CSCI-SHU 381: Recommendation Systems

Final Project Report - Spring 2023

Countering Popularity Bias via Mixed Sampling Guided Regularization

Haoming Liu and Haohai Pang

Supervised by Hongyi Wen

Abstract

Due to the natural long-tail distribution of user-item interactions, recommendation systems tend to favor popular items during prediction, resulting in popularity bias. Previous work has demonstrated that enforcing a direct regularization on the BPR loss can significantly reduce the model bias while maintaining accuracy. However, it fails to achieve satisfactory performance for users with limited interaction histories. To alleviate this problem, this project proposes a systematic mixed sampling strategy to boost the debias performance without sacrificing the accuracy of recommendations, whose efficacy has been shown by the experiments on both synthetic and real-world datasets. The code is publicly available at: `https://github.com/hmdliu/RecSys-SP23`.

Keywords

Popularity Bias; BPR Loss; Regularization; Mixed Sampling

Contents

1 Introduction

Popularity bias is a phenomenon that occurs when popular items are recommended more frequently than unpopular ones, regardless of their actual quality or relevance to the user's interests. It has a significant impact on the fairness of recommendation systems, as it can lead to a narrow and limited set of items being presented to users. This problem is caused by the natural long-tail distribution of item popularity and was further amplified during training, where the popular items dominate most of the training steps.

Various debias methods have been proposed to cope with popularity bias, and most of them can fall into the following three categories: inverse propensity weighting [1], casual intervention [2], and regularization [3, 4]. In particular, the recently proposed regularization approach [3] exhibits remarkable improvements over performance of earlier debias methods. However, it fails to achieve consistent debias performance on users with limited interaction histories. Hence, this project aims to tackle this drawback from the sampling perspective. In particular, we propose a mixed sampling strategy to accommodate the users and items on the end of the tail.

2 Related Works

2.1 Inverse Propensity Weighting

Inverse propensity weighting (IPW) [1] is a statistical method commonly used to address popularity bias. It allocates more weight to under-represented items and less weight to over-represented items, by computing the inverse of each item's probability of being included in the sample and using these weights to adjust the estimation of the treatment effect. This is generally helpful in reducing the effects of model bias and increasing the validity of observational studies.

2.2 Causal Intervention

Causal intervention approaches selectively modifies the popularity of certain items to alleviate popularity bias [5]. It improves the diversity and fairness of recommendations by increasing the popularity of unpopular items and decreasing the popularity of more popular items artificially. Popularity-bias Deconfounding and Adjusting (PDA) [2] is another classical causal intervention approach, which removes the confounding popularity bias in model training and adjusts the recommendation scores with desired popularity bias through causal intervention.

2.3 Personalized Re-ranking

Personalized re-ranking adjusts recommended items according to users' interests and preferences. [6] proposes a re-ranking algorithm that incorporates user history and behavior to improve recommendation accuracy and diversity. The algorithm uses a combination of popularity and diversity metrics to re-rank recommended items and increase the exposure of less popular but relevant items to users. By incorporating personalized re-ranking techniques, the systems can mitigate the impact of prevalence bias and provide more diverse recommendations.

2.4 Regularization

A few recent works introduce regularization terms to deal with the discrepancy between the prediction scores for popular and unpopular items. One representative approach is to regulate the Pearson correlation of item popularity and item score for positive items such that the recommendation scores can be independent of item popularity [4]. Another approach extends the BPR loss [7] and regularizes the score differences between positive and/or negative item pairs [3]. Our work is built upon the latter, and we wish to further mitigate the occurrence of popularity bias under the worst-case scenario, namely users with limited interaction histories.

3 Methods

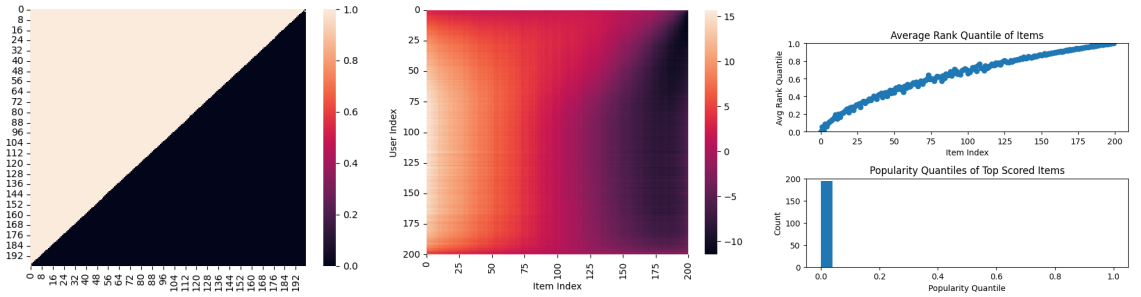
3.1 Synthetic Dataset

Following [3], we construct a synthetic dataset with explicit popularity bias to visualize the debias performance. Specifically, we build a 200×200 user-item interaction matrix R as follows:

$$R[u, i] = \begin{cases} 1, \text{ if } u + i \leq 200 \\ 0, \text{ otherwise} \end{cases}, \tag{1}$$

where u and i are the index of the user and item, respectively. A visualization of the synthetic dataset is shown in Figure 1(a). Overall, the popularity of each item decreases linearly as the item index increases. If we train a MF model using the BPR loss [7], the model will naturally exhibit salient popularity bias. Figure 1(b) plots the prediction heat map for all the user-item pairs, and we can observe that the brighter region (with higher prediction scores) concentrates on items with smaller indices (i.e., the popular items). Figure 1(c) plots the average rank quantile of

the items and the histogram of the popularity quantiles of the top positive items, which illustrates the popularity bias exhibited by the model from different perspectives.



(a) Synthetic Dataset (b) Heatmap (c) Average Ranking & Popularity

Figure 1: Visualizations of the synthetic dataset and baseline model.

3.2 Baseline Models

We adopt a matrix factorization (MF) model trained with the Bayesian Personalized Ranking (BPR) loss [7] as the baseline model. The formulation of the BPR loss is as follows:

$$\ell_{BPR} = -\sum_{u}\sum_{p,n} \log(\hat{y}_{u,p} - \hat{y}_{u,n}), \tag{2}$$

where $u \in U$, $p \in Pos_u$, $n \in Neg_u$, $\hat{y}_{u,i}$ is the prediction score for user-item pair (u, i); U denotes the set for all users; Pos_u denotes the positive item set for user u; Neg_u denotes the negative item set for user u. Note that the L2 regularization term has been omitted for simplicity. Remarkably, the intuition behind BPR loss is to maximize the score differences between positive and negative items (i.e., $\hat{y}_{u,p} - \hat{y}_{u,n}$).

We also introduce the two regularization terms proposed in [3] as our baselines, which are:

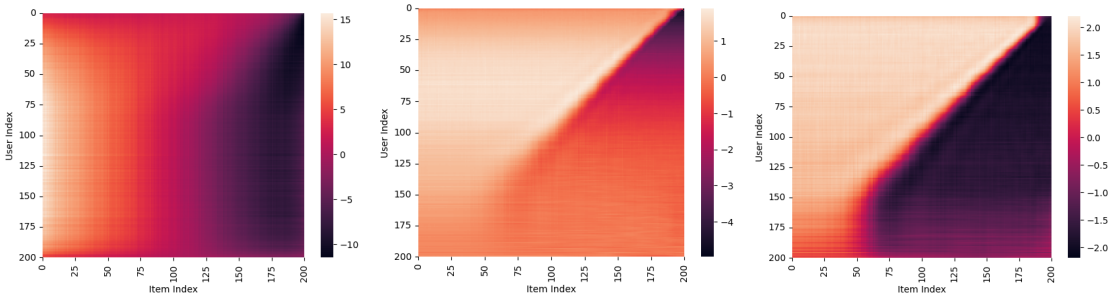
$$\ell_{Pos2Neg2} = -\sum_{u}\sum_{p_1,p_2,n_1,n_2} \log(1 - \tanh|\hat{y}_{u,p_1} - \hat{y}_{u,p_2}|) + \log(1 - \tanh|\hat{y}_{u,n_1} - \hat{y}_{u,n_2}|), \tag{3}$$

$$\ell_{Zerosum} = -\sum_{u}\sum_{p_1,n_1} \log(1 - \tanh|\hat{y}_{u,p_1} + \hat{y}_{u,n_1}|), \tag{4}$$

where $u \in U$; $p_1, p_2 \in Pos_u$; $n_1, n_2 \in Neg_u$. The Pos2Neg2 regularization explicitly minimizes the score differences between positive-positive and negative-negative item pairs; whereas the Zerosum regularization enforces the scores for any positive-negative to be close to zero. For the baselines with debias regularization, the loss function can be formulated as:

$$\ell = \lambda \ell_{BPR} + (1 - \lambda)\ell_{Reg}. \tag{5}$$

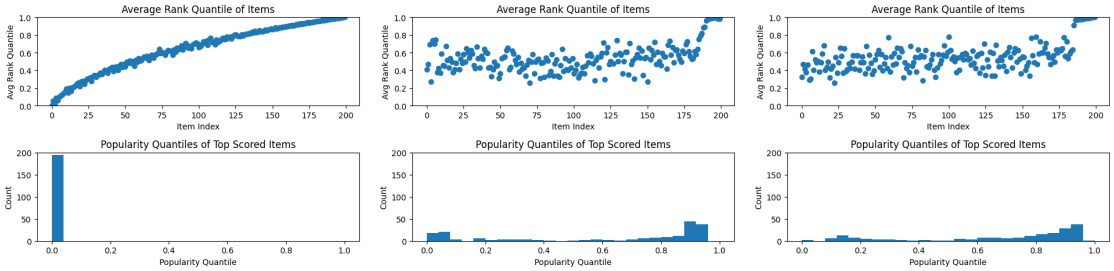
Following [3], we empirically set $\lambda = 0.8$ for the experiments on the synthetic dataset, and we set $\lambda = 0.9$ for the experiments on real-world datasets. In general, a smaller λ indicates a stronger emphasis on reducing the popularity bias.



(a) None (b) Pos2Neg2 (c) Zerosum

Figure 2: Heatmaps of the three baseline models on the synthetic dataset.

Figure 2 presents the qualitative results for the three baselines, and we can clearly observe that both of the regularization terms give remarkable debias performance, as the prediction score matrices are more similar to the ground-truth labels. Then, we plot the average rank quantile of the items and the popularity quantiles of the top positive items in Figure 3, where the correlation between popularity and ranking has been significantly weakened. However, we can also see that even the best performed method, Zerosum, still fails to achieve satisfactory performance on users with larger indices, whose interaction histories are rather limited. We attribute this defect to the sampling strategy, as the users on the tail of the distribution are less likely to be sampled. To cope with this challenge, we propose to adopt a mixed sampling strategy for training.



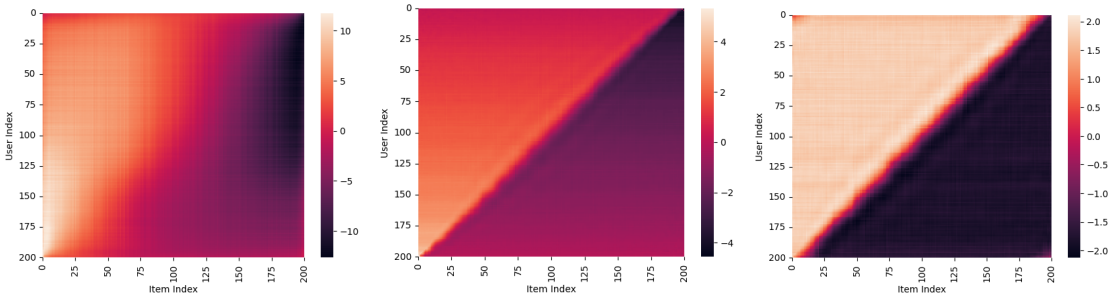
(a) None (b) Pos2Neg2 (c) Zerosum

Figure 3: Average ranking & popularity of the three baseline models.

3.3 Mixed Sampling Strategy

The sampling for the original baselines are performed uniformly over all positive interactions, and this can potentially be problematic on a dataset that follows a long-tail distribution, especially when we are performing debias regularization. In such a case, the users with limited interaction data are unable to obtain sufficient training steps, and the model may thus fail to capture their

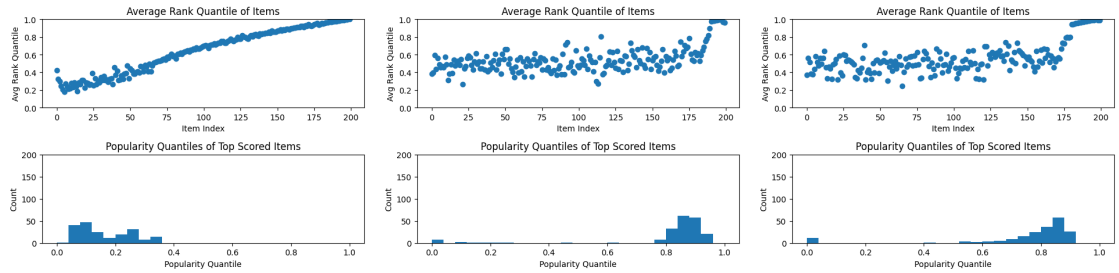
true preferences. An alternative sampling strategy is to perform it uniformly over all the users, meaning that each users is equal likely to be sampled.



(a) None (b) Pos2Neg2 (c) Zerosum

Figure 4: Heatmaps of the baselines with user-uniform sampling.

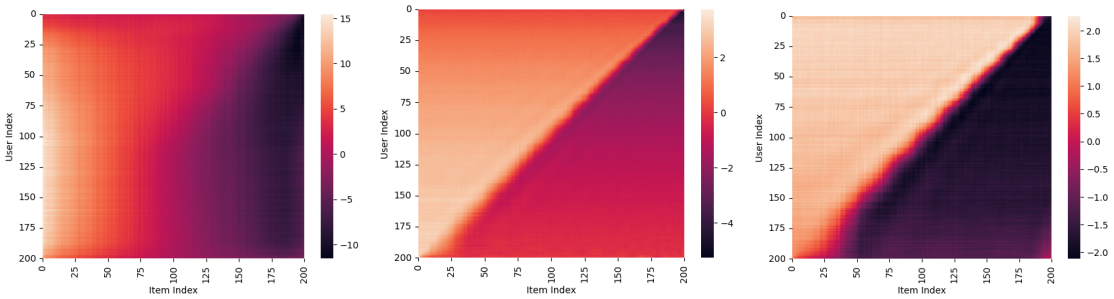
As shown in Figure 4, the user-uniform sampling strategy results in better qualitative results, where the performance on end-of-the-tail users has been lifted up. However, this sampling strategy has its drawback as well. If we look closely at the histograms of the popularity quantiles of the top positive items (lower part of Figure 5), we can observe that the baselines with debias regularization now favor unpopular items over the popular ones. In essence, this is caused by the over-sampling of the end-of-the-tail interactions, as the train stage allocates excessive focus on the users with limited interaction histories. Considering the pros and cons of the two aforementioned sampling strategies, a reasonable sampling strategy should thus be a mixture of both.



(a) None (b) Pos2Neg2 (c) Zerosum

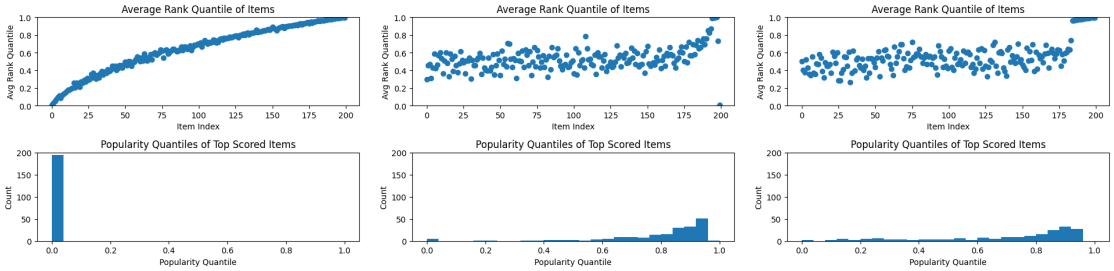
Figure 5: Average ranking & popularity of the baselines with user-uniform sampling.

One natural idea is to formulate a mixed sampling strategy based on epsilon-greedy. Specifically, we randomly generate a number $p \in (0, 1)$ from a uniform distribution and compare it with a threshold ϵ. If $p > \epsilon$, then we pick a user u uniformly over all positive interactions; otherwise, we pick a user u uniformly over all users. In general, a larger ϵ indicates a stronger emphasis on accommodating the end-of-the-tail users, whereas a smaller ϵ enforces the distribution we sampled from to be closer to the original long-tail distribution. With a properly chosen hyper-parameter ϵ, we can achieve an equilibrium of sampling.



(a) None (b) Pos2Neg2 (c) Zerosum

Figure 6: Heatmaps of the baselines with mixed sampling ($\epsilon = 0.2$).



(a) None (b) Pos2Neg2 (c) Zerosum

Figure 7: Average ranking & popularity of the baselines with mixed sampling ($\epsilon = 0.2$).

As shown in Figure 6 and 7, the proposed mixed sampling strategy can be viewed as an interpolation of the two sampling strategies and is able to boost both accuracy and debias performance. Some quantitative results are presented in Section 4.4 and 4.5.

4 Experiments & Results

4.1 Datasets

We perform experiments on two datasets, the synthetic dataset and the MovieLens-1M dataset. The synthetic dataset has been introduced in Section 3.1. MovieLens-1M is a real-world movie rating dataset, consisting of 998,539 interactions between 6,040 users and 3,260 items. We adopt the data pre-processing steps from [3] and filter out users and items with less than 10 interactions.

4.2 Evaluation Metrics

For the experiments on the synthetic dataset, we use `ER` (error rate), `PRI`, and `PopQ@1` as the metrics. The `accuracy` is defined as the average frequency of the positive item being scored higher than the negative item over all positive-negative item pairs, and we have `ER` = (1 - `accuracy`) $\times$ 100%. `PRI` is the Spearman rank correlation coefficient (SRC) between item popularity and

the average ranking quantile, conditioned on the positive items. `PopQ@1` computes the average popularity quantile of the top scoring positive items of each user. For the experiments on the MovieLens-1M dataset, we use `Hit@10`, `NDCG@10`, and `PopQ@1` as the metrics. We perform sampled evaluation and pair each positive test item with 100 test negative items.

4.3 Implementation Details

We split all positive interactions by a ratio of 3:1:1 for training, validation, and testing, respectively. We sample 100 test negative items for each user before the training stage. We use a matrix factorization model, a learning rate of 1e-3, and a batch size of 256 for all experiments. For the loss weighting factor λ, we empirically set $\lambda = 0.8$ on the synthetic dataset and $\lambda = 0.9$ on the MovieLens-1M dataset. For the mix sampling threshold, we set $\epsilon = 0.2$ by default.

4.4 Quantitative Results on the Synthetic Dataset

As shown in Table 1, the positive interaction uniform sampling performs better on the `PRI` and `PopQ@1`, indicating a better debias performance. This is because the distribution we sample from is the original long-tail distribution, which is more suitable for eliminating the popularity bias. In contrast, the user-uniform sampling exhibits better accuracy, whose performance gain can be attributed to the accommodations on the users with limited training data. Furthermore, the performance of the mixed sampling is roughly an interpolation of the other two sampling strategies. With sophisticated hyper-parameter tuning, we can find a properly-chosen ϵ such that both accuracy and debias performance are boosted.

	PosInteraction-Uniform			User-Uniform			Mixed ($\epsilon = 0.2$)		
Baseline	None	Pos2Neg2	Zerosum	None	Pos2Neg2	Zerosum	None	Pos2Neg2	Zerosum
`ER` (%)	0.012	0.022	0.009	0.010	0.001	**0.000**	0.013	0.015	0.002
`PRI`	0.998	**0.388**	0.493	0.992	0.469	0.580	0.998	0.390	0.483
`PopQ@1`	0.002	**0.615**	0.676	0.162	0.815	0.760	0.003	0.799	0.701

Table 1: Quantitative results using different sampling strategies on the synthetic dataset. Ideally, `ER` and `PRI` should be close to 0, `PopQ@1` should be close to 0.5.

4.5 Quantitative Results on the MovieLens-1M Dataset

As shown in Table 2, the Zerosum baseline with the mixed sampling strategy performs the best on the `Hit@10` and `NDCG@10` metrics, indicating a better recommendation accuracy. Similar to that of the synthetic dataset, the Pos2Neg2 baseline with the positive interaction uniform

sampling still performs the best on the PopQ@1 metric, whereas the user-uniform sampling strategy may potentially strengthen the correlation between popularity and item ranking. Overall, the combination of mixed sampling strategy and debias regularization is able to improve the accuracy and debias performance at the same time.

	PosInteraction-Uniform			User-Uniform			Mixed ($\epsilon = 0.2$)		
Baseline	None	Pos2Neg2	Zerosum	None	Pos2Neg2	Zerosum	None	Pos2Neg2	Zerosum
Hit@10	0.662	0.679	0.678	0.670	0.681	0.682	0.662	0.681	**0.684**
NDCG@10	0.396	0.407	0.411	0.408	0.414	0.415	0.399	0.410	**0.417**
PopQ@1	0.365	**0.418**	0.382	0.324	0.320	0.323	0.349	0.395	0.365

Table 2: Quantitative results using different sampling strategies on the MovieLens-1M dataset. Ideally, Hit@10 and NDCG@10 are the larger the better, PopQ@1 should be close to 0.5.

5 Discussions & Future Work

Inspired by the epsilon-greedy algorithm, we propose a mixed sampling strategy that incorporates positive interaction uniform sampling and user-uniform sampling. By combining it with the regularization-based debias approaches, we are able to eliminate popularity bias without sacrificing recommendation accuracy. Besides, the ideal values for PRI and PopQ@1 metrics still worth further discussion. Different from [3], we argue that setting the ideal correlation between the popularity and average ranking of the items to 0 is counter-intuitive. Instead, it should be somewhere between 0 (no correlation) and 1 (strong positive correlation), as the popularity of certain items also entails their higher overall quality. Similarly, the ideal value of the PopQ@1 metric might be slightly less than 0.5. Hence, it is reasonable for them to get more exposure at the serving stage.

The future work of this project mainly lies in exploring other mixed sampling strategies (e.g., introducing a KL divergence term to the loss function).

Acknowledgements

We thank Professor Hongyi Wen for his suggestions on the project. This work was supported through the NYU IT High Performance Computing resources, services, and staff expertise.

References

[1] T. Joachims, A. Swaminathan, and T. Schnabel, "Unbiased learning-to-rank with biased feedback," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, ser. WSDM '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 781–789. [Online]. Available: https://doi.org/10.1145/3018661.3018699

[2] Y. Zhang, F. Feng, X. He, T. Wei, C. Song, G. Ling, and Y. Zhang, "Causal intervention for leveraging popularity bias in recommendation," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 11–20. [Online]. Available: https://doi.org/10.1145/3404835.3462875

[3] W. Rhee, S. M. Cho, and B. Suh, "Countering popularity bias by regularizing score differences," in *Proceedings of the 16th ACM Conference on Recommender Systems*, ser. RecSys '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 145–155. [Online]. Available: https://doi.org/10.1145/3523227.3546757

[4] Z. Zhu, Y. He, X. Zhao, Y. Zhang, J. Wang, and J. Caverlee, "Popularity-opportunity bias in collaborative filtering," ser. WSDM '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 85–93. [Online]. Available: https://doi.org/10.1145/3437963.3441820

[5] T. Wei, F. Feng, J. Chen, Z. Wu, J. Yi, and X. He, "Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, ser. KDD '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 1791–1800. [Online]. Available: https://doi.org/10.1145/3447548.3467289

[6] H. Abdollahpouri, R. Burke, and B. Mobasher, "Managing popularity bias in recommender systems with personalized re-ranking," 2019.

[7] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "Bpr: Bayesian personalized ranking from implicit feedback," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, ser. UAI '09. Arlington, Virginia, USA: AUAI Press, 2009, p. 452–461.