

FairMatch: A Graph-based Approach for Improving Aggregate Diversity in Recommender Systems

Masoud Mansoury*
Eindhoven University of Technology
Eindhoven, the Netherlands
m.mansoury@tue.nl

Himan Abdollahpouri
University of Colorado Boulder
Boulder, USA
himan.abdollahpouri@colorado.edu

Mykola Pechenizkiy
Eindhoven University of Technology
Eindhoven, the Netherlands
m.pechenizkiy@tue.nl

Bamshad Mobasher
DePaul University
Chicago, USA
mobasher@cs.depaul.edu

Robin Burke
University of Colorado Boulder
Boulder, USA
robin.burke@colorado.edu

ABSTRACT

Recommender systems are often biased toward popular items. In other words, few items are frequently recommended while the majority of items do not get proportionate attention. That leads to low coverage of items in recommendation lists across users (i.e. low aggregate diversity) and unfair distribution of recommended items. In this paper, we introduce FairMatch, a general graph-based algorithm that works as a post-processing approach after recommendation generation for improving aggregate diversity. The algorithm iteratively finds items that are rarely recommended yet are high-quality and add them to the users' final recommendation lists. This is done by solving the maximum flow problem on the recommendation bipartite graph. While we focus on aggregate diversity and fair distribution of recommended items, the algorithm can be adapted to other recommendation scenarios using different underlying definitions of fairness. A comprehensive set of experiments on two datasets and comparison with state-of-the-art baselines show that FairMatch, while significantly improving aggregate diversity, provides comparable recommendation accuracy.

CCS CONCEPTS

• **Information systems** → **Personalization**; **Information retrieval diversity**; **Recommender systems**.

KEYWORDS

Recommender Systems, Fairness, Popularity bias, Recommendation coverage, Long-tail, Aggregate diversity

ACM Reference Format:

Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. FairMatch: A Graph-based Approach for

*This author also has affiliation in School of Computing, DePaul University, Chicago, USA, mmansou4@depaul.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

UMAP '20, July 14–17, 2020, Genoa, Italy

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-6861-2/20/07...\$15.00

<https://doi.org/10.1145/3340631.3394860>

Improving Aggregate Diversity in Recommender Systems. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20)*, July 14–17, 2020, Genoa, Italy. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3340631.3394860>

1 INTRODUCTION

Recommender systems are used in a variety of different applications including movies, music, e-commerce, online dating, and many other areas where the number of options from which the user needs to choose can be overwhelming. There are many different metrics to evaluate the performance of the recommender systems ranging from accuracy metrics such as precision, normalized discounted cumulative gain (NDCG), and recall to non-accuracy ones like novelty and serendipity [16]. One of the measures often used to evaluate the effectiveness of a given recommender system is how diverse the list of recommendations given to each user is (aka individual list diversity) [15]. Recommending a diverse list of items is shown to improve user satisfaction as they give a wider range of options to the user [7].

The problem with individual list diversity is that it does not capture the extent to which an algorithm covers a diverse set of items across all users which is an important consideration for many applications. Aggregate diversity [4] is a notion to measure this characteristic of the recommender systems and several algorithms have been proposed for that matter by other researchers [4, 5]. Note that a high *individual list diversity* of recommendations does not necessarily imply *high aggregate diversity*. For instance, if the system recommends to all users the same 10 items that are not similar to each other, the recommendation list for each user is diverse (i.e., high individual list diversity), but only 10 distinct items are recommended to all users (i.e., resulting in low aggregate diversity).

An algorithm with low aggregate diversity could be problematic for several reasons. On the one hand, it concentrates on a limited number of popular items which, in the long run, might negatively affect users' experience in terms of item discovery. Users already know about popular items and recommending them would not add any new information. On the other hand, often items belong to different suppliers and, hence, covering fewer distinct items can indirectly result in an unfair distribution of items across recommendations from the suppliers' perspective. Thus, a low aggregate

diversity in recommendation results would have a negative impact on business success and profit [6, 12].

In this paper, we introduce *FairMatch*, a general graph-based algorithm that works as a post-processing approach after recommendation generation (on top of any existing standard recommendation algorithm) for improving the aggregate diversity. The idea is to generate a list of recommendations with a size larger than what we ultimately want for the final list using a standard recommendation algorithm and then use our *FairMatch* algorithm to build the final list using a subset of items in the original list. In *FairMatch*, the main goal is to improve the visibility of high-quality items that have a low visibility in the original set of recommendations. This is done by iteratively solving *Maximum Flow* problem on a recommendation bipartite graph which is built using the recommendations in the original list (left nodes are recommended items and right nodes are the users). At each iteration, the items that can be good candidates for the final list will be selected and removed from the graph, and the process will continue on the remaining part of the graph.

To show the effectiveness of our *FairMatch* algorithm on improving aggregate diversity and fair visibility of recommended items, we perform a comprehensive set of experiments on recommendation lists of different sizes generated by two standard recommendation algorithms on two publicly available datasets. We intentionally picked two algorithms from two different classes of algorithms (factorization and neighborhood-based models), so our approach is not dependent on any certain type of recommendation algorithms.

Comparison with several state-of-the-art baselines shows that our *FairMatch* algorithm is able to significantly improve the performance of recommendation results in terms of aggregate diversity and long-tail visibility, with a negligible loss in the recommendation accuracy in some cases.

2 RELATED WORK

The concept of aggregate diversity has been studied by many researchers often under different names such as long-tail recommendation [1, 25], Matthew effect [20] and, of course, aggregate diversity [3, 17] all of which refer to the fact that the recommender system should recommend a wider variety of items across all users.

Vargas and Castells in [23] proposed probabilistic models for improving novelty and diversity of recommendations by taking into account both relevance and novelty of target items when generating recommendation lists. In another work [24], they proposed the idea of recommending users to items for improving novelty and aggregate diversity. They applied this idea to nearest neighbor models as an inverted neighbor and a factorization model as a probabilistic reformulation that isolates the popularity components.

Adomavicius and Kwon [4] proposed the idea of diversity maximization using a maximum flow approach. They used a specific setting for the bipartite recommendation graph in a way that the maximum amount of flow that can be sent from a source node to a sink node would be equal to the maximum aggregate diversity for those recommendation lists. In their setting, given the number of users is m , the source node can send a flow of up to m to the left nodes, left nodes can send a flow of up to 1 to the right nodes, and right nodes can send a flow of up to 1 to the sink node. Since the

capacity of left nodes to right nodes is set to 1, thus the maximum possible amount of flow through that recommendation bipartite graph would be equivalent to the maximum aggregate diversity.

A more recent graph-based approach for improving aggregate diversity was proposed by Antikacioglu and Ravi in [5]. They generalized the idea proposed in [4] and showed that the minimum-cost network flow method can be efficiently used for finding recommendation subgraphs that optimizes the diversity. In this work, an integer-valued constraint and an objective function are introduced for discrepancy minimization. The constraint defines the maximum number of times that each item should appear in the recommendation lists and the objective function aims to find an optimal subgraph that gives the minimum discrepancy from the constraint. This work shows improvement in aggregate diversity with a smaller accuracy loss compared to the work in [23] and [24]. Similar to this work, our *FairMatch* algorithm also uses a graph-based approach to improve aggregate diversity. However, unlike the work in [5] which tries to minimize the discrepancy between the distribution of the recommended items and a target distribution, our *FairMatch* algorithm has more freedom in promoting high-quality items with low visibility since it does not assume any target distribution of the recommendation frequency.

3 FAIRMATCH ALGORITHM

We formulate our *FairMatch* algorithm as a post-processing step after the recommendation generation. In other words, we first generate recommendation lists of larger size than what we ultimately desire for each user using any standard recommendation algorithm and use them to build the final recommendation lists. *FairMatch* works as a batch process, similar to that proposed in [28] where all the recommendation lists are produced at once and re-ranked simultaneously to achieve the objective. In this formulation, we produce a longer recommendation list of size t for each user and then, after identifying high-quality items (items closer to the top of the list) with low visibility (i.e. are not recommended frequently) by iteratively solving the maximum flow problem on recommendation bipartite graph, we generate a shorter recommendation list of size n (where $t \gg n$).

Let $G = (I, U, E)$ be a bipartite graph of recommendation lists where I is the set of left nodes (representing items), U is the set of right nodes (representing users), and E is the set of edges between left and right nodes showing an item in the left nodes is recommended to a user in the right nodes in recommendation lists of size t . G is initially a uniformly weighted graph, but we will update the weights for edges as part of our algorithm. We will discuss the initialization and our weighting method in section 3.2.

Given a weighted bipartite graph G , the goal of our *FairMatch* algorithm is to find high-quality items with low visibility and maximizing their visibility as much as possible without a significant loss in accuracy of the recommendations. Visibility is characterized by the degree of the node in the recommendation graph, while accuracy is captured by the rank position of the items in the original recommendation list. We develop our algorithm by extending the approach introduced in [10] to improve the aggregate diversity of the recommender systems.

We use an iterative process to identify the subgraphs of G that contain the highest quality items with low visibility for each user. After identifying a subgraph Γ at each iteration, we remove Γ from G and continue the process of finding subgraphs on the rest of the graph (i.e., G/Γ). We keep track of all the subgraphs as we use them to generate the final recommendations in the last step.

Identifying Γ at each iteration is done by solving a *Maximum Flow* problem (explained in section 3.3) on the graph obtained from the previous iteration. Solving the maximum flow problem returns the left nodes connected to the edges with lower weight (i.e., more relevant items with low visibility) on the graph. After finding those left nodes, we form subgraph Γ by separating identified left nodes and their connected right nodes from G . Finally, $\langle \text{user}, \text{item} \rangle$ pairs in subgraphs are used to construct the final recommendation lists of size n . We will discuss this process in detail in the following sections.

Algorithm 1 shows the pseudocode for FairMatch. Overall, our FairMatch algorithm consists of the following four steps: 1) Graph preparation, 2) Weight computation, 3) Candidate selection, and 4) Recommendation list construction.

Algorithm 1 The FairMatch Algorithm

```

function FAIRMATCH(Recommendations  $R$ , TopN  $n$ , Coefficient  $\alpha$ )
  Build graph  $G = (I, U, E)$  from  $R$ 
  Initialize subgraphs to empty
  repeat
     $G = \text{WeightComputation}(G, R, \alpha)$ 
     $I_C = \text{Push-relabel}(G)$ 
    Initialize subgraph to empty
    for each  $i \in I_C$  do
      if  $\text{label}_i \geq |I| + |U| + 2$  then
        for each  $u \in \text{Neighbors}(i)$  do
          Append  $\langle i, u, e_{iu} \rangle$  to subgraph
        end for
      end if
    end for
    if subgraph is empty then
      break
    end if
    Append subgraph to subgraphs
     $G = \text{Remove subgraph from } G$ 
  until (true)
  Reconstruct  $R$  of size  $n$  based on subgraphs
end function

```

3.1 Graph Preparation

Given long recommendation lists of size t generated by a standard recommendation algorithm, we create a bipartite graph from recommendation lists in which items and users are the nodes (called, respectively, left and right nodes) and recommendations are expressed as edges. Since our FairMatch algorithm is formulated as a maximum flow problem, we also add two nodes, *source* (s_1) and *sink* (s_2). The purpose of having a source and sink node in the maximum flow problem is to have a start and endpoint for the flow going through the graph. We connect s_1 node to all left nodes and also we connect all right nodes to s_2 . Figure 1 shows a sample bipartite graph resulted in this step.

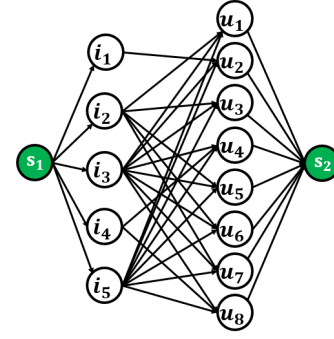


Figure 1: An example of a recommendation bipartite graph of recommendation lists of size 3.

3.2 Weight Computation

Given the bipartite recommendation graph, $G = (I, U, E)$, the task of weight computation is to calculate the weight for edges between the source node and left nodes, left nodes and right nodes, and right nodes and sink node.

For edges between left nodes and right nodes, we define the weights as the weighted sum of item visibility and relevance. The visibility of each item is defined as the degree of the node corresponding to that item (excluding the edge with the source node). Item degree is the number of edges going out from that node connecting it to the user nodes and that shows how often it is recommended to different users. Relevance is based on the rank of the item in the original recommendation list for each user (lower rank is more relevant).

For computing the weight between $i \in I$ and $u \in U$, we use the following equation:

$$w_{iu} = \alpha \times \text{degree}_i + (1 - \alpha) \times \text{rank}_{iu} \quad (1)$$

where degree_i is the number of edges from i to right nodes (i.e., $u \in U$), rank_{iu} is the position of item i in the recommendation list of size t generated for user u , and α is a coefficient to control the trade-off between accuracy and diversity (or visibility).

Note that in equation 1, degree and rank have different ranges. The range for rank is from 1 to t (there are t different positions in the original list) and the range of degree depends on the frequency of the item recommended to the users (the more frequent it is recommended to different users the higher its degree is). Hence, for a meaningful weighted sum, we normalize degree to be in the same range as rank .

Given weights of the edges between $i \in I$ and $u \in U$, w_{iu} , total capacity of I and U would be $C_T = \sum_{i \in I} \sum_{u \in U} w_{iu}$ which simply shows the sum of the weights of the edges connecting left nodes to the right nodes.

For computing the weight for edges connected to the source and sink nodes, first, we equally distribute C_T to left and right nodes. Therefore, the capacity of each left node, $C_{eq}(I)$, and right node, $C_{eq}(U)$, would be as follow:

$$C_{eq}(I) = \left\lceil \frac{C_T}{|I|} \right\rceil, \quad C_{eq}(U) = \left\lceil \frac{C_T}{|U|} \right\rceil \quad (2)$$

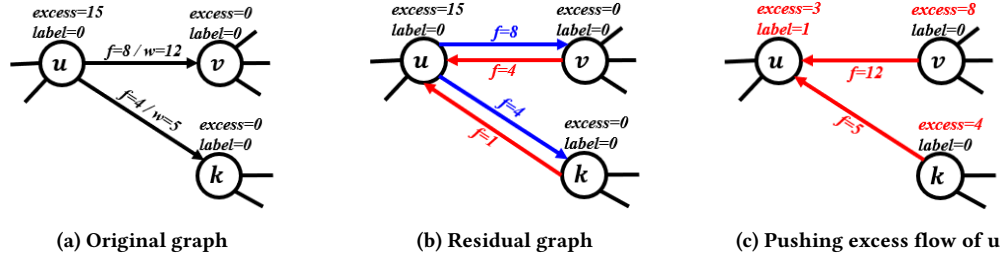


Figure 2: Example of push and relabel operation.

where $\lceil a \rceil$ returns the ceil value of a . Then, based on equal capacity assigned to each left and right nodes, we follow the method introduced in [10] to compute weights for edges connected to source and sink nodes as follow:

$$\forall i \in I, w_{s_1 i} = \left\lceil \min\left(\frac{C_{eq}(I)}{\gcd(C_{eq}(I), C_{eq}(U))}, \frac{C_{eq}(U)}{\gcd(C_{eq}(I), C_{eq}(U))}\right) \right\rceil \quad (3)$$

$$\forall u \in U, w_{us_2} = \left\lceil \frac{C_{eq}(I)}{\gcd(C_{eq}(I), C_{eq}(U))} \right\rceil \quad (4)$$

where $\gcd(C_{eq}(I), C_{eq}(U))$ is the Greatest Common Divisor of the distributed capacity of left and right nodes. Assigning the same weight to edges connected to the source and sink nodes guaranties that all nodes in I and U are treated equally and the weights between them play an important role in our FairMatch algorithm.

3.3 Candidate Selection

The graph constructed in previous steps is ready to be used for solving the maximum flow problem. In a maximum flow problem, the main goal is to find the maximum amount of feasible flow that can be sent from the source node to the sink node through the flow network. Several algorithms have been proposed for solving a maximum flow problem. Well-known algorithms are Ford–Fulkerson [9], Push-relabel [11], and Dinic’s algorithm [8]. In this paper, we use Push-relabel algorithm to solve the maximum flow problem on our bipartite recommendation graph as it is one of the efficient algorithms for this matter.

In push-relabel algorithm, each node will be assigned two attributes: *label* and *excess flow*. The label attribute is an integer value that is used to identify the neighbors to which the current node can send flow. A node can only send flow to neighbors that have lower label than the current node. Excess flow is the remaining flow of a node that can still be sent to the neighbors. When all nodes of the graph have excess flow equals to zero, the algorithm will terminate.

The push-relabel algorithm combines *push* operations that send a specific amount of flow to a neighbor, and *relabel* operations that change the label of a node under a certain condition (when the node has excess flow greater than zero and there is no neighbor with label lower than the label of this node).

Here is how the push-relabel algorithm works: Figure 2 shows a typical graph in the maximum flow problem and an example of push and relabel operations. In Figure 2a, f and w are current flow and weight of the given edge, respectively. In Push-relabel

algorithm, a residual graph, G' , will be also created from graph G . As graph G shows the flow of forward edges, graph G' shows the flow of backward edges calculated as $f_{backward} = w - f$. Figure 2b shows residual graph of graph G in Figure 2a. Now, we want to perform a push operation on node u and send its excess flow to its neighbors.

Given x_u as excess flow of node u , $push(u, v)$ operation will send a flow of amount $\Delta = \min(x_u, f_{uv})$ from node u to node v and then will decrease excess flow of u by Δ (i.e., $x_u = x_u - \Delta$) and will increase excess flow of v by Δ (i.e., $x_v = x_v + \Delta$). After $push(u, v)$ operation, node v will be put in a queue of active nodes to be considered by the push-relabel algorithm in the next iterations and residual graph would be updated. Figure 2c shows the result of $push(u, v)$ and $push(u, k)$ on the graph shown in Figure 2b. In $push(u, v)$, for instance, since u and all of its neighbors have the same label value, in order to perform push operation, first we need to perform relabel operation on node u to increase the label of u by one unit more than the minimum label of its neighbors to guaranty that there is at least one neighbor with lower label for performing push operation. After that, node u can send flow to its neighbors.

Given $x_u = 15$, $f_{uv} = 8$, and $f_{uk} = 4$ in Figure 2b, after performing relabel operation, we can only send the flow of amount 8 from u to v and the flow of amount 4 from u to k . After these operations, residual graph (backward flow from v and k to u) will be updated.

The push-relabel algorithm starts with a "preflow" operation to initialize the variables and then it iteratively performs push or relabel operations until no active node exists for performing operations. Assuming \mathcal{L}_v as the label of node v , in preflow step, we initialize all nodes as follow: $\mathcal{L}_{s_1} = |I| + |U| + 2$, $\mathcal{L}_{i \in I} = 2$, $\mathcal{L}_{u \in U} = 1$, and $\mathcal{L}_{s_2} = 0$. This way, we will be able to send the flow from s_1 to s_2 as the left nodes have higher label than the right nodes. Also, we will push the flow of amount $w_{s_1 i}$ (where $i \in I$) from s_1 to all the left nodes.

After preflow, all of the left nodes $i \in I$ will be in the queue, Q , as active nodes because all those nodes now have positive excess flow. The main part of the algorithm will now start by dequeuing an active node v from Q and performing either push or relabel operations on v as explained above. This process will continue until Q is empty. At the end, each node will have specific label value and the sum of all the coming flows to node s_2 would be the maximum flow of graph G . For more details see [11]

An important question is: *how does the Push-relabel algorithm can find high-quality nodes (items) with low degree (visibility)?*

We answer this question by referring to the example in Figure 2c. In this figure, assume that u has a backward edge to s_1 . Since u has excess flow greater than zero, it should send it to its neighbors. However, as you can see in the figure, u does not have any forward edge to v or k nodes. Therefore, it has to send its excess flow back to s_1 as s_1 is the only reachable neighbor for u . Since s_1 has the highest label in our setting, in order for u to push all its excess flow back to s_1 , it should go through a relabel operation so that its label becomes larger than that of s_1 . Therefore, the label of u will be set to $\mathcal{L}_{s_1} + 1$ for an admissible push.

The reason that u receives high label value is the fact that it initially receives high flow from s_1 , but it does not have enough capacity (the sum of weights between u and its neighbors is smaller than its excess flow. i.e. $8+4 < 15$) to send all that flow to them. In FairMatch, in step 3 (i.e. section 3.3), left nodes without sufficient capacity on their edges will be returned as part of the outputs from push-relabel algorithm and are considered for constructing the final recommendation list in step 4 (i.e. section 3.4). These nodes are the ones that their edges received low weights by equation 1 in step 2 (i.e. section 3.2) because of their low degree (low visibility) and rank (high relevance) on the graph. Therefore, FairMatch aims at promoting those high relevance items with low visibility.

3.4 Recommendation List Construction

In this step, the goal is to construct a recommendation list of size n by the $\langle user, item \rangle$ pairs identified in previous step.

Given a recommendation list of size n for user u , R_u , sorted based on the scores generated by a standard recommendation algorithm, candidate items identified by FairMatch connected to u as I_C , and visibility of each item, i , in recommendation lists of size n as V_i , we use the following process for generating recommendation list for u . First, we sort recommended items in R_u based on their V_i in ascending order. Then, we remove $\min(n, |I_C|)$ from the bottom of sorted R_u and add $\min(n, |I_C|)$ items from I_C to the end of R_u .

This process will ensure that extracted items in the previous step will replace the frequently recommended items meaning that it decreases the visibility of the frequently recommended items and increases the visibility of rarely recommended items to generate a fair distribution on recommended items.

4 EXPERIMENTS

We performed a comprehensive evaluation of the effectiveness of FairMatch in improving aggregate diversity of recommender systems. Our evaluation on two standard recommendation algorithms and comparison to various diversification methods to increase aggregate diversity as baselines on two datasets shows that FairMatch significantly improves item visibility with a negligible loss in the accuracy of recommendations.

4.1 Experimental Setup

Experiments are performed on two publicly available datasets: Epinions and MovieLens. The Epinions dataset was collected from Epinions web site which is an item reviewing system. It is a subset extracted from Epinions dataset in which each user has rated at least 15 items and each item is rated by at least 15 users (i.e. core-15). The MovieLens dataset [14] is movie ratings data and was

Table 1: Statistical properties of datasets

Dataset	#users	#items	#ratings	density
Epinions	5,531	4,287	186,995	0.789%
ML1M	6,040	3,706	1,000,209	4.468%

collected by the GroupLens research group. The characteristics of the datasets are summarized in Table 1.

The initial longer recommendation lists of size t are generated by two well-known recommendation algorithms: list-wise matrix factorization (ListRank) [22] and user-based collaborative filtering (UserKNN) [21]. As mentioned earlier, we chose these two algorithms to cover different approaches in recommender systems: matrix factorization and neighborhood models. We performed grid-search¹ on hyperparameters for each algorithm and selected the results with the highest precision value for our next analysis.

To show the effectiveness of the FairMatch algorithm in improving the aggregate diversity, we compare its performance with two state-of-the-art algorithms and also two simple baselines.

- (1) **FA*IR**. This is the method introduced in [26] and was mentioned in our related work section. The method was originally used for improving group fairness in job recommendation. However, we use this method for improving aggregate diversity in item recommendation. We define protected and unprotected groups as long-tail and short-head items, respectively. For separating short-head from long-tail items, we consider those top items which cumulatively take up $K\%$ of the ratings as the short-head and the rest as long-tail items. For experiments in this paper, we have tried different values of $K \in \{10, 20, 30, 50\}$. Also, we set the other two hyperparameters, proportion of protected candidates in the top n items² and significance level³, to $\{0.25, 0.5, 0.75, 0.95\}$ and $\{0.05, 0.1, 0.15\}$, respectively.
- (2) **Discrepancy Minimization (DM)**. This is the method introduced in [5] and was explained in our related work section. For hyperparameter tuning, we followed the experimental settings suggested by the original paper for our experiments. We set the target degree distribution to $\{1, 5, 10\}$ and relative weight of the relevance term to $\{0.01, 0.5, 1\}$.
- (3) **Reverse**. Given a recommendation list of size t for each user generated by standard recommendation algorithm, in this method, instead of picking the n items from the top, we pick them from the bottom of the list. In this approach, we expect to see an increase in aggregate diversity as we are giving higher priority to the items with lower scores to be picked first. However, the accuracy of the recommendations will decrease as we give higher priority to lower quality items.
- (4) **Random**. Given a recommendation list of size t for each user generated by standard recommendation algorithm, we randomly choose n items from that list and create a final recommendation list for that user. Note that this is different

¹For ListRankMF, we set all regularizers $\in \{0.0001, 0.001, 0.01\}$, iterations $\in \{30, 50, 100\}$, learningrate $\in \{0.0001, 0.001, 0.005, 0.01\}$, and factors $\in \{50, 100, 150, 200\}$. For UserKNN, we set neighbors $\in \{10, 20, 30, 50, 100\}$.

²Based on suggestion from the released code, the range should be in $[0.02, 0.98]$

³Based on suggestion from the released code, the range should be in $[0.01, 0.15]$

Table 2: Comparison of different post-processing techniques on Epinions dataset for recommendation list of size 10.

algorithms	baselines	$t = 20$				$t = 50$				$t = 100$			
		P@10	C@10	G@10	E@10	P@10	C@10	G@10	E@10	P@10	C@10	G@10	E@10
ListRankMF	Standard	0.015	24.4%	0.937	4.35	0.015	24.4%	0.937	4.35	0.015	24.4%	0.937	4.35
	Random	0.010	33.1%	0.882	5.23	0.006	45.0%	0.869	5.97	0.004	53.5%	0.856	6.39
	Reverse	0.005	37.3%	0.839	5.69	0.003	51.9%	0.867	5.58	0.002	60.7%	0.814	6.70
	FA*IR	0.013	28.8%	0.917	4.56	0.009	30.4%	0.934	5.01	0.008	33.8%	0.945	5.01
	DM	0.014	36.9%	0.907	4.45	0.011	56.6%	0.748	5.69	0.010	69.1%	0.680	6.07
	FairMatch	0.014	38.0%	0.884	4.72	0.010	61.4%	0.789	5.96	0.008	77.7%	0.720	6.53
UserKNN	Standard	0.045	46.4%	0.925	5.19	0.045	46.4%	0.925	5.19	0.045	46.4%	0.925	5.19
	Random	0.035	53.6%	0.896	5.63	0.023	65.7%	0.875	6.26	0.016	75.6%	0.831	6.73
	Reverse	0.025	58.2%	0.870	5.96	0.013	73.3%	0.825	6.70	0.008	83.1%	0.753	7.20
	FA*IR	0.044	61.1%	0.868	5.62	0.038	65.0%	0.865	6.22	0.030	65.4%	0.867	6.41
	DM	0.044	64.1%	0.850	5.65	0.041	84.3%	0.732	6.40	0.037	95.4%	0.529	7.18
	FairMatch	0.044	67.0%	0.853	5.72	0.038	90.8%	0.732	6.67	0.029	98.1%	0.580	7.41

Table 3: Comparison of different post-processing techniques on MovieLens dataset for recommendation list of size 10.

algorithms	baselines	$t = 20$				$t = 50$				$t = 100$			
		P@10	C@10	G@10	E@10	P@10	C@10	G@10	E@10	P@10	C@10	G@10	E@10
ListRankMF	Standard	0.152	14.0%	0.916	4.13	0.152	14.0%	0.916	4.13	0.152	14.0%	0.916	4.13
	Random	0.124	17.5%	0.861	4.76	0.089	24.6%	0.834	5.48	0.066	32.2%	0.809	5.97
	Reverse	0.097	19.0%	0.831	4.97	0.055	28.4%	0.786	5.73	0.037	37.9%	0.757	6.20
	FA*IR	0.143	14.2%	0.907	4.26	0.136	14.3%	0.937	4.34	0.128	16.5%	0.949	4.41
	DM	0.148	18.7%	0.850	4.41	0.138	28.4%	0.801	4.76	0.130	38.1%	0.764	5.02
	FairMatch	0.149	19.4%	0.870	4.40	0.138	30.0%	0.836	4.90	0.130	40.2%	0.834	5.10
UserKNN	Standard	0.196	10.7%	0.884	4.37	0.196	10.7%	0.884	4.37	0.196	10.7%	0.884	4.37
	Random	0.163	12.3%	0.836	4.73	0.120	15.9%	0.805	5.29	0.094	19.4%	0.780	5.71
	Reverse	0.130	13.4%	0.791	4.99	0.082	17.7%	0.726	5.63	0.058	22.2%	0.703	6.01
	FA*IR	0.192	11.3%	0.855	4.60	0.181	12.3%	0.869	4.88	0.168	18.0%	0.858	5.19
	DM	0.192	13.8%	0.835	4.63	0.184	19.2%	0.800	4.98	0.180	25.0%	0.780	5.21
	FairMatch	0.193	13.9%	0.863	4.48	0.184	18.6%	0.872	4.69	0.170	23.6%	0.850	5.05

from randomly choosing items from all catalog to recommend to users. The reason we randomly choose the items from the original recommended list of items (size t) is to compare other post-processing and re-ranking techniques with a simple random re-ranking.

FairMatch algorithm only involves one hyperparameter, α , to control the balance between the node degree and relevance. For our experiments we try $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$. A lower value for α indicates more focus on maintaining the accuracy of the recommendations, while a higher value for α indicates more focus on improving aggregate diversity. We also perform a sensitivity analysis to show how α can play an important role in the accuracy-diversity trade-off.

For evaluation, we use the following metrics to measure the effectiveness of each method:

- (1) **Precision** ($P@n$): The fraction of the recommended items shown to the users that are part of the users' profile in the test set.
- (2) **Coverage** ($C@n$): The percentage of items which appear at least once in the recommendation lists.
- (3) **Gini index** ($G@n$): The measure of fair distribution of recommended items. It takes into account how uniformly items appear in recommendation lists. Uniform distribution will have Gini index equal to zero which is the ideal case (lower

Gini index is better). Given all the recommendation lists for users, L , and $p(i_k|L)$ as the probability of the k -th least recommended item being drawn from L calculated as [24]:

$$p(i|L) = \frac{\sum_{u \in U} \mathbb{1}_{i \in L_u}}{\sum_{u \in U} \sum_{j \in L_u} \mathbb{1}_{j \in L_u}} \quad (5)$$

where L_u is the recommendation list for user u . Now, Gini index of L can be computed as:

$$Gini(L) = \frac{1}{|I| - 1} \sum_{k=1}^{|I|} (2k - |I| - 1) p(i_k|L) \quad (6)$$

- (4) **Entropy** ($E@n$): Given the distribution of recommended items, entropy measures the uniformity of that distribution. Uniform distribution has the highest entropy or information gain, thus higher entropy is more desired when the goal is increasing diversity.

$$Entropy(L) = - \sum_{i \in I} p(i|L) \log p(i|L) \quad (7)$$

where $p(i|L)$ is the observed probability value of item i in recommendation lists L .

We performed 5-fold cross validation in our experiments, and we generated recommendation lists of size 10, 20, 50, and 100 for

each user by each recommendation algorithm. Recommendation lists of size 10 are used for evaluating standard recommendation algorithms and longer recommendation lists of size 20, 50, and 100 are used as input for diversification techniques to generate recommendation lists of size 10 as output. Recommendation lists of size 10 generated by each diversification technique are evaluated by aforementioned metrics and their effectiveness is compared. We used *librec-auto* and LibRec 2.0 for running experiments [13, 19].

4.2 Comparative Evaluation

Table 2 and 3 summarize the performance of FairMatch and other baselines on Epinions and MovieLens datasets, respectively. For each metric (ignoring Random and Reverse techniques), the bolded values show the best results and a statistically significant change from the second best baseline with $p < 0.01$.

As mentioned earlier, extensive experiments are performed by each diversification technique with multiple hyperparameter values and for the purpose of comparison, from each of three diversification algorithms (DM, FA*IR, and our FairMatch) the configuration which yields, more or less, the same precision loss is reported. These results enable us to better compare the performance of each technique on improving aggregate diversity while maintaining the same level of accuracy.

Based on experiments on Epinions dataset shown in table 2, FairMatch significantly outperforms all the baselines on various sizes of initial recommendation lists generated by both recommendation algorithms in terms of coverage ($C@10$). The coverage of FairMatch is even higher than the Random and Reverse techniques without losing much accuracy which is indicative of its power in finding high-quality items with minimum visibility. Again, the *Random* algorithm used here is randomly picking n items from the original list and put them in the final list, so it is still possible that many popular items could end up being in the final list. In terms of fair distribution, the same improvement is also consistently observed on entropy. Entropy of FairMatch technique is significantly higher than other techniques in all cases showing that the recommendations generated by FairMatch are fairer and closer to a uniform distribution. However, in terms of Gini index, FairMatch generated comparable results to DM.

Table 3 shows the experimental results in MovieLens dataset. Based on these results, except for *UserKNN* with $t = 50$ and $t = 100$, FairMatch provides higher coverage in all cases which is consistent with the results from Epinions dataset. In terms of entropy and Gini index, FairMatch was outperformed by DM in most of the cases.

It is worth noting that the Gini can be a misleading measure if it is looked at in isolation. For instance, if an algorithm recommends only a few items (low coverage) but does so by recommending each item exactly in an equal proportion, then it will achieve a perfect Gini. However, having a low coverage is not desired and therefore it is more reasonable to look at the coverage and Gini together.

4.3 Accuracy-Diversity Trade-Off

We also investigated the precision and diversity trade-off in our FairMatch algorithm under various settings. Figure 3 shows the experimental results on Epinions (Figure 3a) and MovieLens (Figure 3b) datasets. In these plots, x-axis shows the precision and y-axis

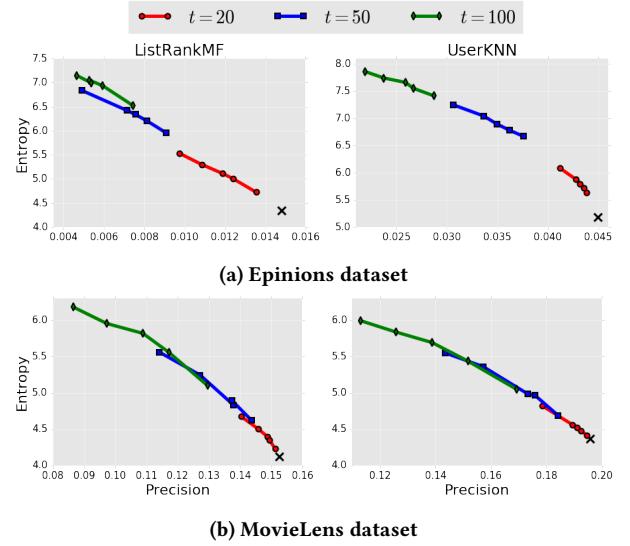


Figure 3: Precision and entropy trade-off in the FairMatch algorithm on Epinions and MovieLens datasets using ListRankMF and UserKNN. The black cross shows the performance of original recommendation lists at size 10.

shows the entropy of the recommendation results at size 10. Similar results are also observed when Gini index or coverage metrics are used as diversity measures. Each point on the plot corresponds to a specific α value and the black cross shows the performance of original recommendation lists at size 10.

Results in Figure 3 show that α plays an important role in controlling the precision-diversity trade-off. As we increase the α value, precision increases, while diversity decreases. According to equation 1, for a higher α value, FairMatch will concentrate more on improving the accuracy of the recommendations, while for lower α value, it will have a higher concentration on improving the diversity of the recommendations.

Also, it can be observed from Figure 3 that for longer initial recommendation lists (i.e., higher values for t), although the diversity of the recommendations increases, the precision decreases. These parameters allow system designers to better control the precision-diversity trade-off.

4.4 Long-tail Coverage Analysis

Recommending more items by a given recommendation algorithm is a desired characteristic. However, it is important to check if the increase in item coverage comes from recommending more long-tail items or it is just covering more popular items. Figure 4 shows the long-tail coverage for different algorithms on Epinions (Figure 4a) and MovieLens (Figure 4b) datasets for different original recommendations sizes t . For these experiments, we specified long-tail items using the technique introduced in [27]. Except for the *UserKNN* on MovieLens dataset which our FairMatch algorithm covers fewer long-tail items than the DM algorithm, in all other cases, the FairMatch algorithm outperforms all other algorithms on both datasets. In fact, on MovieLens, the FairMatch algorithm

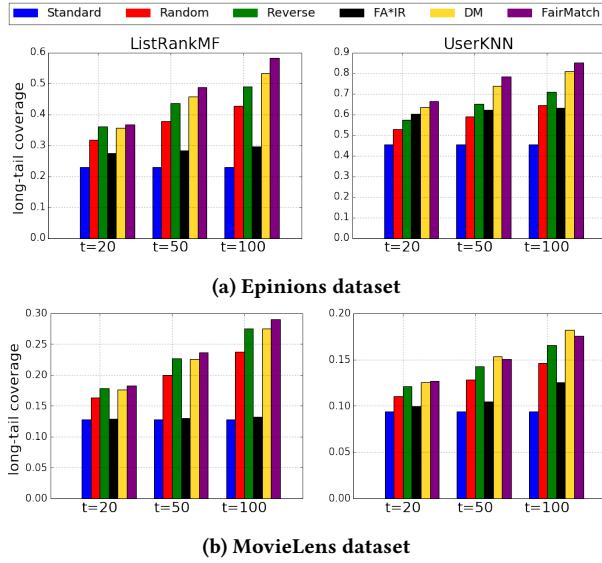


Figure 4: Long-tail coverage of diversification methods on Epinions and MovieLens datasets using ListRankMF and UserKNN.

also beats DM algorithm with a slight margin when the size of the original recommendation is 20. In other words, when the time and space complexity become an issue (larger values for t) and a smaller t is desired then the FairMatch algorithm outperforms every other algorithm in this experiment on both datasets.

4.5 Complexity Analysis

Solving the maximum flow problem is the core computation part of the FairMatch algorithm. We used Push-relabel algorithm as one of the efficient algorithms for solving the maximum flow problem. This algorithm has a polynomial time complexity as $O(V^2E)$ where V is the number of nodes and E is the number of edges in bipartite graph. For other parts of the FairMatch algorithm, the time complexity would be in the order of the number of edges as it mainly iterates over the edges in the bipartite graph.

Since FairMatch is an iterative process, unlike other maximum flow based techniques [4, 5], it requires solving maximum flow problem on the graph multiple times and this could be one limitation of our work. However, except for the first iteration that FairMatch executes on the original graph, at the next iterations, the graph will be shrunk as FairMatch removes some parts of the graph at each iteration. Regardless, the upper-bound for the complexity of FairMatch will be $O(V^3E)$ assuming in each iteration we still have the entire graph (which is not the case). Therefore, the complexity of FairMatch is certainly less than $O(V^3E)$ which is still polynomial.

5 DISCUSSION AND FUTURE WORK

In this section, we discuss the advantages that FairMatch provides on improving the performance of recommender systems. Also, we will discuss possible future work that can be considered for further improvement in FairMatch.

Generalization. In this paper, we studied the ability of FairMatch for improving aggregate diversity and one special case of supplier fairness under the assumption of each item belongs to one supplier (i.e., fair distribution on recommended items). However, FairMatch can be generalized to other definitions of fairness including supplier-side fairness. In this scenario, we can create recommendation bipartite graph between users and suppliers (based on recommended items), and then assign weights to edges based on suppliers' information (e.g., the probability of their items being shown in recommendation results and the quality of their items). At the third step, we can solve the maximum flow problem on this graph to extract high-quality suppliers with unfair visibility in recommendation lists. Finally, we can reconstruct the final recommendation lists by adding high-quality items from those suppliers according to each user's preferences.

Similar settings can also be considered on FairMatch for improving user fairness [18]. Considering the job recommendation domain where the task is recommending jobs to users, FairMatch can be formulated to fairly distribute "good" jobs (e.g. highly-paying jobs) to each group of users based on sensitive attributes (e.g. men and women). We consider these scenarios in our future work.

Flexibility. Another potential interesting improvement on FairMatch is taking into account the item ranking in final recommendation lists. In this paper, we aimed at creating final recommendation lists to include high-quality items with low visibility and we measured it in terms of precision. However, FairMatch allows to consider creating fair ranked lists by modifying the last step (recommendation construction). To do this, given extracted items from step 3 and top n recommendation lists from standard recommendation algorithm, the goal is to find the fair position for extracted items in the top n recommendation.

Finally, weight computation at step 2 also provides flexibility in optimizing FairMatch to capture some other aspects. For instance, considering the popularity of items for computing weights of edges on recommendation bipartite graph may help to further control popularity bias in recommender systems.

6 CONCLUSION

In this paper, we proposed a graph-based approach, FairMatch, for improving the aggregate diversity of recommender systems. FairMatch is a post-processing technique that works on the top of any recommendation algorithm. In other words, it re-ranks the output from the standard recommendation algorithm such that it improves the aggregate diversity of final recommendation lists, while it maintains the accuracy of recommendations. Experimental results on two publicly available datasets showed that the FairMatch algorithm outperforms several state-of-the-art methods in improving aggregate diversity. One of the limitations of our work is that our algorithm does not leverage the information about the popularity of items in rating data. We believe this information could play an important role in further improving aggregate diversity of the final recommendation lists because usually algorithms are biased toward popular items [2] and tackling this bias could increase the number of distinct recommended items, hence higher aggregate diversity. We intend to investigate this limitation in future work.

REFERENCES

- [1] Himan Abdollahpour, Robin Burke, and Bamshad Mobasher. 2017. Controlling Popularity Bias in Learning-to-Rank Recommendation. In *RecSys '17 Proceedings of the Eleventh ACM Conference on Recommender Systems*. 42–46.
- [2] Himan Abdollahpour, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The unfairness of popularity bias in recommendation. In *RecSys Workshop on Recommendation in Multistakeholder Environments (RMSE)*.
- [3] Gediminas Adomavicius and YoungOk Kwon. 2011. Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Transactions on Knowledge and Data Engineering* 24, 5 (2011), 896–911.
- [4] Gediminas Adomavicius and YoungOk Kwon. 2011. Maximizing aggregate recommendation diversity: A graph-theoretic approach. In *Proceedings of the 1st International Workshop on Novelty and Diversity in Recommender Systems (DiveRS 2011)*. 3–10.
- [5] Arda Antikacioglu and R. Ravi. 2017. Post processing recommender systems for diversity. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 707–716.
- [6] Erik Brynjolfsson, Yu Hu, and Duncan Simester. 2011. Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales. *Management Science* 57, 8 (2011), 1373–1386.
- [7] Erik Brynjolfsson, Yu Hu, and Michael D. Smith. 2003. Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. *Management Science* 49, 11 (2003), 1580–1596.
- [8] Efim A. Dinic. 1970. Algorithm for solution of a problem of maximum flow in networks with power estimation. In *Soviet Math. Doklady* 11 (1970), 1277–1280.
- [9] Lester Randolph Ford and Delbert R. Fulkerson. 1956. Maximal flow through a network. *Canadian Journal of Mathematics* 8 (1956), 399–404.
- [10] David García-Soriano and Francesco Bonchi. 2018. Fair-by-design algorithms: matching problems and beyond. *CoRR abs/1802.02562* (2018). [arXiv:1802.02562](https://arxiv.org/abs/1802.02562) [http://arxiv.org/abs/1802.02562](https://arxiv.org/abs/1802.02562)
- [11] Andrew V. Goldberg and Robert E. Tarjan. 1988. A new approach to the maximum-flow problem. *Journal of the ACM (JACM)* 35, 4 (1988), 921–940.
- [12] Daniel G. Goldstein. 2006. Profiting from the long tail. *Harvard Business Review* 84, 6 (2006), 24–28.
- [13] Guibing Guo, Jie Zhang, Zhu Sun, and Neil Yorke-Smith. 2015. LibRec: A Java Library for Recommender Systems. In *UMAP Workshops*.
- [14] F. Maxwell Harper and Joseph A. Konstan. 2015. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5, 4 (2015), 1–19.
- [15] Neil Hurley and Mi Zhang. 2011. Novelty and diversity in top-n recommendation-analysis and evaluation. *ACM Transactions on Internet Technology (TOIT)* 10, 4 (2011), 1–30.
- [16] Marius Kaminskis and Derek Bridge. 2016. Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7, 1 (2016), 1–42.
- [17] Haifeng Liu, Xiaomei Bai, Zhuo Yang, Amr Tolba, and Feng Xia. 2015. Trust-aware recommendation for improving aggregate diversity. *New Review of Hypermedia and Multimedia* 21, 3–4 (2015), 242–258.
- [18] Masoud Mansoury, Himan Abdollahpour, Jessie Smith, Arman Dehpanah, Mykola Pechenizkiy, and Bamshad Mobasher. 2020. Investigating Potential Factors Associated with Gender Discrimination in Collaborative Recommender Systems. In *The Thirty-Third International Flairs Conference*.
- [19] Masoud Mansoury, Robin Burke, Aldo Ordonez-Gauger, and Xavier Sepulveda. 2018. Automating recommender systems experimentation with librec-auto. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, 500–501.
- [20] Judith Möller, Damian Trilling, Natali Helberger, and Bram van Es. 2018. Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society* 21, 7 (2018), 959–977.
- [21] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. GroupLens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*. ACM, 175–186.
- [22] Yue Shi, Martha Larson, and Alan Hanjalic. 2010. List-wise learning to rank with matrix factorization for collaborative filtering. In *Proceedings of the Fourth ACM Conference on Recommender Systems*. ACM, 269–272.
- [23] Saül Vargas and Pablo Castells. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems*. 109–116.
- [24] Saül Vargas and Pablo Castells. 2014. Improving sales diversity by recommending users to items. In *Proceedings of the 8th ACM Conference on Recommender systems*. 145–152.
- [25] Hongzhi Yin, Bin Cui, Jing Li, Junjie Yao, and Chen Chen. 2012. Challenging the long tail recommendation. *arXiv preprint arXiv:1205.6700* (2012).
- [26] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 1569–1578.
- [27] Óscar Celma and Pedro Cano. 2008. From hits to niches? or how popular artists can bias music recommendation and discovery. In *Proceedings of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition*. 1–8.
- [28] Özge Sürer, Robin Burke, and Edward C. Malthouse. 2018. Multistakeholder recommendation with provider constraints. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 54–62.