# Aggregately Diversified Bundle Recommendation via Popularity Debiasing and Configuration-aware Reranking

Hyunsik Jeon, Jongjin Kim, Jaeri Lee, Jong-eun Lee, and U Kang

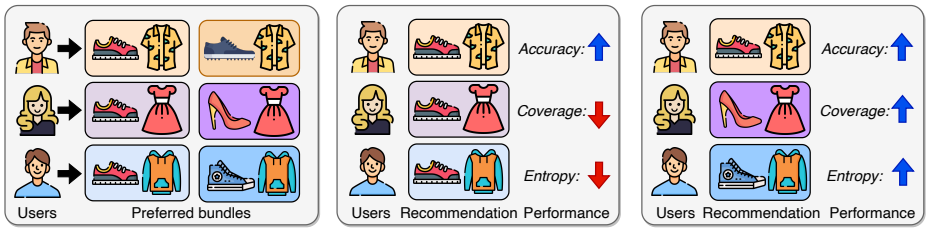
Seoul National University, Seoul, South Korea
{jeon185,j2kim99,jlunits2,kjayjay40,ukang}@snu.ac.kr

**Abstract.** How can we expose diverse items across all users while satisfying their needs in bundle recommendations? Diversified bundle recommendation is a crucial task since it leads to great benefits for both sellers and users. However, there have been no studies on aggregate diversity in bundle recommendation, while they have been intensively studied in item recommendation. Moreover, existing methods of aggregately diversified item recommendation are not fully suitable for bundle recommendation. In this paper, we propose PopCon (Popularity Debiasing and Configuration-aware Reranking), an accurate method for aggregately diversified bundle recommendation. PopCon mitigates the popularity bias of a recommendation model by a popularity-based negative sampling in training process, and maximizes accuracy and aggregate diversity by a configuration-aware reranking algorithm. We show that PopCon provides state-of-the-art performance on real-world datasets, achieving up to 60.5% higher Entropy@5 and 3.92× higher Coverage@5 with comparable accuracies compared to the best competitor.

**Keywords:** Bundle Recommendation · Aggregate Diversity · Popularity Debiasing · Configuration-aware Reranking

## 1 Introduction

*How can we expose diverse items across all users as well as satisfying their needs in bundle recommendations?* Recommender systems [16,10,13,9] have been indispensable techniques in online platforms providing customers with several relevant items from numerous ones [20]. Bundle recommendation aims to suggest sets of items instead of individual ones to users. It has been gaining attention in online platforms due to its advantage of providing items that customers need with one-stop convenience [14]. Furthermore, bundles are ubiquitous in real-world scenarios because they provide effectual marketing strategies (e.g., discount sales) which are appealing to customers [6]. However, traditional bundle recommendation models [18,3,5,6,4,14,8] have focused only on accuracy without paying attention to diversity. Fig. 1 compares the traditional bundle recommendation and an aggregately diversified bundle recommendation. Note that aggregate diversity is measured by the degree of fair exposure of items (i.e., coverage and




(a) Ground-truth preferences of users

(b) Traditional bundle recommendation

(c) Aggregately diversified bundle recommendation

Fig. 1: Illustrative comparison of (b) traditional bundle recommendation and (c) aggregately diversified bundle recommendation when the (a) ground-truth preferences of users are given.

entropy) in recommendation results across all users. As shown in Fig. 1(b), the traditional bundle recommendation, despite achieving high accuracy, results in a low aggregate diversity by recommending bundles that contain a popular item (e.g., the red shoes). On the other side, as shown in Fig. 1(c), the *aggregately diversified bundle recommendation (our task)* further aims to achieve high aggregate diversity by exposing diverse items across all users.

In the last decade, there have been several studies for aggregate diversity in item recommendation. Reranking-based methods [2,11,15,7], which rerank the recommendation results of a trained model to achieve both high accuracy and high aggregate diversity, are the most prevailing approaches in aggregately diversified item recommendation owing to their effectiveness in handling aggregate diversity. However, they are not fully suitable for bundle recommendation due to the following two limitations. First, a bundle recommendation model used as a backbone is easily overfitted to some popular bundles, and thus relying on the backbone model's results inevitably results in sacrificing a lot of accuracies to increase aggregate diversity. Second, they do not consider the configuration of bundles which is pivotal information to address the diversity of item exposure in bundle recommendation.

We propose PopCon (Popularity Debiasing and Configuration-aware Reranking), an accurate method for aggregately diversified bundle recommendation. PopCon consists of two phases, model training and reranking. In the training phase, PopCon trains a bundle recommendation model as a backbone with a popularity-based negative sampling to mitigate the popularity bias of the model. In the reranking phase, PopCon reranks the recommendation result of the models to maximize both accuracy and aggregate diversity. PopCon exploits each bundle's configuration to effectively deal with the aggregate diversity in the reranking phase. The contributions of PopCon are summarized as follows.

- **Problem.** To the best of our knowledge, our work is the first study that focuses on aggregately diversified bundle recommendation, which is of large importance in real-world scenarios.
- **Method.** We propose PopCon, an accurate method for aggregately diversified bundle recommendation. PopCon mitigates the popularity bias of a backbone model via a popularity-based negative sampling and maximizes the accuracy and aggregate diversity by a configuration-aware reranking.

– **Experiments.** Extensive experiments on three real-world datasets show that PopCon provides state-of-the-art performance achieving up to 60.5% higher Entropy@5 and 3.92× higher Coverage@5 with comparable accuracies compared to the best competitor.

## 2 Problem Definition and Related Works

### 2.1 Problem Definition

Bundle recommendation aims to predict sets of items, instead of individual items, that users would prefer. In this work, we focus on aggregate diversity in the bundle recommendation. We give the formal definition of the problem, namely aggregately diversified bundle recommendation, as Problem 1.

*Problem 1 (Aggregately diversified bundle recommendation).* Let $\mathcal{U}$, $\mathcal{I}$, and $\mathcal{B}$ be the sets of users, items, and bundles, respectively. We have matrices of user-bundle interactions, user-item interactions, and bundle-item affiliations which are denoted as $\mathbf{X} = [x_{ub}] \in \mathbb{R}^{|\mathcal{U}|\times|\mathcal{B}|}$, $\mathbf{Y} = [y_{ui}] \in \mathbb{R}^{|\mathcal{U}|\times|\mathcal{I}|}$, and $\mathbf{Z} = [z_{bi}] \in \mathbb{R}^{|\mathcal{B}|\times|\mathcal{I}|}$, respectively. $x_{ub}, y_{ui}, z_{bi} \in \{0, 1\}$ are binary values, indicating an observation or a non-observation of interaction or affiliation. Then, the problem is to recommend a list of $k$ bundles to each user $u$ as $\mathbf{r}_u(k) \subset \{b|b \in \mathcal{B}, x_{ub} = 0\}$, which have not been observed in the user-bundle interactions. The goal is to make $\mathbf{r}_u(k)$ accurate for each user $u$, and to make the overall recommendation results $\mathbf{R}(k) = (\mathbf{r}_1(k), \cdots, \mathbf{r}_{|\mathcal{U}|}(k))$ aggregately diverse.

The aggregate diversity is evaluated for the items in $\mathbf{R}(k)$ by two metrics.

– **Coverage** measures how many different items are contained in the results.

$$Coverage@k = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} app(i, \mathbf{R}(k)), \tag{1}$$

where $app(i, \mathbf{R}(k)) = [i \in \bigcup_{b \in \mathbf{R}(k)} \Omega_b]$ indicates whether item $i$ appears in $\mathbf{R}(k)$. $\Omega_b = \{i|i \in \mathcal{I}, z_{bi} = 1\}$ is the set of bundle $b$'s constituent items. The Iverson bracket $[\cdot]$ returns 1 if the statement is true, 0 otherwise.

– **Entropy** measures how evenly all items appear in the results.

$$Entropy@k = -\sum_{i \in \mathcal{I}} p(i, \mathbf{R}(k)) \log p(i, \mathbf{R}(k)), \tag{2}$$

where $p(i, \mathbf{R}(k)) = \frac{Freq(i, \mathbf{R}(k))}{\sum_{j \in \mathcal{I}} Freq(j, \mathbf{R}(k))}$. $Freq(i, \mathbf{R}(k)) = \sum_{u \in \mathcal{U}} freq(i, \mathbf{r}_u(k))$ where $freq(i, \mathbf{r}_u(k)) = \sum_{b \in \mathbf{r}_u(k)} [i \in \Omega_b]$ indicates item $i$'s frequency in user $u$'s recommended bundles $\mathbf{r}_u(k)$.

### 2.2 Related Works

**Bundle recommendation.** Bundle recommendation aims to recommend a set of items instead of an individual one to users. Existing bundle recommendation methods are mainly divided into matrix factorization-based approaches [18,3,5] and graph learning-based approaches [6,4,14]. BR [18] and EFM [3] jointly factorize user-item and user-bundle interactions to predict unseen user-bundle interactions. DAM [5] further introduces an attention mechanism to effectively learn

bundle embeddings. With the proliferation of graph learning approaches, several studies [6,4,14] formulate the bundle recommendation in a tripartite graph with nodes of users, items, and bundles. BundleNet [6] learns a graph convolutional network to predict interactions between the nodes, while BGCN [4] further decomposes user preferences into item-view and bundle-view to effectively predict the interactions. CrossCBR [14] captures cooperative association between the item-view and bundle-view by a contrastive learning method to improve performance. However, such previous works for bundle recommendation focus only on accuracy. In this work, we further address aggregate diversity which is of great importance but makes the problem more challenging.

**Aggregately diversified recommendation.** Aggregately diversified recommendation aims to increase diversity of recommendations across all users [2,12]. It is important to accomplish high aggregate diversity because it alleviates the long tail problems and maximizes the profit of the sales platform. Most existing methods for aggregately diversified recommendations modify the results of a backbone model to achieve high aggregate diversity since it is difficult to optimize the model both for accuracy and diversity. Kwon et al. [2] rerank the recommendation results of a backbone model based on item popularity and heuristic thresholds of scores. Karakaya et al. [11] replace recommended items with similar ones through a random walk on an item co-occurrence graph. FairMatch [15] finds high-quality but less frequently recommended items in a recommendation list by solving the maximum flow problem. UImatch [7] constrains the limit of each item and solves the matching problem with a greedy strategy. However, there has been no study of aggregate diversity for bundle recommendation, which is crucial in practical scenarios but more challenging to address.

## 3 Proposed Method

In this section, we propose PopCon (Popularity Debiasing and Configuration-aware Reranking) to address the aggregately diversified bundle recommendation.

### 3.1 Overview

We concentrate on the following challenges to achieve high aggregate diversity with comparable accuracy in bundle recommendation.

C1. **Mitigating popularity bias of a backbone model.** A bundle recommendation model easily overfits to some popular bundles. How can we mitigate the popularity bias of the backbone model?

C2. **Fitting two opposite criteria, accuracy and diversity.** It is challenging to fit accuracy and diversity simultaneously since they are opposite criteria. How can we satisfy both opposite criteria?

C3. **Simultaneously considering how many items appear and how evenly items appear.** To achieve high aggregate diversity, we need to consider not only whether items appear or not, but whether items appear evenly. How can we consider both simultaneously?

The main ideas of PopCon are summarized as follows.

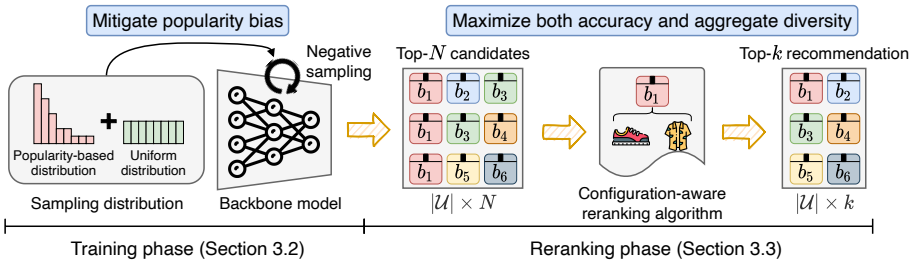
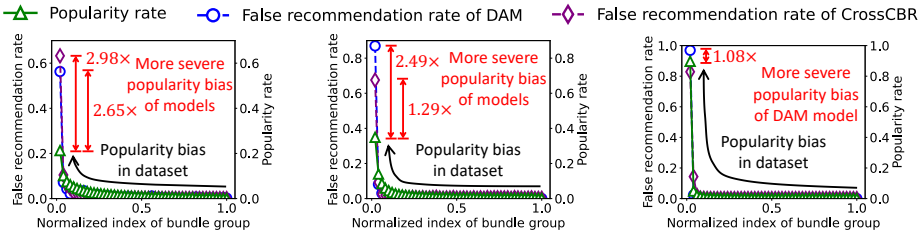



Fig. 2: Overview of PopCon which consists of training and reranking phases.




(a) NetEase (b) Youshu (c) Steam

Fig. 3: Popularity bias of real-world datasets (NetEase, Youshu, and Steam) and that of trained models on them. The datasets entail popularity biases, and the trained models have more severe ones.

I1. **Popularity-based negative sampling.** It mitigates the popularity bias of a backbone model and enables us to effectively leverage the user-bundle relationship scores.

I2. **Accuracy-prioritized coupling.** It enables us to retain high-scored bundles in recommendation results and replace low-scored bundles with more diverse ones.

I3. **Maximizing the gains of coverage and entropy.** It encourages bundles that have not been recommended and that are less recommended to be recommended more.

Fig. 2 shows the overall process of PopCon. PopCon consists of two phases, model training phase and reranking phase. In the training phase, PopCon trains a bundle recommendation model such as DAM [5] or CrossCBR [14] as a backbone while mitigating its popularity bias by a popularity-based negative sampling. In the reranking phase, PopCon selects candidate bundles for each user and reranks the candidates by a configuration-aware reranking algorithm to maximize both accuracy and aggregate diversity.

### 3.2 Training Phase with Popularity Debiasing

The objective of the training phase is to train a model $f(u, b)$ that accurately predicts the score between user $u$ and bundle $b$. We first investigate the popularity bias of traditional models and propose a popularity-based negative sampling to mitigate the popularity bias of the models.

Real-world datasets for bundle recommendation commonly entail popularity bias because of various factors such as exposure mechanisms and public opin-

ions. Accordingly, bundle recommendation models suffer from the popularity bias in their output [1]. Fig. 3 shows the popularity bias of real-world datasets and that of trained models. We train DAM [5] and CrossCBR [14] which are state-of-the-art bundle recommendation models on real-world datasets. For each dataset, we split bundles into 50 groups in the order of their popularity, and sum up the number of incorrect recommendations for each group's bundles in the top-5 recommendation of the model. As shown in the figure, the real-world datasets entail the popularity bias (i.e., long-tail problem [17]) and the trained recommendation models emphasize popular items, showing their vulnerability to the popularity bias. The popularity bias of the model gives incorrect information about user-bundle relationships because popular bundles easily receive high scores regardless of user preferences, and makes it challenging to achieve high aggregate diversity when using the predicted scores in the reranking phase.

We propose a popularity-based negative sampling in training process to mitigate the popularity bias of a backbone model. Assume we have matrices of user-bundle interactions, user-item interactions, and bundle-item affiliations as $\mathbf{X}=[x_{ub}]\in\mathbb{R}^{|\mathcal{U}|\times|\mathcal{B}|}$, $\mathbf{Y}=[y_{ui}]\in\mathbb{R}^{|\mathcal{U}|\times|\mathcal{I}|}$, and $\mathbf{Z}=[z_{bi}]\in\mathbb{R}^{|\mathcal{B}|\times|\mathcal{I}|}$, respectively. $\mathcal{U}$, $\mathcal{B}$, and $\mathcal{I}$ are the sets of users, bundles, and items, respectively. Then, a bundle recommendation model $f$ aims to predict the scores of user-bundle pairs. Specifically, the model $f$ is defined as matrix factorization-based [18,3,5] or graph-based frameworks [6,4,14] to utilize $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$. Then, the model $f$ is trained by minimizing the Bayesian Personalized Ranking (BPR) loss [19] as follows:

$$\sum_{(u,b,b')\in D} -\ln\sigma\left(f(u,b)-f(u,b')\right), \tag{3}$$

where $D=\{(u,b,b')|u\in\mathcal{U},b\in\mathcal{B},b'\in\mathcal{B},x_{ub}=1,x_{ub'}=0\}$, and $\sigma(\cdot)$ is the sigmoid function. In Equation (3), $b$ is a positive sample which user $u$ has interacted with, whereas $b'$ is a negative sample which user $u$ has not interacted with. However, the previous works [18,3,5,6,4,14] sample the negative bundles $b'$ from the uniform distribution although popular bundles are more likely to be picked as positive samples. This makes the model overfit to some popular bundles and causes the popularity bias as in Fig. 3. To mitigate the popularity bias, we increase the probability that popular bundles are selected as negative samples. We propose the probability of sampling negative bundle $b'$ as follows:

$$p(b')=\alpha\frac{freq(b')}{\sum_{j\in\mathcal{B}}freq(j)}+(1-\alpha)\frac{1}{|\mathcal{B}|}, \tag{4}$$

where $freq(j)$ is the number of bundle $j$'s interactions (i.e., number of non-zeros in $\mathbf{X}$'s $j$th column), $\alpha\in[0,1]$ is a balancing hyper-parameter between the popularity-based distribution and the uniform distribution. If $\alpha$ is large, the sampling probability of a bundle is largely affected by its popularity, whereas if $\alpha$ is small, a bundle is selected almost uniformly regardless of its popularity.

### 3.3 Reranking Phase with Configuration-Awareness

The objective of the reranking phase is to maximize both accuracy and aggregate diversity using the trained backbone model $f$. We first select top-$N$

candidate bundles for each user $u$ using the scores $f(u,b) = \hat{x}_{ub} \in \mathbb{R}$ of all bundles $b \in \mathcal{B}$. Then, we rerank the candidate bundles to recommend $k$ bundles ($N \gg k$) for each user. Specifically, we select the most suitable bundle among the candidates for each user and repeat it $k$ times. The main challenge in the reranking phase is to measure which bundle is the best for user $u$ at each time in terms both of accuracy and aggregate diversity.

It is straightforward to select the best bundle using a single criterion: accuracy or aggregate diversity. Assume we consider the candidate bundle $b$ for user $u$ currently. We compare $\sigma(\hat{x}_{ub})$ of each candidate to obtain the best accuracy because it measures how bundle $b$ is appropriate for user $u$. To obtain the best aggregate diversity, we simultaneously measure the gains of coverage and entropy when recommending a bundle and select the one that maximizes it. Specifically, we propose to compare $DivGain(b, \hat{\mathbf{R}}(k)) \in \mathbb{R}$, which considers the appearance of new items and the fair appearance of items, as follows:

$$DivGain(b, \hat{\mathbf{R}}(k)) = \frac{1}{2} CovGain(b, \hat{\mathbf{R}}(k)) + \frac{1}{2} EntGain(b, \hat{\mathbf{R}}(k)), \tag{5}$$

where $DivGain(b, \hat{\mathbf{R}}(k)), CovGain(b, \hat{\mathbf{R}}(k))$, and $EntGain(b, \hat{\mathbf{R}}(k)) \in \mathbb{R}$ denote the gains of aggregate diversity, coverage, and entropy, respectively, and $\hat{\mathbf{R}}(k)$ is the current recommendation results for all users. $CovGain(b, \hat{\mathbf{R}}(k)) \in [0, 1]$ and $EntGain(b, \hat{\mathbf{R}}(k)) \in [-1, 1]$ are measured as the changes of Equations (1) and (2), respectively, when adding a bundle $b$ to the current recommendation result $\hat{\mathbf{R}}(k)$; we obtain $EntGain(b, \hat{\mathbf{R}}(k))$ by dividing the original entropy gain by the maximum entropy so that the resulting value is in $[-1, 1]$.

However, the main difficulty of the reranking is to select the best bundle by measuring the accuracy and aggregate diversity simultaneously. For example, for user $u$, if $\sigma(\hat{x}_{ub}) > \sigma(\hat{x}_{ub'})$ and $DivGain(b, \hat{\mathbf{R}}(k)) < DivGain(b', \hat{\mathbf{R}}(k))$, it is difficult to decide which bundle should be recommended. It is essentially challenging because the accuracy and aggregate diversity are opposite in most cases. For instance, popular bundles usually provide high accuracy scores but less aggregate diversity scores.

**Desired properties.** To deal with this conflict, we propose three desired properties for a measurement function $g(u, b, \hat{\mathbf{R}}(k))$, which is used to select the best bundle $b$ for user $u$ and the current recommendation results $\hat{\mathbf{R}}(k)$.

*Property 1 (Increasing for accuracy).* The function should satisfy $g(u, b, \hat{\mathbf{R}}(k)) \geq g(u, b', \hat{\mathbf{R}}(k))$ if $\sigma(\hat{x}_{ub}) > \sigma(\hat{x}_{ub'})$ and $DivGain(b, \hat{\mathbf{R}}(k)) = DivGain(b', \hat{\mathbf{R}}(k))$.

*Property 2 (Increasing for diversity).* The function should satisfy $g(u, b, \hat{\mathbf{R}}(k)) \geq g(u, b', \hat{\mathbf{R}}(k))$ if $\sigma(\hat{x}_{ub}) = \sigma(\hat{x}_{ub'})$ and $DivGain(b, \hat{\mathbf{R}}(k)) > DivGain(b', \hat{\mathbf{R}}(k))$.

Properties 1 and 2 are essential because they allow fair comparisons for accuracy and aggregate diversity when the other metrics are the same. One candidate measurement function to satisfy both Properties 1 and 2 are as follows.

$$g(u, b, \hat{\mathbf{R}}(k)) = (1 - \beta)\sigma(\hat{x}_{ub}) + \beta DivGain(b, \hat{\mathbf{R}}(k)), \tag{6}$$

where $\beta \in [0, 1]$ is a balancing hyper-parameter. Equation (6) is a weighted sum of the accuracy and aggregate diversity terms to measure two criteria together.

On the other hand, it is also necessary to ensure that bundles that users like a lot are recommended regardless of the gains of aggregate diversity to satisfy the users. This is challenging in our task because accuracy and aggregate diversity are opposite in most cases. Thus, we need to reduce the influence of the gain of aggregate diversity as the accuracy increases. In this regard, we propose a property of accuracy priority as follows.

*Property 3 (Accuracy priority).* The function should satisfy $\frac{\partial g(u,b,\hat{\mathbf{R}}(k))}{\partial DivGain(b,\hat{\mathbf{R}}(k))} < \frac{\partial g(u,b',\hat{\mathbf{R}}(k))}{\partial DivGain(b',\hat{\mathbf{R}}(k))}$ if $\sigma(\hat{x}_{ub}) > \sigma(\hat{x}_{ub'})$.

**Accuracy-prioritized coupling.** We propose a measurement function $g$ that satisfies all the desired properties by prioritizing accuracy as follows.

$$g(u, b, \hat{\mathbf{R}}(k)) = \sigma(\hat{x}_{ub})^{\beta} + (1 - \sigma(\hat{x}_{ub})^{\beta}) DivGain(b, \hat{\mathbf{R}}(k)), \tag{7}$$

where $\beta \geq 1$ is a balancing hyper-parameter. If $\beta$ is small, the recommendation result is highly dependent on accuracy, and if $\beta$ is large, it is highly dependent on aggregate diversity because $\sigma(\hat{x}_{ub}) \in [0, 1]$. We show in Lemmas 1, 2, and 3 that Equation (7) satisfies all the desired properties. In the Lemmas, we denote $\sigma(\hat{x}_{ub})$ as $A(b)$, $DivGain(b, \hat{\mathbf{R}}(k))$ as $D(b)$, and $g(u, b, \hat{\mathbf{R}}(k))$ as $G(b)$ for brevity.

**Lemma 1.** *Equation* (7) *satisfies Property 1.*

*Proof.* If $A(b) > A(b')$ and $D(b) = D(b')$, then $G(b) - G(b') = (A(b)^{\beta} - A(b')^{\beta})(1 - D(b))$. Thus, $G(b) \geq G(b')$ because $A(b)^{\beta} > A(b')^{\beta}$ and $D(b) \leq 1$.

**Lemma 2.** *Equation* (7) *satisfies Property 2.*

*Proof.* If $A(b) = A(b')$ and $D(b) > D(b')$, then $G(b) - G(b') = (1 - A(b)^{\beta})(D(b) - D(b'))$. Thus, $G(b) \geq G(b')$ because $A(b)^{\beta} \leq 1$ and $D(b) > D(b')$.

**Lemma 3.** *Equation* (7) *satisfies Property 3.*

*Proof.* $\frac{\partial G(b)}{\partial D(b)} = 1 - A(b)^{\beta}$. Thus, $\frac{\partial G(b)}{\partial D(b)} < \frac{\partial G(b')}{\partial D(b')}$ if $A(b) > A(b')$.

Note that Equation (6) does not satisfy Property 3 because its $\frac{\partial G(b)}{\partial D(b)}$ is a constant value $\beta$, although it satisfies Properties 1 and 2.

**Reranking algorithm.** We repeat recommending the most suitable bundle among the candidate bundles to each user, $k$ times. Specifically, let the current recommendation results be $\hat{\mathbf{R}}(k) = (\hat{\mathbf{r}}_1(k), \hat{\mathbf{r}}_2(k), \cdots, \hat{\mathbf{r}}_{|\mathcal{U}|}(k))$, where $\hat{\mathbf{r}}_u(k)$ is the current recommendation result for user $u$; $\hat{\mathbf{r}}_u(k)$ for every $u \in \mathcal{U}$ is empty at the initial state. In random order of users $u \in \mathcal{U}$, we add $b' = \arg\max_b g(u, b, \hat{\mathbf{R}}(k))$ to $\hat{\mathbf{r}}_u(k)$ among $u$'s candidate $N$ bundles. We adopt a mini-batch technique that randomly selects $m$ users in every step. We repeat this process $k$ times, and finally obtain the recommendation results $\mathbf{R}(k)$.

Table 1: Summary of bundle recommendation datasets. U, B, and I indicate users, bundles, and items, respectively.

| Dataset | #U | #B | #I | #U-B (dens.) | #U-I (dens.) | #B-I (dens.) | Avg. B size |
|---|---|---|---|---|---|---|---|
| Steam[1] | 29,634 | 615 | 2,819 | 87,565 (0.48%) | 902,967 (1.08%) | 3,541 (0.20%) | 5.76 |
| Youshu[2] | 8,039 | 4,771 | 32,770 | 51,377 (0.13%) | 138,515 (0.05%) | 176,667 (0.11%) | 37.03 |
| NetEase[3] | 18,528 | 22,864 | 123,628 | 302,303 (0.07%) | 1,128,065 (0.05%) | 1,778,838 (0.06%) | 77.80 |

[1] https://github.com/technoapurva/Steam-Bundle-Recommendation
[2] https://github.com/yliuSYSU/DAM
[3] https://github.com/cjx0525/BGCN

# 4 Experiments

In this section, we perform experiments to answer the following questions.

Q1. **Performance Trade-off (Section 4.2).** Does PopCon provide the best trade-off between accuracy and aggregate diversity?
Q2. **Ablation Study (Section 4.3).** How do the main ideas in PopCon help improve the performance?
Q3. **Effects of number of candidates (Section 4.4).** How does the number $N$ of candidates affect the performance of PopCon?

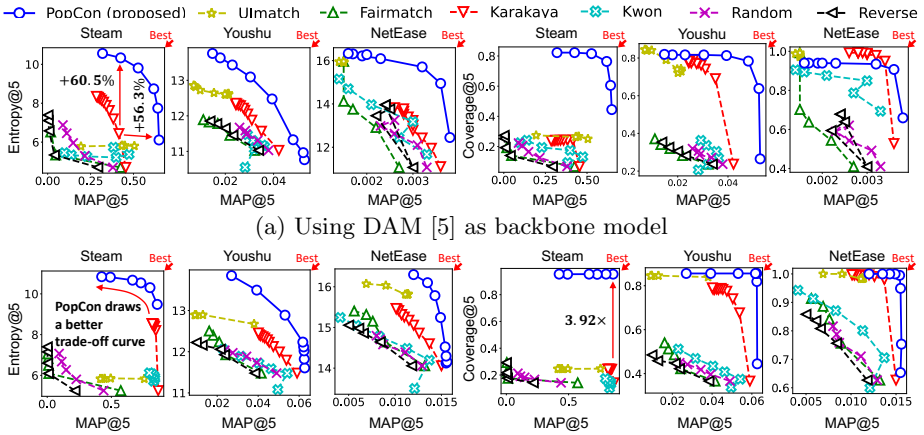
## 4.1 Experimental Setup

**Datasets.** We use three real-world datasets of bundle recommendation as summarized in Table 1. Steam [18] is constructed from Australian Steam community, a video game distribution platform. Youshu [5] is constructed from Youshu, a book review site. Netease [3] is constructed from Netease, a cloud music service.

**Baselines.** We compare PopCon with six baselines of aggregately diversified recommendation. Given a recommendation list of size $N(N > k)$ for each user, Reverse and Random pick bottom-$k$ bundles and random-$k$ bundles, respectively. Kwon [2] heuristically replaces the popular bundles of a recommendation list with unpopular ones. Karakaya [11] replaces bundles in a recommendation list with other bundles through random walk on an item co-occurrence network. Fairmatch [15] handles the maximum flow problem to replace bundles in a recommendation list with other bundles. UImatch [7] assigns capacity of each bundle to be recommended and generates a recommendation list in a greedy manner.

**Backbone models.** We leverage two existing bundle recommendation models, DAM [5] and CrossCBR [14], as backbone models of PopCon and the baselines. DAM and CrossCBR are the state-of-the-art models among matrix factorization-based methods and graph learning-based methods, respectively.

**Evaluation metrics.** We employ leave-one-out protocol [5] where one of each user's interactions is randomly selected for testing. We evaluate the performance in two criteria, accuracy and aggregate diversity. We use mean average precision (MAP@$k$) for the accuracy, and Coverage@$k$ and Entropy@$k$ for the aggregate diversity. MAP@$k$ considers highly ranked bundles more importantly for accuracy. Coverage@$k$ and Entropy@$k$ are explained in Section 2.1. We investigate the trade-off curve between accuracy and aggregate diversity. We set the number $k$ of bundles to 5, which is the most widely used setting.




(a) Using DAM [5] as backbone model

(b) Using CrossCBR [14] as backbone model

Fig. 4: PopCon outperforms baselines in most cases using the two different backbone models (a) DAM [5] and (b) CrossCBR [14].

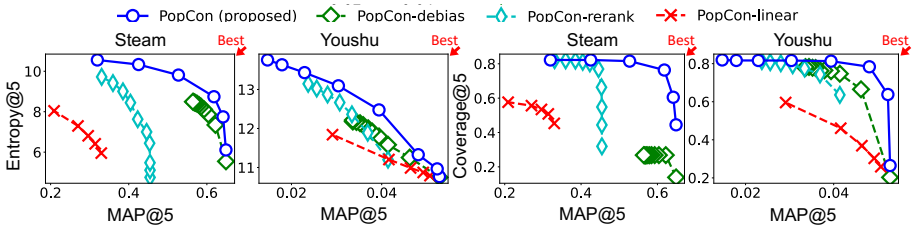
**Hyperparameters.** We set the embedding dimensionality of DAM and CrossCBR to 20. We set the batch size $m$ in the reranking phase to 10. For both DAM and CrossCBR, we set $\alpha$ to 0.1, 0.05, and 0.02 on Steam, Youshu, and NetEase, respectively. In Sections 4.2 and 4.3, we set $N$ to 100, 1,000, and 1,000 on Steam, Youshu, and NetEase, respectively. For each curve, $\beta$ is not a fixed value but controls the trade-off between accuracy and aggregate diversity.

### 4.2 Performance Trade-off (Q1)

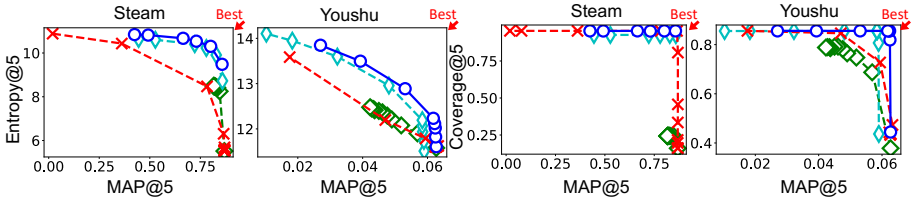
We compare PopCon and baselines on real-world datasets in Fig. 4. As shown in the figure, PopCon outperforms the baselines noticeably, drawing better trade-off curves between accuracy and aggregate diversity than all baselines in most cases. Especially, PopCon using DAM backbone achieves up to 60.5% higher Entropy@5 with comparable MAP@5, and up to 56.3% higher MAP@5 with comparable Entropy@5 compared with the best competitor Karakaya on Steam dataset. Furthermore, PopCon using CrossCBR achieves 3.92× higher Coverage@5 than Karakaya with similar MAP@5 on Steam dataset.

### 4.3 Ablation Study (Q2)

Fig. 5 provides an ablation study that compares PopCon with its three variants PopCon-debias, PopCon-rerank, and PopCon-linear on Steam and Youshu datasets. PopCon-debias adopts the proposed popularity debiasing in the training phase, but utilizes Karakaya in the reranking phase. PopCon-rerank does not adopt the popularity debiasing in the training phase while utilizing the proposed reranking algorithm in the reranking phase. PopCon-linear uses Equation (6) instead of Equation (7) in the reranking phase. As shown in the figure, PopCon outperforms all the variants, which verifies all the main ideas help improve the performance. Especially, PopCon-linear shows a severe performance drop compared with PopCon, justifying the importance of satisfying Property 3 (accuracy priority) in aggregately diversified bundle recommendation.




(a) Usage of DAM [5] as the backbone model




(b) Usage of CrossCBR [14] as the backbone model

Fig. 5: All the main ideas of PopCon help improve the performance.

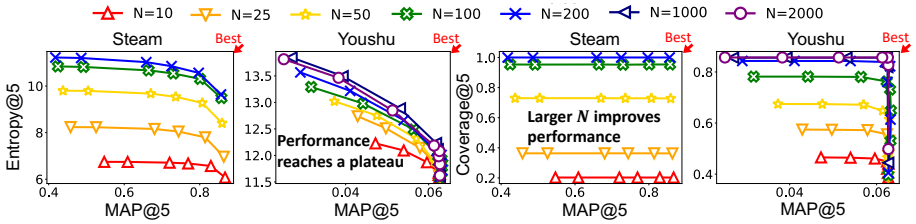



Fig. 6: The performance improves as $N$ increased and reaches a plateau eventually. CrossCBR is used as the backbone of PopCon.

### 4.4 Effects of Number of Candidates (Q3)

Fig. 6 shows the effects of the number $N$ of candidates for the performance of PopCon using CrossCBR on Steam and Youshu datasets. We set $N$ up to 200 on Steam dataset because Steam contains much fewer amount of bundles than Youshu. As shown in the figure, Entropy@5 and Coverage@5 are significantly improved as $N$ increased, and finally reaches a plateau. Thus, we set $N$ to 100 and $1,000$ on Steam and Youshu, respectively, since they provide sufficient high performance despite being far lower than the total number of bundles.

## 5 Conclusion

In this paper, we propose PopCon, an accurate method for aggregately diversified bundle recommendation. PopCon mitigates the popularity bias of a backbone model using a popularity-based negative sampling, and reranks the recommendation results of the backbone model by a configuration-aware reranking algorithm to simultaneously maximize accuracy and aggregate diversity. PopCon provides the state-of-the-art performance in aggregately diversified bundle recommendation, achieving up to 60.5% higher Entropy@5 and 3.92× higher Coverage@5 with comparable accuracies compared to the best competitor.

## Acknowledgments

This work was supported by Jung-Hun Foundation. The Institute of Engineering Research and ICT at Seoul National University provided research facilities for this work. U Kang is the corresponding author.

## References

1. Abdollahpouri, H., Burke, R., Mobasher, B.: Controlling popularity bias in learning-to-rank recommendation. In: RecSys (2017)
2. Adomavicius, G., Kwon, Y.: Improving aggregate recommendation diversity using ranking-based techniques. IEEE Trans. Knowl. Data Eng. (2012)
3. Cao, D., Nie, L., He, X., Wei, X., Zhu, S., Chua, T.: Embedding factorization models for jointly recommending items and user generated lists. In: SIGIR (2017)
4. Chang, J., Gao, C., He, X., Jin, D., Li, Y.: Bundle recommendation with graph convolutional networks. In: SIGIR (2020)
5. Chen, L., Liu, Y., He, X., Gao, L., Zheng, Z.: Matching user with item set: Collaborative bundle recommendation with deep attention network. In: IJCAI (2019)
6. Deng, Q., Wang, K., Zhao, M., Zou, Z., Wu, R., Tao, J., Fan, C., Chen, L.: Personalized bundle recommendation in online games. In: CIKM (2020)
7. Dong, Q., Xie, S., Li, W.: User-item matching for recommendation fairness. IEEE Access (2021)
8. Jeon, H., Jang, J.G., Kim, T., Kang, U.: Accurate bundle matching and generation via multitask learning with partially shared parameters. Plos one (2023)
9. Jeon, H., Kim, J., Yoon, H., Lee, J., Kang, U.: Accurate action recommendation for smart home via two-level encoders and commonsense knowledge. In: CIKM. ACM (2022)
10. Jeon, H., Koo, B., Kang, U.: Data context adaptation for accurate recommendation with additional information. In: BigData (2019)
11. Karakaya, M.Ö., Aytekin, T.: Effective methods for increasing aggregate diversity in recommender systems. Knowl. Inf. Syst. (2018)
12. Kim, J., Jeon, H., Lee, J., Kang, U.: Diversely regularized matrix factorization for accurate and aggregately diversified recommendation. In: PAKDD (2023)
13. Koo, B., Jeon, H., Kang, U.: Accurate news recommendation coalescing personal and global temporal preferences. In: PAKDD (2020)
14. Ma, Y., He, Y., Zhang, A., Wang, X., Chua, T.: Crosscbr: Cross-view contrastive learning for bundle recommendation. In: KDD (2022)
15. Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., Burke, R.: Fairmatch: A graph-based approach for improving aggregate diversity in recommender systems. In: UMAP (2020)
16. Park, H., Jung, J., Kang, U.: A comparative study of matrix factorization and random walk with restart in recommender systems. In: BigData (2017)
17. Park, Y., Tuzhilin, A.: The long tail of recommender systems and how to leverage it. In: RecSys (2008)
18. Pathak, A., Gupta, K., McAuley, J.J.: Generating and personalizing bundle recommendations on *Steam*. In: SIGIR (2017)
19. Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L.: BPR: bayesian personalized ranking from implicit feedback. In: UAI (2009)
20. Zhang, S., Yao, L., Sun, A., Tay, Y.: Deep learning based recommender system: A survey and new perspectives. ACM Comput. Surv. (2019)