Bootstrapped Personalized Popularity for Cold Start Recommender Systems

Iason Chaimalas
University College London
London, United Kingdom
iason.chaimalas.20@ucl.ac.uk

Duncan Martin Walker
British Broadcasting Corporation
London, United Kingdom

Edoardo Gruppi
University College London
London, United Kingdom

Benjamin Richard Clark
British Broadcasting Corporation
London, United Kingdom

Laura Toni
University College London
London, United Kingdom
l.toni@ucl.ac.uk

ABSTRACT

Recommender Systems are severely hampered by the well-known Cold Start problem, identified by the lack of information on new items and users. This has led to research efforts focused on data imputation and augmentation models as predominantly data preprocessing strategies, yet their improvement of cold-user performance is largely indirect and often comes at the price of a reduction in accuracy for warmer users. To address these limitations, we propose Bootstrapped Personalized Popularity (B2P), a novel framework that improves performance for cold users (directly) and cold items (implicitly) via popularity models personalized with item metadata. B2P is scalable to very large datasets and directly addresses the Cold Start problem, so it can complement existing Cold Start strategies. Experiments on a real-world dataset from the BBC iPlayer and a public dataset demonstrate that B2P (1) significantly improves cold-user performance, (2) boosts warm-user performance for bootstrapped models by lowering their training sparsity, and (3) improves total recommendation accuracy at a competitive diversity level relative to existing high-performing Collaborative Filtering models. We demonstrate that B2P is a powerful and scalable framework for strongly cold datasets.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; *Collaborative search*; *Personalization.*

KEYWORDS

recommender systems, collaborative filtering, cold-start problem, popularity modelling, high-order personalization

ACM Reference Format:
Iason Chaimalas, Duncan Martin Walker, Edoardo Gruppi, Benjamin Richard Clark, and Laura Toni. 2023. Bootstrapped Personalized Popularity for Cold Start Recommender Systems. In *Seventeenth ACM Conference on Recommender Systems (RecSys '23), September 18–22, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3604915.3608820

RecSys '23, September 18–22, 2023, Singapore, Singapore
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0241-9/23/09.
https://doi.org/10.1145/3604915.3608820

1 INTRODUCTION

The information overload on consumers in our digitized society [21, 39] is alleviated by recommender systems that learn and curate user preferences [25]. Particularly, the top-k task of recommending the k most relevant items to consumers finds widespread use across e-commerce, movie-streaming and news platforms to name a few, and provides billions of dollars in value to companies [16], while enhancing user experience. Within top-k recommendation, longstanding and highly successful models are Collaborative Filtering (CF) and hybridisations with user and item content metadata [22, 40]. Such techniques range from methods based on Deep Neural Networks (DNN) [37] to lower-complexity solutions mainly classed as Neighborhood, Graph-based, Matrix Factorization and Full-Rank, which perform competitively in public datasets [11, 13, 14, 37]. However, while these CF recommenders inherently deal with highly sparse user-interaction data [17], their accuracy severely deteriorates in the case of extreme data sparsity known as Cold Start.

The Cold Start problem describes the accuracy drop of recommender systems when faced with a large proportion of users with very few or even zero training interactions (*cold users*), or items with very low exposure among users (*cold items*). Because of the dynamic nature of real platforms, with new items released and a considerable inflow of new and infrequent users, this Cold Start problem is highly pertinent to production recommender systems. Recent recommenders designed to improve Cold Start performance center on data imputation [31, 32, 44, 46], by inferring the missing interactions, and model training augmentation, by creating extra synthetic users and items [5] or varying training regularization [6, 42]. However, these approaches have a number of limitations. First, most Cold Start recommenders are applied over an existing 'CF Backbone' model to increase its accuracy in Cold Start settings, instead of holistically approaching a top-k task consisting of mixed cold and non-cold users. For instance, imputation methods essentially address sparsity instead of directly tackling the Cold Start problem. Moreover, the accuracy gain for cold users often entails the sacrifice of accuracy for non-cold users [8, 19, 42]. Lastly, there is a widespread lack of evaluation on *diversity* metrics for Cold Start models, despite diverse recommendations being significant in retaining cold users and promoting cold items; diversity is key for user satisfaction in real systems [7, 15, 27] and for the prevention of filter bubbles [41, 43].

The limitations of current Cold Start recommenders lead us to propose a novel and principled *framework* called *Bootstrapped Personalized Popularity* (B2P). It directly and holistically addresses the Cold Start problem, so it can be treated as a comprehensive CF Backbone to complement existing Cold Start methods. B2P addresses cold users with popularity modelling that is personalized via our novel diversity-boosting method *Metadata Infusion.* Finally, B2P also applies Metadata Infusion to existing CF models, and bootstraps them with the personalized popularity model. Therefore, our proposal of B2P herein is driven by the following Research Questions (RQs):

- RQ1: How does popularity modelling perform for cold and warm users relative to established CF methods in terms of rank-accuracy and diversity?
- RQ2: To what extent does Metadata Infusion impact model rank-accuracy and diversity?
- RQ3: Does B2P outperform established CF methods in rank-accuracy on cold datasets with competitive diversity?

In presenting B2P, we begin by demonstrating the statistical optimality of popularity modelling for new users and then extend it to non-new cold users. Next, we introduce Metadata Infusion to increase the popularity model's diversity across item coverage and popularity. Finally, we rigorously bootstrap personalized popularity for cold users with a Metadata-Infused CF model for warm users, which implicitly reduces sparsity for the bootstrapped CF model. Therefore, we simultaneously leverage metadata and implicit sparsity reduction – as in content-based and imputation-based Cold Start approaches – while offering a new and principled outlook on the treatment of cold users and cold items.

2 RELATED WORKS

We begin by reviewing current research on popularity analysis and high-order similarity in CF and content-based hybrids. We thus highlight both the inspirations for B2P and its contribution in overcoming previous efforts' limitations.

2.1 Popularity Modelling of User Behaviour

Beyond their use as an extreme-case baseline that recommends to all users the most popular items [13], popularity methods have been studied to explore the impact of active users and popular items on cold users [20, 45]. However, they have known limitations of non-personalization, exacerbation of popularity biases via unfair exposure of a fraction of available items [26], and unstable accuracy in dynamic platforms. Hence, the popularity bias has also been widely studied by proposing post-processing re-ranking approaches [1] or models that optimize novel metrics [29, 34] for the fairer representation of low-popularity cold items. Also, Inverse Propensity Weighting and other de-biasing strategies focus on more robust metrics against biases in the policy with which the data was collected [7]. For instance, popularity is biased by past exposure of items, leading to biased models. Overall, these efforts avoid the issue of cold users.

Inspired by the impact of popularity on user preference [46] and item exposure [26, 34], we propose a novel popularity-driven model conditioned – and hence weakly personalized – on the level of user interaction, which we demonstrate as optimal for the pure Cold Start case of new users. We then further personalize it for all users with the novel Metadata Infusion technique (introduced in Section 2.2) that directly addresses cold users and cold items.

2.2 High-Order Similarity

Top-k recommenders typically model the *similarity* between users and items – whether via user- or item-content metadata in common or via collaborative models on the observed user-item interactions – to make predictions.

Firstly, content-based approaches can be useful in a cold dataset with little available interaction information to exploit collaboratively. However, they can promote a 'filter bubble' [22], which is particularly detrimental for cold users in real platforms. They are also limited in cases of scarce metadata [4, 37]. Secondly, in the pure CF case, the high-order directional similarity of an item interaction conditional on an entire set of past interactions would yield optimal-accuracy recommendations. Yet because of the computational intractability and risk of overfitting to the observed user-item interactions, the major CF families typically model only the first-order similarity of an item *separately* conditioned on each prior item interaction of a user. This has been modelled by Neighborhood correlation heuristics [40], user-item Graph traversal [10, 30], or Full-Rank constrained optimization [28, 35]. However, these first-order similarity models unrealistically assume that a user's past interactions are *independent* [2, 3]. Moreover, estimates of true similarity have high uncertainty for cold data, decreasing model accuracy. For instance, the Neighborhood models and Full-Rank EASE$^{\mathrm{R}}$ estimate learnt similarities in the form of Gram matrices [35], where estimation uncertainty increases in cold datasets.

Motivated by methods that learn second-order similarity conditioned on *pairs* of past item interactions and increase model accuracy [9, 38], we design Metadata Infusion as a novel technique that merges second-order modelling with principled hybridization via content metadata. First, it requires minimal metadata to perform powerfully. Second, Metadata Infusion utilizes known metadata so it scales to all possible item pairs or even higher permutations, unlike learnt second-order modelling [9, 38]. Third, Metadata Infusion applied to an existing CF model mixes accuracy-centered first-order similarity with second-order content similarity, which increases personalization and item coverage.

3 BACKGROUND THEORY

3.1 Notation

The typical CF problem is the determination of the probability for users in set $\mathcal{U}$ to interact with items in set $\mathcal{I}$ in the testing phase, given the users' interactions with items in model training. The set of new users in testing is $\mathcal{U}_{\text{new}} \subseteq \mathcal{U}$. The training interactions of each user $u \in \mathcal{U}$ are $X \in \mathbb{R}^{|\mathcal{U}|\times|\mathcal{I}|}$, where each row $\mathbf{x}_u \in \mathbb{R}^{1\times|\mathcal{I}|}$ has elements 0 for non-interactions or the rating given to an item by u (1 in implicit feedback). Similarly, the true testing interactions of each user u are $\mathbf{y}_u \in \mathbb{R}^{1\times|\mathcal{I}|}$. While random data-splitting is prevalent [11, 13, 24, 35, 38], it unrealistically inflates evaluation metrics [47], so we use temporal splitting. Thus, for testing timestep T, $\mathbf{x}_u \,\forall\, u \in \mathcal{U}$ occur at $t < T$ and $\mathbf{y}_u \,\forall\, u \in \mathcal{U}$ occur at $t = T$, where both are drawn from a temporally-variant data-generating distribution $\mathcal{D}_T$: $\mathbf{x}_u, \mathbf{y}_u \sim \mathcal{D}_T(\mathcal{U})$.

3.2 Markov Random Fields and Auto-Normality

The spatial stochastic interaction sampling of items in $\mathcal{I}$ by users in $\mathcal{U}$ can be modelled as a Markov Random Field (MRF) [2, 3], wherein items $i \in \mathcal{I}$ are co-dependent Random Variable nodes and users' observed training interactions are samples from the MRF. Then, a CF recommender is an autoregressor that uses the sample drawn by some $u \in \mathcal{U}$ during $t < T$ to predict the most likely items to be sampled at T by u. Assuming the MRF is Gaussian [2, 23], the Auto-Normal parametrization estimates the expectation of sampling some item i at $t = T$, given $\mathbf{x}_u = \{x_j\}_{j=1}^{|\mathcal{I}|}$ observed over $t < T$, as $\mathbb{E}[X_i | x_j, j \neq i] = \mu_i + \sum_{j=1}^{|\mathcal{I}|} \beta_{i,j}(x_j - \mu_j)$, where $\beta_{i,j} = p(i|j)$ is the first-order similarity. The Auto-Normal parametrization simplifies high-order similarity by assuming the items in the MRF are sampled independently by each u, and it directly leads to EASE$^{\text{R}}$ [35] by assuming $\mu_i = 0 \,\forall\, i \in \mathcal{I}$ and uniform MRF variance across items [3, 36].

4 BOOTSTRAPPED PERSONALIZED POPULARITY

We now present B2P, with our contributions structured as (1) rigorous popularity modelling for cold users, (2) Metadata Infusion inspired from work on high-order similarity [9, 28, 35, 36, 38], and (3) principled total bootstrapping.

4.1 Popularity Methods for Cold Users

4.1.1 Optimal Recommendations for New Users. In pure Cold Start for new users $u_{\text{new}} \in \mathcal{U}_{\text{new}}$, we have $\mathbf{x}_{un}, \mathbf{y}_{un} \sim \mathcal{D}_T(\mathcal{U}_{\text{new}})$ where $\mathbf{x}_{un} = \vec{0}$. Then, in the absence of content metadata, $\mathcal{D}_T(\mathcal{U}_{\text{new}})$ can be equivalently described by the test-time distribution of normalized popularity $\mathcal{P}_T(i) \,\forall\, i \in \mathcal{I}$ among $\mathcal{U}_{\text{new}}$ at $t = T$, where $\sum_{i \in \mathcal{I}} \mathcal{P}_T(i) = 1$.

We prove that inferring expected behaviour of $\mathcal{U}_{\text{new}}$ from $\mathcal{P}_T$ is optimal by framing purely-cold CF as *multi-label Supervised Classification.* Here, a recommender model $f \in \mathcal{F}$ produces j-wide recommendation lists $f(\mathbf{x}_{un})_{:j} = [f(\mathbf{x}_{un})_1, \ldots, f(\mathbf{x}_{un})_j]$ for u_{new}, while the observed interactions of u_{new} in testing $t = T$ are contained in the item set $rel(u_{\text{new}}) = \{i \,|\, [\mathbf{y}_{un}]_i \neq 0\}_{i=1}^{|\mathcal{I}|}$. Among $u_{\text{new}} \in \mathcal{U}_{\text{new}}$, the Bayes Optimal Classifier, when reworded from loss minimization to maximization of Mean Average Precision (MAP)[1] at cutoff k [12], is $f^* = \underset{f \in \mathcal{F}}{\operatorname{argmax}}\, \mathbb{E}_{\mathcal{D}_T(\mathcal{U}_{\text{new}})}[\text{MAP}(f(\mathbf{x}_{un}))@k]$. Expanding MAP [48] and taking the expectation over all $u_{\text{new}} \in \mathcal{U}_{\text{new}}$,

$$\begin{aligned} f^* &= \underset{f \in \mathcal{F}}{\operatorname{argmax}} \sum_{j=1}^{k} \frac{|rel(u_{\text{new}}) \cap f(\mathbf{x}_{un})_{:j}|}{j} p(f(\mathbf{x}_{un})_j | \mathbf{x}_{un}) \\ &= \underset{f \in \mathcal{F}}{\operatorname{argmax}} \sum_{j=1}^{k} \frac{|rel(u_{\text{new}}) \cap f(\mathbf{x}_{un})_{:j}|}{j} \mathcal{P}_T(f(\mathbf{x}_{un})_j). \end{aligned} \tag{1}$$

Therefore, recommendations $f(\mathbf{x}_{un})_{:k}$ with maximum total empirical popularity at $t = T$ optimize recommendation accuracy for new users, on average. However, the test-time $\mathcal{P}_T$ among $\mathcal{U}_{\text{new}}$ is unknown in training, so it must be estimated via observed proxies $\mathcal{P}_{t<T}$, with incurred bias. Via simple and scalable zero-order Euler approximants, this is achieved by forming $\mathcal{P}_{t<T}$ from the interactions of some user subset $\tilde{\mathcal{U}}$ with modelling $\mathcal{D}_T(\mathcal{U}_{\text{new}}) \approx \mathcal{D}_{t<T}(\tilde{\mathcal{U}})$. With hypothesis testing and offline experiments in Section 5.2, we find that the most empirically-effective zero-order Euler estimator for the test-time $\mathcal{P}_T$ among users on a biased dataset is to form $\mathcal{P}_{t<T}$ from the interactions of users in subset $\tilde{\mathcal{U}} = \mathcal{U}_{T-1,\text{new}}$. This subset contains users at $t = T-1$ with no previous interactions, who are thus new at this timestep.

4.1.2 Extending Popularity to Cold Users. *Cold users* are both purely-new users and non-new but infrequent users. For new users $\mathcal{U}_{\text{new}}$ at test-time $t = T$, our popularity method that approximates optimal-MAP recommendations predicts item scores as $P = \text{pop}(X_{t_{-1},n=0})$, where $X_{t_{-1},n=0} \sim \mathcal{D}_{T-1}(\mathcal{U}|n = 0)$ is the interactions matrix for new training users at $T-1$ with $n = 0$ past interactions before $T-1$, and $\text{pop}(\cdot)$ ranks items by their popularity $\mathcal{P}$ in matrix $(\cdot)$. We now generalize this as our *Popularity Model* $P = \text{pop}(X_{t_{-1},n})$ for the interactions $X_{t_{-1},n}$ of cold test-time users with $n \leq \theta$ past interactions up to an empirical or learnt cut-off $\theta \geq 0$. Here, for each stratum of users in testing with n past interactions in $t < T$, the Popularity Model would recommend the k most popular items at $T-1$ among users in training with n past interactions in $t < T-1$. Indeed, popularity and preference are entangled in $p(i|\mathbf{x}_u)$ [46], particularly for cold users [20]. Therefore, the Popularity Model is a rigorous and principled approximation of high-order similarity for highly-cold users. However, it is starkly non-diverse and becomes inaccurate for warm users with $n > \theta$.

4.2 Metadata Infusion

In the framing of CF as sampling from an MRF, generalising Auto-Normality to significantly higher orders better approximates true high-order similarity but increases computational complexity at marginal performance gains [38]. Thus, we model second-order similarity using the triplet extension to Auto-Normality scaled by hyperparameter κ_C,

$$P_{u,i} = p(X_i | X_{u,:}) = \sum_{j \neq i} X_{u,j} B_{j,i} + \kappa_C \sum_{\substack{j<k \\ j,k \neq i}} X_{u,j} X_{u,k} \left(\sum_{m=1}^{M} \vec{v}_m^{(i)} \vec{\gamma}_m^{(j)} \vec{\gamma}_m^{(k)} \right) \tag{2}$$

which predicts the similarity score of each $i \in \mathcal{I}$ at $t = T$ given all pairs $j, k \in \mathcal{I}_{-i}$ sampled at $t < T$ for each $u \in \mathcal{U}$, as $P = X\hat{B} + \kappa_C \langle XC \rangle$. Matrix $\langle XC \rangle \in \mathbb{R}^{|\mathcal{U}| \times |\mathcal{I}|}$ contains the summed elementwise products of latent factors $\vec{v}^{(i)}, \vec{\gamma}^{(j)}, \vec{\gamma}^{(k)}$ for all $i, j, k \in \mathcal{I}$ and all $u \in \mathcal{U}$. Here, element $\langle XC \rangle_{u,i}$ indicates the second-order similarity of item i to the pairwise j, k interaction history of user u, over a learnt or pre-defined latent space. Indeed, $\langle XC \rangle$ is *learnable*, as pursued in [9, 33, 38]. However, learning every second-order triplet is unscalable, so recent works threshold the learnt triplets to those formed with the most popular 40000 j, k pairs among all users [38].

[1] MAP is our primary optimization target in this work as it is a widespread and reliable rank-accuracy metric both in BBC and the broader research space.

Instead, we propose a novel treatment of latent factors as explicit M-wide metadata that characterize $i, j, k \in \mathcal{I}$, such that *Metadata Infusion* occurs in predictions P. Since $\vec{v}^{(i)}, \vec{\gamma}^{(j)}, \vec{\gamma}^{(k)}$ are known instead of learnt, this approach easily scales to modelling all possible pairs j, k. Further, while a learnt $\langle XC\rangle$ requires an iterative algorithmic solution [9, 38], Metadata Infusion maintains an efficient closed-form solution like EASE$^\mathrm{R}$, of the form:

$$\begin{aligned}
&\mathcal{L}(\Psi) = ||X - XB - \kappa_C\langle XC\rangle||_F^2 + \lambda_B||B||_F^2 + 2\gamma^T \mathrm{diag}(B)\\
\implies &\frac{\partial\mathcal{L}}{\partial B} = 0 \implies -2X^T(X - X\hat{B} - \kappa_C\langle XC\rangle) + 2\lambda_B\hat{B} + 2\gamma^T \odot \mathbf{I} = 0\\
&\implies \hat{B} = \mathbf{I} - \hat{P}(\kappa_C X^T\langle XC\rangle + \mathrm{diagMat}(\tilde{\gamma}))
\end{aligned}$$

where $\hat{P} \triangleq (X^T X + \lambda_B\mathbf{I})^{-1}$ for sufficiently large λ_B, $\tilde{\gamma} \triangleq \lambda_B\vec{1} + \gamma$ and $\mathrm{diagMat}(\cdot)$ defines a zero square matrix with vector $(\cdot)$ in the main diagonal. Importantly, $X^T X + \lambda_B\mathbf{I}$ is always full-rank for $\lambda_B \neq 0$, such that $\exists\hat{P}$ and the solution for $\hat{B}$ always holds. Then, constraining the Lagrangian multipliers with $\mathrm{diag}(B) = 0$ yields the closed form of $\hat{B}$:

$$\hat{B} = \mathbf{I} - \hat{P}\left(\kappa_C X^T\langle XC\rangle + \mathrm{diagMat}\left(\vec{1} \oslash \mathrm{diag}(\hat{P}) - \kappa_C\mathrm{diag}(X^T\langle XC\rangle)\right)\right). \tag{3}$$

4.3 Personalized Popularity (2P)

We extend the Popularity Model to *Personalized Popularity* (2P) via Metadata Infusion: $P = \mathrm{pop}(X_{t_{-1},n}) + \kappa_C\langle X_{t_{-1},n}C\rangle$. In computing $\langle X_{t_{-1},n}C\rangle$, pairs j, k are drawn from the interactions at $t = T - 1$ among only users with n interactions in $t < T - 1$. This personalizes the Popularity Model because $\langle XC\rangle_{u,i}$ are scores of item similarity for each u based on the metadata in common between each test item i and all past interaction pairs of u. For sparse binary X, which is a characteristic of implicit-feedback systems, the Gram matrix $G = X^T X$ in Eq. 3 and in EASE$^\mathrm{R}$ [35] is a co-occurence matrix, which should have sufficiently large elements G_{ij} to estimate $\hat{B}$ with low error. However, increasing the proportion of cold users in X means lower G_{ij} and higher error. Then, there must be a cutoff $\theta = \theta^*$, below which the empirical MRF sampling popularities extended from their optimality for new users outperform EASE$^\mathrm{R}$ and Eq. 3.

4.4 Overall Model: Bootstrapped Personalized Popularity

Personalized Popularity (2P) models the empirical popularity of sampling from $\mathcal{D}_T$ by assuming users of the same interaction level behave similarly on average. This probabilistic similarity is shown to be optimal for new users and outperforms $P = X\hat{B} + \kappa_C\langle XC\rangle$ for cold users with up to $n \leq \theta^*$, where θ^* depends on the dataset and evaluation metrics. However, second-order statistical similarity in Metadata-Infused EASE$^\mathrm{R}$ $P = X\hat{B} + \kappa_C\langle XC\rangle$ becomes more accurate for $n > \theta^*$. Broadly, Metadata Infusion is highly flexible and applicable to CF models that predict a scores matrix P to compute the top-k recommendations, including our Popularity Model in Section 4.1.2 and EASE$^\mathrm{R}$.

Thus, we form our general B2P *framework* for treating cold datasets by bootstrapping 2P with a Metadata-Infused CF model as a weighted sum at each user stratum $\mathcal{U}_n \in \mathcal{U}$ with n train-time interactions over $t < T$. Therefore, B2P joins exploitation with popularity exploration at each $\mathcal{U}_n$, catering to stratum-average user preferences for recommendation diversity [27]. In this paper, we implement the bootstrap between 2P and Metadata-Infused EASE$^\mathrm{R}$ via a binary switch given by Heaviside $H(\theta^*)$. This particular configuration of the general B2P is *B2P Binary EASE^R* (B2P-BE):

$$\begin{aligned}
\underset{B}{\mathrm{argmin}} \sum_n \Big(||H(\theta^*)\,(X - XB) + (1 - H(\theta^*))\mathrm{pop}(X_{t_{-1},n})&\\
- \kappa_C\langle XC\rangle||_F^2 + \lambda_B||B||_F^2\Big).&
\end{aligned} \tag{4}$$

We selected EASE$^\mathrm{R}$ due to its closed-form scalability and higher rank-accuracy than other considered baselines on the BBC iPlayer data; however, B2P is flexible and implementable with various other CF models. Figure 1 visualizes B2P-BE.

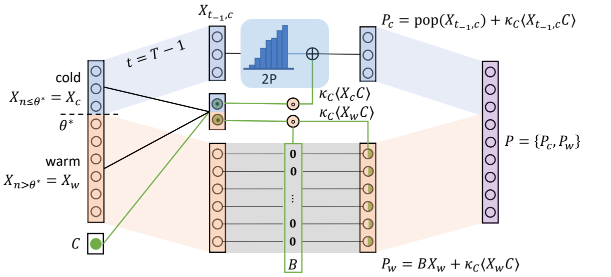
$X_{t_{-1},c}$
$t = T - 1$
2P
$P_c = \mathrm{pop}(X_{t_{-1},c}) + \kappa_C\langle X_{t_{-1},c}C\rangle$
cold
$X_{n\leq\theta^*} = X_c$
θ^*
$\kappa_C\langle X_c C\rangle$
$\kappa_C\langle X_w C\rangle$
warm
$X_{n>\theta^*} = X_w$
$P = \{P_c, P_w\}$
0
0
⋮
0
0
C
B
$P_w = BX_w + \kappa_C\langle X_w C\rangle$

Figure 1: B2P-BE architecture with binary-switch bootstrap between 2P for cold users, and Metadata-Infused EASE$^\mathrm{R}$ for warm users. EASE$^\mathrm{R}$ is represented in the grey sub-diagram with self-similarity of items from input X_w to output P_w constrained to zero.

5 EVALUATION

5.1 Experimental Set-up

Two datasets were used in the experiments for the RQs, with characteristics after pre-processing given as follows:

- BBC iPlayer (iPlayer; *Private*): 896,311 users and 2,766 items with 3,450,175 interactions.
- MovieLens 20 Million (ML-20M; *Public*): 136,677 users and 20,720 items with 9,991,282 interactions *(binarized [2])*.

Both datasets are split temporally between training and testing [47]; this split exhibits *Weak Generalization* as the split data is disjoint in terms of user-item interactions but not in terms of users. The iPlayer dataset consists of interactions in a real programme-streaming platform for 31 consecutive days in April-March 2022. It is recent and dynamic, with a non-static item catalog, and it is highly cold: 35.8% of test-set users are new on the test day ($|\mathcal{U}_\mathrm{new}| = 321{,}118$) and over 70% of the 575,193 non-new $\mathcal{U} - \mathcal{U}_\mathrm{new}$ users interact with at most 5 items in training, which is the minimum interaction level in ML-20M [18]. Thus, we use the iPlayer dataset to demonstrate

[2] ML-20M is an explicit-feedback dataset; binarization treats ratings 1-3 as non-interactions and ratings 4-5 as interactions, leading to a reduction from 20M explicit interactions to 10M implicit interactions.

the significant gains of B2P on a Cold setting, and show that Metadata Infusion in B2P also generalizes well to warmer datasets like ML-20M. We measure rank-accuracy with MAP, NDCG, MRR and HR, and diversity across catalog span and popularity with Item Coverage (Cov) and Gini Index (Gini). The top-k cutoff is 20, which is applicable to recommendation batches shown in real streaming platforms. Finally, we benchmark B2P against leading high-performance models [13, 14] across the CF families: CF and hybrid ItemKNN, graph-based P^3_α and RP^3_β [10, 30], Full-Rank EASE^R [35] and Neural Mult-VAE [24].

5.2 Popularity Modelling (RQ1)

In analysing zero-order Euler proxies for popularity distribution $\mathcal{P}_T$ among $\mathcal{U}_\mathrm{new}$, we evaluate the Popularity Model with multiple subsets $\tilde{\mathcal{U}}$ on the iPlayer dataset, as shown in Table 1. Subset $\mathcal{U}_{t_i:t_f}$ considers interactions of users on timestep range $[t_i, t_f]$, and $\mathcal{U}_{t_i:t_f,\mathrm{new}}$ considers new users at each timestep. Indeed, via the Kolmogorov-Smirnov (KS) Test, we find a significant difference in the popularity distribution of new users in the last training timestep $\mathcal{U}_{T-1,\mathrm{new}}$ to other subsets $\tilde{\mathcal{U}}$, $KS(1746, 1532) = 0.0983, p < 0.001$. Other subsets are not significantly different at $\alpha = 0.05$. This suggests the effectiveness of $\mathcal{U}_{T-1,\mathrm{new}}$ in modelling $\mathcal{P}_T \sim \mathcal{D}_T(\mathcal{U}_\mathrm{new})$, which is experimentally validated in Table 1.

Thus, $\mathcal{U}_{T-1,\mathrm{new}}$ is the most accurate from the considered zero-order Euler approximants for new and existing users, but it has near-zero diversity, as do the other samples in the Popularity Model. This is shown in the Diversity comparison with EASE^R in Figure 2. Moreover, Figure 2 shows the empirical cutoff as $\theta^* = 2$ interactions in the iPlayer dataset: the Popularity Model outperforms EASE^R in terms of MAP@20 for users with $n \in \{0, 1, 2\}$, which covers over 59.8% of all users $\mathcal{U}$ in the iPlayer dataset. Hence, the Popularity Model is strongly performant in extreme Cold Start settings. Moreover, its issue of non-personalization and near-zero diversity is moderated by Metadata Infusion in RQ2.

5.3 Metadata Infusion (RQ2)

While learning the item latent factors in Eq. 2 directly improves model *accuracy* in [9, 38], we observe in the iPlayer and ML-20M datasets that using explicit metadata as the latent factors improves model *diversity* for a non-significant trade-off in accuracy. We use metadata of M *classes – genres* in our specific datasets – that categorize items $i \in \mathcal{I}$ with Boolean latent factor $\vec{\gamma}^{(i)}_m = 1$, $m \in \{\mathbb{N}^+ | m \leq M\}$, if i is described by the m^th class. In the iPlayer dataset, $M = 127$ and only 36 of 2766 available items have no associated metadata ($\vec{\gamma} = \vec{0}$), while in ML-20M $M = 20$ and zero items have $\vec{\gamma} = \vec{0}$. Hence, minimal item-based metadata can represent most items here, which avoids the content-based over-reliance criticized in [4, 5]. This renders Metadata Infusion viable for academic and industrial applications.

On the iPlayer dataset, Figure 2 plots the Accuracy-Diversity trade-off that results from Metadata Infusion on the Popularity Model (yielding 2P) and on EASE^R (yielding $P = X\hat{B} + \kappa_C \langle XC \rangle$). Increasing κ_C weighs class exploration more heavily over learnt similarity or empirical popularity, such that accuracy is traded for diversity. This metadata-driven exploration is valuable in production systems, where users have varying levels of preference of recommendation diversity [27]. Moreover, Figure 2 shows that Metadata Infusion most affects cold users, both in the highly-significant diversity improvement and the comparatively non-significant accuracy decrease.

Table 2 gives the performance of the Metadata-Infused EASE^R at the optimal trade-off of maximum diversity gain for minimum accuracy loss. Metadata Infusion is very valuable in the warm ML-20M, but its Accuracy-Diversity trade-off is less favorable than in the colder iPlayer dataset. This further supports that Metadata Infusion is most effective in Cold Start. Indeed, in 2P with cold users, Cov@20 increased by 1340% for only 17% drop in MAP@20 (Figure 2, $\kappa_C = 5$).

5.4 Overall Results for B2P (RQ3)

The benchmarking of B2P-BE on testing users $\mathcal{U} - \mathcal{U}_\mathrm{new}$ in the iPlayer dataset is presented in Table 3. While we also use ML-20M to demonstrate the wide applicability of Metadata Infusion and favorable generalization between cold and warm datasets, we only benchmark B2P on the iPlayer dataset, since it reflects a realistic and dynamic Cold Start setting. $\mathcal{U}_\mathrm{new}$ are omitted in view of a fair benchmarking, since B2P is capable of approximating optimal new-user recommendations whereas baseline models are architecturally unable to make recommendations for new users. Overall, B2P-BE outperforms all high-performing baselines in accuracy metrics at competitive diversity, which satisfies RQ3.

B2P is highly tractable, being limited only by the scalability of the bootstrapped baseline model if $\langle XC \rangle$ fits in memory. Indeed, the Metadata Infusion of $\langle XC \rangle$ into the computation of score matrix P is batch-vectorizable addition. Also, caching latent factor elementwise products and batch-computing each row $\langle XC \rangle_{u,:}$ leads to complexity $O(n^2_\mathrm{avg} M |\mathcal{U}|)$ in computing $\langle XC \rangle$ with M-wide metadata and n_avg average interactions per user.

Further, while existing Cold Start models are applied to a CF backbone trained on all users' interactions, B2P handles cold-user recommendations with 2P so it reduces the data sparsity of the bootstrapped baseline by only training it on warm users $n > \theta^*$. This implicitly boosts performance by avoiding the Cold Start problem for the bootstrapped model. Finally, B2P can act as the CF Backbone to complement other Cold Start methods.

6 CONCLUSION

In this work we proposed the novel, efficient, scalable and principled framework B2P for directly addressing the Cold Start problem. We support B2P by deriving the optimality of popularity-based recommendations for new users, their valid extension to cold users, and their personalization via the novel diversity-boosting method of Metadata Infusion. B2P outperforms current high-performing CF and hybrid recommenders in accuracy across both cold and warm users, and maintains highly competitive diversity on a real, dynamic and large-scale Cold Start iPlayer dataset. Future work could explore higher-order popularity proxies and learnt bootstrapping configurations to extend B2P.

Table 1: Performance of different training user subsets $\tilde{\mathcal{U}}$ for new-user recommendations on iPlayer dataset. Cutoff $k = 20$. Results are shown when evaluating against the testing interactions of new testing users $\mathcal{U}_{\text{new}}$, and of non-new testing users $\mathcal{U}_{nn} = \mathcal{U} - \mathcal{U}_{\text{new}}$.

Train Subset	Evaluation on $\mathcal{U}_{\text{new}}$						Evaluation on $\mathcal{U}_{nn}$	
	MAP	NDCG	MRR	HR	Cov	Gini	MAP	Cov
$\mathcal{U}$	0.0304	0.0567	0.0334	0.1627	0.0072	0.0076	0.0334	**0.0163**
$\mathcal{U}_{1:T-1,\text{new}}$	0.0301	0.0560	0.0330	0.1603	**0.0094**	**0.0077**	0.0301	0.0098
$\mathcal{U}_{T-1}$	0.0617	0.1171	0.0678	**0.3279**	0.0072	0.0076	0.0620	0.0090
$\mathcal{U}_{T-1,\text{new}}$	**0.1056**	**0.1509**	**0.1150**	0.3195	0.0072	0.0076	**0.1051**	0.0083

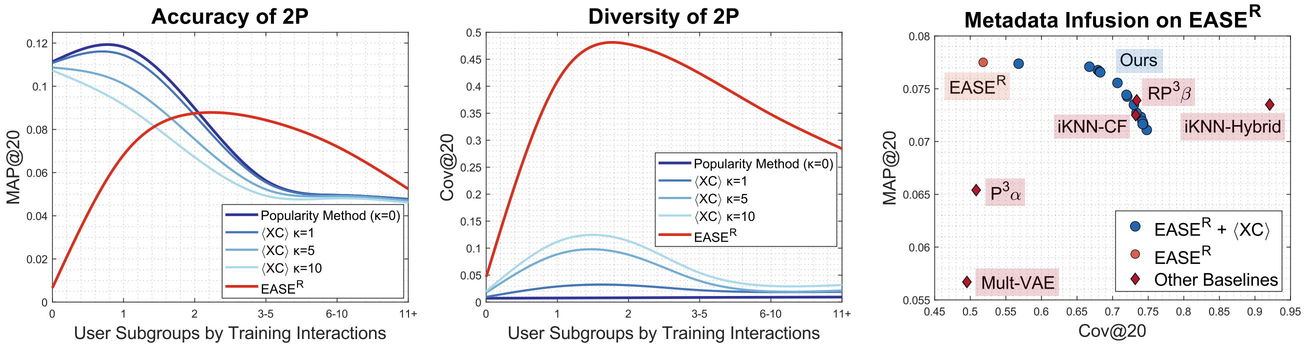

Figure 2: Left and Middle plots show, respectively, the MAP@20 and Cov@20 per stratum $\mathcal{U}_n$ for 2P with varying κ_C, relative to EASE$^{\text{R}}$. Right plot shows MAP@20 and Cov@20 for Metadata-Infused EASE$^{\text{R}}$ with varying κ_C, relative to EASE$^{\text{R}}$ and other baselines (*iKNN denotes ItemKNN*). All model instances in all plots are evaluated on the iPlayer dataset.

Table 2: Evaluation of EASE$^{\text{R}}$ with Metadata Infusion on iPlayer dataset ($\kappa_C = 0.5$) and ML-20M ($\kappa_C = 1$) datasets. Cutoff $k = 20$. New test users $\mathcal{U}_{\text{new}}$ are omitted from evaluation. Percentages show increase or decrease in metrics due to Metadata Infusion.

Dataset	Model	MAP	NDCG	MRR	HR	Cov	Gini
iPlayer	EASE$^{\text{R}}$	**0.0775**	**0.1195**	**0.0857**	**0.2843**	0.5181	0.0382
	EASE$^{\text{R}}$ + $\kappa_C\langle XC\rangle$	$0.0769^{-0.8\%}$	0.1181	0.0850	0.2800	$\mathbf{0.6670}^{+28.7\%}$	$\mathbf{0.0456}^{+19.4\%}$
ML-20M	EASE$^{\text{R}}$	**0.0524**	**0.1093**	**0.1420**	**0.4761**	0.1557	0.0185
	EASE$^{\text{R}}$ + $\kappa_C\langle XC\rangle$	$0.0514^{-1.9\%}$	0.1077	0.1392	0.4722	$\mathbf{0.1780}^{+14.3\%}$	$\mathbf{0.0199}^{+7.6\%}$

Table 3: Evaluation of B2P-BE against baseline models on iPlayer dataset omitting new test users. Cutoff $k = 20$.

Model	MAP	NDCG	MRR	HR	Cov	Gini
ItemKNN-CF	0.0725	0.1106	0.0803	0.2591	0.7325	0.0344
ItemKNN-Hybrid	0.0735	0.1119	0.0815	0.2615	**0.9208**	0.0387
P$^3\alpha$	0.0654	0.1025	0.0724	0.2461	0.5083	**0.0555**
RP$^3\beta$	0.0739	0.1123	0.0818	0.2622	0.7339	0.0419
EASE$^{\text{R}}$	0.0775	0.1197	0.0858	0.2842	0.5181	0.0382
Mult-VAE	0.0567	0.0947	0.0630	0.2449	0.4953	0.0573
B2P-BE	**0.0869**	**0.1310**	**0.0922**	**0.3014**	0.5803	0.0348

ACKNOWLEDGMENTS

Part of this research has been funded by the EPSRC Impact Acceleration Account (IAA) award (Grant Reference EP/R511638/1).

REFERENCES

[1] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2019. Managing Popularity Bias in Recommender Systems with Personalized Re-ranking. In *FLAIRS*. AAAI Press, Sarasota, FL, USA, 413–418.

[2] Julian Besag. 1975. Statistical Analysis of Non-Lattice Data. *Journal of the Royal Statistical Society. Series D (The Statistician)* 24, 3 (9 1975), 179–195. https://doi.org/10.2307/2987782

[3] Julian Besag. 1977. Efficiency of Pseudolikelihood Estimation for Simple Gaussian Fields. *Biometrika* 64, 3 (12 1977), 616–618. https://doi.org/10.2307/2345341

[4] Dong-Kyu Chae, Jin-Soo Kang, Sang-Wook Kim, and Jaeho Choi. 2019. Rating Augmentation with Generative Adversarial Networks towards Accurate Collaborative Filtering. In *The World Wide Web Conference*. ACM, New York, NY, USA, 2616–2622. https://doi.org/10.1145/3308558.3313413

[5] Dong-Kyu Chae, Jihoo Kim, Duen Horng Chau, and Sang-Wook Kim. 2020. AR-CF: Augmenting Virtual Users and Items in Collaborative Filtering for Addressing Cold-Start Problems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 1251–1260. https://doi.org/10.1145/3397271.3401038

[6] Hung-Hsuan Chen and Pu Chen. 2019. Differentiating Regularization Weights – A Simple Mechanism to Alleviate Cold Start in Recommender Systems. *ACM Transactions on Knowledge Discovery from Data* 13, 1 (2 2019), 1–22. https://doi.org/10.1145/3285954

[7] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and Debias in Recommender System: A Survey and Future Directions. *ACM Transactions on Information Systems* 41, 3 (7 2023), 1–39. https://doi.org/10.1145/3564284

[8] Weiyu Cheng, Yanyan Shen, Yanmin Zhu, and Linpeng Huang. 2018. DELF: A Dual-Embedding based Deep Latent Factor Model for Recommendation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, California, 3329–3335. https://doi.org/10.24963/ijcai.2018/462

[9] Evangelia Christakopoulou and George Karypis. 2014. HOSLIM: Higher-Order Sparse LInear Method for Top-N Recommender Systems. In *Advances in Knowledge Discovery and Data Mining*. Springer, Tainan, Taiwan, 38–49. https://doi.org/10.1007/978-3-319-06605-9_4

[10] Colin Cooper, Sang Hyuk Lee, Tomasz Radzik, and Yiannis Siantos. 2014. Random walks in recommender systems. In *Proceedings of the 23rd International Conference on World Wide Web*. ACM, New York, NY, USA, 811–816. https://doi.org/10.1145/2567948.2579244

[11] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 101–109. https://doi.org/10.1145/3298689.3347058

[12] Luc Devroye, László Györfi, and Gábor Lugosi. 1996. *A Probabilistic Theory of Pattern Recognition*. Vol. 31. Springer-Verlag, New York. 1–20 pages.

[13] Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. 2021. A Troubling Analysis of Reproducibility and Progress in Recommender Systems Research. *ACM Transactions on Information Systems* 39, 2 (1 2021), 1–49. https://doi.org/10.1145/3434185

[14] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2021. Methodological Issues in Recommender Systems Research (Extended Abstract). In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Yokohama, Japan, 4706–4710. https://doi.org/10.24963/ijcai.2020/650

[15] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. 2010. Beyond accuracy. In *Proceedings of the fourth ACM conference on Recommender systems*. ACM Press, New York, New York, USA, 257–260. https://doi.org/10.1145/1864708.1864761

[16] Carlos A. Gomez-Uribe and Neil Hunt. 2015. The Netflix Recommender System. *ACM Transactions on Management Information Systems* 6, 4 (12 2015), 1–19. https://doi.org/10.1145/2843948

[17] Guibing Guo. 2013. Improving the Performance of Recommender Systems by Alleviating the Data Sparsity and Cold Start Problems. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*. AAAI Press, Beijing, China, 3217–3218.

[18] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets. *ACM Transactions on Interactive Intelligent Systems* 5, 4 (12 2015), 1–19. https://doi.org/10.1145/2827872

[19] Zan Huang, Hsinchun Chen, and Daniel Zeng. 2004. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. *ACM Transactions on Information Systems* 22, 1 (1 2004), 116–142. https://doi.org/10.1145/963770.963775

[20] Noor Ifada, Ummamah, and Mochammad Kautsar. 2020. Hybrid popularity model for solving cold-start problem in recommendation system. In *Proceedings of the 5th International Conference on Sustainable Information Engineering and Technology*. ACM, New York, NY, USA, 40–44. https://doi.org/10.1145/3427423.3427425

[21] Sheena S. Iyengar and Mark R. Lepper. 2000. When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology* 79, 6 (12 2000), 995–1006. https://doi.org/10.1037/0022-3514.79.6.995

[22] Hyeyoung Ko, Suyeon Lee, Yoonseo Park, and Anna Choi. 2022. A Survey of Recommendation Systems: Recommendation Models, Techniques, and Application Fields. *Electronics* 11, 1 (1 2022), 141. https://doi.org/10.3390/electronics11010141

[23] Steffen L. Lauritzen. 1996. *Graphical Models*. Clarendon Press, Oxford, UK.

[24] Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. 2018. Variational Autoencoders for Collaborative Filtering. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, Geneva, Switzerland, 689–698. https://doi.org/10.1145/3178876.3186150

[25] Shoshana Loeb and Douglas Terry. 1992. Information filtering. *Commun. ACM* 35, 12 (12 1992), 26–28. https://doi.org/10.1145/138859.138860

[26] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback Loop and Bias Amplification in Recommender Systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. ACM, New York, NY, USA, 2145–2148.

[27] Sean M. McNee, John Riedl, and Joseph A. Konstan. 2006. Being accurate is not enough. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1097–1101. https://doi.org/10.1145/1125451.1125659

[28] Xia Ning and George Karypis. 2011. SLIM: Sparse Linear Methods for Top-N Recommender Systems. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*. IEEE Computer Society, Washington DC, USA, 497–506. https://doi.org/10.1109/ICDM.2011.134

[29] Ruilin Pan, Chuanming Ge, Li Zhang, Wei Zhao, and Xun Shao. 2020. A New Similarity Model Based on Collaborative Filtering for New User Cold Start Recommendation. *IEICE Transactions on Information and Systems* E103-D, 6 (6 2020), 1388–1394. https://doi.org/10.1587/transinf.2019EDP7258

[30] Bibek Paudel, Fabian Christoffel, Chris Newell, and Abraham Bernstein. 2016. Updatable, Accurate, Diverse, and Scalable Recommendations for Interactive Applications. *ACM Transactions on Interactive Intelligent Systems* 7, 1 (12 2016), 1–34. https://doi.org/10.1145/2955101

[31] Yongli Ren, Gang Li, Jun Zhang, and Wanlei Zhou. 2012. The efficient imputation method for neighborhood-based collaborative filtering. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, New York, NY, USA, 684–693. https://doi.org/10.1145/2396761.2396849

[32] Yongli Ren, Gang Li, Jun Zhang, and Wanlei Zhou. 2013. AdaM. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, New York, NY, USA, 628–635. https://doi.org/10.1145/2492517.2492565

[33] Steffen Rendle. 2010. Factorization Machines. In *2010 IEEE International Conference on Data Mining*. IEEE, New York, NY, USA, 995–1000. https://doi.org/10.1109/ICDM.2010.127

[34] Harald Steck. 2011. Item popularity and recommendation accuracy. In *Proceedings of the fifth ACM conference on Recommender systems*. ACM, New York, NY, USA, 125–132. https://doi.org/10.1145/2043932.2043957

[35] Harald Steck. 2019. Embarrassingly Shallow Autoencoders for Sparse Data. In *The World Wide Web Conference*. ACM, New York, NY, USA, 3251–3257. https://doi.org/10.1145/3308558.3313710

[36] Harald Steck. 2019. Markov Random Fields for Collaborative Filtering. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates, Red Hook, NY, USA, 5473–5484.

[37] Harald Steck, Linas Baltrunas, Ehtsham Elahi, Dawen Liang, Yves Raimond, and Justin Basilico. 2021. Deep Learning for Recommender Systems: A Netflix Case Study. *AI Magazine* 42, 3 (11 2021), 7–18. https://doi.org/10.1609/aimag.v42i3.18140

[38] Harald Steck and Dawen Liang. 2021. Negative Interactions for Improved Collaborative Filtering: Don't go Deeper, go Higher. In *Proceedings of the 15th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 34–43. https://doi.org/10.1145/3460231.3474273

[39] Stuart Jeffries. 2015. Why too much choice is stressing us out.

[40] Xiaoyuan Su and Taghi M. Khoshgoftaar. 2009. A Survey of Collaborative Filtering Techniques. *Advances in Artificial Intelligence* 2009, 4 (1 2009), 2. https://doi.org/10.1155/2009/421425

[41] Wenlong Sun, Sami Khenissi, Olfa Nasraoui, and Patrick Shafto. 2019. Debiasing the Human-Recommender System Feedback Loop in Collaborative Filtering. In *Companion Proceedings of The 2019 World Wide Web Conference*. ACM, New York, NY, USA, 645–651. https://doi.org/10.1145/3308560.3317303

[42] Maksims Volkovs, Guangwei Yu, and Tomi Poutanen. 2017. DropoutNet: Addressing Cold Start in Recommender Systems. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Accociates, Red Hook, NY, USA, 4964–4973. https://doi.org/10.5555/3295222.3295249

[43] Sanne Vrijenhoek, Gabriel Bénédict, Mateo Gutierrez Granada, Daan Odijk, and Maarten De Rijke. 2022. RADio – Rank-Aware Divergence Metrics to Measure Normative Diversity in News Recommendations. In *Proceedings of the 16th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 208–219. https://doi.org/10.1145/3523227.3546780

[44] Qinyong Wang, Hongzhi Yin, Hao Wang, Quoc Viet Hung Nguyen, Zi Huang, and Lizhen Cui. 2019. Enhancing Collaborative Filtering with Generative Augmentation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, New York, NY, USA, 548–556. https://doi.org/10.1145/3292500.3330873

[45] Xueting Wang, Yiwei Zhang, and Toshihiko Yamasaki. 2020. Earn More Social Attention: User Popularity Based Tag Recommendation System. In *Companion Proceedings of the Web Conference 2020*. ACM, New York, NY, USA, 212–216. https://doi.org/10.1145/3366424.3383543

[46] An Zhang, Jingnan Zheng, Xiang Wang, Yancheng Yuan, and Tat-Seng Chua. 2023. Invariant Collaborative Filtering to Popularity Distribution Shift. In *Proceedings of the ACM Web Conference 2023*. ACM, New York, NY, USA, 1240–1251. https://doi.org/10.1145/3543507.3583461

[47] Wayne Xin Zhao, Zihan Lin, Zhichao Feng, Pengfei Wang, and Ji-Rong Wen. 2022. A Revisiting Study of Appropriate Offline Evaluation for Top-N Recommendation Algorithms. *ACM Transactions on Information Systems* 41, 2 (12 2022), 1–41. https://doi.org/10.1145/3545796

[48] Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, Yingqian Min, Zhichao Feng, Xinyan Fan, Xu Chen, Pengfei Wang, Wendi Ji, Yaliang Li, Xiaoling Wang, and Ji-Rong Wen. 2022. RecBole 1.1.1 Documentation: recbole.evaluator.metrics. https://recbole.io/docs/recbole/recbole.evaluator.metrics.html?highlight=metrics#module-recbole.evaluator.metrics