-interest disentanglement. Third, the item partition task achieves less improvement in REMI than MIND and ComiRec, which may be attributed to the REMI partially alleviating the interest collapse issue with the routing regularization.

5.5 Hyper-Parameter Study (RQ4)

There are several key parameters for the proposed method DisMIR, including the trade-off coefficient λ, interest number F, sampled sizes N_v in contrastive loss, and number of partition groups K. We thus investigate how these parameters influence the performance of DisMIR. As depicted in Figure 4, we assess DisMIR's performance by varying the values of these hyper-parameters. For the trade-off coefficient λ and interest number F, we observe that the highest performance is achieved when trade-off coefficients $\lambda = [0.01, 0.1, 1.0]$ and interest numbers $F = [4, 8, 6]$ for Gowalla Retail, Rocket, and Amazon Books datasets, respectively. For real-world applications, we suggest adopting a grid search strategy as a practical approach to select the optimal hyper-parameters for the trade-off coefficient and interest number. For the sampled sizes N_v in contrastive loss, we observe that the highest performance is typically achieved when $N_v = 100$. Therefore, we suggest setting $N_v = 100$ for implementation. For the number of partition groups K, we observe that the performance of DisMIR improves with larger K, but the increase slows down when $K > 64$. Therefore, considering both effectiveness and efficiency, we suggest setting $K = 64$ for implementation.

6 CONCLUSION

In this paper, we introduce a generic disentangled multi-interest representation learning method for sequential recommendation, which contributes to the field of multi-interest preference learning both theoretically and practically. To effectively disentangle the representation of multi-interests, we propose a proactive items partition based on global users' behavior patterns. In addition, we establish its connection to the spectral cluster to prove its effectiveness in alleviating item-level and facet-level collapse issues, and provide an efficient way for its optimization practically. Experiments on three real-world datasets show that the proposed DisMIR not only consistently outperforms state-of-the-art methods, but also can flexibly integrate with existing multi-interest models as a plugin to enhance their performances.

7 ACKNOWLEDGE

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISGAward No: AISG3-RP-2022-031). This research is partially supported by the Agency for Science, Technology and Research (A*STAR) under its RIE 2025 - Industry Alignment Fund - Pre Positioning (IAF-PP) funding scheme (Project No: M23L4a0001). This work is also partially supported by the MOE AcRF Tier 1 funding (RG13/23).

REFERENCES

[1] Ting Bai, Jian-Yun Nie, Wayne Xin Zhao, Yutao Zhu, Pan Du, and Ji-Rong Wen. 2018. An attribute-aware neural attentive model for next basket recommendation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1201–1204.

[2] Yukuo Cen, Jianwei Zhang, Xu Zou, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. Controllable multi-interest framework for recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2942–2951.

[3] Zheng Chai, Zhihong Chen, Chenliang Li, Rong Xiao, Houyi Li, Jiawei Wu, Jingxu Chen, and Haihong Tang. 2022. User-aware multi-interest learning for candidate matching in recommenders. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1326–1335.

[4] Jianxin Chang, Chen Gao, Yu Zheng, Yiqun Hui, Yanan Niu, Yang Song, Depeng Jin, and Yong Li. 2021. Sequential recommendation with graph neural networks. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 378–387.

[5] Gaode Chen, Xinghua Zhang, Yanyan Zhao, Cong Xue, and Ji Xiang. 2021. Exploring periodicity and interactivity in multi-interest framework for sequential recommendation. *arXiv preprint arXiv:2106.04415* (2021).

[6] Hong Chen, Yudong Chen, Xin Wang, Ruobing Xie, Rui Wang, Feng Xia, and Wenwu Zhu. 2021. Curriculum disentangled recommendation with noisy multi-feedback. *Advances in Neural Information Processing Systems* 34 (2021), 26924–26936.

[7] Wanyu Chen, Pengjie Ren, Fei Cai, Fei Sun, and Maarten De Rijke. 2021. Multi-interest diversification for end-to-end sequential recommendation. *ACM Transactions on Information Systems (TOIS)* 40, 1 (2021), 1–30.

[8] Xu Chen, Hongteng Xu, Yongfeng Zhang, Jiaxi Tang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2018. Sequential recommendation with user memory networks. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 108–116.

[9] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*. 191–198.

[10] Tim Donkers, Benedikt Loepp, and Jürgen Ziegler. 2017. Sequential user-based recurrent neural network recommendations. In *Proceedings of the eleventh ACM conference on recommender systems*. 152–160.

[11] Yingpeng Du, Hongzhi Liu, and Zhonghai Wu. 2021. Modeling multi-factor and multi-faceted preferences over sequential networks for next item recommendation. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II 21*. Springer, 516–531.

[12] Taeyoung Hahn, Myeongjang Pyeon, and Gunhee Kim. 2019. Self-routing capsule networks. *Advances in neural information processing systems* 32 (2019).

[13] Ruining He and Julian McAuley. 2016. Fusing similarity models with markov chains for sparse sequential recommendation. In *2016 IEEE 16th international conference on data mining (ICDM)*. IEEE, 191–200.

[14] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 639–648.

[15] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939* (2015).

[16] Haoji Hu, Xiangnan He, Jinyang Gao, and Zhi-Li Zhang. 2020. Modeling personalized item frequency information for next-basket recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1071–1080.

[17] Jin Huang, Wayne Xin Zhao, Hongjian Dou, Ji-Rong Wen, and Edward Y Chang. 2018. Improving sequential recommendation with knowledge-enhanced memory networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*. 505–514.

[18] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*. IEEE, 197–206.

[19] Chao Li, Zhiyuan Liu, Mengmeng Wu, Yuchi Xu, Huan Zhao, Pipei Huang, Guoliang Kang, Qiwei Chen, Wei Li, and Dik Lun Lee. 2019. Multi-interest network with dynamic routing for recommendation at Tmall. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 2615–2623.

[20] Jian Li, Jieming Zhu, Qiwei Bi, Guohao Cai, Lifeng Shang, Zhenhua Dong, Xin Jiang, and Qun Liu. 2022. MINER: multi-interest matching network for news recommendation. In *Findings of the Association for Computational Linguistics: ACL 2022*. 343–352.

[21] Qingfeng Li, Huifang Ma, Wangyu Jin, Yugang Ji, and Zhixin Li. 2024. Multi-Interest Network with Simple Diffusion for Multi-Behavior Sequential Recommendation. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*. SIAM, 734–742.

[22] Danyang Liu, Yuji Yang, Mengdi Zhang, Wei Wu, Xing Xie, and Guangzhong Sun. 2022. Knowledge Enhanced Multi-Interest Network for the Generation of Recommendation Candidates. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 3322–3331.

[23] Xiaolong Liu, Liangwei Yang, Zhiwei Liu, Xiaohan Li, Mingdai Yang, Chen Wang, and S Yu Philip. 2023. Group-Aware Interest Disentangled Dual-Training for Personalized Recommendation. In *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 393–402.

[24] Yaokun Liu, Xiaowang Zhang, Minghui Zou, and Zhiyong Feng. 2024. Attribute Simulation for Item Embedding Enhancement in Multi-interest Recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 482–491.

[25] Chen Ma, Liheng Ma, Yingxue Zhang, Jianing Sun, Xue Liu, and Mark Coates. 2020. Memory augmented graph neural networks for sequential recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 5045–5052.

[26] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. 2019. Learning disentangled representations for recommendation. *Advances in neural information processing systems* 32 (2019).

[27] Jianxin Ma, Chang Zhou, Hongxia Yang, Peng Cui, Xin Wang, and Wenwu Zhu. 2020. Disentangled self-supervision in sequential recommenders. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 483–491.

[28] Chang Meng, Ziqi Zhao, Wei Guo, Yingxue Zhang, Haolun Wu, Chen Gao, Dong Li, Xiu Li, and Ruiming Tang. 2023. Coarse-to-fine knowledge-enhanced multi-interest learning framework for multi-behavior recommendation. *ACM Transactions on Information Systems* 42, 1 (2023), 1–27.

[29] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*. 811–820.

[30] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. *Advances in neural information processing systems* 30 (2017).

[31] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1441–1450.

[32] Gábor J Székely and Maria L Rizzo. 2009. Brownian distance covariance. *The annals of applied statistics* (2009), 1236–1265.

[33] Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. 2007. Measuring and testing dependence by correlation of distances. (2007).

[34] Zhiquan Tan, Yifan Zhang, Jingqin Yang, and Yang Yuan. 2023. Contrastive Learning Is Spectral Clustering On Similarity Graph. *arXiv preprint arXiv:2303.15103* (2023).

[35] Yu Tian, Jianxin Chang, Yanan Niu, Yang Song, and Chenliang Li. 2022. When multi-level meets multi-interest: A multi-grained neural model for sequential recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1632–1641.

[36] Nhu-Thuat Tran and Hady W Lauw. 2023. Multi-Representation Variational Autoencoder via Iterative Latent Attention and Implicit Differentiation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 2462–2471.

[37] Hadrien Van Lierde, Tommy WS Chow, and Guanrong Chen. 2019. Scalable spectral clustering for overlapping community detection in large-scale networks. *IEEE Transactions on Knowledge and Data Engineering* 32, 4 (2019), 754–767.

[38] Ulrike Von Luxburg. 2007. A tutorial on spectral clustering. *Statistics and computing* 17 (2007), 395–416.

[39] Shoujin Wang, Liang Hu, Yan Wang, Longbing Cao, Quan Z Sheng, and Mehmet Orgun. 2019. Sequential recommender systems: challenges, progress and prospects. In *28th International Joint Conference on Artificial Intelligence, IJCAI 2019*. International Joint Conferences on Artificial Intelligence, 6332–6338.

[40] Xin Wang, Hong Chen, Yuwei Zhou, Jianxin Ma, and Wenwu Zhu. 2022. Disentangled representation learning for recommendation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 1 (2022), 408–424.

[41] Xin Wang, Hong Chen, and Wenwu Zhu. 2021. Multimodal disentangled representation for recommendation. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[42] Xiang Wang, Hongye Jin, An Zhang, Xiangnan He, Tong Xu, and Tat-Seng Chua. 2020. Disentangled graph collaborative filtering. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 1001–1010.

[43] Yong Wang, Yuan Jiang, Yi Wu, and Zhi-Hua Zhou. 2011. Spectral clustering on multiple manifolds. *IEEE Transactions on Neural Networks* 22, 7 (2011), 1149–1161.

[44] Yifan Wang, Suyao Tang, Yuntong Lei, Weiping Song, Sheng Wang, and Ming Zhang. 2020. Disenhan: Disentangled heterogeneous graph attention network for recommendation. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 1605–1614.

[45] Yuling Wang, Xiao Wang, Xiangzhou Huang, Yanhua Yu, Haoyang Li, Mengdi Zhang, Zirui Guo, and Wei Wu. 2023. Intent-aware recommendation via disentangled graph contrastive learning. In *Proceedings of the Thirty-Second International*

Joint Conference on Artificial Intelligence. 2343–2351.
[46] Zhibo Xiao, Luwei Yang, Wen Jiang, Yi Wei, Yi Hu, and Hao Wang. 2020. Deep multi-interest network for click-through rate prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2265–2268.
[47] Yueqi Xie, Jingqi Gao, Peilin Zhou, Qichen Ye, Yining Hua, Jae Boum Kim, Fangzhao Wu, and Sunghun Kim. 2023. Rethinking multi-interest learning for candidate matching in recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 283–293.
[48] Chengfeng Xu, Pengpeng Zhao, Yanchi Liu, Victor S Sheng, Jiajie Xu, Fuzhen Zhuang, Junhua Fang, and Xiaofang Zhou. 2019. Graph contextualized self-attention network for session-based recommendation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 3940–3946.
[49] Haochao Ying, Fuzhen Zhuang, Fuzheng Zhang, Yanchi Liu, Guandong Xu, Xing Xie, Hui Xiong, and Jian Wu. 2018. Sequential recommender system based on hierarchical attention network. In *IJCAI International Joint Conference on Artificial Intelligence*.
[50] Xu Yuan, Dongsheng Duan, Lingling Tong, Lei Shi, and Cheng Zhang. 2021. Icai-sr: Item categorical attribute integrated sequential recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1687–1691.
[51] Shengyu Zhang, Lingxiao Yang, Dong Yao, Yujie Lu, Fuli Feng, Zhou Zhao, Tat-seng Chua, and Fei Wu. 2022. Re4: Learning to re-contrast, re-attend, re-construct for multi-interest recommendation. In *Proceedings of the ACM Web Conference 2022*. 2216–2226.

8 APPENDIX

Table 6: Symbols and notations

Notation	Description
M, N	The number of users and items.
$\mathcal{U}$, $\mathcal{I}$	The whole set of users and items.
$\mathcal{E}_u$	The user u's recent κ sequential behaviors.
F	The the number of users' multi-interests.
K	The the number of item partition groups.
$\boldsymbol{m}_f$	The f-th interest of the user's preferences.
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	The item-confidence graph has nodes $\mathcal{V}$ representing item entities and edges $\mathcal{E}$ representing their relations.
S_{ij}	The confidence of item-to-item edges measured by frequency patterns of users' behaviors.
w_{ik}	The partition weight of the item i belongs to the group k.

Theorem 1. (Non-overlapped spectral clustering) The combinatorial optimization problem in Equation (3) is equivalent to spectral clustering based on the item-confidence graph,

$$OPT \propto -\sum_{k=1}^{K} Cut(\mathcal{A}_k, \bar{\mathcal{A}}_k) + N \cdot entropy(|\mathcal{A}_k|/N),$$

where $\{\mathcal{A}_1, \cdots, \mathcal{A}_K\}$ denote the item partition groups which satisfy $\cup_k \mathcal{A}_k = \mathcal{I}$ and $\mathcal{A}_{k_1} \cap \mathcal{A}_{k_2} = \emptyset$. $\bar{\mathcal{A}}_k = \mathcal{I} - A_k$ denote the complement set of $\bar{\mathcal{A}}_k$. $Cut(A_i, A_j)$ calculates the price (total weights) of cutting off all edges between $\mathcal{A}_{k_1}$ and nodes $\mathcal{A}_{k_2}$.

Proof. For the first term in Equation (3), we have

$$\begin{aligned}
&\sum_{i\neq j}\sum_{k=1}^{K} w_{ik}\cdot w_{jk}\cdot S_{ij} = \sum_{k=1}^{K}\sum_{i\neq j} w_{ik}\cdot w_{jk}\cdot S_{ij} \\
&\overset{(a)}{=} \sum_{k=1}^{K}\sum_{i,j\in\mathcal{A}_k} S_{ij} = \sum_{k=1}^{K} Cut(\mathcal{A}_k, \mathcal{A}_k) \\
&\overset{(b)}{=} \sum_{k=1}^{K}[Cut(\mathcal{I}, \mathcal{A}_k) - Cut(\mathcal{A}_k, \bar{\mathcal{A}}_k)] \\
&= Cut(\mathcal{I}, \mathcal{I}) - N\log N - \sum_{k=1}^{K}[Cut(\mathcal{A}_k, \bar{\mathcal{A}}_k)] \\
&\propto -\sum_{k=1}^{K} Cut(\mathcal{A}_k, \bar{\mathcal{A}}_k),
\end{aligned}$$

where Step (a) is established because of $w_{ik}\cdot w_{jk} = 1$ only and if only the pair of items i and j are partitioned into the same group k. Step (b) is established because of the linearity of the cut-off operation, i.e., $Cut(A - B, C) = Cut(A, C) - Cut(B, C)$ where $B \subseteq A$.

For the section term in Equation (3), we have

$$\begin{aligned}
&\sum_i \log\sum_j\sum_{k=1}^{K} w_{ik}\cdot w_{jk} = \sum_i \log\sum_{k=1}^{K}\sum_j w_{ik}\cdot w_{jk} \\
&\overset{(c)}{=} \sum_i \log\sum_{k=1}^{K} |\mathcal{A}_k|\cdot\mathbb{I}(i\in\mathcal{A}_k) = \sum_{k=1}^{K}\sum_{i\in\mathcal{A}_k}\log|\mathcal{A}_k| \\
&= \sum_{k=1}^{K}|\mathcal{A}_k|\log|\mathcal{A}_k| = N\cdot\sum_{k=1}^{K}\frac{|\mathcal{A}_k|}{N}\log|\mathcal{A}_k| \\
&= N\cdot\sum_{k=1}^{K}\frac{|\mathcal{A}_k|}{N}\log\frac{|\mathcal{A}_k|}{N} + N\cdot\sum_{k=1}^{K}\frac{|\mathcal{A}_k|}{N}\log N \\
&\propto -N\cdot entropy(\frac{|\mathcal{A}_k|}{N}),
\end{aligned}$$

where Step (c) is established because $\sum_j w_{ik}\cdot w_{jk} = |\mathcal{A}_k|$ if $i\in\mathcal{A}_k$ else $\sum_j w_{ik}\cdot w_{jk} = 0$.

Combining the two terms in Equation (3), we have

$$OPT \propto -\sum_{k=1}^{K} Cut(\mathcal{A}_k, \bar{\mathcal{A}}_k) + N\cdot entropy(|\mathcal{A}_k|/N). \tag{7}$$

□

Theorem 2. (Overlapped spectral clustering) Maximizing Equation (4) with $w_{ik}\in\mathbb{R}$ is equivalent to conducting overlapped spectral clustering based on the item-confidence graph, i.e.,

$$OPT = -tr(\boldsymbol{W}^\top\boldsymbol{L}\boldsymbol{W}) + \sum_i \log\frac{\exp(\sum_k w_{i,k}\cdot w_{i,k})}{\sum_j\exp(\sum_k w_{i,k}\cdot w_{j,k})},$$

where $L = I - S$ denote the Laplacian matrix of matrix $\boldsymbol{S}$. It is equivalent to doing spectral clustering with a repulsion regularization.

Proof.

$$\begin{aligned}
&Eq.(4) = tr((A_{mn})_{K\times K}) - \sum_i\log\sum_j\exp(\sum_{k=1}^{K} w_{ik}\ w_{jk}) \\
&= tr\left(\boldsymbol{W}^\top(\boldsymbol{S}-\boldsymbol{I})\boldsymbol{W}\right) + tr(\boldsymbol{W}^\top\boldsymbol{I}\boldsymbol{W}) - \sum_i\log\sum_j\exp(\sum_{k=1}^{K} w_{ik}\cdot w_{jk}) \\
&= -tr(\boldsymbol{W}^\top(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{W}) + \sum_i\log\frac{\exp(\sum_k w_{i,k}\cdot w_{i,k})}{\sum_j\exp(\sum_k w_{i,k}\cdot w_{j,k})},
\end{aligned}$$

where $A_{mn} = \sum_{i,j} w_{im}w_{jn}S_{ij}$ denote the m-the row and the n-th column of the matrix $(A_{mn})_{K\times K}$. $\boldsymbol{I}$ and $tr(\cdot)$ denote the identity matrix and the trace operation. □

Lemma 4.1 The first term in the item partition problem can be formulated as the expectation of $\mathbb{P}(\cdot|\boldsymbol{S})$ about $\boldsymbol{Z}$:

$$\sum_{i\neq j}\sum_{k=1}^{K} w_{ik}\cdot w_{jk}\cdot S_{ij} = \mathbb{E}_{\mathbb{P}(\boldsymbol{Y}|\boldsymbol{S})}\log\left(\Omega(\boldsymbol{Y})\cdot\prod_{Y_{ij}=1} Z_{ij}\right),$$

where $Z_{ij}=\exp(\sum_k w_{ik}w_{jk})$.

Proof.

$$\begin{aligned}
&\sum_{i\neq j}\sum_{k=1}^{K} w_{ik}\cdot w_{jk}\cdot S_{ij} = \sum_{i\neq j} S_{ij}\cdot\log\left(e^{\sum_{k=1}^{K} w_{ik}\cdot w_{jk}}\right)\\
&\overset{(a)}{=} \sum_{i\neq j}\mathbb{E}_{\mathbb{P}(\boldsymbol{Y}|\boldsymbol{S})}[\mathbb{I}(Y_{ij}=1)]\cdot\log Z_{ij}\\
&\overset{(b)}{=} \mathbb{E}_{\mathbb{P}(\boldsymbol{Y}|\boldsymbol{S})}\left(\sum_{i\neq j}\mathbb{I}(Y_{ij}=1)\log Z_{ij}\right)=\mathbb{E}_{\mathbb{P}(\boldsymbol{Y}|\boldsymbol{S})}\left(\sum_{Y_{ij}=1}\log Z_{ij}\right)\\
&\overset{(c)}{=} \mathbb{E}_{\mathbb{P}(\boldsymbol{Y}|\boldsymbol{S})}\log\left(\prod_{Y_{ij}=1} Z_{ij}\right)+\mathbb{E}_{\mathbb{P}(\boldsymbol{Y}|\boldsymbol{S})}\log\left(\Omega(\boldsymbol{Y})\right)\\
&= \mathbb{E}_{\mathbb{P}(\boldsymbol{Y}|\boldsymbol{S})}\log\left(\Omega(\boldsymbol{Y})\cdot\prod_{Y_{ij}=1} Z_{ij}\right),
\end{aligned}$$

where Step (a) is established according to $\mathbb{E}_{\mathbb{P}(\boldsymbol{Y}|\boldsymbol{R})}(\mathbb{I}(Y_{ij}=1))=R_{ij}$; Step (b) is established because Z_{ij} is independent of Y; Step (c) is established according to $\mathbb{E}_{\mathbb{P}(\boldsymbol{Y}|\boldsymbol{R})}\log(\Omega(\boldsymbol{Y}))=0$. □

Lemma 4.2 Maximize the second term in the item partition problem is equivalent to maximizing the aggregation of graphs whose nodes have exactly one out-going edge:

$$\sum_i\log\sum_j\exp(\sum_{k=1}^{K} w_{ik}\cdot w_{jk}) = -\log\left(\sum_{Y\in\mathcal{Y}}\Omega(\boldsymbol{Y})\cdot\prod_{Y_{ij}=1} Z_{ij}\right).$$

Proof.

$$\begin{aligned}
&-\sum_i\log\sum_j\exp(\sum_{k=1}^{K} w_{ik}\cdot w_{jk}) = -\sum_i\log\sum_j Z_{ij}\\
&= -\left(\log(\sum_j Z_{1j})+\cdots+\log(\sum_j Z_{Nj})\right)\\
&= -\log\left((\sum_j Z_{1j})\cdots(\sum_j Z_{Nj})\right) = -\log\left(\sum_{\pi\in\kappa(1,N)}\prod_i Z_{i\pi_i}\right)\\
&= -\log\left(\sum_{Y\in\mathcal{Y}}\Omega(\boldsymbol{Y})\cdot\prod_{\mathbb{I}(Y_{ij}=1)} Z_{ij}\right),
\end{aligned}$$

where $\kappa(1,N)$ denotes the permutation set that contains all permutations of $[1,\cdots,N]$. □

Theorem 3. (Approximation) Maximizing the proposed partition problem is equivalent to optimizing the cross-entropy loss between the two MRFs:

$$\max \text{OPT} \Leftrightarrow \max \mathbb{E}_{\mathbb{P}(\boldsymbol{Y}|\boldsymbol{S})}\log(\mathbb{P}(\boldsymbol{Y}|\boldsymbol{Z})),$$

where $\boldsymbol{S}$ and $\boldsymbol{Z}$ describe the relations in the item-confidence MRF and the item-partition MRF, respectively.

Proof. According to Lemma (4.1) and Lemma (4.2), maximizing the partition goals is equivalent to

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}(\boldsymbol{Y}|\boldsymbol{S})}\log\left(\Omega(\boldsymbol{Y})\prod_{Y_{ij}=1} Z_{ij}\right)-\log\left(\sum_{Y\in\mathcal{Y}}\Omega(\boldsymbol{Y})\prod_{Y_{ij}=1} Z_{ij}\right)\\
&= \mathbb{E}_{\mathbb{P}(\boldsymbol{Y}|\boldsymbol{S})}\log\left(\frac{\Omega(\boldsymbol{Y})\prod_{Y_{ij}=1} Z_{ij}}{\sum_{Y\in\mathcal{Y}}\Omega(\boldsymbol{Y})\prod_{Y_{ij}=1} Z_{ij}}\right)\\
&= \mathbb{E}_{\mathbb{P}(\boldsymbol{Y}|\boldsymbol{S})}\log(\mathbb{P}(\boldsymbol{Y}|\boldsymbol{Z})).
\end{aligned}$$

□

Theorem 4. (Contrastive learning) The proposed partition goals in Equation (3) and Equation (4) can be solved in a contrastive learning manner, i.e.,

$$\begin{aligned}
\mathcal{L}_{Partition} &= \mathbb{E}_{\mathbb{P}(\boldsymbol{Y}|\boldsymbol{S})}\log(\mathbb{P}(\boldsymbol{Y}|\boldsymbol{Z}))\\
&= \sum_i \mathbb{E}_{\mathbb{M}(\boldsymbol{Y_i}|\boldsymbol{S_i})}\log\frac{\exp(\sum_k w_{ik}w_{jk})}{\sum_{j'}\exp(\sum_k w_{ik}w_{j'k})},
\end{aligned}$$

where Y_i is an one-hot encoding vector with the j-th element $Y_{ij}=1$.

Proof.

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}(\boldsymbol{Y}|\boldsymbol{S})}\log(\mathbb{P}(\boldsymbol{Y}|\boldsymbol{Z})) \overset{(a)}{=} \mathbb{E}_{\mathbb{P}(\boldsymbol{Y}|\boldsymbol{S})}\log\left(\prod_i\mathbb{M}(\boldsymbol{Y_i}|\boldsymbol{Z_i})\right)\\
&= \sum_i\mathbb{E}_{\mathbb{P}(\boldsymbol{Y}|\boldsymbol{S})}\log\left(\mathbb{M}(\boldsymbol{Y_i}|\boldsymbol{Z_i})\right) = \sum_i\mathbb{E}_{\prod_i[\mathbb{M}(\boldsymbol{Y_i}|\boldsymbol{S_i})]}\log\left(\mathbb{M}(\boldsymbol{Y_i}|\boldsymbol{Z_i})\right)\\
&\overset{(b)}{=} \sum_i\mathbb{E}_{\mathbb{M}(\boldsymbol{Y_i}|\boldsymbol{S_i})}\log(\mathbb{M}(\boldsymbol{Y_i}|\boldsymbol{Z_i})) = \sum_i\mathbb{E}_{\mathbb{M}(\boldsymbol{Y_i}|\boldsymbol{S_i})}\log\frac{Z_{ij}}{\sum_{j'}Z_{ij'}}\\
&= \sum_i\mathbb{E}_{\mathbb{M}(\boldsymbol{Y_i}|\boldsymbol{S_i})}\log\frac{\exp(\sum_k w_{ik}w_{jk})}{\sum_{j'}\exp(\sum_k w_{ik}w_{j'k})}.
\end{aligned}$$

Step (a) replaces the $\mathbb{P}(\cdot)$ with $\mathbb{M}(\cdot)$ according to Property 2 in MRFs; Step (b) is obvious established because variables $(\boldsymbol{Y_1},\cdots,\boldsymbol{Y_N})$ are independent with each other. □