Introducing a Novel Framework for Recognizing Social Media Recommenders Under Absent Recommendations and a First Graph Neural Network-Based Implementation

Sabrina Guidotti
University of Milan-Bicocca
Milan, Italy
s.guidotti2@campus.unimib.it

Gregor Donabauer
University of Regensburg
Regensburg, Germany
Gregor.Donabauer@ur.de

Davide Taibi
CNR
Palermo, Italy
davide.taibi@itd.cnr.it

Giuseppe Vizzari
University of Milan-Bicocca
Milan, Italy
giuseppe.vizzari@unimib.it

Udo Kruschwitz
University of Regensburg
Regensburg, Germany
Udo.Kruschwitz@ur.de

Dimitri Ognibene
University of Milan-Bicocca
Milan, Italy
dimitri.ognibene@unimib.it

ABSTRACT

Recommender algorithms shape public discourse on social media, often intensifying polarization and accelerating misinformation. Their opacity and integration within social networks make assessing their true impact exceptionally difficult.

To make these systems more explainable, we introduce the concept of Social Media Recommenders Recognition under Absent Recommendations (SM-ARR), which accounts for the complexity induced by platforms not publicly releasing interaction logs and algorithmic details, which hinders detecting specific misbehaviors, e.g., dark patterns. We also present a proof-of-concept implementation, SM-ARR-G (Social Media Automatic Recommender Recognition through Graph Neural Networks), a Graph Neural Network (GNN) based framework designed to identify which recommendation algorithm is shaping interactions during a given period, without requiring access to the platform's internal recommendation logs. SM-ARR-G learns to forecast user actions by combining their past behavior with candidate "infospheres", simulated content exposure patterns generated by alternative recommender models. The candidate that produces the most accurate generalization performance is taken as the best explanation of the recommender shaping the observed dynamics.

Our evaluation draws on the DBLP-Citation-Network V14 dataset, chosen for its scale and structural richness as a proxy for social media data. While not a typical social media graph, it offers many of the same relational patterns, dense connectivity, influence dynamics, and recommendation-like exposure. By embedding multiple recommender algorithms into this environment, we create controlled yet varied scenarios. This enables us to assess how consistently SM-ARR-G can identify the hidden recommender across different conditions, without relying on access to proprietary platform data.

Our experiments demonstrate that our SM-ARR-G approach is promising and can reliably detect the hidden recommender while highlighting how alternative algorithms alter user interaction patterns. While further improvements are necessary (e.g., handling unknown or evolving recommenders), we expect the framework to complement existing audit strategies, broaden the possibilities for recommender assessment, and improve explainability of social media dynamics.

Complete code is available at https://github.com/DimNeuroLab/academic_network_project.

KEYWORDS

Social Media, Recommender Systems, Graph Neural Networks, Algorithm Auditing, Transparency, User Behavior Modeling

1 INTRODUCTION

The ethical and societal implications of social media recommender systems have drawn considerable attention. Research indicates that engagement-focused algorithms can inadvertently spread misinformation [31] and amplify polarization through filter bubbles, echo chambers, and backfire effects [2, 3, 9, 19]. Additionally, studies highlight a paradox: engagement-oriented algorithms increase platform usage but decrease user satisfaction and well-being [1]. Collectively, these findings indicate that social media algorithms and their specific parametrization can simultaneously boost usage and degrade user welfare, potentially intensifying polarization and societal tensions. While some experimental research has nuanced the 'personalization-polarization' hypothesis, suggesting that algorithmic effects on political polarization and misinformation exposure may be significantly more limited than previously assumed [11, 20], recent contributions bring back the important role played by social media and their algorithms in affective polarization and attitudinal changes [8, 23, 29]. Regardless of the specific causal link, the public perception that 'black-box' algorithms manipulate discourse remains a powerful driver for accountability. This underscores the urgent need for transparency and accountability in recommender algorithms [21, 22].

Various auditing techniques have been developed to examine these systems' biases and societal impacts. Black-box auditing, which analyzes user interactions without internal system access, has revealed biases that can amplify polarizing content [24], particularly on platforms like YouTube [30, 32]. Fairness-focused audits have also shown that recommendation systems may disproportionately impact certain groups, often perpetuating societal biases [26]. These insights have fueled calls for transparency mechanisms to better understand and control algorithmic influence [4].

Efforts toward explainable algorithms aim to clarify content recommendations and support user trust [5]. Automated auditing tools,

leveraging AI, are being explored to regularly monitor and detect biases in recommendation algorithms [27], though this approach raises ethical concerns of its own [18].

Regulatory frameworks, such as the General Data Protection Regulation (GDPR), enforce transparency and uphold users' "right to explanation", pushing for accountability in algorithmic processes [10]. Simulation studies also help illuminate recommendation impacts, demonstrating how repeated interactions can amplify biases and influence long-term user behavior [7, 15, 30]. While there are previous studies that use simulations to understand the impact of recommender systems on social media, in this paper we attempt to reverse the process and propose a simulation-based, likelihood-driven framework to recognize the recommender underlying the observed user behaviors.

Despite the availability of real-world social media datasets (e.g., Facebook, Twitter, Twitch [6, 17, 25]), most lack critical temporal information, recommendation metadata, and user–user connection dynamics. In contrast, many large-scale academic network datasets are periodically updated and expose richer interaction histories, enabling more robust evaluation and prediction of recommendation system impact.

In summary, current research underscores the importance of accessible, fair, and explainable recommendation algorithms, alongside regulatory measures to uphold ethical standards. Our work takes an initial step in this direction by applying Graph Neural Networks to academic networks as a proxy for identifying recommender algorithms active on social media. To do that, we aim to complement traditional recommender audits (which could only work on newly collected data, for a short time, and with a limited number of users) by indirectly inferring the characteristics of hidden recommendation systems through user behavior modeling. Rather than directly auditing these systems, which often require black-box or white-box access to the recommender algorithm, our approach tentatively estimates the influence of different hypothetical recommenders on user interactions by analyzing variations in model loss under each hypothesis. Despite limitations of our approach, such as computational complexity, we offer an initial strategy to approximate the likelihood of the recommender model active on a social media platform, even when direct auditing is infeasible.

2 METHODOLOGY

Problem statement: Social Media Recommenders Recognition under Absent Recommendations (SM-ARR). We formalize *SM-ARR* as the task of inferring, from observed user interactions alone and *without* access to platform recommendation logs, which candidate recommender R most plausibly generated the dynamics. Concretely, for each hypothesized R we condition a user–behavior predictor on the corresponding simulated *infosphere* and evaluate out-of-sample negative log-likelihood; the R yielding the lowest test loss is selected as the recognized recommender. This decision rule defines the recognition problem independently of any particular model family or training recipe. In what follows we instantiate this problem with a proof-of-concept implementation, **SM-ARR-G**, a GNN-based framework operationalizing the above criterion.

2.1 SM-ARR-G: a GNN-based proof-of-concept implementation.

Detecting which recommender system drives user behavior on social media, without access to the platform's proprietary recommendation logs, is challenging. Guidotti *et al.* [12] showed that injecting a *simulated* recommender into a GNN-based user-behavior predictor markedly changes its test loss, thereby revealing how well that recommender explains the data. Motivated by this insight, we assert that, among a set of candidate recommenders, the one yielding the *lowest* test loss (i.e., the best generalization performance) is the most plausible explanation of the observed dynamics.

Likelihood formulation. Let $D = \{(x_i, y_i)\}_{i=1}^{N}$ be the sequences of past actions x_i and corresponding next actions y_i for N users, and let R denote a hypothesized recommender algorithm. Training a GNN with parameters θ under hypothesis R yields the test loss

$$\mathcal{L}(\theta; D, R) \;=\; -\sum_{i=1}^{N} \log P_\theta\big(y_i \,\big|\, x_i, R\big), \tag{1}$$

which is precisely the negative log-likelihood (NLL) of the observed interactions under R. Comparing $\mathcal{L}$ across candidate recommenders therefore amounts to a likelihood-based model-selection procedure that identifies the recommender most consistent with the data.

Eq. (1) is the NLL used as our empirical objective. We use it both for training and for *model selection across recommenders*; i.e., $R^\star = \arg\min_R \mathcal{L}(\hat{\theta}_R; D_{\text{test}}, R)$, where $\hat{\theta}_R$ are parameters trained under hypothesis R.

Notation.

- $D = \{(x_i, y_i)\}_{i=1}^{N}$ represents the dataset of N observed user interactions.
- $x_i = (u_i, v_i)$ denotes a pair of users, where u_i and v_i are engaging on the platform (e.g., exposed to each other's content or work and potentially collaborating).
- y_i denotes the observed outcome of this interaction (e.g., whether u_i co-authors a paper with v_i).
- θ represents the parameters of the user behavior model to be learned.
- R denotes a hypothesis about the recommender system's behavior (e.g., the algorithm or ranking strategy it employs).

Data availability challenge. Testing our method would require datasets that contain both real user interactions and the platform's recommender logic. However such datasets are rare. We therefore *simulate* social-media interactions under several known recommenders, but, unlike traditional simulations that use simplistic, uniform user rules, we first *learn* a realistic user model from data. In particular, to ameliorate the bias introduced by the (unknown) recommender we aim to uncover, we use the method presented in [12] that introduces several paths in the simulated infosphere connected to the actual future users actions (see *sorted-hidsight* results).

Workflow. In summary, our approach follows these steps:

(1) **Learn a recommender-neutral user model** (RNU) from historical data.
(2) **Generate synthetic datasets** by pairing the RNU with each simulated recommender.

(3) **Train user models** on every synthetic dataset under competing recommender hypotheses and compare their test losses to identify the true generator.

This pipeline operationalizes the hypothesis that the likelihood-based criterion of Eq. (1) indeed pinpoints the recommender system that governs observed social-media dynamics.

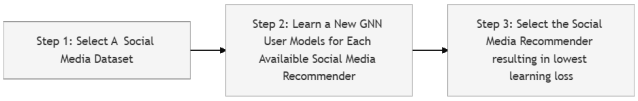

Figure 1: Schema to select the most likely recommender system for a recommender model

2.2 Training of a Recommender-Neutral User Model

In this section, we present our approach for creating a *recommender-neutral user model* (RNU) for social media interactions. This model is crucial, as observed user behavior on social media is influenced by an unknown recommender system. Our goal is to develop a model of user behavior that minimizes dependence on the hidden recommender and better reflects the user's intrinsic preferences, interactions, and information state.

We aim to extract the probability model of observed interactions y between agents U and V, given their simulated recommendations (or "infosphere") r_u and r_v and their observed histories or states s_u and s_v:

$$P(y(u,v) \mid r_u, r_v, s_u, s_v).$$

2.2.1 Marginalizing Over the Hidden Recommender. To account for the unknown influence of the hidden recommender on user actions, we propose to marginalize over potential recommender behaviors. Since we lack explicit recommendations and a conditional model of user behavior given these recommendations, we employ a parameterized user model. This model estimates the probability of a user u interacting with content y based on their interaction history and hypothetical recommendations r_u and r_v that might have been received. To account for the unknown influence of the hidden recommender on user actions, we employ a latent-variable marginal likelihood. We fit the parameters θ of the user model from logs where recommendation exposures are absent by maximizing the log-likelihood of observed actions:

$$\sum_i \log P_\theta(y_i \mid s_{u,i}, s_{v,i}) \tag{2}$$

Formally, we define the probability of an interaction y given only the observed states (s_u, s_v) by marginalizing over the latent recommendations (r_u, r_v) and the hypothesized recommender R:

$$P_\theta(y \mid s_u, s_v) = \int P(R) \left[\iint P_\theta(y \mid s_u, s_v, r_u, r_v) P(r_u, r_v \mid R, s_u, s_v)\, dr_u\, dr_v \right] dR \tag{3}$$

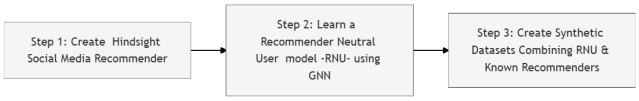

Figure 2: Schema to Generate Synthetic Social Media Datasets with Selected Recommenders

In this expression: Here, $P_\theta(y \mid s_u, s_v, r_u, r_v)$ is the parameterized user model; $P(r_u, r_v \mid R, s_u, s_v)$ denotes the probability of recommendation r given hypothesis R; and $P(R)$ represents the prior over plausible recommenders.

Since the true recommender is unknown, we assume a uniform prior over a range of plausible recommenders R. However, marginalizing over all possible recommenders remains computationally infeasible due to the combinatorial complexity of algorithms and configurations. Moreover, simulating interactions across numerous user states and recommendations compounds the computational cost. Consequently, we explore alternative methods to approximate this marginalization, enabling us to develop a user model that is less dependent on the hidden recommender's influence.

2.2.2 Hindsight Model of the Recommender. As an alternative approach, we propose using a *hindsight predictive model* of the recommender system when training the user model. This method considers the user's actual future behavior to generate recommendations and avoids the need for simulating multiple recommenders. The approach closely follows that proposed in [12].

This results in several advantages: First of all, the method reduces the dependency on unknown recommenders, as it incorporates both actual interactions and noisy recommendations (which we also have full control of). In addition, it reduces the computational complexity by using the actual future behavior.

3 EXPERIMENTS

Metrics. We evaluate recognition using: **NLL** on a held-out test split (column "Loss" in tables; lower is better), and **edge classification accuracy** on the same split (column "Accuracy"; higher is better). We report accuracy strictly as a supplementary diagnostic, excluding it from both model selection and primary evaluation.

3.1 Datasets

As mentioned in the beginning, the few real-world social media datasets that are available come with significant limitations, such as a restricted scope or outdated information. Because of that, we use the DBLP-Citation-network v14 dataset [28] from AMiner as a proxy for a social network. This dataset integrates data from sources such as DBLP, ACM, and MAG to provide a comprehensive view of academic publications and their citation relationships. While not a typical social media graph, academic citation networks share many of the same relational patterns, dense connectivity, influence dynamics, and recommendation-like exposure.

We chose this specific dataset because it is the most recent release (2023) and offers a balanced scale of nodes and edges, making it well-suited for our experiments. Specifically, it includes 5,259,858 paper nodes and 36,630,661 citation edges. Each paper contains

additional metadata, such as authors, venues, and topics, which can be represented as additional node types in a heterogeneous social network.

3.2 Synthetic Data Generation

We generate synthetic user interaction data by simulating user behavior under specific, known recommender systems defining the users' infospheres.

3.2.1 Infospheres Descriptions. We evaluate various types of infospheres to assess how accurately these underlying simulated recommendations can be predicted in our experimental setting.

(1) **No Infosphere:** In this case, no recommender system is applied. The results reflect the baseline scenario where only the real network data is considered, without any enhancements or expansions of the co-author network.
(2) **Hindsight Infosphere:** We adopt this infosphere definition from prior research [12]. It is based on a seedgraph which is a directed graph composed of paths associated with each author. Each of the paths traces the shortest connection from an element in the author's history in year $y+1$ back to the graph in year y. This structure is enhanced by adding alternative branches as noise to improve realism. For a more detailed description we refer to [12], the paper we adopted this infosphere type from.
(3) **Top Paper:** This infosphere includes the n most popular papers until year y, with different values of n.
(4) **Top Paper * Topic:** This variant selects top papers both globally and within the m most-used topics by each author up to year y (e.g., [n=5, m=2]).
(5) **LightGCN:** We used RecBole with the LightGCN model [13], a state of the art collaborative filtering recommender based on optimized GNNs to learn embedding of both users and items.
(6) **NAIS (used in selected experiments):** We employed the Neural Attentive Item Similarity model (NAIS) [14] as an additional baseline, a neural collaborative filtering approach that extends item-based recommendation with attention mechanisms.

3.2.2 Data Generation. Once the RNU and the recommender are in place, generating interactions may seem straightforward. However, this step involves significant computational complexity due to the large scale of social media networks.

To generate the number of co-authors in year $y+1$, we used the actual co-author count labels as input. This choice allowed us to better control the complexity of the simulated interactions while ensuring consistency with observed collaboration patterns. By anchoring the generation process to real-world values, we maintained a high degree of realism across different infospheres.

For each infosphere, we compute the synthetic ground truth by simulating a sparse and symmetric connectivity matrix among authors, based on their historical and contextual information. The generation process accounts for personalized upper bounds on node degree, enforces symmetric co-authorship links, and leverages learned pairwise connection probabilities derived from author embeddings. The detailed algorithm used to perform this simulation is available in the shared repository.

For timing reasons, the computation of this synthetic ground truth was performed only on authors who will publish papers in year $y+1$ (in this case 2020).

3.3 Applying the Recognition Method

Using the generated synthetic interaction datasets $D(R)$, we apply the approach presented in [12] training from scratch a new user model GNN $UM(D(R), R')$ for each recommender R'. This allows to compare the likelihood for models learned under matching hypotheses, i.e. $L(\theta; D, UM(D(\hat{R}), \hat{R}))$, with that of models learned with mismatching hypotheses $L(\theta; D, UM(D(R''), R'))$ with $R' \neq R''$.

The GNN model employed is based on Heterogeneous Graph Transformers (HGT) [16], which leverage Transformer layers to process heterogeneous graph data. The architecture includes two subsequent HGT layers followed by a fully connected layer, which is used to predict the existence of an interaction between a pair of nodes. Each node type has an associated initial linear transformation, followed by HGT-based message passing, batch normalization, and dropout regularization (set to 0.3) to enhance model stability. Additionally, embeddings are used for node types without features, such as authors and topics.

For these experiments, the authors were randomly divided into five distinct folds, using four of them for training and the fifth for evaluation. The same test fold is consistently used across all experiments to ensure comparability. Training relies on binary cross-entropy loss applied to a link prediction task over author co-authorship. To prevent overfitting, early stopping is applied with a patience of five epochs, and the best model according to the validation loss is saved.

We optimize using Adam with a OneCycleLR scheduler (cosine annealing). Gradients are clipped to 2.0, and batch size is 1024. Input graphs are treated as undirected, employing random negative sampling to balance the classification task.

The implementation code is provided in the shared repository.

The reported results refer to the best validation model obtained before early stopping.

4 RESULTS

To evaluate the robustness of our model in inferring the characteristics of different recommenders, we generated four synthetic datasets, each corresponding to a different recommendation strategy. Importantly, three distinct methodologies were employed to generate the ground truth used during training, each designed to reflect varying levels of alignment between the training data and the underlying recommender logic.

In the first methodology, the ground truth was created using embeddings derived from the hindsight infosphere (i.e., hindsight embeddings). Connections were then sampled randomly, without enforcing any ranking based on likelihood. The hindsight infosphere recommender benefits from access to information not typically available at simulation time, and thus, while its performance is superior, it cannot be considered a viable real-world recommender.

In the second methodology, connections were not sampled randomly but instead ranked by decreasing probability based on the

hindsight infosphere. The top-ranked connections were then selected. This further amplified the advantage of the hindsight infosphere by ensuring that only highly likely connections, according to the hindsight model, were used.

Finally, in the third methodology, we employed embeddings from the recommender-specific infosphere. These were obtained by keeping the hindsight infosphere's weights but swapping in the target infosphere at the final step. The connections were again ranked by decreasing probability, and the most probable were chosen. This methodology introduced a shift that, while still leveraging the strong modeling of the hindsight embeddings, resulted in a reduction in performance for the hindsight infosphere model.

We report in the following section the accuracies and losses of the models trained on these datasets, each under different recommender assumptions. *In italic the generating infosphere*, and in **bold** the lowest loss and highest accuracy.

4.1 Hindsight Infosphere Embeddings + Random Sampling

Infosphere	Params	Accuracy	Loss
HINDSIGHT INFOSPHERE	**5**	**0.9905**	**0.0426**
NO INFOSPHERE	N/A	0.8494	0.3109
TOP PAPER	10	0.5579	0.6924
TOP PAPER	50	0.8249	0.3619
TOP PAPER * TOPIC	[5,2]	0.8366	0.3589
TOP PAPER * TOPIC	[5,10]	0.8519	0.3343
LightGCN	N/A	0.8661	0.2913

Table 1: Accuracy and loss for *Hindsight Infosphere* ground truth (hindsight embeddings + random sampling).

Infosphere	Params	Accuracy	Loss
HINDSIGHT INFOSPHERE	5	0.8295	0.3872
NO INFOSPHERE	N/A	0.7329	0.4769
TOP PAPER	10	0.6529	0.5446
TOP PAPER	50	0.6636	0.5457
TOP PAPER * TOPIC	***[5,2]***	***0.8836***	***0.2868***
TOP PAPER * TOPIC	[5,10]	0.8168	0.3446
LightGCN	N/A	0.6821	0.5169

Table 2: Accuracy and loss for *Top Paper* × *Topic [5,2]* ground truth (hindsight embeddings + random sampling).

In the case of the synthetic dataset generated with the Top Paper recommender reported in table 3, the results are influenced by the fact that the generated ground truth is mainly composed of the history of authors who will publish in year y+1. This is due to a limitation on the maximum number of co-authors, which in this scenario matches the actual number of co-authors. As a result, the hindsight model has an advantage, as it has knowledge of which authors will publish. Since the real value is the actual number of co-authors, the top-paper selection in the infosphere tends to be less effective, as it limits the number of collaborations an author with high citation counts can have relative to their citation impact. If a model focused on citation-based predictions had been attempted, the effects would likely differ.

Infosphere	Params	Accuracy	Loss
HINDSIGHT INFOSPHERE	**5**	**0.9315**	**0.1602**
NO INFOSPHERE	N/A	0.8295	0.3471
TOP PAPER	**10**	**0.8301**	**0.3534**
TOP PAPER	50	0.8073	0.4041
TOP PAPER * TOPIC	[5,2]	0.7893	0.4223
TOP PAPER * TOPIC	[5,10]	0.8285	0.3396
LightGCN	N/A	0.8419	0.3170

Table 3: Accuracy and loss for *Top Paper (10)* ground truth (hindsight embeddings + random sampling).

Infosphere	Params	Accuracy	Loss
HINDSIGHT INFOSPHERE	5	0.9750	0.7337
NO INFOSPHERE	N/A	0.9297	0.1923
TOP PAPER	10	0.8983	0.2652
TOP PAPER	50	0.9149	0.2181
TOP PAPER * TOPIC	[5,2]	0.8909	0.2137
TOP PAPER * TOPIC	[5,10]	0.9164	0.2557
LightGCN	***0.9971***	***N/A***	***0.0134***

Table 4: Accuracy and loss for *LightGCN* ground truth (hindsight embeddings + random sampling).

The results in Table 1 show that the model trained with the hindsight infosphere ground truth achieves near-perfect performance, substantially outperforming all other candidate recommenders, clearly due to the advantages that this infosphere presents. Table 2 shows that when the dataset is generated with the Top Paper × Topic [5,2] logic, the corresponding infosphere yields the highest accuracy and lowest loss, with similarly strong results also for the hindsight one. In the Top Paper (10) scenario (Table 3), the hindsight model again holds a clear advantage, indicating its privileged access to future publication information, which is also due to the limitation previously described regarding the composition of this infosphere. Finally, the LightGCN-specific dataset (Table 4) confirms that the LightGCN-based ground truth is best recapitulated by the LightGCN model itself. The markedly higher accuracy observed for the LightGCN-specific dataset stems from the inherent differences between recommender families. Heuristic, popularity-based recommenders (like Top Paper) tend to suggest the same items to many users, which quickly saturates the maximum co-authorship slots enforced by our simulation constraints. This leads to lower distinctiveness in the synthetic ground truth. In contrast, embedding-based models like LightGCN distribute recommendations more diversely across the graph, providing a stronger signal for the GNN to learn and subsequently recognize the underlying generator.

Infosphere	Params	Accuracy	Loss
HINDSIGHT INFOSPHERE	***5***	***0.9824***	***0.0654***
NO INFOSPHERE	N/A	0.5443	0.7218
TOP PAPER	10	0.8056	0.4284
TOP PAPER	50	0.7993	0.4206
TOP PAPER * TOPIC	[5,2]	0.8594	0.3948
TOP PAPER * TOPIC	[5,10]	0.7523	0.5320
LightGCN	N/A	0.8681	0.3916

Table 5: Accuracy and loss for *Hindsight Infosphere* ground truth (hindsight embeddings + descending-probability sorting).

Infosphere	Params	Accuracy	Loss
HINDSIGHT INFOSPHERE	**5**	**0.9686**	**0.1025**
NO INFOSPHERE	N/A	0.5941	0.6750
TOP PAPER	10	0.8614	0.5151
TOP PAPER	50	0.8027	0.6917
TOP PAPER * TOPIC	**[5,2]**	**0.7492**	**0.5469**
TOP PAPER * TOPIC	[5,10]	0.8430	0.3694
LightGCN	N/A	0.6905	0.5612

Table 6: Accuracy and loss for *Top Paper* × *Topic [5,2]* ground truth (hindsight embeddings + descending-probability sorting).

Infosphere	Params	Accuracy	Loss
HINDSIGHT INFOSPHERE	**5**	**0.9590**	**0.1284**
NO INFOSPHERE	N/A	0.8592	0.3787
TOP PAPER	**10**	**0.8084**	**0.4577**
TOP PAPER	50	0.8728	0.2959
TOP PAPER * TOPIC	[5,2]	0.7678	0.5033
TOP PAPER * TOPIC	[5,10]	0.7553	0.5400
LightGCN	N/A	0.5863	0.5986

Table 7: Accuracy and loss for *Top Paper (10)* ground truth (hindsight embeddings + descending-probability sorting).

Infosphere	Params	Accuracy	Loss
HINDSIGHT INFOSPHERE	**5**	**0.9530**	**0.1402**
NO INFOSPHERE	N/A	0.6401	0.6355
TOP PAPER	10	0.8621	0.3347
TOP PAPER	50	0.9007	0.2830
TOP PAPER * TOPIC	[5,2]	0.8417	0.4370
TOP PAPER * TOPIC	[5,10]	0.9017	0.2870
LightGCN	**N/A**	**0.9517**	**0.1944**

Table 8: Accuracy and loss for *LightGCN* ground truth (hindsight embeddings + descending-probability sorting).

4.2 Hindsight Infosphere Embeddings + Probability-Based Sorting

Under descending-probability sorting using the embeddings of the hindsight infosphere, an additional advantage of the hindsight infosphere consistently emerges. When ground truth follows Top Paper × Topic [5,2] logic (Table 6), the hindsight model continues to dominate, eclipsing the true generation strategy. The Top Paper (10) dataset (Table 7) similarly favors the hindsight model, underscoring the impact of hindsight-derived sorting. Notably, in the LightGCN-sorted scenario (Table 8), the hindsight infosphere model performs on par with LightGCN, indicating that probability-based sampling using just the hindsight embedding can blur the distinction between generation strategies.

4.3 Recommender-Specific Infosphere Embeddings + Probability-Based Sorting

Infosphere	Params	Accuracy	Loss
HINDSIGHT INFOSPHERE	**5**	**0.8972**	**0.2386**
NO INFOSPHERE	N/A	0.8340	0.4013
TOP PAPER	10	0.7651	0.4480
TOP PAPER	50	0.7114	0.5325
TOP PAPER * TOPIC	**[5,2]**	**0.8136**	**0.4965**
TOP PAPER * TOPIC	[5,10]	0.7700	0.5080
LightGCN	N/A	0.7049	0.5613

Table 9: Accuracy and loss for *Top Paper* × *Topic [5,2]* ground truth (specific-infosphere embeddings + descending-probability sorting).

Infosphere	Params	Accuracy	Loss
HINDSIGHT INFOSPHERE	**5**	**0.9247**	**0.2211**
NO INFOSPHERE	N/A	0.8584	0.4140
TOP PAPER	**10**	**0.8715**	**0.3109**
TOP PAPER	50	0.8646	0.3511
TOP PAPER * TOPIC	[5,2]	0.8136	0.4615
TOP PAPER * TOPIC	[5,10]	0.7527	0.5602
LightGCN	N/A	0.8331	0.3710

Table 10: Accuracy and loss for *Top Paper (10)* ground truth (specific-infosphere embeddings + descending-probability sorting).

Infosphere	Params	Accuracy	Loss
HINDSIGHT INFOSPHERE	5	0.9596	0.1410
NO INFOSPHERE	N/A	0.9013	0.3293
TOP PAPER	10	0.9198	0.2729
TOP PAPER	50	0.7658	0.5446
TOP PAPER * TOPIC	[5,2]	0.7601	0.5596
TOP PAPER * TOPIC	[5,10]	0.9024	0.3645
LightGCN	***N/A***	***0.9792***	***0.0449***

Table 11: Accuracy and loss for *LightGCN* ground truth (specific-infosphere embeddings + descending-probability sorting).

For the specific-infosphere embedding experiments, Table 9 shows that even when using recommender-aligned embeddings, the hindsight model attains the highest accuracy and lowest loss on the Top Paper × Topic [5,2] dataset, though the matching infosphere remains competitive. In the Top Paper (10) setting (Table 10), the hindsight infosphere again marginally outperforms the true generator, suggesting residual benefits from hindsight knowledge. Finally, Table 11 confirms that the LightGCN embeddings paired with sorting may lead to a greater advantage for the former compared to the previous case when contrasted with the hindsight one.

4.4 5-Fold Cross Validation

To assess robustness and variance, 5-fold cross-validation was performed across all configurations using the *recommender-specific infosphere embeddings + probability-based sorting* methodology. The *Hindsight* infosphere ground truth, previously omitted because its results coincide with earlier experiments using Hindsight embeddings, is included here for completeness. A *NAIS*-based infosphere was also considered as an additional baseline.

For each ground truth, the mean and standard deviation of Accuracy, Loss, and F1-score across folds are reported.

Infosphere	Params	Accuracy ($\mu \pm \sigma$)	Loss ($\mu \pm \sigma$)	F1 ($\mu \pm \sigma$)
HINDSIGHT	***5***	***0.9800 ± 0.0022***	***0.0860 ± 0.0095***	***0.9802 ± 0.0021***
NO INFOSPHERE	N/A	0.6206 ± 0.1310	0.6113 ± 0.1520	0.6694 ± 0.1375
TOP PAPER	10	0.7363 ± 0.1271	0.5557 ± 0.1937	0.6129 ± 0.3108
TOP PAPER * TOPIC	[5,2]	0.7484 ± 0.0795	0.5346 ± 0.0967	0.7786 ± 0.0531
LightGCN	N/A	0.8308 ± 0.0310	0.4250 ± 0.0739	0.8314 ± 0.0362
NAIS	N/A	0.7515 ± 0.1273	0.5619 ± 0.2237	0.7951 ± 0.0671

Table 12: 5-fold cross validation results for *HINDSIGHT* ground truth (recommender-specific infosphere embeddings + probability-based sorting).

Infosphere	Params	Accuracy ($\mu \pm \sigma$)	Loss ($\mu \pm \sigma$)	F1 ($\mu \pm \sigma$)
HINDSIGHT	**5**	**0.8450 ± 0.1732**	**0.2854 ± 0.1964**	**0.8806 ± 0.1077**
NO INFOSPHERE	N/A	0.7698 ± 0.0450	0.5066 ± 0.0624	0.7986 ± 0.0260
TOP PAPER	**10**	**0.8356 ± 0.0427**	**0.3937 ± 0.0918**	**0.8254 ± 0.0615**
TOP PAPER * TOPIC	[5,2]	0.7526 ± 0.0058	0.5532 ± 0.0087	0.7760 ± 0.0040
LightGCN	N/A	0.7506 ± 0.0981	0.4943 ± 0.0600	0.7972 ± 0.0546
NAIS	N/A	0.8346 ± 0.0140	0.4575 ± 0.0355	0.8413 ± 0.0064

Table 13: 5-fold cross validation results for *TOP PAPER (10)* ground truth (recommender-specific infosphere embeddings + probability-based sorting).

The results from the 5-fold cross-validation experiments (Tables 12, 13, 14, 15, and 16) confirm the same overall trends observed in the previous evaluations. Across all ground truths, the corresponding recommender-specific infosphere generally achieves the best or near-best performance, indicating that the model can correctly identify the underlying recommender from user interactions alone. Occasional deviations, such as the slightly superior performance of the *Hindsight* infosphere, are consistent with its access to information not available during realistic simulation time. Taken together, these results further support the robustness of the proposed approach and its ability to generalize across different recommender configurations.

Infosphere	Params	Accuracy ($\mu \pm \sigma$)	Loss ($\mu \pm \sigma$)	F1 ($\mu \pm \sigma$)
HINDSIGHT	**5**	**0.9101 ± 0.0237**	**0.2500 ± 0.0537**	**0.9136 ± 0.0188**
NO INFOSPHERE	N/A	0.7031 ± 0.1230	0.5393 ± 0.0934	0.6335 ± 0.2944
TOP PAPER	10	0.6477 ± 0.0955	0.5934 ± 0.0836	0.6246 ± 0.1158
TOP PAPER * TOPIC	**[5,2]**	**0.8350 ± 0.0384**	**0.4023 ± 0.1062**	**0.8302 ± 0.0441**
LightGCN	N/A	0.7898 ± 0.0640	0.4491 ± 0.0999	0.7850 ± 0.0874
NAIS	N/A	0.7812 ± 0.0565	0.4575 ± 0.0826	0.8117 ± 0.0366

Table 14: 5-fold cross validation results for *TOP PAPER * TOPIC [5;2]* ground truth (recommender-specific infosphere embeddings + probability-based sorting).

Infosphere	Params	Accuracy ($\mu \pm \sigma$)	Loss ($\mu \pm \sigma$)	F1 ($\mu \pm \sigma$)
HINDSIGHT	5	0.9536 ± 0.0310	0.1553 ± 0.0947	0.9560 ± 0.0275
NO INFOSPHERE	N/A	0.8453 ± 0.0305	0.4282 ± 0.0534	0.8608 ± 0.0222
TOP PAPER	10	0.7558 ± 0.1778	0.4769 ± 0.2042	0.7017 ± 0.2403
TOP PAPER * TOPIC	[5,2]	0.8508 ± 0.0604	0.4123 ± 0.1102	0.8569 ± 0.0506
LightGCN	**N/A**	**0.9836 ± 0.0113**	**0.0900 ± 0.0579**	**0.9839 ± 0.0108**
NAIS	N/A	0.9638 ± 0.0226	0.1452 ± 0.0897	0.9643 ± 0.0228

Table 15: 5-fold cross validation results for *LightGCN* ground truth (recommender-specific infosphere embeddings + probability-based sorting).

Infosphere	Params	Accuracy ($\mu \pm \sigma$)	Loss ($\mu \pm \sigma$)	F1 ($\mu \pm \sigma$)
HINDSIGHT	5	0.9547 ± 0.0189	0.1424 ± 0.0475	0.9564 ± 0.0171
NO INFOSPHERE	N/A	0.8826 ± 0.0352	0.3017 ± 0.0505	0.8894 ± 0.0292
TOP PAPER	10	0.8515 ± 0.0735	0.3861 ± 0.1535	0.8439 ± 0.0720
TOP PAPER * TOPIC	[5,2]	0.7081 ± 0.1850	0.5033 ± 0.1375	0.5797 ± 0.3610
LightGCN	N/A	0.9873 ± 0.0100	0.0570 ± 0.0289	0.9874 ± 0.0097
NAIS	**N/A**	**0.9900 ± 0.0046**	**0.0372 ± 0.0059**	**0.9900 ± 0.0046**

Table 16: 5-fold cross validation results for *NAIS* ground truth (recommender-specific infosphere embeddings + probability-based sorting).

4.5 Test Performance on a Single Partition

To further assess the stability and generalization of our inference framework, we conducted additional experiments using the same fold configuration as in the previous setting. In this case, training was performed on a single fold, while the same held-out fold , specifically, fold 4 , was consistently used as the test set across all runs to maintain comparability. Table 18 presents the performance metrics obtained when fold 4 serves as the test set. Additionally, Table 17 reports the mean accuracy and loss, along with their standard deviations, calculated over the remaining folds combined, thus providing further insight into model robustness. These evaluations were conducted using the final experimental setup, which includes *specific-infosphere embeddings* combined with *descending-probability sorting*.

This setup reflects a more realistic scenario in which full access to the network is not feasible, common in large-scale or dynamic graphs. In many real-world applications, it is only possible to observe a subset of the data. By evaluating under such constrained conditions, we aim to assess the model's generalization capabilities when only partial information is available.

These targeted experiments confirm that the LightGCN-based model consistently attains the highest average accuracy and lowest

Infosphere	Par.	Acc.	Acc. SD	Loss	Loss SD
HIND. INFOSPHERE	5	0.9292	0.0206	0.1716	0.0294
NO INFOSPHERE	N/A	0.6632	0.1977	0.5830	0.1784
TOP PAPER	10	0.7113	0.1430	0.5606	0.1942
TOP PAPER * TOPIC	[5,2]	0.7345	0.0965	0.4968	0.0885
LightGCN	***N/A***	***0.9845***	***0.0008***	***0.0804***	***0.0292***

Table 17: Test performance metrics for *LightGCN* ground truth (single fold train).

Infosphere	Par.	Acc.	Acc. SD	Loss	Loss SD
HIND. INFOSPHERE	5	0.9296	0.0223	0.1728	0.0304
NO INFOSPHERE	N/A	0.6628	0.2006	0.6169	0.2105
TOP PAPER	10	0.7165	0.1538	0.5664	0.2179
TOP PAPER * TOPIC	[5,2]	0.7459	0.0876	0.5016	0.0950
LightGCN	***N/A***	***0.9858***	***0.0019***	***0.0805***	***0.0293***

Table 18: Fold4 performance metrics for *LightGCN* ground truth (single fold train).

loss, whether evaluated on the aggregate of three held-out folds or solely on fold 4. Notably, its standard deviations remain an order of magnitude smaller than those of the alternative recommenders, indicating markedly lower sensitivity to fold-to-fold variation. The hindsight infosphere also exhibits strong performance and greater stability compared to citation-based strategies, accompanied by modest standard deviations.

5 DISCUSSION

Overall, the model trained with the recommender matching that used at dataset generation time achieved the highest likelihood. The only notable exception is the hindsight infosphere model, which consistently performs better, even though it is not a valid recommender in practical terms, as it relies on inaccessible information during simulation.

These findings suggest that the model can capture key characteristics of the underlying recommender system from user interactions alone. The recommender-neutral model trained using the hindsight infosphere (see Sec. 2.2) remains sensitive to the recommender used at generation time, producing coherent and internally consistent synthetic outcomes across scenarios.

6 CONCLUSION

By testing the method proposed by Guidotti et al. [12] on synthetic datasets with known underlying recommenders, we show that it can partially infer the nature of hidden recommender systems from user interaction data. This validation provides preliminary evidence of the approach's potential to generate flexible, recommender-neutral user models. Overall, these results represent an initial step toward studying how recommender mechanisms may influence behavior within social platforms and indicate several directions for refinement and further investigation.

These results suggest that, even when only user interactions are available, it is possible to infer certain characteristics of the underlying recommender. Additionally, the model's sensitivity to different inputs and its ability to reproduce training data indicate that the proposed Recommender Neutral User model is suitable for simulating the impact of various recommenders in realistic social media settings.

Our approach extends existing audits that are only working in certain settings and fall short when recommenders are changing over time. We instead show a direction to detect recommenders from users behavior only during the actual operation which may allow more reliable results. While these are promising results, several limitations need to be addressed as we will outline below. However, we still introduce an initial attempt to recognize social media recommenders, an idea that can provide substantial improvements on social media societal impacts.

7 LIMITATIONS AND FUTURE WORK

While promising in recognizing hidden recommenders and modeling user behavior, our method is subject to limitations that should be addressed in future work.

(1) **Synthetic vs. Real Data:** Synthetic datasets used in our study may not fully capture the complexity of real-world user behavior and systems.
(2) **Assumption of Known Recommenders:** Real-world recommenders are more diverse than our predefined set. Future work should evaluate additional infosphere types.
(3) **Scalability:** Our approach is computationally expensive due to full dataset processing. This burden grows significantly with the number of potential recommenders and larger real-world datasets.
(4) **Assumptions in User Behavior Modeling:** The balance between intrinsic preferences and recommender influence may be more complex than our RNU model accounts for, particularly regarding the recommender's impact on interaction history.
(5) **Overfitting to Simulation Parameters:** Despite adding noise to the generation process, the method might still overfit to specific properties of the synthetic data.

ACKNOWLEDGMENTS

This research was supported by the Italian Ministry of University and Research under Grant No. 2023-NAZ-0206, PsyFuture – Dipartimento di Eccellenza 2023-2027 and by Volkswagen Foundation OpenUp Grant Ref. 9E530 Developing an Artificial Social Childhood (ASC).

This paper has been published as an Extended Abstract at AAMAS 2026. The final authenticated version is available online at: https://doi.org/10.65109/YMXQ7472

REFERENCES

[1] Hunt Allcott, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. 2020. The welfare effects of social media. *American economic review* 110, 3 (2020), 629–676.
[2] Mohamed Basel Almourad, John McAlaney, Tiffany Skinner, Megan Pleya, and Raian Ali. 2020. Defining digital addiction: Key features from the literature. *Psihologija* 53, 3 (2020), 237–253.
[3] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can

increase political polarization. *Proceedings of the National Academy of Sciences* 115, 37 (2018), 9216–9221.

[4] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173951

[5] Nicholas Diakopoulos. 2016. Accountability in algorithmic decision making. *Commun. ACM* 59 (01 2016), 56–62. https://doi.org/10.1145/2844110

[6] Christian G. Fink, Nathan Omodt, Sydney Zinnecker, and Gina Sprint. 2023. A Congressional Twitter network dataset quantifying pairwise probability of influence. *Data in Brief* 50 (2023), 109521. https://doi.org/10.1016/j.dib.2023.109521

[7] Daniel Fleder, Kartik Hosanagar, and Andreas Buja. 2010. Recommender systems and their effects on consumers: the fragmentation debate. In *Proceedings of the 11th ACM Conference on Electronic Commerce* (Cambridge, Massachusetts, USA) *(EC '10)*. Association for Computing Machinery, New York, NY, USA, 229–230. https://doi.org/10.1145/1807342.1807378

[8] Germain Gauthier, Roland Hodler, Philine Widmer, and Ekaterina Zhuravskaya. 2026. The political effects of X's feed algorithm. *Nature* (18 feb 2026). https://doi.org/10.1038/s41586-026-10098-2

[9] Nabeel Gillani, Ann Yuan, Martin Saveski, Soroush Vosoughi, and Deb Roy. 2018. Me, My Echo Chamber, and I: Introspection on Social Media Polarization. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 823–831. https://doi.org/10.1145/3178876.3186130

[10] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine* 38, 3 (2017), 50–57.

[11] Andrew M. Guess, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, Sandra González-Bailón, Edward Kennedy, Young Mie Kim, David Lazer, Devra Moehler, Brendan Nyhan, Carlos Velasco Rivera, Jaime Settle, Daniel Robert Thomas, Emily Thorson, Rebekah Tromble, Arjun Wilkins, Magdalena Wojcieszak, Beixian Xiong, Chad Kiewiet de Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud, and Joshua A. Tucker. 2023. How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science* 381, 6656 (2023), 398–404. https://doi.org/10.1126/science.abp9364 arXiv:https://www.science.org/doi/pdf/10.1126/science.abp9364

[12] Sabrina Guidotti, Gregor Donabauer, Simone Somazzi, Udo Kruschwitz, Davide Taibi, and Dimitri Ognibene. 2025. Modeling Social Media Recommendation Impacts Using Academic Networks: A Graph Neural Network Approach. In *Recommender Systems for Sustainability and Social Good*, Ludovico Boratto, Allegra De Filippo, Elisabeth Lex, and Francesco Ricci (Eds.). Springer Nature Switzerland, Cham, 63–72.

[13] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, YongDong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 639–648. https://doi.org/10.1145/3397271.3401063

[14] Xiangnan He, Zhankui He, Jingkuan Song, Zhenguang Liu, Yu-Gang Jiang, and Tat-Seng Chua. 2018. NAIS: Neural Attentive Item Similarity Model for Recommendation. *IEEE Transactions on Knowledge and Data Engineering* 30, 12 (Dec. 2018), 2354–2366. https://doi.org/10.1109/tkde.2018.2831682

[15] Kartik Hosanagar, Daniel Fleder, Dokyun Lee, and Andreas Buja. 2014. Will the global village fracture into tribes? Recommender systems and their effects on consumer fragmentation. *Management Science* 60, 4 (2014), 805–823.

[16] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous Graph Transformer. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 2704–2710. https://doi.org/10.1145/3366423.3380027

[17] Julian McAuley and Jure Leskovec. 2012. Learning to discover social circles in ego networks. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1* (Lake Tahoe, Nevada) *(NIPS'12)*. Curran Associates Inc., Red Hook, NY, USA, 539–547.

[18] Ivy Munoko, Helen L Brown-Liburd, and Miklos Vasarhelyi. 2020. The ethical implications of using artificial intelligence in auditing. *Journal of business ethics* 167, 2 (2020), 209–234.

[19] Dimitar Nikolov, Diego FM Oliveira, Alessandro Flammini, and Filippo Menczer. 2015. Measuring online social bubbles. *PeerJ computer science* 1 (2015), e38.

[20] Brendan Nyhan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, Annie Chen, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, Sandra González-Bailón, Andrew Guess, Edward Kennedy, Young Kim, David Lazer, Neil Malhotra, Devra Moehler, and Joshua Tucker. 2023. Like-minded sources on Facebook are prevalent but not polarizing. *Nature* 620 (07 2023), 1–8. https://doi.org/10.1038/s41586-023-06297-w

[21] Dimitri Ognibene, Gregor Donabauer, Emily Theophilou, Sathya Buršić, Francesco Lomonaco, Rodrigo Wilkens, Davinia Hernández-Leo, and Udo Kruschwitz. 2023. Moving beyond benchmarks and competitions: towards addressing social media challenges in an educational context. *Datenbank-Spektrum* 23, 1 (2023), 27–39.

[22] Dimitri Ognibene, Rodrigo Wilkens, Davide Taibi, Davinia Hernández-Leo, Udo Kruschwitz, Gregor Donabauer, Emily Theophilou, Francesco Lomonaco, Sathya Bursic, Rene Alejandro Lobo, et al. 2023. Challenging social media threats using collective well-being-aware recommendation algorithms and an educational virtual companion. *Frontiers in Artificial Intelligence* 5 (2023), 654930.

[23] Tiziano Piccardi, Martin Saveski, Chenyan Jia, Jeffrey Hancock, Jeanne L. Tsai, and Michael S. Bernstein. 2025. Reranking partisan animosity in algorithmic social media feeds alters affective polarization. *Science* 390, 6776 (2025), eadu5584. https://doi.org/10.1126/science.adu5584 arXiv:https://www.science.org/doi/pdf/10.1126/science.adu5584

[24] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira. 2020. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 131–141. https://doi.org/10.1145/3351095.3372879

[25] Benedek Rozemberczki, Carl Allen, and Rik Sarkar. 2019. Multi-scale Attributed Node Embedding. arXiv:1909.13021 [cs.LG]

[26] Piotr Sapiezynski, Avijit Ghosh, Levi Kaplan, Aaron Rieke, and Alan Mislove. 2022. Algorithms that "Don't See Color": Measuring Biases in Lookalike and Special Ad Audiences. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) *(AIES '22)*. Association for Computing Machinery, New York, NY, USA, 609–616. https://doi.org/10.1145/3514094.3534135

[27] Jakub Simko, Matus Tomlein, Branislav Pecher, Robert Moro, Ivan Srba, Elena Stefancova, Andrea Hrckova, Michal Kompan, Juraj Podrouzek, and Maria Bielikova. 2021. Towards Continuous Automatic Audits of Social Media Adaptive Behavior and its Role in Misinformation Spreading. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (Utrecht, Netherlands) *(UMAP '21)*. Association for Computing Machinery, New York, NY, USA, 411–414. https://doi.org/10.1145/3450614.3463353

[28] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. ArnetMiner: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Las Vegas, Nevada, USA) *(KDD '08)*. Association for Computing Machinery, New York, NY, USA, 990–998. https://doi.org/10.1145/1401890.1402008

[29] H. Holden Thorp and Valda Vinson. 2024. Context matters in social media. *Science* 385, 6716 (2024), 1393–1393. https://doi.org/10.1126/science.adt2983 arXiv:https://www.science.org/doi/pdf/10.1126/science.adt2983

[30] Matus Tomlein, Branislav Pecher, Jakub Simko, Ivan Srba, Robert Moro, Elena Stefancova, Michal Kompan, Andrea Hrckova, Juraj Podrouzek, and Maria Bielikova. 2021. An Audit of Misinformation Filter Bubbles on YouTube: Bubble Bursting and Recent Behavior Changes. In *Proceedings of the 15th ACM Conference on Recommender Systems* (Amsterdam, Netherlands) *(RecSys '21)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3460231.3474241

[31] Antonela Tommasel and Filippo Menczer. 2022. Do Recommender Systems Make Social Media More Susceptible to Misinformation Spreaders?. In *Proceedings of the 16th ACM Conference on Recommender Systems* (Seattle, WA, USA) *(RecSys '22)*. Association for Computing Machinery, New York, NY, USA, 550–555. https://doi.org/10.1145/3523227.3551473

[32] Zeynep Tufekci. 2018. YouTube, the great radicalizer. *The New York Times* 10, 3 (2018), 2018.