ALGORITHMIC DRIFT: A SIMULATION FRAMEWORK TO STUDY THE EFFECTS OF RECOMMENDER SYSTEMS ON USER **PREFERENCES**

A PREPRINT

Erica Coppolillo* University of Calabria **ICAR-CNR**

erica.coppolillo@unical.it

Simone Mungari

University of Calabria **ICAR-CNR** Revelis s.r.l.

simone.mungari@unical.it

Ettore Ritacco

University of Udine ettore.ritacco@uniud.it Francesco Fabbri

Spotify francescof@spotify.com Marco Minici

University of Pisa **ICAR-CNR**

marco.minici@icar.cnr.it

Francesco Bonchi CENTAI

francesco.bonchi@centai.eu

Giuseppe Manco **ICAR-CNR**

giuseppe.manco@icar.cnr.it

September 26, 2024

ABSTRACT

Digital platforms such as social media and e-commerce websites adopt Recommender Systems to provide value to the user. However, the social consequences deriving from their adoption are still unclear. Many scholars argue that recommenders may lead to detrimental effects, such as biasamplification deriving from the feedback loop between algorithmic suggestions and users' choices. Nonetheless, the extent to which recommenders influence changes in users leaning remains uncertain. In this context, it is important to provide a controlled environment for evaluating the recommendation algorithm before deployment. To address this, we propose a stochastic simulation framework that mimics user-recommender system interactions in a long-term scenario. In particular, we simulate the user choices by formalizing a user model, which comprises behavioral aspects, such as the user resistance towards the recommendation algorithm and their inertia in relying on the received suggestions. Additionally, we introduce two novel metrics for quantifying the algorithm's impact on user preferences, specifically in terms of drift over time. We conduct an extensive evaluation on multiple synthetic datasets, aiming at testing the robustness of our framework when considering different scenarios and hyper-parameters setting. The experimental results prove that the proposed methodology is effective in detecting and quantifying the drift over the users preferences by means of the simulation. All the code and data used to perform the experiments are publicly available².

Keywords Simulation Framework · Evaluation Metrics · Recommender Systems · Algorithmic Drift

^{*}E. Coppolillo and S. Mungari equally contributed to the paper.

²https://anonymous.4open.science/r/AlgorithmicDrift-D553

1 Introduction

Nowadays, people adopt social media as a fundamental medium to share and consume information. With a growing amount of available content, recommender systems represent a viable way to help users navigate large volumes of information by suggesting content that the user may like. As a downside, recommendation algorithms have also been blamed for detrimental effects such as echo chambers [Pariser, 2011] and opinion polarization [Cho et al., 2020], which in turn may lead to pernicious phenomena such as misinformation spreading [Del Vicario et al., 2016], fragmentation [Sunstein, 2018], and radicalization [Sunstein, 1999]. As a natural consequence, measuring the potential side-effects of recommender systems in the emergence of these issues is attracting much interest, with scholars proposing data-driven analysis [Bakshy et al., 2015, De Francisci Morales et al., 2021, Cinelli et al., 2021], model-based methods [Minici et al., 2022], and simulation studies [Fabbri et al., 2020, Santos et al., 2021, Cinus et al., 2022, Tommasel and Menczer, 2022].

Despite this growing interest, the challenge is still open in the current literature in analyzing and measuring how the algorithm effects on users may evolve over time. Although several models have been proposed [Chaney et al., 2018, Bountouridis et al., 2019, Yao et al., 2021, Coppolillo et al., 2024a], these studies focus on specific topics, such as behavioral homogeneity or items diversity. Indeed, there is still lack of a unified and generalizable methodology which allows to track users preferences' evolution over time, and that is orthogonal to the kind of drift to quantify. Further, most of the existing simulation environments follow a simplistic behavioral model, providing limited flexibility in representing different behavioral patterns.

To fill this gap, in this paper, we propose a stochastic model for characterizing time-based interactions between users and recommendation algorithms in operational environments. In our scenario, users interact with various items on a platform, such as media content, news articles, or videos. The platform collects data on user preferences and utilizes this information to generate personalized recommendations. These recommendations are typically provided by ranking items based on their predicted relevance to the user [Agarwal et al., 2019].

An important aspect of our formulation consists in accounting for the "feedback loop" effect described in the literature [Chaney et al., 2018], where user choices are influenced by the recommendations provided, and the algorithm relies on the user's past interactions rather than aligning with their true interests. Therefore, to accurately simulate user choices, it is essential to characterize user behavior by considering several factors, such as their potential *resistance* to recommendations (i.e., the autonomous selection of an item from the entire catalog) and their *inertia* in following the provided suggestions. Starting from this, the research questions we aim to address are the following: *Can we quantify the effect of the recommendation algorithm in altering user preferences over the long term? How do different behavioral users pattern affect the influence of the recommender?*

To answer these questions, we propose a controlled simulated environment where mimicking user-recommender system interactions, along with two novel metrics to evaluate whether and to what extent the recommender system contributes to drifting users' initial preferences. Following from these assumptions, we introduce the concept of "algorithmic driff", in order to characterize how the recommendation algorithm contributes in changing user leanings. In practice, the simulation model starts from an initial group of heterogeneous user preferences, from which the recommender system induces initial transition probabilities between item categories. Notably, besides (blindly) following the provided recommendations, our model allows users to either completely ignore the suggestions and pick an item autonomously from the catalog (resistance), or still examine the recommendation list but choose the item still relying on their own preferences (inertia). Further, the user's choice can be influenced by exogenous hence unpredictable factors (randomness), that can lead to spurious interactions. By supporting different behavioral patterns in the users choices, we provide a comprehensive framework that allows for measuring the effects of the recommendation algorithms under different conditions. Notably, the metrics we propose enable such measurements, by tracking and quantifying changes according to predefined criteria.

As a paradigmatic example of such criteria, in our experiments we assume that the items available on the platform can be categorized as either *harmful* or *neutral*. Simultaneously, users are classified into three categories based on their interactions with these items: *non-radicalized*, *semi-radicalized*, or *radicalized*. This classification is determined by the proportion of harmful interactions they have exhibited. Further, we focus on a collaborative filtering-based scenario, where we consider the implicit connections that can be induced by the collaborative nature of the algorithm, hence exploiting the bridging nature of some users to propagate influences on the adoptions. As a result, the proposed framework allows us to quantify the changes in user leanings.

Our contributions can be summarized as follows:

• We define the concept of *algorithmic drift* and introduce two metrics for evaluating the impact of recommender systems on user behavior changes in the long term.

- We model a novel simulation framework and standard methodology to study the effect of any recommendation algorithm in a completely model-driven setting.
- We conduct an extensive analysis to assess the effectiveness of our methodology and its robustness across different scenarios.

The rest of the paper is structured as follows. Section 2 illustrates the current literature in the context of recommendation. In Section 3, we formally describe the proposed simulation framework. Section 4 presents the results of the experimental evaluation. Finally, Section 5 concludes the paper, depicting some pointers for future work.

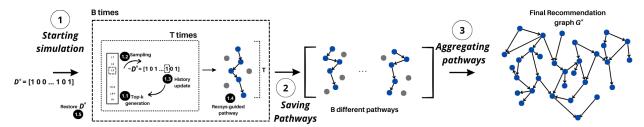


Figure 1: Framework overview. Considering the initial interactions of each user u, we start the simulation (1) by invoking the trained recsys model B times. In each of these iterations, we get the top-k recommended items (1.1) and we sample an item (1.2), thus updating the history of u (1.3). We repeat the process T times, in order to build a tree whose nodes are the items sampled (1.4), of depth T. Intuitively, each tree represents a pathway of the user u guided by the recsys. After each iteration, we restore the initial user history (1.5), and the process restarts. At the end of the B iterations, we obtain T different trees, representing each pathway (2). Hence, we aggregate all the trees (3) and we obtain the final recommendation graph G^u .

2 Related work

The contribution of this work spans on a variety of research fields, whose literature is explored as follows.

Simulation-based studies. We can observe a growing research line studying the long-term effects of recommender systems through simulation models. Badami et al. [2018] propose a simulation framework for studying polarization phenomena in a context in which users are binarized into two homogeneous groups.

Chaney et al. [2018] claim that recommenders try to homogenize user behavior without increasing the utility of their suggestions. They argue that the most recent recommendation systems are actually trained over preferences that are already biased by previous recommendations, thus triggering a feedback loop where a suggestion algorithm tries to fit the recommendation history instead of matching the real interests of the users.

Bountouridis et al. [2019] propose a simulation framework for studying the impact of recommender systems in drifting users towards different content topics, thus mainly focusing on items diversity and users serendipity. Similarly, Yao et al. [2021] depicts a simulated environment to assess the actual contribution of recommenders in modifying users habits, particularly focusing on the phenomenon of popularity bias.

Fabbri et al. [2022], Cinus et al. [2022], Santos et al. [2021] show how people-recommenders can exacerbate social media critical issues, such as polarization, misinformation, and pre-existing inequalities in user communities. The authors define a simulation model that highlights the following recommender systems effects: (i) the growth of exposure inequalities at the individual level, strengthening the "rich-get-richer" effect Fabbri et al. [2022], and (ii) the formation of polarized communities and echo chambers within homophilic groups Cinus et al. [2022], Santos et al. [2021]. In a similar vein, de Arruda et al. [2021] devised a model for simulating changes in users' opinions in social networks, showing potential detrimental effects such as polarization and echo chambers formations.

Tommasel and Menczer [2022] focus on the misinformation spread in social networks due to the recommender system capability of influencing the network topology. First, they simulate changes in user misinformation-spreading behavior and define counterfactual scenarios. Then, they simulate an opinion dynamic model over the derived network to estimate how recommendations affect the influence of users spreading misinformation.

We would like to emphasize how our work draws inspiration from the discussed literature but substantially differs from it. Indeed, we are not interested in explaining or assessing detrimental phenomena induced by recommender systems, but we aim at providing a controlled environment where the algorithm can be tested before deployment, by simulating its interaction with the users in the long term.

Data-driven studies. Ribeiro et al. [2020] and Haroon et al. [2022] show that YouTube recommendations lead to the genesis of "rabbit holes", i.e., a journey toward increasingly radicalized contents that can exaggerate user bias, belief, and opinion. Ribeiro et al. [2020] analyze videos on channels belonging to four political leanings. The authors observe that some center channels serve as gateways to push users to far-right ideology, thus showing evidence of radicalization phenomena where users migrate from milder to extreme ideologies. On the other hand, Haroon et al. [2022]'s analysis is conducted by exploiting sock puppets (i.e., brand new users with ad-hoc history) which are actively monitored while using YouTube, in order to discover and quantify the emergence of ideological biases. Similarly, Phadke et al. [2022] provide empirical modeling for analyzing the radicalization phenomenon over conspiracy theory discussions in Reddit. As a result, users on increasing conspiracy engagement pathways progress successively through various radicalization stages.

Once again, our proposal shows strong differences with respect to this literature. One of our main contributions is the definition of general quantitative indexes to measure users' opinion drift induced by several CF recommender systems.

Model-based approaches. User behavior deviation from natural evolution has been widely studied in the literature, still remaining an open challenge. Rastegarpanah et al. [2019] define a strategy to build *antidote data* to feed models aiming at promoting the fairness. To this purpose, authors define a metric to measure user homogeneity as the average variance of the preferences on items. Ramaciotti Morales and Cointet [2021] model social graphs and user belief as time-dependent processes, where ideological changes and network topology co-evolution in time is monitored. Their idea is to start from a real social network (e.g., Twitter) and apply changes guided by people-recommenders: thus, drawing inspiration from Degroot [1974], Dandekar et al. [2013], they apply the Duclos-Esteban-Ray polarization measure Duclos et al. [2004] to estimate how recommendation-driven conformities would form. Minici et al. [2022] propose a probabilistic generative model that discovers echo-chambers in a social network by introducing a Generalized Expectation Maximization algorithm, that clusters users according to how information spread among them. Chang and Ugander [2022] define a counterfactual model called "organic model", which is used to compare user outcomes influenced by a recommender against the outcomes under the natural evolution of the users' choices. The main finding of this work states that the recommender and organic model dramatically differ from each other, thus highlighting the influence of the recommendation.

Differently, our proposal is to formalize a standard methodology, to measure deviations (triggered by any target recommender system) from user normal evolution, that exhaustively navigates the user choice space, instead of considering only one pathway that a user can choose.

3 Simulation Framework

The core of our approach consists in devising the simulation framework and a set of metrics for measuring the potential algorithmic drift that a target recommendation algorithm may induce. In the following, we describe such contributions.

3.1 Preliminaries

Let U (resp. I) be the set of users (resp. items). We consider an implicit-feedback matrix $\mathcal{D} \in \{0,1\}^{|U| \times |I|}$ of user-item interactions. \mathcal{D}_{ui} is equal to 1 if user u has interacted with item i, and 0 otherwise. With an abuse of notation, we denote by $\mathcal{D}_u \subseteq I$ the set of all items i such that $\mathcal{D}_{ui} = 1$. Also, $(u, i) \in \mathcal{D}$ whenever $\mathcal{D}_{ui} = 1$.

We assume a labeling function $\ell: I \to \{c_1, \dots, c_N\}$ that tags an item i as belonging to a category among c_1, \dots, c_N , thus partitioning I into N disjoint subsets $I_i = \{i \in I | \ell(i) = c_i\}, j = 1, \dots, N$.

We adopt a recommender system $\mathcal{R}_{\theta}: U \times I \to [0,1]$, parameterized by θ and previously trained on \mathcal{D} . Given an unseen user-item interaction (u,i), i.e., $D_{ui}=0$, $\mathcal{R}_{\theta}(u,i)$ estimates how likely the user u interacts with an item i once it is exposed to it. Such a likelihood can be exploited to build a ranked recommendation list for a user u, formed by unseen items and sorted by decreasing likelihood.

3.2 Organic Model

To assess the effect of recommender algorithms in terms of preference alteration, it is crucial to have a counterfactual based on an organic model — i.e., assuming users and items interact without any \mathcal{R}_{θ} . For this purpose, we adopt the organic model proposed by Chang and Ugander [2022]. Assuming to know user and item features, ρ_u and α_i , respectively, preferences are generated according to the measure $\|\rho_u - \alpha_i\|$. However, users don't have full knowledge of the actual features α_i for each item i and can only access samples through a Normal distribution $\hat{\alpha_i} \sim \text{Normal}(\alpha_i, \Sigma)$, with $\Sigma = 0.5 \cdot \Sigma_{\text{item}}$ and $\Sigma_{\text{item}} = \alpha^T \alpha$ being the empirical covariance of the item features. Chang and Ugander [2022]

also propose to use an interpolation between the sample $\hat{\alpha}_i$ and the mean of all the samples:

$$\hat{\alpha_i}^S = \xi \hat{\alpha_i} + (1 - \xi) \left(\frac{1}{|I|} \sum_{j \in I} \hat{\alpha_j} \right)$$
 (1)

We use the shrunken estimate using a value of $\xi = 0.4$ since it was empirically determined to be optimal. However, similar results are obtained by putting the value of $\xi = 0$, thus disregarding a global effect in the noisy estimate.

3.3 User Behaviour

Our simulation model mimics how user preferences and recommender system co-evolve in a long-term scenario. In each step, the user selects an item to interact with, by either resorting to the algorithm recommendations, or by autonomously examining the whole catalog. Consequently, the algorithm adapts the next recommendations relying on these new interactions: this tight relationship between users and recommender system models the well-known concept of feedback loop [Chaney et al., 2018]. Further, our framework aims at simulating the user navigation across the whole choice space, instead of a single pathway. For this reason, the procedure consists of T iterative trials, each of which is repeated for B independent rounds. We impose that, in each of the B rounds, a user cannot interact with an item more the once. Within the simulation, for each user u, we instantiate an item-item matrix $S^u \in \mathbb{N}^{|I| \times |I|}$. Intuitively, $S^u_{i,j}$ counts how many times u shifted from item i to item j in two consecutive rounds of the simulation, as explained in the following.

At each step t, The recommender system provides a list $L_t \subset I$ of k items to the user. The latter behaves according to the following process: if their *resistance* to the suggestions of the recommender system (quantified as γ) is high, the user may (stochastically) select an item from the whole catalog, by either relying on their own preferences, or picking it randomly. The random factor η (which we assume to be very low in practice) models the influence of exogenous hence uncontrollable factors that can guide the user's choice in that round. For instance, the user may select a spurious item under the influence of a friend's suggestion, by relying on advertisements, or even by mistake.

Alternatively, if the user is not highly resistant to the algorithm, they may examine the recommendation list and pick an item. This can occur by either totally relying on the system (high *inertia*), or under the influence of their own interests (low *inertia*). Here, the core idea is that the user's choice may be influenced by its own preferences and beliefs, or conversely, they may completely rely on the recommender system, thus resembling the trust bias phenomenon [Agarwal et al., 2019].

Therefore, the probability for user u to select item $i \in I$ at step t is given by:

$$P_t(i|u) = \gamma \cdot P_t^I(i|u) + (1-\gamma) \cdot P_t^L(i|u)$$
(2)

where

$$P_t^I(i|u) = \eta \cdot \frac{1}{|I|} + (1 - \eta) \cdot P^o(i|u)$$
(3)

$$P_t^L(i|u) = \begin{cases} \delta \cdot P_t^s(i|u) + (1-\delta) \cdot P^o(i|u) & \text{if } i \in L_t \\ 0 & \text{otherwise} \end{cases}$$
 (4)

Here, $P_t^s(i|u)$ represents the probability of selecting the item based on the score provided by the recommendation algorithm, while $P^o(i|u)$ is the probability of picking it by following the natural preferences of the user. We define them as:

$$P_t^s(i|u) = \frac{\mathcal{R}_{\theta}(i)}{\sum_{j \in L_t} \mathcal{R}_{\theta}(j)}$$
 (5)

$$P^{o}(i|u) = \frac{\|\rho_{u} - \hat{\alpha_{i}}^{S}\| - \min_{j \in L_{t}}(\|\rho_{u} - \hat{\alpha_{j}}^{S}\|)}{\max_{j \in L_{t}}(\|\rho_{u} - \hat{\alpha_{j}}^{S}\|) - \min_{j \in L_{t}}(\|\rho_{u} - \hat{\alpha_{j}}^{S}\|)}$$
(6)

The linear combination of these two components is weighted in Equation 2 using the *inertia* parameter $\delta \in [0,1]$. Intuitively, if $\delta = 1$, the user blindly follows the recommendations provided at each round, resembling the trust bias phenomenon [Agarwal et al., 2019]; conversely, if $\delta = 0$, the user selects the item based on their preferences only, i.e., by following the organic model; finally, if $0 < \delta < 1$, the choice is conditioned by both the user's interests and the recommendation score. We denote the selected item as i^* .

Algorithm 1 Users-Recsys Interaction Process

```
1: Input: Number of rounds T, user set U, item set I, history dataset \mathcal{D}, user matrices \mathcal{S}^{u_1}, \ldots, \mathcal{S}^{u_{|U|}}, recommender system \mathcal{R}_{\theta},
     resistance factor \gamma, random factor \eta, inertia factor \delta
     Output: G^{u_1},\ldots,G^{u_{|U|}}
 3: for b \in \{1, ..., B\} do
                                                                                                                                                         \triangleright B independent trials
 4:
           \hat{\mathcal{D}} \leftarrow \mathcal{D}
                                                                                                                                     ▶ Revert the original history dataset
 5:
           for t \in \{1, \ldots, T\} do
 6:
                for u \in U do
 7:
                     \hat{I}^u \leftarrow \emptyset
 8:
                     j \leftarrow \text{None}
                                                                                                                                     ▶ Initialize previously selected item
 9:
                     if Bernoulli(\gamma) then
                                                                                                                                                          \triangleright Resistance factor \gamma
10:
                           if Bernoulli(\eta) then
                                                                                                                                                             \triangleright Random factor \eta
                                 Pick i^* uniformly from I
11:
12:
13:
                                Pick i^* with probability P^o(i|u), i \in I

    ▷ Sampling from catalog based on preferences

14:
                           end if
                      else
15:
                           L = [i_1, i_2, \dots, i_k] \leftarrow \mathcal{R}_{\theta}(\hat{\mathcal{D}})
                                                                                                                                                   \triangleright Top-k recommendations
16:
17:
                           if Bernoulli(\delta) then

ightharpoonup Inertia factor \delta
                                Pick i^* with probability P_t^s(i|u), i \in L
                                                                                                               > Sampling from list based on recommender scores
18:
19:
                                Pick i^* with probability P^o(i|u), i \in L
20:
                                                                                                                            > Sampling from list based on preferences
21:
                           end if
22:
                      end if
23:
                      \hat{\mathcal{D}}_u \leftarrow \hat{\mathcal{D}}_u \cup \{i^*\}

    □ Updating the history dataset

24:
                      if j is not None then
25:
                           \mathcal{S}_{j,i^*}^u \leftarrow \mathcal{S}_{j,i^*}^u + 1
                                                                                                                                                  ▶ Updating the user matrix
26:
                      end if
                     \begin{array}{l} \hat{I}_u \leftarrow \hat{I}_u \cup \{i^*\} \\ j \leftarrow i^* \end{array}
27:
28:
29:
                end for
30:
           end for
31: end for
32: \hat{\mathcal{S}}^{u_1}, \dots, \hat{\mathcal{S}}^{u_{|U|}} \leftarrow \text{Normalize}(\mathcal{S}^{u_1}, \dots, \mathcal{S}^{u_{|U|}})
                                                                                                         33: G^{u_1}, \ldots, G^{u_{|U|}} \leftarrow (\hat{I}_1^u, \mathcal{S}^{u_1}), \ldots, (\hat{I}^{u_{|U|}}, \mathcal{S}^{u_{|U|}})
```

Within the modeled scenario, at each round, the choice of i^* is tracked within the dataset \mathcal{D}_u , which is used by the recommender algorithm to provide new suggestions. This also results in increasing by 1 the value S_{j,i^*}^u , where j represents the item that u selected in the previous round.

Notably, the matrix S^u represents a blueprint of how the preferences of a user u co-evolve together with a recommender system \mathcal{R}_{θ} from a long-term perspective. Starting from S^u , in fact, we are able to generate a probabilistic graph $G^u = (\hat{I}^u, \hat{S}^u)$, where $\hat{I}^u \subset I$ is the subset of items for which u had at least one interaction during the simulation, and \hat{S}^u being the row-normalized stochastic shifting matrix. Note that G^u captures the effects of the recommender model \mathcal{R}_{θ} in the long term, as it exhibits transition probabilities and hence ultimately the likelihood of users drifting their leanings. The final output of our simulation process is hence the probabilistic graph G^u , retrieved for each user u. We summarize the whole procedure in Algorithm 1 and illustrate the overall flow in Figure 1. In our experiments, we use B=50 and T=100.

As a final note, we model the user-algorithm feedback loop by feeding \mathcal{R}_{θ} with the updated matrix $\hat{\mathcal{D}}$ at each step t (line 16 of Algorithm 1). Despite other modeling choices can be considered (e.g., retraining the recommender at each step), we here do not investigate the impact of such alternative implementations, this being orthogonal to our study and beyond the scope of this paper.

3.4 Evaluation

Through our simulation framework, we are interested in studying the way user preferences evolve in the long term. We refer to this phenomenon as "algorithmic drift", and we further introduce two novel metrics in order to quantify it.

Algorithmic Drift Score. As aforesaid, it can be expressed as the tendency of the recommendation algorithm to alter user preferences after multiple interactions. In other words, assuming that content in a platform can be tagged as

belonging to a given category, the drift represents the deviation of the initial users preferences towards other kinds of content, after interacting with the recommendation algorithm (\mathcal{R}_{θ}).

To quantify the extent of this phenomenon induced by the recommender system, we define the Algorithmic Drift Score (ADS) over the probabilistic graph G^u of a user u. Such graph-based metric is an adaptation of the Random Walk Controversy Score (RWC), proposed by Garimella et al. [2017] to quantify the radicalization of user opinions in an online social network. In our case, this metric describes whether the user is more inclined to encounter and remain into pathways of items belonging to a given target category \tilde{c} ($i \in I_{\tilde{c}}$), starting from items belonging to a different category ($i \in I_i$, $j \neq t$ target).

Let us consider a user u, and its probabilistic graph G^u . The measure of algorithmic drift $ADS(G^u) \in [-N+1,1]$ can be hence defined as:

$$ADS(G^{u}) = \prod_{i=1}^{N} Pr(I_{\tilde{c}}|I_{i}) - \sum_{\substack{j=1\\j \neq \tilde{c}}}^{N} \prod_{k=1}^{N} Pr(I_{j}|I_{k})$$
(7)

where Pr(X|Y) is the probability that a random walk ends on an element belonging to the set X, given that the walk starts on an element of the set Y. Intuitively, the larger is the value of $ADS(G^u)$, the more likely the user u will consume content belonging to the target category. Vice-versa, the lower value, the higher the probability that user u follows pathways of content from different categories.

Delta Target Consumption. To quantify the change in user consumption of a target category before and after interacting with the recommendations, we introduce the *Delta Target Consumption (DTC)* rate. Let us consider a user u, their historical set of interactions \mathcal{D}_u , and the set of simulated interactions \hat{I}_u . Then,

$$DTC(u) = \frac{|I_{\tilde{c}} \cap (\hat{I}_u \cup \mathcal{D}_u)|}{|\hat{I}_u \cup \mathcal{D}_u|} - \frac{|I_{\tilde{c}} \cap \mathcal{D}_u|}{|\mathcal{D}_u|}$$
(8)

representing the variation of the relative frequency of harmful items after the interaction summarized in I_u . The larger the value of DTC(u), the more the recommender \mathcal{R}_{θ} is shifting the user's choices towards items belonging to the target category \tilde{c} ($i \in I_{\tilde{c}}$) with the respect to their initial preferences.

Intuitively, a positive value for $ADS(G^u)$ and/or DTC(u) implies that u's preferences have been driven toward more content tagged as \tilde{c} ; vice versa, if $ADS(G^u)$ and/or DTC(u) < 0, u has interacted with less items in $I_{\tilde{c}}$. Finally, a value equal to 0 means that the preferences of u have not been altered in the long term with respect to the target category.

4 Experiments

In this section, we study how the simulation framework and the proposed metrics act when dealing with different behavioral users patterns. Specifically, we aim at empirically addressing the following research questions:

- **RQ1.** Can the proposed methodology (simulation framework and related metrics) effectively quantify drifts in users' leaning due to recommendation algorithms?
- **RO2.** How robust is such an approach to different behavioral users patterns, i.e., resistance, inertia, and randomness?

Motivating Use Case. For the experimental evaluation of our methodology, as a practical use case, we devise a scenario where items are categorized as either harmful or neutral, i.e., $i \in \{I_h, I_n\}$, and users are divided into three sub-populations: non-radicalized, semi-radicalized, and radicalized, according to the percentage of harmful content they exhibit. Items are here intended as harmful in a broad sense, without referring to any specific semantic field. In other words, they can spread to whatsoever kind of content (noxious, explicit, inappropriate), e.g., violence, misinformation, pornography, and hate speech. In this respect, radicalized users can also be considered as prone to a high consumption rate of harmful content.

Therefore, in this setting, the objective is evaluating the algorithmic drift induced over users w.r.t. the *harmful* category, i.e., quantifying the radicalizing pathways encountered by non-radicalized users, after interacting with the algorithm in the long-term. In order to do this, Equation 7 can be easily adapted by fixing the target category as *h*, as follows:

$$ADS(G^u) = Pr(I_h|I_h) \cdot Pr(I_h|I_n) - Pr(I_n|I_n) \cdot Pr(I_n|I_h)$$
(9)

Notably, the ADS score can now assume values in the range [-1, 1], these being the corresponding preference poles of non-radicalized and radicalized users, respectively. Therefore, intuitively, the higher the drift effect, the more the

value will shift toward the opposite pole. In other words, in presence of significant drift, the ADS computed over a non-radicalized user will approach 1, and vice-versa (the ADS computed over a radicalized user will approach -1).

We can similarly tailor the definition of DTC in Equation 8 (target = h) in order to trace the variation of user consumption in terms of harmful content. Intuitively, the higher the rate, the more harmful content the user will consume in the long term.

4.1 Data Generation

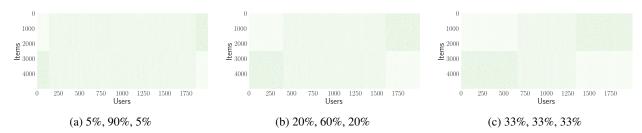


Figure 2: Users' preferences matrix in different samples, varying the proportion of Non-/Semi-/Radicalized users. Assuming that the first (resp. last) $\frac{|I|}{2}$ items are labelled as "neutral" (resp. "harmful"), we impose the Non-Radicalized (resp. Radicalized) users prefer neutral (resp. harmful) items. Conversely, we assume semi-radicalized users span from all items' categories, exhibiting common preferences with the two polarized communities.

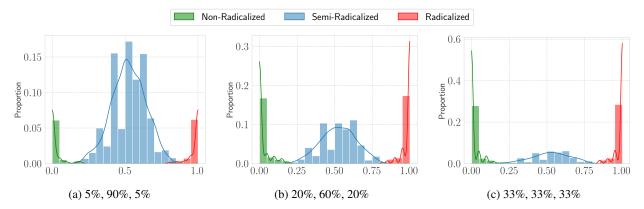


Figure 3: Harmful distribution of Non-/Semi-/Radicalized users, varying the population proportion in the sample. The X-axis represents the harmful percentage in users history, while the Y-axis shows the percentage of users in the dataset.

To validate the effectiveness of our framework, we employ synthetically generated data, which allows us to reproduce different scenarios with great flexibility at low cost [Leskovec et al., 2010, Smith et al., 2017, Chaney et al., 2018]. Specifically, we rely on a synthetic procedure introduced by Coppolillo et al. [2024b], where users and items latent features are sampled from a Dirichlet distribution, while Long-Tail distributions are employed to mimic items' popularity and users' engagement. Finally, the likelihood of an interaction between a user u and an item i is proportional to the dot product of the two latent vectors, combined with popularity of i and engagement of u.

The aforesaid procedure provides great customizability in terms of both data distributions and users/items group partition. First, we impose users and items to follow power-law patterns, with parameters $\alpha=2.2$ and $\alpha=2.0$, respectively. Further, as aforesaid, we split items into two categories: harmful and neutral, respectively, and users into non-radicalized, semi-radicalized, and radicalized, according to the following assumptions: a non-radicalized user shows a percentage of harmful interactions in the range [0,0.2]; a semi-radicalized user, a percentage in the range [0.2,0.8); a radicalized user, a percentage in the range [0.8,1].

Finally, we ensure that the polarized communities (i.e., non-radicalized and radicalized users) share a certain amount of interactions with the semi-radicalized sub-population, by using the parameters setting as suggested by Coppolillo et al. [2024b].

To validate the robustness of our methodology, we generate three data samples exhibiting an increasing portion of *semi-radicalized* users: (i) 5%, 90%, 5%, (ii) 20%, 60%, 20%, and (iii) 33%, 33%, respectively for non-radicalized,

semi-radicalized and radicalized users. In all the samples, items are equally distributed between harmful and neutral. We fixed the number of users to 2000, and $|\mathcal{D}_u| \geq 20$.

The final result in terms of users' preferences and items distribution are depicted in Figures 2 and 3, respectively. Notably, despite the preference sets of non-radicalized and radicalized are mostly disjoint, they indeed share interests with the semi-radicalized group, which is supposed to act as a sort of "bridge" between the two sub-populations. Therefore, the core intuition is that an higher portion of semi-radicalized users in the sample will trigger the collaborative filtering nature of the recommender algorithm, thus increasing the probability of polarized sub-groups of modifying their initial preferences. In other words, we expect that, the higher the portion of semi-radicalized users in the sample, the stronger the "algorithmic drift" effect observed.

4.2 Settings

For the experiments, we adopt RecVAE, a collaborative filtering algorithm based on Variational Auto Encoder (VAE) and developed by Shenbin et al. [2020]. The model has been implemented using the Recbole library [Zhao et al.]. We trained the recommender system for 100 epochs setting a hidden dimension equal to 512. To obtain training, validation, and test sets, the dataset has been split with a ratio of 80/10/10. All the code and data used to perform the experiments are publicly available³.

4.3 Results

Varying population proportion. First, we investigate the impact of the semi-radicalized sub-population in the sample, which (by construction) is supposed to act as a sort of "bridge" between the two polarized communities, by triggering the collaborative filtering nature of the underlying algorithm, and thus fostering the drift effect. To assess if the proposed framework and metrics are able to capture this effect, we fix the resistance (γ) and inertia (δ) parameter across the three sub-populations, and compare the results with the organic model, i.e., the natural evolution of users preferences without the intervention of any recommender system. In particular, we set $\gamma=0.1$ and $\delta=0.5$. Figures 4 and 5 show the results in terms of Algorithmic Drift Score (ADS) and Delta Target Consumption (DTC), respectively.

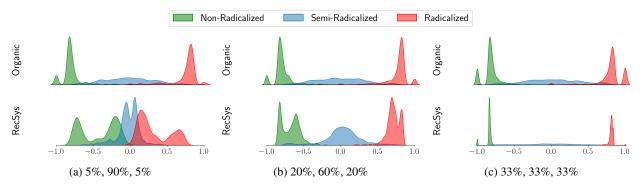


Figure 4: Algorithmic Drift Score (ADS) computed over users graphs by varying the proportion of the starting population (Non-/Semi-/Radicalized %).

As we can see from the plots, while the organic model produces values distributions that are almost constant across the different samples, when employing the recommender system, both our metrics are able to capture the drift induced over the users preferences. Indeed, as expected, when the portion of semi-radicalized users increases, this effect is significantly more prominent, while it is basically absent when the three populations exhibit the same percentage, aligning with the organic distributions.

Notably, the two metrics provides complementary information with respect to the target category: while the DTC rate quantifies the consumption increment in the final user history, the ADS provides a probabilistic perspective computed over the interactions graph.

Impact of *resistance* and *inertia*. Further, we conduct a more fine-grained analysis on the effects of varying the user behavioral factors γ (resistance) and δ (inertia).

As mentioned in Sections 1 and 3.3, we denote the resistance as the user hesitancy in relying on the recommender: the higher, the more prone the user is to autonomously select an item from the catalog; conversely, the inertia models

³https://anonymous.4open.science/r/AlgorithmicDrift-D553

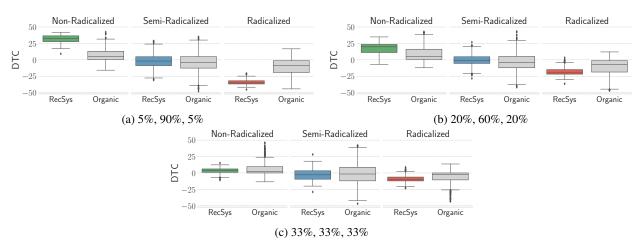
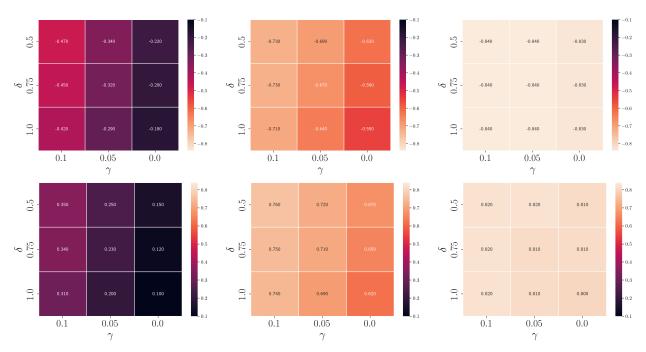


Figure 5: Delta Target Consumption (DTC), expressed in percentage, computed by varying the proportion of the starting population (Non-/Semi-/Radicalized %).

the user tendency in following the recommendations: the higher, the more trust the user shows in the algorithm when choosing the item, ignoring their own preferences.

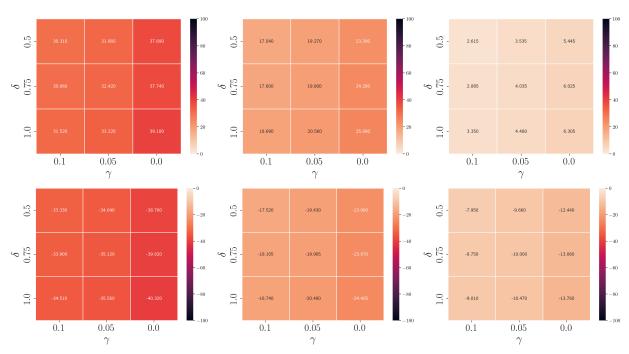


(a) Population proportion 5%, 90%, 5% (b) Population proportion 20%, 60%, 20% (c) Population proportion 33%, 33%, 33%

Figure 6: Algorithmic Drift Score (ADS) induced by the recommendation algorithm over non-radicalized (top row) and radicalized (bottom row) by varying the proportion of the starting population.

Following from these assumptions, we indeed expect that when γ decreases (low resistance), and δ increases (high inertia), the drift effect is more prominent; vice-versa, when γ is high or δ is low, the deviations of users preferences in the long run are negligible.

For the experiments, we vary δ in the range [0.5, 0.75, 1.0], and γ in the range [0.0, 0.05, 0.1]. The results are depicted in Figures 6 and 7, in terms of ADS and DTC, respectively. In each figure, the heatmaps in the top-row show the median value of the corresponding metric computed over non-radicalized users, while the heatmaps in the bottom-row refer to the radicalized population. Further, each column indicates a different proportion of the original sample, in terms of non-, semi-, and radicalized users.



(a) Population proportion 5%, 90%, 5% (b) Population proportion 20%, 60%, 20% (c) Population proportion 33%, 33%, 33%

Figure 7: Delta Target Consumption (DTC), in percentage, computed over non-radicalized (top row) and radicalized (bottom row) users by varying the population proportion, the resistance (γ) and inertia (δ) parameters.

Two considerations can be made on the results. First, as in the previous set of experiments, the portion of semi-radicalized users has great impact on the final results: the higher the population, the darker the grid, i.e., the more prominent is the deviation of the initial users preferences. Secondly, fixed a sample proportion, the devised drift effect is increasingly evident when going from $\delta=0.5, \gamma=0.1$ (top-left) to $\delta=1.0, \gamma=0.0$ (bottom-right), thus perfectly reflecting our intuition.

Increasing choice randomness. Finally, we aim at evaluating the impact of the random parameter η in the user selections, resembling exogenous factors like a friend's suggestion or a misclick. Since we assume random factors being very low in practice, for this set of experiments we set η spanning in the range [0.01, 0.03, 0.05, 0.1]. We further fix $\gamma=0.1$ and $\delta=1.0$, and the population proportion to be equal 20%, 60%, 20%. Intuitively, the spurious interactions introduced by means of randomness should not alter the user preferences in the long term, who will indeed follow their own intrinsic preferences. Figure 8 shows the experimental results in terms of ADS (top-row) and DTC (bottom-row), by varying the η parameter, and comparing to the same setting with no randomness ($\eta=0$), as well as to the organic model.

Two considerations can be here made, regarding the different effects captured by the two metrics. Indeed, while DTC devises a slight increment of harmful (resp., neutral) consumption by non-radicalized (resp., radicalized) users, this is due to the fact that, the higher η , the higher the probability the user picks an item belonging to the opposite category w.r.t. their initial history. In other words, if the user belongs to the non-radicalized community, the majority of their initial interactions consists of neutral items, thus increasing the probability of (randomly) selecting content tagged as harmful (and vice-versa, given a radicalized user). Remind in fact that, in each of the B independent rounds, we assume an user cannot interact with the same item twice.

This, however, does not necessarily imply a drift in their natural preferences, i.e., an higher probability in encountering and remaining in harmful (resp., neutral) pathways: indeed, the users distributions computed in terms of ADS are not affected by increasing the η parameter, thus showing no alteration in terms of users preferences in the long term, as we expected.

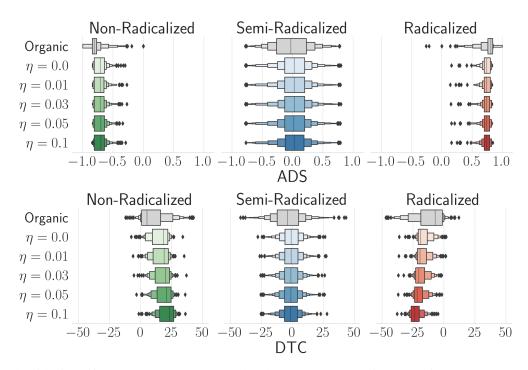


Figure 8: Algorithmic Drift Score (ADS) (top-row) and Delta Target Consumption (DTC, in percentage) (bottom-row) computed varying the random factor η , comparing with the organic model.

5 Conclusions and Future Work

In this paper, we proposed a novel stochastic model for studying potential deviations of users' preferences due to the influence of recommendation systems in the long term. We denote this phenomenon as "algorithmic drift", and we introduce two novel metrics, namely Algorithmic Drift Score and Delta Target Consumption, in order to quantify it. Further, our framework provides great flexibility in representing user behaviors throughout the simulation process, by modeling behavioral patterns such as user resistance to recommendations, inertia in following the provided suggestions, and choice randomness due to exogenous hence uncontrollable factors.

Our main contributions can be indeed summarized as follows: (i) the definition of the *algorithmic drift* concept and the introduction of two novel metrics in order to quantify it; (ii) the implementation of a stochastic model for analyzing the impact of recommender systems in the long term; and (iii) an extensive evaluation through a practical use-case based on a collaborative-filtering algorithm, showing the model's capabilities across different scenarios. The ultimate result is a robust controlled environment for evaluating the recommendation algorithm before deployment.

The proposed framework is amenable for further extensions in many different directions. First, we assume the items catalog to be fixed, while a more dynamic setting could be considered where novel items are continuously introduced. Also, the proposed user model does not take into account contextualization. In the depicted use case, items are categorized as either harmful or neutral. However, more realistic scenarios can embrace situations where items are tagged as harmful depending on the user features or the recommendation context. Moreover, radicalization and harmfulness can also be considered according to specific ideological axes upon which users and items can be aligned. A final line of further investigation is the adaptation of the proposed methodology to study other typical weaknesses that can occur in a recommendation setting, such as popularity bias and/or diversity and serendipity.

Acknowledgements

This work was partially supported by: (i) SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU; (ii) MUR on D.M. 351/2022, PNRR Ricerca, CUP H23C22000440007, and (iii) MUR on D.M. 352/2022, PNRR Ricerca, CUP H23C22000550005.

References

- Aman Agarwal, Xuanhui Wang, Cheng Li, Mike Bendersky, and Marc Najork. Addressing trust bias for unbiased learning-to-rank. In *Proceedings of the 2019 World Wide Web Conference*, pages 4–14, 2019.
- Mahsa Badami, Olfa Nasraoui, and Patrick Shafto. Prcp: Pre-recommendation counter-polarization. In *Proceedings* of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2018, Volume 1: KDIR, Seville, Spain, September 18-20, 2018, pages 280–287, 2018.
- Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.
- Dimitrios Bountouridis, Jaron Harambam, Mykola Makhortykh, Mónica Marrero, Nava Tintarev, and Claudia Hauff. Siren: A simulation framework for understanding the effects of recommender systems in online news environments. FAT* '19. Association for Computing Machinery, 2019. doi:10.1145/3287560.3287583.
- Allison June-Barlow Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, pages 224–232, 2018.
- Serina Chang and Johan Ugander. To recommend or not? A model-based comparison of item-matching processes. In *Proceedings of the Sixteenth International AAAI Conference on Web and Social Media, ICWSM 2022, Atlanta, Georgia, USA, June 6-9, 2022*, pages 55–66, 2022.
- Jaeho Cho, Saifuddin Ahmed, Martin Hilbert, Billy Liu, and Jonathan Luu. Do search algorithms endanger democracy? an experimental investigation of algorithm effects on political polarization. *Journal of Broadcasting & Electronic Media*, 64(2):150–172, 2020.
- Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. The echo chamber effect on social media. *Proc. Natl. Acad. Sci. USA*, 118(9):e2023301118, 2021.
- Federico Cinus, Marco Minici, Corrado Monti, and Francesco Bonchi. The effect of people recommenders on echo chambers and polarization. In *Proceedings of the Sixteenth International AAAI Conference on Web and Social Media, ICWSM 2022, Atlanta, Georgia, USA, June 6-9, 2022*, pages 90–101, 2022. URL https://ojs.aaai.org/index.php/ICWSM/article/view/19275.
- Erica Coppolillo, Giuseppe Manco, and Aristides Gionis. Relevance meets diversity: A user-centric framework for knowledge exploration through recommendations. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 490–501, New York, NY, USA, 2024a. Association for Computing Machinery. ISBN 9798400704901. doi:10.1145/3637528.3671949. URL https://doi.org/10.1145/3637528.3671949.
- Erica Coppolillo, Simone Mungari, Ettore Ritacco, and Giuseppe Manco. Genrec: A flexible data generator for recommendations. *CoRR*, abs/2407.16594, 2024b. doi:10.48550/ARXIV.2407.16594.
- Pranav Dandekar, Ashish Goel, and David T. Lee. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences*, 110(15):5791–5796, 2013.
- Henrique Ferraz de Arruda, Felipe Maciel Cardoso, Guilherme Ferraz de Arruda, Alexis R. Hernandez, et al. Modeling how social network algorithms can influence opinion polarization. *CoRR*, 2021. URL https://arxiv.org/abs/2102.00099.
- Gianmarco De Francisci Morales, Corrado Monti, and Michele Starnini. No echo in the chambers of political interactions on reddit. *Scientific reports*, 11(1):1–12, 2021.
- Morris H. Degroot. Reaching a consensus. Journal of the American Statistical Association, 69(345):118–121, 1974.
- Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, et al. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559, 2016.
- Jean-Yves Duclos, Joan Esteban, and Debraj Ray. Polarization: Concepts, measurement, estimation. *Econometrica*, 72 (6):1737–1772, 2004.
- Francesco Fabbri, Francesco Bonchi, Ludovico Boratto, and Carlos Castillo. The effect of homophily on disparate visibility of minorities in people recommender systems. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, pages 165–175, 2020. URL https://ojs.aaai.org/index.php/ICWSM/article/view/7288.
- Francesco Fabbri, Maria Luisa Croci, Francesco Bonchi, and Carlos Castillo. Exposure inequality in people recommender systems: The long-term effects. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 194–204, 2022.

- Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Quantifying controversy in social media, 2017.
- Muhammad Haroon, Anshuman Chhabra, Xin Liu, Prasant Mohapatra, Zubair Shafiq, and Magdalena Wojcieszak. Youtube, the great radicalizer? auditing and mitigating ideological biases in youtube recommendations. *arXiv* preprint arXiv:2203.10666, 2022.
- Jure Leskovec, Deepayan Chakrabarti, Jon Kleinberg, Christos Faloutsos, and Zoubin Ghahramani. Kronecker graphs: An approach to modeling networks. *J. Mach. Learn. Res.*, 2010.
- Marco Minici, Federico Cinus, Corrado Monti, Francesco Bonchi, and Giuseppe Manco. Cascade-based echo chamber detection, 2022. URL https://arxiv.org/abs/2208.04620.
- E Pariser. The filter bubble: How the new personalized web is changing what we read and how we think. The Penguin Press, 2011.
- Shruti Phadke, Mattia Samory, and Tanushree Mitra. Pathways through conspiracy: The evolution of conspiracy radicalization through engagement in online conspiracy discussions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 770–781, 2022.
- Pedro Ramaciotti Morales and Jean-Philippe Cointet. Auditing the effect of social network recommendations on polarization in geometrical ideological spaces. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, page 627–632. Association for Computing Machinery, 2021.
- Bashir Rastegarpanah, Krishna P. Gummadi, and Mark Crovella. Fighting fire with fire: Using antidote data to improve polarization and fairness of recommender systems. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 231–239. Association for Computing Machinery, 2019.
- Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. Auditing radicalization pathways on youtube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 131–141, 2020.
- Fernando P. Santos, Yphtach Lelkes, and Simon A. Levin. Link recommendation algorithms and dynamics of polarization in online social networks. *Proc. Natl. Acad. Sci. USA*, 118(50):e2102141118, 2021.
- Ilya Shenbin, Anton Alekseev, Elena Tutubalina, Valentin Malykh, and Sergey I Nikolenko. Recvae: A new variational autoencoder for top-n recommendations with implicit feedback. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 528–536, 2020.
- Matthew Smith, Laurent Charlin, and Joelle Pineau. A sparse probabilistic model of user preference data. In *Advances in Artificial Intelligence*, 2017.
- Cass R Sunstein. The law of group polarization. *University of Chicago Law School, John M. Olin Law & Economics Working Paper*, (91), 1999.
- Cass R Sunstein. # Republic: Divided democracy in the age of social media. Princeton University Press, 2018.
- Antonela Tommasel and Filippo Menczer. Do recommender systems make social media more susceptible to misinformation spreaders? In *RecSys* '22: Sixteenth ACM Conference on Recommender Systems, Seattle, WA, USA, September 18 23, 2022, pages 550–555, 2022.
- Sirui Yao, Yoni Halpern, Nithum Thain, Xuezhi Wang, Kang Lee, Flavien Prost, Ed H. Chi, Jilin Chen, and Alex Beutel. Measuring recommender system effects with simulated users, 2021. URL https://arxiv.org/abs/2101.04526.
- Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, et al. Recbole: Towards a unified, comprehensive and efficient framework for recommendation algorithms. In CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 5, 2021, pages 4653–4664.