Recommendation fairness and where to find it: An empirical study on fairness of user recommender systems

Antonela Tommasel
Aarhus University, Department of Computer Science
DIGIT Aarhus University Centre for
Digitalisation, Big Data and Data Analytics, Denmark
ISISTAN, CONICET-UNCPBA, Argentina
Email: antonela.tommasel@isistan.unicen.edu.ar

Ira Assent
Aarhus University, Department of Computer Science
DIGIT Aarhus University Centre for
Digitalisation, Big Data and Data Analytics, Denmark
Email: ira@cs.au.dk

***Abstract*—Recommender systems play a crucial role in how users consume information and establish new social relations. However, different factors (such as the data collection process, the designed recommendation model, or even the interpretation of findings) could make recommenders (unintendedly) prone to biases, favouring certain user groups or items, thus resulting in unfair outcomes. Recommenders also face fairness criticism for inducing filter bubbles, echo chambers, and, more generally, facilitating opinion manipulation. In this work, we study the impact of user recommender systems on fairness. To this end, we carry out a user recommendation task on a politically polarized Twitter data collection. Then, we evaluate how the different politically aligned user groups experience recommendation quality. Finally, we explore causal models to identify data and model-related features that could affect the fairness of recommender outcomes. Our study shows that political alignment is associated with the unfairness of recommenders affecting not only the relevance of recommendations, but also their diversity and the resulting interaction patterns.**

I. Introduction

While recommendation models have reached remarkable maturity in terms of performance in many domains, concerns have been raised regarding their fairness, posing an obstacle to further integrating recommenders into society [1, 2]. Unfairness has often been associated with a non-uniform distribution of performance among different groups and systematically favouring over-represented groups while at the same time hurting the under-represented ones [3], potentially resulting in discrimination and reduced visibility of these groups and limited recommendation diversity [3, 2]. Such groups are often based on sensitive attributes like gender, race, ethnicity, and religion, which are also protected by anti-discrimination laws. Although not often considered, political alignment might also be sensitive, not only because discrimination based on it may be forbidden but also because it can significantly affect how people perceive consensus and mobilize around policy issues [4], shaping not only the beliefs and decisions of one individual but even the entire society [5].

Interactions are the cornerstone of social media, driving their features and functionality [6, 7]. User recommenders (also known as friend/people recommenders or link prediction) aim to infer which new user interactions are likely to be observed in the future [8]. They have become the most effective mechanism for the creation or reinforcement of interactions [7, 9, 10]. The social implications of recommendations are one of the main things differentiating user and item recommenders [11]. For example, before interacting with someone (especially in symmetric interactions, like on *Facebook*), one often considers how the other person would perceive this action and whether they would reciprocate it. These dynamics can also be obstacles to accepting recommendations, even when they are relevant [11]. In these recommenders, historical biases towards majorities prevent minorities from being high-reach social influencers, affecting network dynamics [2]. These unfairness patterns could be even more significant in polarized networks, in which unfairness could contribute to strengthening phenomena like echo chambers [12]. For example, users exhibiting a certain minority political orientation might receive recommendations tailored for another political orientation, degrading user experience and inducing users to solely interact with a set of already known users, reinforcing polarization.

While there is an increasing interest in recommendation fairness, user recommenders have received relatively less attention. The few existing studies have focused on fairness to gender and age groups. In this work, we contribute a study of *whether user recommender tasks are affected by (un)fairness concerning users' political alignment*. First, we study *whether different politically aligned groups receive recommendations of different quality*. To this end, we generate recommendations and define scenarios of how they would alter the structure of social interactions. We consider recommenders based on various underlying mechanisms that are commonly used in the literature. Then, we evaluate the fairness towards user groups in terms of relevance, diversity/novelty, induced patterns of interactions, and divergences.

Second, we *study which features contribute to recommendation (un)fairness*. There are different reasons why recommenders might exhibit seemingly unfair behavior[5]. For example, unfairness can arise during data collection, processing,

979-8-3503-2445-7/23/$31.00

and analysis [3, 13], in the recommender itself, or even when evaluating the recommendations and interpreting the findings [13]. A correction step could be applied to recommenders when aiming for fairness. However, without knowing which is the cause of (un)fairness, corrections might not have the desirable effect [14]. We complement our analysis by means of causal models to help understand how data features can "propagate" through recommenders and affect their outcomes.

We conduct an experimental analysis of a politically polarized *Twitter* data collection. Our study shows that political alignment can induce unfairness of recommenders, affecting not only the relevance of recommendations but also their diversity and the resulting interaction patterns. Recommenders show different degrees of unfair treatment, which vary according to the evaluation dimension under analysis. For example, in general, neural-based recommenders showed a better balance than traditional recommenders in terms of the relevance and calibration of recommendations. This does not only condition how groups are treated in the short term but can also have unwanted consequences in the long term, strengthening filter bubbles and reinforcing polarization.

II. Related Works

Even with no intention to discriminate and without access to sensitive attributes, recommenders can inadvertently create bias [14]. Fairness and bias have long been topics of interest [5]. Most existing works have focused on popularity bias in item recommendation, with gender and age as sensitive attributes [15, 16]. For example, Lesota et al. [16] studied the divergence of popularity distributions between users' profiles and recommended items. The analysis showed that the techniques suffering popularity bias the least were not necessarily the worst performing. Ekstrand et al. [15] analyzed whether collaborative recommenders showed differences in treatment regarding age and gender.

In the news domain, Fernández et al. [17] studied popularity bias combined with misinformation amplification, finding a relation between both phenomena. Liu et al. [4] studied how users' news preferences in association with users' political alignment could influence news recommendation diversity (as a proxy for filter bubbles). The analysis was based on user-simulated data and included content-based and collaborative filtering recommenders. Results showed a differentiated impact on diversity based on users' political alignment and the strengthening of filter bubbles. Despite sharing the same sensitive attribute as our study, their focus was on short-term item recommendation and considered users as isolated entities, disregarding the contagion effect that neighbours could have on item preferences and consumption.

As regards user recommenders, Fabbri et al. [10] focused on how recommenders affected user exposure (i.e., how often users from a certain group are recommended). Results showed that minority and homophilic gender groups could be disproportionately over-exposed, while non-homophilic groups were under-exposed, reducing diversity. Cinus et al. [18] simulated opinion spreading over recommendations to show that recommenders can reinforce homophilic bias, reducing the diversity of recommendations and strengthening filter bubbles. Nonetheless, the effect of recommenders was negligible when the initial simulated network was already polarized into segregated communities.

In summary, despite user recommenders' effects on network evolution [9, 12], they have comparatively received less attention in the literature. Moreover, they have only focused on gender and age as sensitive attributes, disregarding the effect that other behavioural characteristics might have on recommendation fairness. Unlike gender and age attributes that can be easily discretized[1] without losing information, political alignment and behavioural attributes might be associated with continuous values organized in a spectrum with blurry limits, which might lead to noisy discretizations and losing information. Works mostly focused on analyzing exposure, disregarding recommendations' relevance (and other characteristics). In addition, most works primarily focus on assessing how (a reduced set of) evaluation metrics behave for each group, without studying the role of data and model-induced features on performance, and thus, (un)fairness. In contrast, we evaluate recommendations based on complementary points of view and leverage causal models to analyze the role of data and model features in recommendation outcomes.

III. Task and methods

Recommendations could be affected by different unfairness types [5] during data collection, in the recommender itself, or during result interpretation. Different notions of fairness in the literature reflect normative ideas about how a recommender system should be or behave [5]. In particular, this study focuses on *group fairness*, which requires that the identified groups are treated equally. On the other hand, *individual fairness* requires that the individuals in each group are treated equally. This implies that *group fairness* might not translate to *individual fairness*, as it would be possible that groups are treated fairly in terms of group-aggregated metrics, but some individuals in a group are treated better than others in the same group [5]. These effects could be related to the differences in users belonging to the same group and how such inter-group differences compare across groups, which in turn can also affect group (un)fairness [19].

We study whether user recommenders are affected by (un)fairness patterns concerning users' political alignment. We defined two research questions: **RQ1**: *How do user recommenders contribute to group (un)fairness?* We study whether different groups receive recommendations of significantly different quality. **RQ2**: *Which features contribute to unfairness?* We study how user and model features correlate with unfairness patterns and recommender quality. We define a 4-step pipeline (Figure 1) tackling the modelling of input data, the recommendation models, the outcomes of a user recommendation task, and how each aspect may contribute to (un)fairness. First, we collect data and define the network

[1]Note that all analyzed works refer to gender as a binary attribute.

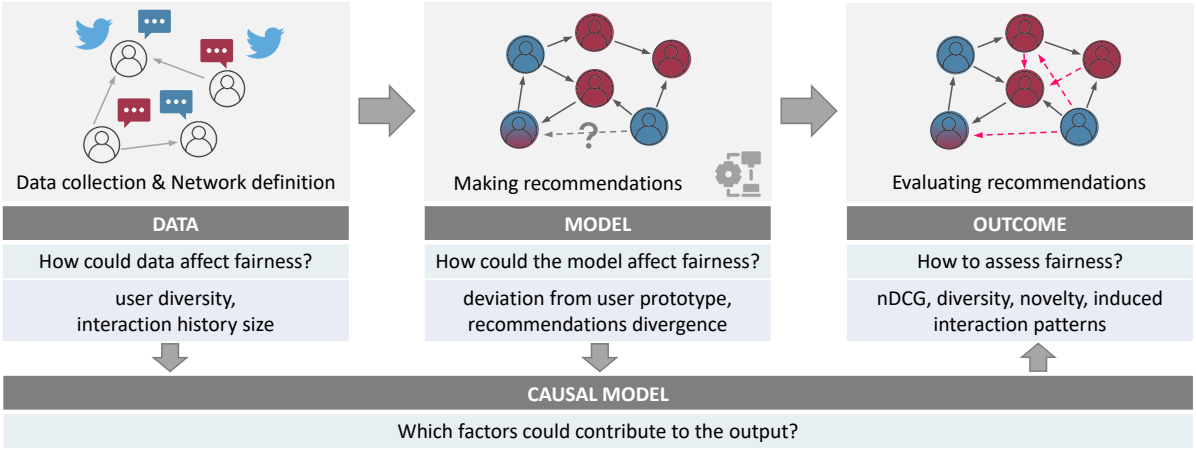

Fig. 1: Phases of the proposed pipeline

of user interactions. Data was obtained from conversations on *Twitter* around the `Obamacare` topic and included users' political alignment. Second, we selected the recommenders to evaluate. We considered recommenders that are commonly used in the literature and respond to different underlying principles (e.g., topology, content, matrix factorization). Third, we defined the aspects and metrics to evaluate the recommenders. We selected metrics covering the relevance, diversity, induced interactions, and divergence of recommendations. Finally, based on data characteristics, recommender metrics and outcomes, we performed a causal analysis to identify the data and model features that could affect the fairness of recommendations. Each pipeline step is described as follows.

A. *Data collection and network definition*

Assessing fairness requires building on appropriate data. Data collections often do not include sensitive attributes and require adaptations, which may not fully reflect the complexity of actual sensitive data but allow mimicking recommendation evaluation [5]. By contrast, our data is based on the politically aligned `Obamacare`[2] [20] collection, including tweets[3] related to the `#obamacare` and `#aca` hashtags, shared between May 2008 and October 2017 and belonging to 8,164 users. The collection also included the estimated political alignment [21] of 5,671 users. Alignments were defined as float scores ranging between -3 and 3, with negative values associated with Democrats and positive values with Republicans. To align with the `Obamacare` topic, alignments were inverted, and positive scores were associated with democrats (referred to as $\texttt{group}_+$, as they were pro `Obamacare`), while negative ones with republicans (referred to as $\texttt{group}_-$, as they were against `Obamacare`). Group size was imbalanced, with 29% of users belonging to $\texttt{group}_+$ and the remaining 71% to $\texttt{group}_-$.

Social media platforms generally allow two types of relations between users: structural and conversational. Structural relations (e.g., the *Twitter* follower/followee or *Facebook* Friends) are associated with already known people from other environments (e.g., work, family, school) or people with shared interests, with a tendency to be reciprocal [6] and long-term. Conversely, conversational interactions focus on the shared and consumed content, which acts as the root of the interaction, and thus are more dynamic and independent from whether users are also structurally related, providing a more timely reflection of users' interests. Conversational networks have been commonly used in the literature to study not only user recommenders [22], but also filter bubbles, echo chambers, and misinformation and political information propagation [23, 24]. Given the dynamic content-based nature of social relations [25], we focused on the conversational network. We built a graph in which nodes represent users and edges the possible tweet actions (reply, mention, quote, or retweet). Edges were directed to represent the flow of conversation [22, 23, 24], i.e., there is an edge from user A to user B if A replied to/mentioned/quoted B, while there is an edge from A to B if B retweet A.

To ensure that each user can be both the source and destination of social interactions and have sufficient interaction history, recommendations target users in the largest connected component with at least 10 interactions. The final collection comprised 2,726 users and over 6 million tweets. Table I summarizes the statistics of the analyzed users[4]. From the total number of shared tweets, on average, 419 ($\pm$ 432) triggered interactions. On average, users interacted with 127 ($\pm$108) unique users, $\texttt{group}_-$ interacted with 111 ($\pm$ 104) unique users, while $\texttt{group}_+$ with 107 ($\pm$ 94). Note that the total and unique number of average interactions, and shared tweets are similar for both groups. No statistically significant differences were observed when comparing the distribution of the number of interactions and shared tweets of groups. The network was temporally split into training and test. User interactions were sorted according to their date, and the first 80% were used for the training set, while the remaining 20% as test set.

[2]The original data collection can be obtained upon request from [20].

[3]Tweets were retrieved using a public tool: https://github.com/knife982000/FakingIt

[4]All processed data and derived models are available at: https://github.com/tommantonela/recommendation-fairness.

	all	group$_+$	group$_-$
# users	2,726	674	2,052
# tweets	6,041,500	1,646,769	4,394,731
avg. total tweets per user	1231 ±1445	1316 ±1492	1200 ±1438
avg. mentions per user	873.49 ±840.85	866.16 ±774.47	875.9 ±861.7
avg. replies per user	127.81 ±223.83	134.77 ±208.36	125.53 ±228.69
avg. retweets per user	293.8 ±352.59	312.28 ±348.01	287.73 ±353.96
avg. total interactions per user	1295.1 ±1216.32	1313.2 ±1141.54	1289.15 ±1240.1
% total interactions with same group	80.32 ±23.81	50.21 ±22.67	90.21 ±13.74
correlation of user and total interactions political alignment	0.75	0.32	0.67

TABLE I: Data collection details

Network topology and homophily (i.e., the tendency of similar users to interact with each other) can affect network characteristics and thus alter the effect of recommenders [12]. We assessed user interactions based on their political alignment using the same metrics as in [20, 26]. For users in group$_-$, on average, 90% of interactions were with others in the same group. Conversely, users in group$_+$ interacted with users with a broader range of alignment scores, accounting, on average, for 50% of interactions with users from the same group. For both groups, retweets and replies were the most common interactions with users from the same group. The same tendency was observed when considering users' unique interactions. Although user behaviour differed according to user alignment, alignments were positively correlated in all cases. The highest correlations (for both groups) were found for retweets, followed by mentions. These observations confirmed the homophilic tendency of users, which could indicate filter bubbles or echo chambers.

We assessed the heterogeneity (or diversity) of users in each group. To this end, we computed user profile divergence (also known as atypicality [27]), defined as the divergence between user preferences of group interactions (i.e., the ratio of interactions with each group) and the prototype user interactions. Similar to the heuristic aggregation strategies commonly used in group recommendation techniques [28, 29], the user prototype is defined as the average ratio of interactions with users in each group. This definition can be extended to groups, where the prototype of each group only averages the ratio of users belonging to such group. In addition, diversity is often measured based on content profiles [30, 31]. Then, we also considered prototypes built based on the normalized content distance between users and their interactions in each group. Content distance was computed as the Euclidean distance of the Word2Vec user tweets representation. Then, the final group distance was defined as the average distance for the users in such group. In both cases, divergence was measured based on a symmetric Kullback-Leibler divergence [16].

We observed significant differences in divergence distribution for both groups, with larger effects for content divergence. On average, divergences were larger and more dispersed for group$_-$ (up to one magnitude order higher), while group$_+$ presented a more homogeneous composition. Given these differences, it is important to analyze recommendations in terms of both divergence towards user preferences and inter-group divergences.

B. Making recommendations

In a traditional user recommendation setting, the task is to predict the likelihood of users interacting in the future so that users receive the top-k recommendations according to some predicted scores.

We study recommenders based on different underlying mechanisms. First, a traditional non-personalized **popularity** baseline based on recommending the most popular users (i.e., those with the largest number of interactions in the past). Second, three traditional recommenders: i) **topological**, we considered commonly used metrics based on neighbourhood overlap (e.g., Jaccard similarity, Adamic-Adar, Resource Allocation) [8] and, based on preliminary evaluations, we selected common neighbours, ii) **content-based** in which users were recommended based on the cosine similarity of their representations given by the centroid of the Word2Vec[5] vector of their shared tweets, iii) **Friends-of-Friends** (FoF) in which users at a 2-hop distance were recommended based on their popularity. This recommender is based on the triadic closure principle, which states that friends of friends also tend to become friends [33].

Finally, given that any recommender can be used for user recommendation if the adjacency matrix (i.e., the matrix of user-user interactions) acts as a rating matrix (i.e., the matrix of user-item interactions) [34], four state-of-the-art item recommenders were also included: i) **ImplicitMF** [35], a matrix factorization technique tailored for implicit feedback settings. We also analyzed diversity-oriented techniques (e.g., [36, 37, 38]) based on re-ranking this recommender, but results were not significantly different, ii) **MultVAE** [39], a neural generative model based on variational autoencoders and multinomial conditional likelihood; and iii) **NeuralMF** [40], a deep learning collaborative filtering model fusing matrix factorization and multi-layer perception.

All evaluations were performed over the same temporal data partitions. From the obtained recommendation rankings, a cut-off threshold was defined to recommend the top-10 users, as almost half of the users had 10 or more interactions in their test set (i.e., for nearly half of the users it would be possible to achieve perfect relevance scores), and, in a real-world evaluation it would represent a reasonable number of items to be effectively shown to the user [41]. Recommendations were deemed correct if they appeared in the test set. When available, original implementations were used. Parameters were optimized based on the procedures described in the original studies. For each recommender, we selected the configuration that achieved the highest results.

C. Association-based evaluation of recommendations

Following RQ1, we want to assess whether the quality of recommendations varies across groups. The simplest approach to evaluate fairness is to consider association-based notions to assess the discrepancy of metrics between the whole population and the identified sensitive groups [42, 5]. The perfor-

[5]Despite being static embeddings, they have been shown to achieve competitive results when compared to more complex dynamic embeddings (such as BERT), while being faster and requiring fewer resources [32].

mance of recommenders was evaluated based on four types of metrics: i) relevance, ii) diversity/novelty, iii) interactivity, and iv) recommendation divergences or calibration [43].

First, for **relevance**, we selected precision and nDCG (normalized Discounted Cumulative Gain). Second, for **diversity** (i.e., differences between the recommended users) and **novelty** (i.e., differences between the recommended and the already known users), we used intra-list dissimilarities [30]. Enhancing diversity/novelty has also been associated with increased fairness and reduced biases [22], as they can measure how recommendations help users in outing from their known community, thus broadening their social experiences [22]. Dissimilarities were computed as the Euclidean distance of the average Word2Vec representation of users' tweets. To assess how network dynamics are altered due to recommendations, we also analyzed the **characteristics of the induced recommendation network** (i.e., the augmentation of the base training graph with the recommended interactions). These metrics act as proxies for analyzing information propagation phenomena, such as filter bubbles and echo chambers. In this regard, we considered: i) the *ratio of interactions* with users of the same group; ii) *clustering coefficient*, a density metric that does not only assess how users are connected but also how actively they could communicate and influence others [44]; iii) *Random Walk Controversy Score* (RWC) [45], which is based on PageRank and computes the probability of a random user being exposed to users from the other group. The higher the score, the lower the likelihood of groups interacting, and the closer the score to 0, the higher the likelihood of users interacting with either group. The metric is not influenced by group size or the total degree of users in each group [18].

Finally, we analyzed the **calibration** of recommendations, which can be defined as the divergence between users' historical interaction preferences and the received recommendations [16, 43] (i.e., how well recommendations reflect users' preferences) by means of the Kullback-Leibler divergence. A well-calibrated recommender will reflect users' preferences in the appropriate proportion, which can positively impact user experience [46]. Considering that atypicality can affect calibration [27], we also assessed the atypicality of the recommendations for a user regarding the prototype recommendations (i.e., the average distributions of recommendations) for their group. In both cases, we followed the considerations and definitions introduced in Section III-A.

D. *Causal fairness analysis*

Our goal here is to analyze how input data and model-induced features correlate to recommender quality and unfairness patterns. The descriptive nature of the association-based evaluation (which primarily focuses on assessing the statistical discrepancy between the metrics for the identified groups) cannot explain the causal relations between features and model outcomes [42]. However, this causal relation between the model outcome and the sensitive features could be the source of unfairness [42]. Causal fairness analysis can leverage prior knowledge about the world and help understand how features "propagate" through the model and affect recommendation quality [42, 5]. We considered causal models based on the combination of the following features:

• **Data induced features.** First, the number of users' historical interactions as (according to Ekstrand et al. [15]) it could contribute to the quality of recommendations, given that the larger the number of training items, the stronger the basis for the recommendations. Second, to assess the relation between the sensitive attribute and the outcome, users' political alignment, as it could be one of the primary sources of (un)fairness. Third, the user profile atypicality, as defined in Section III-A.

• **Recommender induced features.** We included the atypicality of user recommendations in relation to the prototypical recommendations, as defined in Section III-C.

We fitted multiple regression models to capture the relation between the defined combinations of features and recommendation outcome. Each user was considered an observation, and the dependent variable was defined as the recommendation outcome in terms of nDCG [15]. For each user, the average nDCG score across all recommenders was computed to capture an overall idea of how difficult it might be to make recommendations to each user [15].

Linear models rely on certain assumptions that information retrieval metrics (like nDCG) might not directly meet [47]. For example, independence of observations, homoscedasticity, normality, and linearity of the empirical mean. Despite linear models possibly being resilient to violations of their assumptions, departing from them might reduce the effectiveness of analyses [47]. There have been several proposals in the literature to transform information retrieval metrics and bring them closer to the assumptions of linear models [47]. Nonetheless, an alternative is to use Generalized Linear Models (GLMs). These models are a generalization of traditional linear models that relax the assumptions about data distribution and shape. Then, under GLMs, data is no longer required to follow a normal distribution or to have constant variance [47].

Given that the dependent variable is numeric with a continuous distribution, following Faggioli et al. [47][6], we chose a Gaussian distribution with the same domain as nDCG. After preliminary evaluations, we chose the log link function to represent the relation between the expectation of the outcome and the observations[7].

IV. Analysis

This section presents the results of the conducted analyses to answer the proposed research questions.

A. *Recommendation evaluation*

Table II presents the recommendation results[8]. Normally, metrics are averaged over all users, providing a simple aggregated score of recommenders' effectiveness [15]. However,

[6]As implemented by the statsmodels package: https://www.statsmodels.org/stable/index.html

[7]We empirically confirmed the choice of the log link function via the Akaike Information Criterion (AIC) and the likelihood ratio test.

[8]More results can be found in the repository.

		popularity	Friend-of-Friends	topology	content-based	ImplicitMF	MultVAE	NeuralMF	base graph
precision	all	0.123 ±0.195	0.125 ±0.179	0.133 ±0.209	0.045 ±0.109	0.117 ±0.192	0.125 ±0.18	0.054 ±0.121	-
	group$_+$	0.13 ±0.192	0.147 ±0.187	0.148 ±0.211	0.048 ±0.104	0.142 ±0.214	0.14 ±0.18	0.064 ±0.125	-
	group$_-$	0.121 ±0.197	0.117** ±0.176	0.128 ±0.208	0.044 ±0.111	0.108** ±0.183	0.12 ±0.18	0.051 ±0.119	-
nDCG	all	0.328 ±0.371	0.342 ±0.356	0.331 ±0.36	0.143 ±0.272	0.293 ±0.354	0.328 ±0.345	0.144 ±0.25	-
	group$_+$	0.356 ±0.37	0.396** ±0.361	0.361 ±0.355	0.154 ±0.27	0.347** ±0.374	0.372** ±0.353	0.171** ±0.269	-
	group$_-$	0.318 ±0.37	0.323 ±0.352	0.32 ±0.362	0.139 ±0.273	0.275 ±0.345	0.313 ±0.341	0.135 ±0.242	-
Content diversity	all	0.108 ±0.065	0.116 ±0.074	0.108 ±0.081	0.109 ±0.08	0.102 ±0.067	0.125 ±0.081	0.106 ±0.068	0.221 ±0.278
	group$_+$	0.11 ±0.062	0.12 ±0.074	0.118** ±0.096	0.116** ±0.084	0.108** ±0.068	0.133** ±0.083	0.116** ±0.071	0.212 ±0.251
	group$_-$	0.107 ±0.066	0.114 ±0.075	0.104 ±0.075	0.107 ±0.079	0.1 ±0.066	0.122 ±0.08	0.102 ±0.067	0.224 ±0.286
Content novelty	all	0.204 ±0.094	0.173 ±0.03	0.162 ±0.043	0.166 ±0.051	0.164 ±0.028	0.175 ±0.032	0.16 ±0.028	0.183 ±0.045
	group$_+$	0.215** ±0.104	0.176 ±0.033	0.166** ±0.047	0.168 ±0.057	0.167** ±0.03	0.18** ±0.035	0.167** ±0.031	0.187 ±0.04
	group$_-$	0.201 ±0.09	0.172 ±0.029	0.161 ±0.041	0.165 ±0.049	0.162 ±0.027	0.173 ±0.031	0.158 ±0.026	0.181 ±0.046
Ratio Group	all	0.504 ±0.425	0.515 ±0.41	0.527 ±0.415	0.546 ±0.421	0.578 ±0.408	0.516 ±0.406	0.526 ±0.38	0.537 ±0.41
	group$_+$	0.069 ±0.079	0.136 ±0.132	0.226 ±0.23	0.331 ±0.316	0.376 ±0.305	0.155 ±0.151	0.233 ±0.168	0.29 ±0.269
	group$_-$	0.647** ±0.395	0.639** ±0.394	0.626** ±0.415	0.616** ±0.427	0.644 ±0.416	0.634** ±0.394	0.622** ±0.382	0.618 ±0.416
Clustering Coefficient	all	0.178 ±0.094	0.135 ±0.084	0.132 ±0.085	0.118 ±0.081	0.143 ±0.095	0.158 ±0.094	0.149 ±0.092	0.111 ±0.075
	group$_+$	0.187** ±0.104	0.141** ±0.097	0.138 ±0.098	0.123 ±0.094	0.149** ±0.11	0.164 ±0.107	0.158** ±0.104	0.116 ±0.087
	group$_-$	0.175 ±0.091	0.133 ±0.08	0.13 ±0.08	0.116 ±0.076	0.141 ±0.09	0.157 ±0.09	0.146 ±0.087	0.109 ±0.071
Random Walk Controversy	all	0.236	0.254	0.269	0.283	0.278	0.252	0.251	0.277
	group$_+$	0.353	0.359	0.397	0.393	0.403	0.372	0.365	0.391
	group$_-$	0.877	0.887	0.865	0.882	0.868	0.874	0.879	0.877
Calibration Ratio Group	all	2.631 ±5.479	1.434 ±3.813	1.253 ±3.075	1.775 ±4.031	1.585 ±3.986	1.289 ±3.459	0.189 ±0.587	3.139 ±7.049
	group$_+$	6.426** ±8.575	2.595** ±6.13	1.246 ±4.468	1.882 ±6.248	1.626 ±6.188	2.014** ±5.491	0.052 ±0.072	3.881 ±8.296
	group$_-$	1.333 ±2.926	1.037 ±2.462	1.255 ±2.423	1.738 ±2.91	1.572 ±2.873	1.042 ±2.35	0.236** ±0.672	2.886 ±6.552
Calibration Content	all	1.504 ±3.09	1.01 ±2.656	1.01 ±2.747	0.16 ±0.795	0.541 ±2.037	0.708 ±2.232	1.703 ±3.38	0.271 ±1.282
	group$_+$	1.67** ±3.117	0.673 ±1.989	0.416 ±1.544	0.13 ±0.629	0.161 ±0.819	0.457 ±1.611	0.044 ±0.16	0.156 ±0.919
	group$_-$	1.448 ±3.079	1.125** ±2.84	1.212** ±3.025	0.169 ±0.844	0.671 ±2.296	0.793** ±2.402	2.27** ±3.75	0.31 ±1.383
Rec. Atyp. Ratio Group	all	0.072 ±0.079	0.077 ±0.083	0.099 ±0.11	0.149 ±0.143	0.16 ±0.145	0.087 ±0.092	0.062 ±0.08	0.151 ±0.143
	group$_+$	0.086** ±0.088	0.079 ±0.082	0.112** ±0.131	0.161** ±0.175	0.157** ±0.167	0.082 ±0.088	0.067 ±0.082	0.159 ±0.156
	group$_-$	0.068 ±0.075	0.076 ±0.083	0.095 ±0.102	0.144 ±0.131	0.161 ±0.137	0.089 ±0.093	0.061 ±0.08	0.149 ±0.139
Rec. Atyp. Content	all	0.229 ±0.228	0.207 ±0.211	0.208 ±0.216	0.193 ±0.207	0.203 ±0.205	0.2 ±0.204	0.143 ±0.198	0.19 ±0.186
	group$_+$	0.282** ±0.246	0.183 ±0.188	0.12 ±0.142	0.096 ±0.135	0.089 ±0.128	0.157 ±0.167	0.038 ±0.058	0.138 ±0.164
	group$_-$	0.212 ±0.22	0.214** ±0.217	0.237** ±0.228	0.225** ±0.217	0.241** ±0.212	0.214** ±0.213	0.178** ±0.215	0.207 ±0.19

TABLE II: Recommendation results comparison for $k = 10$. Stat. significance: **p $< .01$

this could mask important differences in how user groups interact with the recommender. Thus, we report the average scores (and the corresponding standard deviation) for each metric over three sets of users: all, group$_+$, and group$_-$. Diversity and novelty for the base graph were computed considering the test set as the recommended users. All metrics assume that all recommendations will be accepted, which, although it might not reflect actual user behaviour [22], allows assessing how recommendations would shape the future network.

a) ***Relevance metrics****:* In general, recommenders achieved low precision and moderate nDCG. *ImplicitMF*, *FoF* and *topology* achieved the best results, while *content-based*, and *NeuralMF* the worst. *Content-based* low results could be explained by the topically-focused nature of data, by which users might naturally show high similarities, making it challenging to identify the interesting users. Some differences were statistically significant (with an α of 0.01), favouring group$_+$. This agrees with Ekstrand et al. [15], who also reported better relevance results for the least represented group.

b) ***Diversity metrics****:* Diversity was lower than that of the base graph, indicating the homogeneity of recommendations. *ImplicitMF* and *MultVAE* showed a good trade-off between relevance and diversity/novelty, helping users expand their interests. *Content-based* achieved low diversity/novelty, which is expected considering it recommends the most similar users. As the Table shows, some group differences are statistically significant. *Popularity* increased novelty compared to the original network, which could be caused by users already following the popular users in their own group and thus being recommended users in the other group, which might not be strictly similar to the already known ones.

c) ***Interactivity metrics****:* Differences for Ratio Group were statistically significant. While recommendations foster users in group$_+$ to interact with users in the other group, users in group$_-$ were encouraged to interact with each other. This could be an effect of the larger number of users in group$_-$, which induces recommenders to learn their patterns of interaction and interest. Differences observed for group$_-$ are slim, while group$_+$ showed a higher variability. For group$_+$, *popularity* and *MultVAE* achieved the lowest values, while *topology* achieved the largest ones. *Popularity* values reinforced the idea that as users already followed the popular users in their group, they are recommended popular users from the other group. Regarding *topology*, as recommendations span over users' neighbourhoods, which mainly belong to the same group, the reinforcement of those interactions is expected. When compared to the base graph, for group$_+$, Ratio Group values had up to a magnitude order difference for *popularity*, while some neural-based recommenders even increased such ratio. In general, the neighbourhood composition of users in group$_-$ did not show significant changes, which is also reflected in the correlation between political alignment and the average alignment of recommendations. Correlations for group$_+$ were lower (and even non-existent) than for group$_-$, which was expected as they received recommendations from the other group. Correlations for both groups were strengthened for *topology*, *content-based*, and *ImplicitMF*.

Most differences in clustering coefficient were not statistically significant. In general, scores were similar to that of the base graph, and slightly higher for group$_+$. This implies that users in group$_+$ might occupy central roles in the network, while users in group$_-$ might not be well connected.

Regarding RWC, the likelihood of users in $\texttt{group}_-$ of interacting with the same group was slightly increased when compared to the base graph. Although recommendations for $\texttt{group}_-$ tended to consolidate, they were, in general, the target of recommendations for $\texttt{group}_+$. This indicates that $\texttt{group}_-$ might be polarized, while $\texttt{group}_+$ showed a higher openness. While a decrement in polarity (as RWC defines) might seem positive, it also implies a higher mixture of user interactions, which might expose users to other contradictory or divergent points of view and thus ends up having the opposite effect, contributing to the filter bubble phenomenon [48]. These observations confirm that evaluation should focus on multiple and complementary dimensions.

d) ***Calibration metrics****:* Regarding calibration, in general, while Ratio Group showed higher values for $\texttt{group}_+$, content showed them for $\texttt{group}_-$. This confirms that $\texttt{group}_+$ was often induced to interact with $\texttt{group}_-$ while at the same time keeping a more homogeneous distribution of content similarities. Recommenders had differentiated effects over calibrations and groups. Depending on the alternative, for $\texttt{group}_-$, *ImplicitMF*, *NeuralMF* and *content-based* showed low or high calibration. For $\texttt{group}_+$, the same effect was observed for *popularity*, *FoF*, *NeuralMF* and *content-based*. Overall, the neural-based recommenders showed the lowest values (i.e., high calibration) for both groups, implying a good matching between the user preferences and recommendation characteristics.

Finally, statistically significant differences were observed for recommendation atypicality. While for $\texttt{group}_+$, the largest values were observed for *popularity* (content) and *content-based* (Ratio Group), those for $\texttt{group}_-$ were observed for *ImplicitMF*. These recommenders were among the best in relevance, reinforcing the importance of multiple dimensions for evaluating recommendations. *NeuralMF* achieved the lowest scores for both groups. For them, profile atypicality (Section III-A) did not seem to cause a disparity in recommendations. This could imply that all users were receiving recommendations with similar characteristics (i.e., focusing on the "more common" users), which might negatively affect users with interests that are not strictly aligned with their group [5]. Then, although recommendations might seem fair, other aspects of user experience might be degrading.

In summary, statistically significant differences were observed for both groups and most analyzed metrics, indicating unfairness in group treatment. Recommenders showed different degrees of unfairness, which varied depending on the metric and group under analysis. They were shown to affect the relevance of recommendations in favour of one of the groups and other network qualities, such as diversity/novelty, or the likelihood of being exposed to users from the other group. Besides unfair treatment of groups, user recommenders may contribute to extreme polarization.

B. Causal analysis

We tested the variance information factor (VIF) to determine the existence of variable multi-collinearity. None of the VIFs exceeded the level of 5 [49], indicating that multi-collinearity was not a concern. Table III summarizes the statistics and coefficients for different representative models[9]. Models were compared based on their Akaike Information Criterion (AIC) and log-likelihood ratio.

The worst fit for models with one feature was observed for training size, followed by political alignment. As expected, given the original distribution of interactions, training size was shown to not affect nDCG [15]. Political alignment had a small positive effect on nDCG, showing that despite recommenders being unaware of it, it might have been inadvertently carried into the models by other features. Models including both profile atypicality features improved AIC. Ratio Group and content atypicality had opposite effects and different magnitudes. This might imply that political alignment could be more important than content similarity for estimating recommendation quality. On the other hand, when solely considering the recommendation atypicality features, AIC Was slightly reduced, but none of the features showed a significant effect on nDCG. The sole consideration of model-induced features did not seem to contribute to or condition the quality of said recommendations.

Combining political alignment with profile or recommendation atypicality positively affected AIC. Despite being based on the same attribute, atypicalities had opposite effects. While content profile atypicality contributed negatively to nDCG, recommendation atypicality contributed positively. Political alignment had a different impact on both combinations, indicating the existence of cross-effects among features and the need to continue exploring more complex models. Finally, including all features achieved the best AIC. The largest positive effects were observed for profile atypicality Ratio Group and recommendation atypicality content. Despite being statistically significant, the effect of political alignment had one order of magnitude less than the other features.

Finally, we also analyzed models where each individual recommender acted as the dependent variable. For *NeuralMF* and *content-based* recommenders, the AIC was better than for the averaged scores. The worst results were observed for *popularity*, with AICs up to a triple higher than for the averaged nDCG. These differences highlight the task difficulty and the differentiated effect on the outcome of model-induced features and profile and recommender atypicalities. In most cases, the best model included almost every defined feature (except for training size), and achieved the worst results when individually considering features (in particular, training size and political alignment). The only exception was content-based, in which the worst model included all features, while the model only considering political alignment was among the best ones.

[9]The rest of the analyzed models can be found in the companion repository.

		political alignment	training size	political alignment + training size	profile atypicality	rec atypicality	political alignment + training size + profile atypicality	political alignment + training size + rec atypicality	all features
political alignment		0.035***± 0.009	-	0.036***± 0.009	-	-	0.025**± 0.009	0.058***± 0.011	0.049***± 0.011
training size		-	0.0± 0.0	0.0± 0.0	-	-	0.0± 0.0	0.0± 0.0	0.0± 0.0
profile atypicality	Ratio group	-	-	-	0.79***± 0.19	-	0.65***± 0.196	-	0.653**± 0.211
	Content	-	-	-	-0.571***± 0.157	-	-0.446*± 0.175	-	-0.47**± 0.176
rec atypicality	Ratio group	-	-	-	-	0.579± 0.39	-	-0.147± 0.388	-0.464± 0.392
	Content	-	-	-	-	0.079± 0.163	-	0.609**± 0.19	0.653***± 0.188
AIC		548.405	558.500	546.238	540.543	537.088	538.079	537.088	529.366
log-likelihood		-272.2	-277.25	-270.12	-267.27	-263.54	-264.04	-263.54	-257.68
null log-likelihood		-278.83	-278.82	-278.84	-278.87	-278.9	-278.9	-278.9	-278.98

TABLE III: Regression coefficients for predicting recommendations quality as nDCG. Columns indicate the combination of features included in each evaluated model. A "-" indicates that the feature (row) was not included in the model (column). Stat. significance: *p $<$ .05, **p $<$ .01, ***p $<$ .001

In summary, the analysis showed different contributions of the data and model-induced features to the outcome. When considered independently, profile atypicality had a more significant effect on nDCG than recommender atypicality. Group ratio was more informative than user' political alignment, highlighting the importance of network topology on recommender outcome. Finally, we found cross-effects among the defined features.

V. Conclusions

We study how user recommenders affect fairness and its consequences on network dynamics based on users' political alignment. The study also highlights how (un)fairness could create more polarized environments, strengthening filter bubbles/echo chambers.

The study acknowledges some limitations that could be addressed in future works. First, while we include major approaches, our study does not cover all recommenders and data possibilities. Data had a strong topic focus, which might have affected interaction dynamics and recommendations [12]. The study should be replicated using other collections of varying sizes and topical focus to enable the analysis of conflicting aspects of fairness principles, such as users who engage in harmful behaviors (e.g., misinformation spreading). Second, exploring personalized fairness notions could also be necessary, where sensitive features are treated independently based on users' fairness needs and merits [1, 42]. Third, the analysis focused on a static network snapshot, neglecting the feedback-loop effect that multiple runs could have on network dynamics. It would be beneficial to study whether the observed effects are reinforced over time and to develop strategies for mitigating them. Fourth, we considered a binary political alignment, which may not capture the complex political spectrum and user preferences over different topics [4]. Finally, causal models could be extended to include counterfactual analysis and explain recommendations, which would enhance transparency, trustworthiness and user satisfaction [2].

Ethics statement

Research is based on publicly available *Twitter* data (now *X*) initially collected and tagged by third parties. User identities were not included in the analysis. In compliance with *Twitter/X* TOS, the shared graphs only include Ids. The analysis aims to show the unintentional effects of recommenders on fairness and indirectly on network dynamics. However, other biases may stem from the data collection and the strategy used for computing political alignment that should be considered before using any result from this study in any real-world setting.

References

[1] Y. Deldjoo, V. W. Anelli, H. Zamani, A. Bellogin, and T. Di Noia, "A flexible framework for evaluating user and item fairness in recommender systems," *User Modeling and User-Adapted Interaction*, vol. 31, no. 3, pp. 1–55, 2021.

[2] J. Chen, H. Dong, X. Wang, F. Feng, M. Wang, and X. He, "Bias and debias in recommender system: A survey and future directions," *ACM Transactions on Information Systems*, vol. 41, no. 3, pp. 1–39, 2023.

[3] M. D. Ekstrand, A. Das, R. Burke, F. Diaz *et al.*, "Fairness in information access systems," *Foundations and Trends® in Information Retrieval*, vol. 16, no. 1-2, pp. 1–177, 2022.

[4] P. Liu, K. Shivaram, A. Culotta, M. A. Shapiro, and M. Bilgic, "The interaction between political typology and filter bubbles in news recommendation algorithms," in *Proceedings of the Web Conference 2021*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 3791–3801.

[5] Y. Deldjoo, D. Jannach, A. Bellogin, A. Difonzo, and D. Zanzonelli, "Fairness in recommender systems: research landscape and future directions," *User Modeling and User-Adapted Interaction*, pp. 1–50, 2023.

[6] I. Guy, *People Recommendation on Social Media*. Cham: Springer International Publishing, 2018, pp. 570–623. [Online]. Available: https://doi.org/10.1007/978-3-319-90092-6_15

[7] E. M. Daly, W. Geyer, and D. R. Millen, "The network effects of recommending social connections," in *Proceedings of the Fourth ACM Conference on Recommender Systems*, ser. RecSys '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 301–304. [Online]. Available: https://doi.org/10.1145/1864708.1864772

[8] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *Proceedings of the twelfth international conference on Information and*

knowledge management. New York, NY, USA: Association for Computing Machinery, 2003, pp. 556–559.

[9] J. Su, A. Sharma, and S. Goel, "The effect of recommendations on network structure," in *Proceedings of the 25th International Conference on World Wide Web*, ser. WWW '16. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2016, p. 1157–1167. [Online]. Available: https://doi.org/10.1145/2872427.2883040

[10] F. Fabbri, F. Bonchi, L. Boratto, and C. Castillo, "The effect of homophily on disparate visibility of minorities in people recommender systems," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14. Atlanta, Georgia, USA: AAAI Press, 2020, pp. 165–175.

[11] J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy, "Make new friends, but keep the old: Recommending people on social networking sites," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 201–210. [Online]. Available: https://doi.org/10.1145/1518701.1518735

[12] F. Fabbri, M. L. Croci, F. Bonchi, and C. Castillo, "Exposure inequality in people recommender systems: the long-term effects," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16. Atlanta, Georgia, USA: AAAI Press, 2022, pp. 194–204.

[13] A. Olteanu, C. Castillo, F. Diaz, and E. Kıcıman, "Social data: Biases, methodological pitfalls, and ethical boundaries," *Frontiers in big data*, vol. 2, p. 13, 2019.

[14] A.-A. Stoica, C. Riederer, and A. Chaintreau, "Algorithmic glass ceiling in social networks: The effects of social recommendations on network diversity," in *Proceedings of the 2018 World Wide Web Conference*, ser. WWW '18. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018, p. 923–932. [Online]. Available: https://doi.org/10.1145/3178876.3186140

[15] M. D. Ekstrand, M. Tian, I. M. Azpiazu, J. D. Ekstrand, O. Anuyah, D. McNeill, and M. S. Pera, "All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness," in *Conference on fairness, accountability and transparency*, PMLR. New York, NY, USA: PMLR, 2018, pp. 172–186.

[16] O. Lesota, A. Melchiorre, N. Rekabsaz, S. Brandl, D. Kowald, E. Lex, and M. Schedl, "Analyzing item popularity bias of music recommender systems: are different genders equally affected?" in *Proceedings of the 15th ACM Conference on Recommender Systems*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 601–606.

[17] M. Fernández, A. Bellogín, and I. Cantador, "Analysing the effect of recommendation algorithms on the amplification of misinformation," *CoRR*, vol. abs/2103.14748, 2021.

[18] F. Cinus, M. Minici, C. Monti, and F. Bonchi, "The effect of people recommenders on echo chambers and polarization," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16. Atlanta, Georgia, USA: AAAI Press, 2022, pp. 90–101.

[19] A. Ghosh, L. Genuit, and M. Reagan, "Characterizing intersectional group fairness with worst-case comparisons," in *Artificial Intelligence Diversity, Belonging, Equity, and Inclusion*. PMLR, 2021, pp. 22–34.

[20] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis, "Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship," in *Proceedings of the 2018 world wide web conference*. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018, pp. 913–922.

[21] P. Barberá, J. T. Jost, J. Nagler, J. A. Tucker, and R. Bonneau, "Tweeting from left to right: Is online political communication more than an echo chamber?" *Psychological science*, vol. 26, no. 10, pp. 1531–1542, 2015.

[22] J. Sanz-Cruzado and P. Castells, "Enhancing structural diversity in social networks by recommending weak ties," in *Proceedings of the 12th ACM conference on recommender systems*. New York, NY, USA: Association for Computing Machinery, 2018, pp. 233–241.

[23] F. Pierri, C. Piccardi, and S. Ceri, "Topology comparison of twitter diffusion networks effectively reveals misleading information," *Scientific reports*, vol. 10, no. 1, p. 1372, 2020.

[24] ——, "A multi-layer approach to disinformation detection in us and italian news spreading on twitter," *EPJ Data Science*, vol. 9, no. 1, p. 35, 2020.

[25] P. Cogan, M. Andrews, M. Bradonjic, W. S. Kennedy, A. Sala, and G. Tucci, "Reconstruction and analysis of twitter conversation graphs," in *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*. New York, NY, USA: Association for Computing Machinery, 2012, pp. 25–31.

[26] W. Cota, S. C. Ferreira, R. Pastor-Satorras, and M. Starnini, "Quantifying echo chamber effects in information spreading over political communication networks," *EPJ Data Science*, vol. 8, no. 1, p. 35, 2019.

[27] M. Yuksekgonul, L. Zhang, J. Zou, and C. Guestrin, "Beyond confidence: Reliable models should also quantify atypicality," in *ICLR 2023 Workshop on Pitfalls of limited data and computation for Trustworthy ML*, Kigali, Rwanda, 2023.

[28] T. De Pessemier, S. Dooms, and L. Martens, "Comparison of group recommendation algorithms," *Multimedia tools and applications*, vol. 72, pp. 2497–2541, 2014.

[29] E. Pitoura, K. Stefanidis, and G. Koutrika, "Fairness in rankings and recommendations: an overview," *The VLDB Journal*, vol. 31, no. 3, pp. 1–28, 2022.

[30] J. Sanz-Cruzado, S. M. Pepa, and P. Castells, "Structural

novelty and diversity in link prediction," in *Companion Proceedings of the The Web Conference 2018*. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018, pp. 1347–1351.

[31] D. Sacharidis, "Diversity and novelty in social-based collaborative filtering," in *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, ser. UMAP '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 139–143. [Online]. Available: https://doi.org/10.1145/3320435.3320479

[32] S. Arora, A. May, J. Zhang, and C. Ré, "Contextual embeddings: When are they worth it?" in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 2650–2663. [Online]. Available: https://aclanthology.org/2020.acl-main.236

[33] P. W. Holland and S. Leinhardt, "Transitivity in structural models of small groups," *Comparative group studies*, vol. 2, no. 2, pp. 107–124, 1971.

[34] J. Hannon, M. Bennett, and B. Smyth, "Recommending twitter users to follow using content and collaborative filtering approaches," in *Proceedings of the Fourth ACM Conference on Recommender Systems*, ser. RecSys '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 199–206. [Online]. Available: https://doi.org/10.1145/1864708.1864746

[35] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *8th IEEE international conference on data mining*. Pisa, Italy: IEEE Computer Society, 2008, pp. 263–272.

[36] T. Di Noia, V. C. Ostuni, J. Rosati, P. Tomeo, and E. Di Sciascio, "An analysis of users' propensity toward diversity in recommendations," in *Proceedings of the 8th ACM Conference on Recommender Systems*. New York, NY, USA: Association for Computing Machinery, 2014, pp. 285–288.

[37] Q. Grossetti, C. Du Mouza, and N. Travers, "Community-based recommendations on twitter: avoiding the filter bubble," in *Web Information Systems Engineering–WISE 2019: 20th International Conference, Hong Kong, China, January 19–22, 2020*, Springer. Berlin, Heidelberg: Springer-Verlag, 2019, pp. 212–227.

[38] S. Vargas and P. Castells, "Exploiting the diversity of user preferences for recommendation," in *Proceedings of the 10th conference on open research areas in information retrieval*. Lisbon, Portugal: LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2013, pp. 129–136.

[39] D. Liang, R. G. Krishnan, M. D. Hoffman, and T. Jebara, "Variational autoencoders for collaborative filtering," in *Proceedings of the 2018 world wide web conference*. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018, pp. 689–698.

[40] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proceedings of the 26th international conference on world wide web*. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2017, pp. 173–182.

[41] P. Cremonesi, Y. Koren, and R. Turrin, "Performance of recommender algorithms on top-n recommendation tasks," in *Proceedings of the fourth ACM conference on Recommender systems*. New York, NY, USA: Association for Computing Machinery, 2010, pp. 39–46.

[42] Y. Li, H. Chen, S. Xu, Y. Ge, and Y. Zhang, "Towards personalized fairness based on causal notion," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 1054–1063.

[43] K. Lin, N. Sonboli, B. Mobasher, and R. Burke, "Calibration in collaborative filtering recommender systems: A user-centered analysis," in *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, ser. HT '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 197–206. [Online]. Available: https://doi.org/10.1145/3372923.3404793

[44] Z. Wan, Y. Mahajan, B. W. Kang, T. J. Moore, and J.-H. Cho, "A survey on centrality metrics and their network resilience analysis," *IEEE Access*, vol. 9, pp. 104 773–104 819, 2021.

[45] K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis, "Quantifying controversy on social media," *Trans. Soc. Comput.*, vol. 1, no. 1, jan 2018. [Online]. Available: https://doi.org/10.1145/3140565

[46] H. Steck, "Calibrated recommendations," in *Proceedings of the 12th ACM Conference on Recommender Systems*, ser. RecSys '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 154–162. [Online]. Available: https://doi.org/10.1145/3240323.3240372

[47] G. Faggioli, N. Ferro, and N. Fuhr, "Detecting significant differences between information retrieval systems via generalized linear models," in *Proceedings of the 31st ACM international conference on information & knowledge management*. New York, NY, USA: Association for Computing Machinery, 2022, pp. 446–456.

[48] F. Zuiderveen Borgesius, D. Trilling, J. Möller, B. Bodó, C. H. De Vreese, and N. Helberger, "Should we worry about filter bubbles?" *Internet Policy Review. Journal on Internet Regulation*, vol. 5, no. 1, 2016.

[49] S. Menard, "Collinearity. applied logistic regression analysis second edition (quantitative applications in the social sciences)," 2001.