

# TikTok and the Art of Personalization: Investigating Exploration and Exploitation on Social Media Feeds

## Abstract

Recommendation algorithms for social media feeds often function as black boxes from the perspective of users. We aim to detect whether social media feed recommendations are personalized to users, and we also aim to characterize the extent and factors contributing to personalization in these feeds. We introduce a general framework to examine a set of social media feed recommendations for a user as a temporal graph. We label vertices in the graph as the result of exploration vs. exploitation of the user's interests on the part of the recommendation algorithm and introduce a set of metrics to capture the extent of personalization across user timelines. We apply our framework to a real TikTok dataset and validate our results using a baseline dataset generated from automated TikTok bots, as well as a randomized baseline. We also investigate the extent to which factors such as video viewing duration, liking, and following drive the personalization of content on TikTok. Our results demonstrate that our framework produces intuitive and explainable results, and can be used to audit and understand personalization in social media feeds.

## 1 Introduction

In recent years, researchers have raised concerns that excessive personalization of algorithmic social media feeds can potentially trap people in filter bubbles and echo chambers (TikTok 2020). This can lead to a variety of harms ranging from driving young users to depression and self-harm to creating highly polarized, radicalized, and ideologically fragmented societies (Ribeiro et al. 2020; Ledwich and Zaitsev 2019; Tufekci 2018). Consequently, it is important that we have methods and tools to systematically audit and investigate the extent or level of personalization of users' social media feeds. Recently passed EU legislation, the Digital Services Act (DSA) (European Commission 2023) calls for audits of personalized feeds offered by social media sites.

At the same time, social media feeds are significantly changing; from social-network-oriented feeds that are highly influenced by users' social connections on the platform (made popular by Facebook, Twitter, etc.), to algorithmic-oriented social media feeds that leverage a recommendation algorithm to infer the users' interests and recommend content (e.g., the ones currently popular on YouTube and TikTok). We are also witnessing an extremely popular trend, centered around short format (30-60 seconds) videos in algorithmic social media feeds. TikTok is the exemplar of this new trend in social media. A distinguishing and widely talked about feature of the TikTok social media platform is its personalized algorithmic `For You` feed

that generates content recommendations without relying on users' social network connections (like Facebook's or Twitter's feeds) or users' queries (like traditional Google video or YouTube search). The combination of short format video content, paired with an algorithmically driven feed has the potential to cause greater adverse effects on users, since the recommendation algorithm has to make many short format content recommendations over a small time period. Overall, we argue that there is a pressing need to investigate and audit personalized social media feeds, focusing on short format videos, to understand the extent to which these social media feeds are personalized and the possible effects this personalization can have on users.

Our work focuses on analyzing and auditing TikTok's personalized social media feed that combines both short-format content and algorithmic recommendations. We propose and validate a framework that allows us to model and assess the extent of user personalization on users' social media feeds. Particularly, given a set of social media feed attributes that include content, user, and engagement attributes, we design and implement a framework to assess which video recommendations are the result of personalization (i.e., *exploit recommendations*) and which are not (i.e., *explore recommendations*). We validate and demonstrate the applicability of our framework on a dataset of real traces from TikTok users collected by (Zannettou et al. 2023), as well as other baselines, including traces obtained from automated accounts on TikTok and a randomized baseline. Overall, we focus on the following research questions:

- **RQ1:** Given a sequence of content recommendations from a user's feed, how can we detect which recommendations are the result of personalization.
- **RQ2:** How can we characterize the extent of personalization in individual users' feeds on TikTok? Also, does the extent of personalization vary across users, and what factors affect user personalization?

**Contributions.** There are two types of contributions in this work: First, the proposed framework is flexible and can be used to audit any personalized social media feeds, including those of other short format video sites like Instagram and YouTube Shorts and those of traditional content feed sites like Facebook and Twitter. Also, we believe our framework can be used as part of algorithmic transparency and auditing systems that aim to provide feedback to users on how personalized their feeds are and the underlying reasons for getting recommended specific content.

Second, our methods applied to TikTok's `For You` feed shed light on its recommendation algorithm. We find, for example, that the algorithm exploits users' interests in between 30% and 50% of all recommended videos in the first thou-

sand videos of users’ tenure on TikTok. We introduce the notion of a *personalization score* and observe that this score can indeed estimate the extent of personalization in users’ feeds. We also examine personalization factors on TikTok, and our results show that liking and following are the primary drivers of content personalization for users.

**Broader Perspective and Ethical Considerations.** We obtained approval from our institution’s Ethical Board Review Committee before conducting any data collection/analysis. We emphasize that the real-world traces dataset, obtained by previous work by Zannettou et al. (2023), was collected after obtaining explicit consent from the participants before data donation. Additionally, the video metadata collection focuses on publicly accessible videos at the time of data collection (i.e. we do not have any data about deleted videos or videos from uploaders with private accounts). When conducting our data analysis, we follow standard ethical guidelines (Rivers and Lewis 2014) like not attempting to de-anonymize the participants from their traces or their demographic information and performing our analysis on aggregate. Stepping back, we believe our work ultimately benefits end-users of social media platforms by supporting transparency and algorithmic auditing. While there may be some negative impacts on platforms via such audits, which might result in platform changes (e.g., more or less aggressive personalization, less addictive feeds) that not all users like, we believe this work provides a net benefit.

## 2 Background & Related Work

TikTok is a social media platform with increasing popularity; in 2022, TikTok was the most downloaded mobile application worldwide with 672M downloads (Statista 2023). TikTok heavily relies on two features that make the platform stand out from other social media platforms; short-format videos and an algorithmic recommender system that offers an endless stream of video recommendations to users. Users on TikTok can upload short-format videos with a maximum length of 10 minutes (Landsberg 2022). TikTok offers many traditional social networking features such as: 1) users can follow other users, thus forming a social network, and 2) users can engage with a video by liking it, commenting on it, sharing it, or marking it as a favorite video. Also, TikTok offers two video streams; the “For You” and “Following” streams. The former shows videos based on TikTok’s recommendation algorithm, while the latter shows a stream of videos from uploaders that the user follows. In this work, we focus on the “For You” video stream as our main goal is to demystify how the recommendation algorithm explores and exploits users’ interests. Recently, reports have emerged (e.g., (Macgowan 2020; Smith 2021)) that indicate that TikTok’s recommendation algorithm is very effective in inferring users’ interests and making video recommendations that the users eventually like or engage with. Motivated by these reports, previous research, and investigative journalistic work studies the TikTok algorithm to understand algorithmic recommendations and whether the algorithm drives people toward recommendation paths consisting of problematic content.

**Previous Work.** A substantial body of work leverages automated accounts and qualitative analyses to understand and analyze TikTok’s algorithmic recommendations. Boeker and Urman (2022) explore the effect of various features on the TikTok algorithm personalization. Using automated accounts (i.e., bots), they investigate how features such as the like/follow features, region, language, and how much time users spent on specific content, affect the extent of personalization. Overall, they find that the follow feature exerts the most influence on the extent of personalization, followed by the like feature. Journalistic investigative efforts from the Wall Street Journal (Staff 2021) created over 100 automated accounts to understand what features (i.e., follow, like, share, watch time) affect algorithmic recommendations on TikTok. Their investigation finds that one of the signals alone, the watch time (i.e., how much time a user spends on a video), provides a strong signal to the algorithm and substantially affects algorithmic recommendations. The journalistic investigation shows that for automated accounts that expressed an interest in problematic content such as depression or sad content (as determined by the watch time the user spent on video with such content) often lead users down an algorithmic rabbit hole of problematic content. Klug et al. (2021) undertake a mixed-methods analysis of TikTok’s algorithm by performing 28 qualitative interviews with TikTok users and analyzing a dataset of 300K videos to confirm insights obtained from their interviews. They find that videos with high engagement (e.g., the number of comments, likes, and shares) are more likely to be recommended by the algorithm. At the same time, they find that using very popular hashtags (e.g., #foryou) does not increase the likelihood of a video being recommended by the algorithm. Bandy and Diakopoulos (2020) explore the role of TikTok’s recommendation algorithm in promoting call-for-action videos with a case study on the Tulsa rally. Their analysis shows that the amplification of call-for-action videos is not systematic and is rather likely due to the videos having an increased engagement. Lee et al. (2022) undertake a qualitative analysis of 24 interviews with TikTok users, aiming to explore how algorithmic personalization affects users’ perceptions. Their work highlights that TikTok users can identify parts of their identity via the algorithmic recommendation of the “For You” page and that their behavior can shape the algorithm’s ability to reflect their diverse interests. Simpson and Semaan (2021) perform an interview study by recruiting 16 LGBTQ+ TikTok users, aiming to understand these users’ interactions and encounters on TikTok; they find that TikTok’s algorithm creates contradictory identity spaces that reaffirm LGBTQ+ identities while simultaneously violating intersections of user identities.

**Remarks.** Previous work attempted to demystify algorithmic recommendations by heavily relying on traces obtained exclusively from automated accounts. While these previous efforts yield important insights, we argue that traces from automated accounts may lack the authenticity and diversity of traces from real users. Also, previous work attempted to understand algorithmic recommendations through the lens of users’ responses (i.e., performing qualitative analysis on user interviews), however, these efforts are limited as user

self-reports might introduce biases or discrepancies in the results (Parry et al. 2021; Verbeij et al. 2021; Ernala et al. 2020). In our work, we focus on overcoming these challenges and analyzing algorithmic recommendations through the lens of traces from real TikTok users. At the same time, we complement our analysis with traces from automated accounts that act as baselines for our analyses. Overall, to the best of our knowledge, our work is the first that focuses on understanding the interplay between exploration and exploitation in TikTok recommendations by proposing a framework that can be applied to other algorithmic recommendation-powered social media feeds.

### 3 Datasets

This section describes the two datasets used in our analysis. We use a dataset that includes traces from 347 real TikTok users and a dataset that includes traces from automated bots.

#### 3.1 Dataset from Real Users

Our dataset, including traces from real TikTok users, is based on previous work by Zannettou et al. (2023). Here, we briefly discuss how Zannettou et al. (2023) collected the dataset, and we briefly discuss the dataset. The authors relied on the EU’s General Data Protection Regulation (GDPR), particularly the right of access by data subjects, which describes that a user has the right to request and get access to all the information that a social media platform collects and processes about them. They implemented a privacy-preserving data donation system and recruited 347 TikTok users that donated their TikTok traces. Each trace provides a comprehensive view of the user’s actions on TikTok, including the user’s viewing history, like history, search history, follow history, etc. (see (Zannettou et al. 2023) for a comprehensive list of all the fields included in the datasets provided by TikTok after each user’s access to data request ). Additionally, for each video referenced in the traces from real-world TikTok users, the authors collected additional metadata, as the traces from TikTok include only video identifiers. To do this, they used an unofficial Python wrapper for the TikTok API (Teather 2022), which allowed them to collect metadata for each video, including the video description, the video hashtags, statistics about the video, etc. Overall, the dataset includes 4.9M videos viewed 9.2M times by 347 recruited TikTok users. Note, that only 4.1M videos have associated video metadata (the rest of the videos were either deleted or the uploader made their account private, by the time of the data collection). We refer the readers to the previous work by Zannettou et al. (2023) for a comprehensive description of the data donation system, their recruitment efforts, and the demographic information of the recruited users.

#### 3.2 Dataset from Automated Bots

We complement our real-world user traces dataset with a bot dataset, which we refer to as `simulated-bot`, that includes social media feeds that a set of automated accounts receive on TikTok. Our dataset consists of traces generated

Table 1: Overview of bot configurations for obtaining the bot traces. The table shows the probabilities of the bots watching a video till the end (Watch), skipping a video (Skip), liking a video (Like), and following the video creator (Follow).

	$\mathcal{P}(\text{Watch})$	$\mathcal{P}(\text{Skip})$	$\mathcal{P}(\text{Like})$	$\mathcal{P}(\text{Follow})$
Bot 1	0	1	0	0
Bot 2	1	0	0	0
Bot 3	0.5	0.5	0	0
Bot 4	1	0	0.5	0.5
Bot 5	0.5	0.5	0.5	0.5

by five automated bots; for each bot, we create a new TikTok account, and we use a random date of birth and a unique email address. No other personal information, such as gender or location is provided when creating the accounts. Each account is controlled by an automated bot with a pre-defined policy that dictates the bot’s behavior. Table 1 reports the policies of our bots. In particular, *Bot1* skips all videos without liking any video or following any account, *Bot2* watches all videos until their end without liking any video or following any account, *Bot3* watches videos till the end with a probability of 0.5 without liking any video or following any account. We also include some bots that like videos and follow accounts with a probability of 0.5 (see *Bot4* and *Bot5*). The difference between these two bots is that *Bot4* watches all videos until their end, while *Bot5* watches videos until their end with a probability of 0.5. We implement the bots and their policies using the PlayWright framework (Playwright 2023). Each bot visits TikTok’s For You feed via TikTok’s Web interface and watches 1,000 videos according to the bot’s pre-defined policy. Also, we perform a video metadata collection for each video that a bot encounters using an unofficial Python wrapper for the TikTok API (Teather 2022). The video metadata collection is done in real-time, which ensures that we obtain video metadata for all videos in our dataset. We run our bot dataset collection between December 28, 2022, and January 17, 2023.

### 4 RQ1: Detecting User Personalization

This section presents our modeling framework for detecting videos resulting from user personalization.

#### 4.1 Social Media Feed Attributes

We begin by describing three broad categories of data attributes that we use: *content*, *user*, and *engagement*. These attributes are readily extracted from most social media feed data and form the basis of our modeling framework and analysis to investigate the extent of user personalization in a social media feed. Note that the attributes we describe are not an exhaustive list of all the possible attributes but an intuitive and comprehensive set we curate that is readily applicable to most social media recommendation feeds.

**Content Attributes.** Given a set of  $N$  chronologically ordered recommendations  $R = \{r_1, r_2, r_3, \dots, r_N\}$  each recommendation item  $r_i$  has a set of attributes that provide details about the content of that recommendation. Here we out-

line the content-specific attributes we consider in our model and experiments.

- **Recommendation Index:** For any set of  $N$  chronologically ordered recommendations  $R = \{r_1, r_2, r_3, \dots, r_N\}$ , the recommendation index,  $i \in \mathbf{Z}^+$  (i.e.,  $1 \leq i \leq N$ ) is the index of each recommendation item in sequential order.
- A set of  $M$  hashtags included in the recommendation’s description  $H = \{h_1, h_2, h_3, \dots, h_M\}$ .
- A user,  $u_C$  that created the content that is recommended. For example, on user-generated video platforms, this corresponds to the video uploader.

**User Attributes.** Each user,  $u$  has a set of attributes that pertain to macroscopic user behaviors such as the accounts that are followed, and the user’s topics of interest.

- A set of  $K$  other users,  $U_F = \{u_1, u_2, u_3, \dots, u_K\}$  that are followed.
- A set of  $X$  interests,  $I_P = \{h_1, h_2, h_3, \dots, h_X\}$  as determined by the most popular hashtags in all recommendations made to user  $u$ .
- A set of  $Y$  interests,  $I_D = \{h_1, h_2, h_3, \dots, h_Y\}$  as determined by hashtags that user  $u$  specifies directly with the platform.

**Engagement Attributes.** For every recommendation item - user pair  $(r_i, u)$ , we have a set of attributes that provide information about user  $u$ ’s engagement with the recommendation item,  $r_i$ . We use the subscript  $r_i, u$  to denote the engagement attributes of the recommendation item - user pair.

- **Timestamp:** The timestamp,  $\theta_{r_i, u}$  when the recommended item  $r_i$  was viewed by the user  $u$ .
- **Engagement Duration:** The duration (in seconds),  $t_{r_i, u}$  the user  $u$  engaged with (watched for video items) the recommended item  $r_i$  for.
- **Liking:** We use a boolean,  $\text{liked}_{r_i, u}$  to denote whether the user  $u$  liked the recommended item  $r_i$ , and if  $\text{liked}_{r_i, u} = \text{True}$ , we record the timestamp the item was liked.
- **Following:** We use a boolean,  $\text{followed}_{r_i, u}$  to denote whether the content creator of recommended item  $r_i$  is followed by the user  $u$ , and if  $\text{followed}_{r_i, u} = \text{True}$ , we record the timestamp the creator was followed.
- **Sharing:** We use a boolean,  $\text{shared}_{r_i, u}$  to denote whether the recommended item  $r$  was shared by the user  $u$ , and if  $\text{shared}_{r_i, u} = \text{True}$ , we record the timestamp the item was shared.
- **Favoriting:** We use a boolean,  $\text{favorited}_{r_i, u}$  to denote whether the recommended item  $r_i$  was favorited by the user  $u$ , and if  $\text{favorited}_{r_i, u} = \text{True}$ , we record the timestamp the item was favorited.

We use these content, user, and engagement attributes to specify our modeling framework to detect user personalization in social media feed recommendations.

## 4.2 Exploitation vs. Exploration Framework

Given a user  $u$  and a set of  $N$  chronologically ordered recommendations  $R_u = \{r_1, r_2, r_3, \dots, r_N\}$ , we aim to label each of these recommendation items to demystify which recommendations are the result of user personalization (exploitation) and which are not (exploration). We define an *ex-*

*ploitation* recommendation as a recommendation that is personalized based on the user’s inferred interests or previous actions. On the other hand, we define an *exploration* recommendation as a recommendation that is *not* the result of user personalization, but due to the recommendation algorithm trying to explore if a user might like a specific – and often new or different – topic.

We assume that certain items recommended to a user are related to prior user actions or items recommended to that user. We leverage the content, user, and engagement attributes previously specified to determine whether (and the extent to which) a recommended item  $r_i$  for user  $u$  is related to previously recommended items  $r_j$  in  $u$ ’s feed, where  $\theta_{r_j, u} < \theta_{r_i, u}$ . We model a user  $u$ ’s trace of viewing history as a temporally evolving graph  $G = (V, E)$ , where each vertex corresponds to an item-user pair  $(r_i, u)$ , with its associated timestamp  $\theta_{r_i, u}$ . We use edges in the graph to represent the “relatedness” of vertices, which we compute using the user, content, and engagement attributes outlined previously. Then, we evaluate the extent (or degree) of personalization of each vertex i.e., the degree of personalization of each recommended item at a specific point in time, by analyzing the degree of the vertex and the weight of edges connected to that vertex.

Below, we detail our technique for establishing the temporally evolving graph for a social media feed user, using features that can be constructed from the feed attributes. We then use these features to specify a labeling method to classify all vertices as exploration or exploitation recommendations in the user’s temporal graph. We also delineate a set of metrics that help us better evaluate and understand the extent of personalization in our framework.

**Global and Local Features.** We define a set of *local* and *global* features,  $F = \{f_1, f_2, \dots, f_n\}$ , derived from the attributes available in a user’s recommended item timeline that allow us to classify recommendations as exploitation or exploration recommendations. The set of features we specify is not an exhaustive list of all possible features but an intuitive set we curate to model the “relatedness” of items in most social media feeds. Note that the set of features is readily generalizable to other social media recommendation feeds. Additionally, our framework is flexible, as it allows the addition or removal of features based on the use case.

We use these features,  $F$  to evaluate the connectedness of vertices in each user’s temporal graph by specifying an *activation* condition for each feature; i.e., if the condition specified by the feature is satisfied for a vertex in our graph, then we mark the corresponding vertex as *activated*. Our framework defines features such that the activation condition corresponds to the vertex labeled as an *exploit* recommendation. We describe a set of local and global features applicable to a (TikTok) video recommendation feed in Table 2.

**Local features** model relationships between vertices within a specific *temporal window* of size  $W$ . For example, for a vertex  $(r_i, u)$ , the local feature `likes_hashtag_local` considers preceding items  $r_j$  in the temporal window  $W$ ,  $r_j \in \{r_{i-1}, \dots, r_{i-W}\}$ , and *activates* the vertex  $r_i$  if  $\text{liked}_{r_j, u} = \text{True}$  and  $r_j$  has a

hashtag in common with  $r_i$ . In other words, in this example, we label a recommendation item,  $r_i$  as *exploit* if the user liked a previous item (within the temporal window) that shares a hashtag with  $r_i$ .

**Global features** capture general attributes/factors that capture user personalization at a macroscopic scale for a particular user. For example, for a vertex  $(r_i, u)$ , the global feature `following_global` considers all previously recommended items  $r_j$ , such that  $\theta_{r_j, u} < \theta_{r_i, u}$  in  $u$ 's feed that have `following_{r_j, u} = True`, and *activates* the vertex  $r_i$  if  $r_j$  has a hashtag in common with  $r_i$ . In other words, in this example, we label a recommendation item,  $r_i$  as *exploit* if a user followed the recommended item's creator anytime *before* they were recommended  $r_i$ .

**Labeling Recommendations.** For each recommended item - user pair, that corresponds to vertex  $(r_i, u)$  in our temporal graph  $G$ , we consider a set of  $d$  features  $F = \{f_1, f_2, \dots, f_d\}$ , and tag the vertex as follows.

- We label vertex  $(r_i, u)$  as an *Exploit* recommendation if any of the local or global features,  $F = \{f_1, f_2, \dots, f_d\}$  for that recommendation item satisfy the activation condition for  $(r_i, u)$  to be activated.
- If *none* of the features  $F = \{f_1, f_2, \dots, f_d\}$  used result in the vertex  $(r_i, u)$  being activated, then that vertex is marked as an *Explore* recommendation.

**Personalization Metrics.** We define three metrics in our analysis which serve as quantitative measures to study the extent of personalization in a social media feed. Additionally, we also use these metrics to evaluate our results when applying the framework to real data. We assume the recommendation labeling method outlined above using a set of features  $F = \{f_1, f_2, \dots, f_d\}$  to label each item-user vertex,  $(r_i, u)$  in user  $u$ 's temporal graph,  $G$  as *exploit* or *explore*.

1. **User exploit fraction.** Given a set of recommendations for a certain user,  $u$  a recommendation index,  $i$  and a window  $W$ , we define the user's *exploit* fraction at recommendation index  $i$  to be the fraction of items in the window  $W$  that are marked as *exploit* recommendations. Consider items  $r_j \in \{r_{i-1}, \dots, r_{i-W}\}$  recommended to user  $u$ , and let  $N_{ui}$  correspond to the number of  $r_j$ 's that are labelled *exploit*. Then for the recommendation index,  $i$  we denote the user exploit fraction as  $\alpha_{ui}$ .

$$\alpha_{ui} = \frac{N_{ui}}{W}$$

2. **Mean exploit fraction.** Given a set of  $m$  users, their corresponding recommendation feeds, and a recommendation index,  $i$ , another quantity of interest is the mean user exploit fraction, which we define as the arithmetic mean of the exploit fractions of all  $m$  users in the set, at recommendation index  $i$ . We denote the mean user exploit fraction by  $\bar{\alpha}_i$ .

$$\bar{\alpha}_i = \frac{1}{m} \sum_{u=1}^m \alpha_{ui}$$

3. **Personalization score.** We introduce the concept of a personalization score, where the core idea is to ascertain "how personalized" a specific recommendation item  $r_i$  is

for a user  $u$ . Given an item-user vertex,  $(r_i, u)$ , and  $m$  total user timelines, we estimate the extent of personalization,  $\rho(r_i, u)$ . This is achieved by calculating the number of user timelines in which the recommended item  $r_i$ , when inserted at recommendation index  $i$ , would have the same label (*Explore* or *Exploit*) as it did for user  $u$ . Assuming that item  $r_i$  is marked *Exploit* (or *Explore*) and has the same label *Exploit* (or *Explore*) in  $k$  other user timelines when inserted at recommendation index  $i$ , we then define the personalization score for vertex  $(r_i, u)$  as:

$$\rho(r_i, u) = 1 - \frac{k}{m}$$

The personalization score is between 0 and 1 and uses all  $m$  user timelines to estimate the extent to which item  $r_i$  is personalized for user  $u$ . It's important to note that the personalization score is dependent on the  $m$  user timelines and is not a symmetric measure, in general. Intuitively we expect items labeled *exploit* to have a higher personalization score, assuming they were tailored recommendations, since these items would have a lower probability of being marked as *exploit* in other user timelines. In contrast, we expect the personalization score of items labeled *explore* to be low, since these items are likely to also be marked as *explore* in other user timelines.

**Framework Specification.** Here, we describe the parameters used to fully specify our modeling framework for a social media recommendation feed. Note that this framework is flexible and can be adapted to be instantiated on various social media feed data. In our subsequent experimental evaluation, we will perform an evaluation on real TikTok data.

1. A set of  $m$  users for whom we have data about their timeline of recommended items with the content, user, and engagement attributes specified.
2. The sample size,  $N$ . We include the first  $N$  chronologically ordered recommendation items from a user's feed.
3. A *temporal window* of size  $W$ , that measures the local features by considering user behavior up to  $W$  recommendation items in the past.
4. The interest radii,  $X$  and  $Y$ , which define how many of the top interests of each user we consider.
5. A set of  $d$  features,  $F = \{f_1, f_2, \dots, f_d\}$  that have been pre-selected to model user personalization in the feed.

### 4.3 Feature Selection and Framework Evaluation

An important challenge when studying whether recommendations result from personalization is identifying a set of meaningful features that contribute to this goal without introducing substantial noise. Motivated by this, in this section, we discuss how we use randomization techniques to perform feature selection and evaluate our framework. Below, we describe our randomization approach and baselines, feature selection methodology, and evaluation.

**Baselines.** Intuitively, in a randomized set of  $N$  recommendations  $\{r_1, r_2, r_3, \dots, r_N\}$ , there is no user personalization since the recommended items are (ideally) unrelated to each other. However, our features could be activated due to inherent noise in the data - for example, several random

Table 2: Summary of social media feed features for a sample item  $r_i$  with creator  $u_C$ , and temporal window  $W$ .

Feature	Activation condition for vertex $(r_i, u)$ to be marked <i>exploit</i>
generic_hashtag_local	$r_i$ has a hashtag in common with any preceding item in the temporal window $W$ , $r_j \in \{r_{i-1}, \dots, r_{i-W}\}$ .
generic_creator_local	$r_i$ has the same creator, $u_C$ as any preceding item in the temporal window $W$ , $r_j \in \{r_{i-1}, \dots, r_{i-W}\}$ .
likes_hashtag_local	$r_i$ has a hashtag in common with any preceding item in the temporal window $W$ , $r_j \in \{r_{i-1}, \dots, r_{i-W}\}$ and $\text{liked}_{r_j, u} = \text{True}$ .
likes_creator_local	$r_i$ has the same creator, $u_C$ as any preceding item in the temporal window $W$ , $r_j \in \{r_{i-1}, \dots, r_{i-W}\}$ and $\text{liked}_{r_j, u} = \text{True}$ .
watched_hashtag_local	$r_i$ has a hashtag in common with any preceding item in the temporal window $W$ , $r_j \in \{r_{i-1}, \dots, r_{i-W}\}$ and $t_{r_j, u} \geq 100\%$ .
watched_creator_local	$r_i$ has the same creator, $u_C$ as any preceding item in the temporal window $W$ , $r_j \in \{r_{i-1}, \dots, r_{i-W}\}$ and $t_{r_j, u} \geq 100\%$ .
shares_hashtag_local	$r_i$ has a hashtag in common with any preceding item in the temporal window $W$ , $r_j \in \{r_{i-1}, \dots, r_{i-W}\}$ and $\text{shared}_{r_j, u} = \text{True}$ .
shares_creator_local	$r_i$ has the same creator, $u_C$ as any preceding item in the temporal window $W$ , $r_j \in \{r_{i-1}, \dots, r_{i-W}\}$ and $\text{shared}_{r_j, u} = \text{True}$ .
favoriteVideos_hashtag_global	$r_i$ has at least one hashtag in common with <i>any</i> prior item $r_j$ in $u$ 's feed that has $\text{favorited}_{r_j, u} = \text{True}$ .
favoriteVideos_creator_global	$r_i$ has a creator in common with <i>any</i> prior item $r_j$ in $u$ 's feed that has $\text{favorited}_{r_j, u} = \text{True}$ .
following_global	$r_i$ has at least one hashtag in common with <i>any</i> prior item $r_j$ in $u$ 's feed that has $\text{following}_{r_j, u} = \text{True}$ .
inferred_interests_global	$r_i$ has at least one hashtag in common with $I_P = \{h_1, h_2, h_3, \dots, h_X\}$ , the $Y$ most popular hashtags in all recommendations made to $u$ .

videos might have certain common hashtags. Consequently, we use a randomized baseline to select the features that minimize the noise as measured by randomized data. Also, we use a second baseline to evaluate our framework's results and explanatory power. We describe the two baselines here.

1. **Randomization by recommendation index.** We use  $m$  real user traces to create our first randomized baseline; for each recommendation index, we randomly permute all items at that index across all user timelines. We thus obtain  $m$  randomized user traces corresponding to randomized timelines derived from real user data. We use these  $m$  randomized user traces as a baseline for feature selection. In practice, we repeat the above randomization several times, and report results averaged across all the random samples. We refer to this randomized baseline dataset as *index-randomized*.
2. **Automated bot traces.** We generate automated bot timelines using the bots described in Section 3.2. To simulate different, simple user behaviors we create several different bots, and characterize their behaviors in terms of the probabilities associated with various user actions. We use these bot traces to help validate the explanatory power of our results, by comparing our results in terms of personalization metrics with those of the automated bots. Intuitively, we expect these traces to display a higher degree of personalization than the randomized baseline, but a lower degree than real user data. We refer to the bot baseline dataset as *simulated-bot*.

**Feature Selection.** We outline a procedure for selecting features to be used in our model. We assume we have a set of  $n$  features  $F' = \{f_1, f_2, \dots, f_n\}$ , and the goal is to choose

the top- $d$  features  $F = \{f_1, f_2, \dots, f_d\}$  to label each item-user vertex,  $(r_i, u)$  in user  $u$ 's temporal graph as *exploit* or *explore*. We define the *signal-noise ratio* of a feature (or set of features) to be the ratio of the feature's mean exploit fraction in real user data to the mean exploit fraction in the randomized user traces. We use the signal-noise ratio to measure feature importance in the following feature selection process.

1. We first examine the recommendation item and user data, and consider the different content, user, and engagement attributes to compile a comprehensive list of  $n$  potential global and local features,  $F' = \{f_1, f_2, \dots, f_n\}$  for labeling recommendation items. For reference, we specify a set of local and global features applicable to a (TikTok) video recommendation feed in Table 2.
2. We consider each feature  $f_a \in F'$  individually and label each item-user vertex,  $(r_i, u)$  as *exploit* or *explore* using only this feature  $f_a$ . We then observe the mean exploit fraction,  $\alpha_i$  for each recommendation index  $i \in 1, \dots, N$ . We repeat the same experiment for the  $m$  randomized user timelines in *index-randomized*, since the mean exploit fraction in the randomized timelines captures the noise floor level of feature  $f_a$ .
3. Once we complete step 2 for each feature, we rank the features in descending order of their signal-noise ratio and note their mean exploit fraction in *index-randomized*. From these feature rankings, we choose the top  $d' \leq n$  features with the highest signal-noise ratio that are below a suitable noise threshold,  $\tau$  in *index-randomized*. The noise threshold,  $\tau$  can be chosen by inspecting the distribution of the mean

exploit fraction in `index-randomized` across all features. Here, we aim to filter out the noisiest features and proceed with the top  $d'$  features that best capture the “relatedness” of vertices in the users’ temporal graphs.

4. For each of the  $d'$  subsets of  $d' - 1$  features we label each item-user vertex,  $(r_i, u)$  as exploit or explore if any of the features in the subset considered are satisfied. Again, we observe the mean exploit fraction,  $\alpha_i$  for each recommendation index  $i \in 1, \dots, N$  and repeat the same for `index-randomized`. We then rank the subsets by signal-noise ratio and recursively remove features until we have a set of top- $d$  features  $F = \{f_1, f_2, \dots, f_d\}$ . Note that we don’t specify  $d$  in our procedure, but choose a  $d$  that gives the best trade-off between model complexity and signal-noise ratio.

**Evaluating Results.** We leverage the bot dataset, `simulated-bot` generated from automated bots to evaluate our framework and the efficacy of the model. For each recommendation index, we compare the mean exploit fraction from our framework with the mean exploit fraction in `simulated-bot`. Since the automated bots make randomized choices with independent probabilities, intuitively we expect our metrics to reveal a lower level of personalization associated with the bot-simulated data.

#### 4.4 Experimental Setup for TikTok

Here, we describe how we configure our framework to assess user personalization on TikTok video recommendation social feeds. First, we describe how we preprocess the TikTok dataset, and then we report how we instantiate our framework and label video recommendations on TikTok.

**Preprocessing TikTok Recommendations.** We performed data preprocessing steps on the real users’ TikTok video traces to model the sequence of recommended item-user pairs. Specifically, we filter out generic hashtags, we cluster hashtags into “topics” using word embeddings and clustering techniques, as well as extract the most popular hashtags for each user that act as the global interests of the user. We elaborate on our preprocessing steps below.

**(1) Filter generic hashtags.** On TikTok, there are certain “generic” hashtags that creators add to almost all videos in an attempt to influence the recommendation algorithm. Such generic hashtags include `#foryoupage`, `#fyp`, and `#viral`. These hashtags do not provide meaningful information and will likely affect our labeling of recommended videos as exploit or explore. Therefore, it is paramount to remove these common hashtags and ensure that our framework considers only meaningful hashtags. We follow a similar method to (Boeker and Urman 2022), inspect individual timelines, and filter out the most common hashtags for each user. Removing these generic hashtags helps ensure that videos are only related via meaningful hashtags specific to the videos’ content.

**(2) Word2Vec hashtag clustering.** We use a heuristic fuzzy clustering method and Word2Vec (Mikolov et al. 2013) similarity to cluster sets of hashtags into similar “topics” to combine hashtags into comprehensive and concise groups and enable better matching of hashtags in practice. To do

this, we first train a Word2Vec model, using Continuous Bag of Words, on all the video descriptions referenced in the entire dataset of real user traces (see Section 3.1). For training the Word2Vec model, we exclude words/hashtags that appear less than 10 times in the entire dataset and use a context window of 7.<sup>1</sup> Then, we use the cosine similarities of the Word2Vec embeddings of all the hashtags in our dataset. We then iteratively assign hashtags to the nearest cluster or create a new cluster based on the similarity of that hashtag with pre-existing clusters. For example the hashtags `{#bieber, #biebertiktok, #belieber, #bieberforever}` were all clustered into the same hashtag group. We use cluster centers to represent all hashtags in a particular cluster.

**(3) Extract popular hashtags.** For each user,  $u$  we perform a term frequency-inverse document frequency (TF-IDF) analysis on the set of all hashtags corresponding to all videos recommended to that user (after removing the generic hashtags). Then, we compute the top- $k$  hashtags (topics) of interest, as determined by the most relevant hashtags across all recommendations made to that user.

**Instantiating the Framework for TikTok.** We use our framework with the following parameters.

- From our TikTok dataset of 347 users, we filter out all users with fewer than 1000 recommendations, and consequently consider video feed recommendation data for  $m = 220$  users, such that we have  $N = 1000$  chronologically ordered video recommendations  $R_u = \{r_1, r_2, r_3, \dots, r_{1000}\}$  for each user  $u$ .
- We use a temporal window of size  $W = 50$ , and an interest radius  $X = Y = 25$  such that we consider the top-25 hashtags for each user in both the popular hashtag set  $I_P = \{h_1, h_2, h_3, \dots, h_{25}\}$  and the specified hashtag set  $I_D = \{h_1, h_2, h_3, \dots, h_{25}\}$ . We performed a sensitivity analysis to tune these parameter values for our TikTok dataset.
- Using the feature selection technique outlined previously, we select a set of  $d = 7$  features that have the best signal-noise ratio and obtained the following feature set of features  $F = \{\text{generic\_creator\_local}, \text{likes\_hashtag\_local}, \text{likes\_creator\_local}, \text{watched\_hashtag\_local}, \text{watched\_creator\_local}, \text{favoriteVideos\_hashtag\_global}, \text{following\_global}\}$ .

**Labeling TikTok Videos.** Using the framework we instantiated for the TikTok dataset, we create  $m = 220$  temporal graphs, with vertices in each user’s graph corresponding to the user-video pairs for that users. Then we label each vertex,  $(r_i, u)$  based on the methodology presented above and the set of  $d = 7$  features.

#### 4.5 Results and Discussion

In this section we evaluate and discuss the results upon applying our framework to the real video recommendation feeds from the TikTok dataset, `real-tiktok`. We use

<sup>1</sup>We increase the context window to 7 instead of the default 5, because on TikTok several videos include a long list of hashtags.

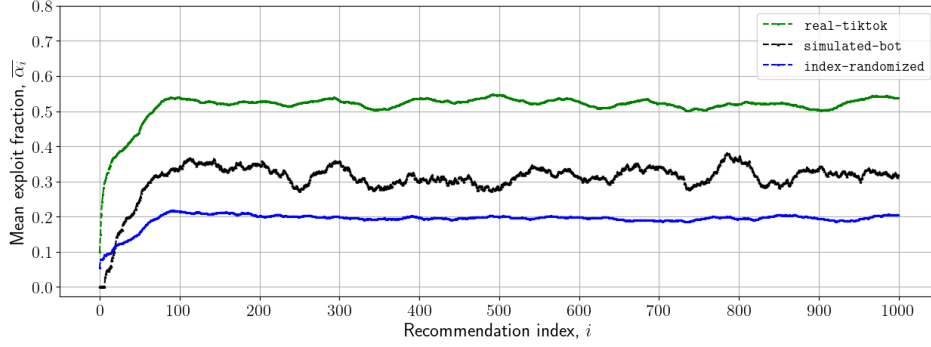


Figure 1: Mean user exploit fraction,  $\bar{\alpha}_i$  for real Tiktok users and baseline datasets.

our framework and label all videos in the  $m = 220$  user timelines as exploit or explore recommendations and we compute each user’s exploit fraction for each recommendation index between  $i \in 1, \dots, 1000$ . Then, we aggregate our analysis across the  $m = 220$  individual users and obtain the mean user exploit fraction, averaged across all users in the TikTok dataset, for each recommendation index. Also, we repeat the same procedure for the randomized baseline, *index-randomized* and the bot traces, *simulated-bot*. We visualize the mean user exploit fraction for all three datasets in Figure 1.

First, we observe that the mean exploit fraction,  $\bar{\alpha}_i$  for both *real-tiktok* and both baselines, initially increases steadily for the first few videos and then stabilizes for recommendation index  $i > 100$ . We attribute the steady increase to (a) the temporal window  $W$ , since the first few videos recommended to a user do not have a full window of past videos and hence exhibit a lower exploit fraction; and (b) potentially to TikTok’s algorithm ability to infer user interests and behavior, hence exploiting the users’ interests to a greater extent. The mean user exploit fraction is relatively stable for recommendation indices with  $i > 100$ . For *index-randomized*, this implies that the level of noise captured remains constant over time which is an expected result since there is no reason for the user exploit fraction to increase or decrease (on average) in a randomized temporal graph. This is also the expected result for *simulated-bot* since the automated bot traces represent timelines derived from random user behaviors. For the real user timelines, the stability of  $\bar{\alpha}_i$  indicates that the TikTok recommendation algorithm tries to recommend videos to users with a relatively constant level of personalization.

**Comparison with Baselines.** We validate our results using data from the automated bot traces, *simulated-bot*. We observe that the mean exploit fraction,  $\bar{\alpha}_i \geq 50\%$ , for *real-tiktok* dataset across most of the 1000 recommendation indices and differs significantly from that of both baseline traces. In contrast, we observe that the mean exploit fraction of the baselines is on average 31% for *simulated-bot* and 20% for *index-randomized*. Since the bots perform user behaviors randomly, intuitively we expect these timelines to have a lower degree of personal-

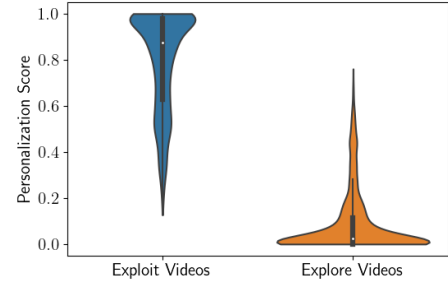


Figure 2: Distribution of Personalization scores for Explore and Exploit labeled videos.

ization than real users. This is evident from Figure 1, where we observe that the automated bot traces have only about 60% of exploit videos compared to real user traces. Under the assumptions of our framework, we observe that the TikTok algorithm attempts to personalize (exploit) a little over half the videos recommended to users. Accounting for the noise level of around 20% in *index-randomized*, we conclude that TikTok’s algorithm exploits real users’ interests in between 30% and 50% of all recommended videos in the first thousand videos of the users’ tenure on TikTok.

**Distribution of Personalization Scores.** We compute the personalization score for the labeled videos in *real-tiktok*, and plot the distribution of the scores in Figure 2. We observe that most exploit videos have a high personalization score (with a mean of 0.83), indicating that these videos are indeed specifically targeted and personalized to the users they are recommended to. Thus, when our framework identifies a vertex  $(r_i, u)$  as an *exploit* recommendation, we can be confident that the exploit recommendation is actually the result of personalization in relation to the other users’ temporal graphs. On the other hand, we observe that most videos that are labeled explore have a much lower personalization score (with a mean of 0.08). This indicates that these videos are likely not personalized to the users they are recommended to since they are also often labeled as explore recommendations for other users as well.

We also investigate the distribution of personalization scores for exploit/explore videos with respect to the viewing



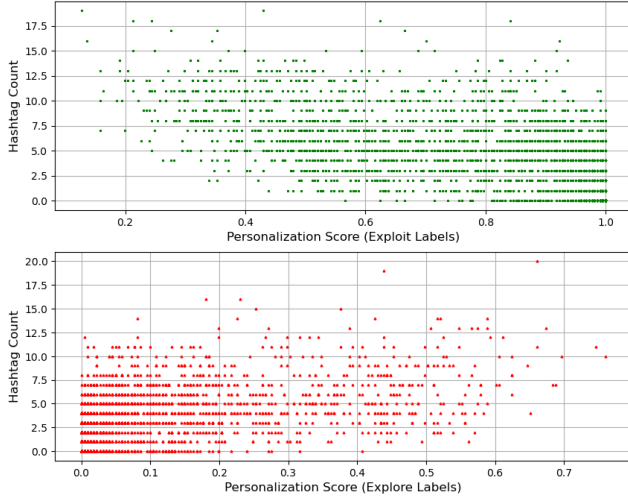


Figure 3: Personalization scores vs. hashtag counts.

duration, video popularity, and hashtag counts. We observe no correlation between personalization scores and video viewing and video popularity (we omit the figures due to space constraints). However, we observe that videos with fewer hashtags tend to have a higher personalization score (see Figure 3). This makes intuitive sense, since videos with more hashtags are often more generic, and hence harder to personalize, whereas videos with fewer hashtags often have more specific content that could be readily personalized to certain users.

Overall, our results which include the distribution of personalization scores for the labeled videos, along with the significantly higher mean user exploit fraction observed for real timelines over the randomized timelines, validate that our framework can be used to estimate the level of personalization in a user’s recommendation feed.

## 5 RQ2: TikTok Personalization Factors

Thus far, we observed that within the set of real users, `real-tiktok` different users have different levels of personalization in their feeds, as measured by their user exploit fractions across their temporal graphs. In this section, we aim to understand better the primary factors that drive personalization on TikTok.

We examine two distinct groups of users and attempt to characterize the extent of personalization observed in the content recommendations in each user group. We again use the mean exploit fraction,  $\bar{\alpha}_i$ , as our estimator of content personalization. For our analysis, we choose to analyze the two groups of users that correspond to the top quartile (TQ) and bottom quartile (BQ) of the mean user exploit fraction, as computed in Section 4 on the `real-tiktok` dataset.

### 5.1 Experimental Setup

We first follow the experimental setup described in Section 4.4 to instantiate the model on the `real-tiktok` dataset. Then, we run our framework to label videos as ex-

Table 3: Factors influencing user personalization on TikTok. We report mean values for the BQ and TQ groups,  $p$ -values for the  $t$  tests, and the impact level of each factor.

Personalization Factor	BQ	TQ	$p$ -value	Impact
Watch Percentage	84%	93%	0.03	Medium
Early Skip Rate	0.09	0.11	0.14	Low
Fraction Liked	0.03	0.14	$10^{-5}$	High
Fraction from Following	0.02	0.3	$10^{-14}$	High

ploit or explore. For each user, we calculate the user’s exploit fraction, which is simply the number of exploit videos divided by the total videos recommended to that user, i.e.,  $N = 1000$ . Next, we create two groups of users: 1) **Top Quartile, TQ**: users that are in the top quartile of the mean user exploit fraction (40 users with a mean user exploit fraction of 0.74); and 2) **Bottom Quartile, BQ**: users that are in the bottom quartile of the mean user exploit fraction (40 users with a mean user exploit fraction of 0.31).

For each user group, we investigate the effect of different personalization factors. We select these personalization factors based on prior related work by Boeker and Urman (2022) and the social media feed attributes available in our TikTok dataset. For each factor, we compare the distribution of that factor between the two user groups. We consider: 1) **Video watch percentage**: The mean video viewing duration percentage across all videos recommended to a user. 2) **Early skip rate**: The fraction of videos in the user’s timeline skipped over very early, i.e., within 1 second. 3) **Fraction liked**: The fraction of liked videos in the user’s timeline. 4) **Fraction from following**: The fraction of videos in the user’s timeline that were uploaded by a creator that the user was following.

Due to the way that our user groups are constructed, they have different levels of personalization. By comparing the distribution of personalization factors between the two groups, we aim to analyze the importance of each factor on user personalization.

### 5.2 Results

We compare the distributions of each personalization factor across both the BQ and TQ user groups and discuss our observations here. Table 3 shows the different personalization factors studied in this section, and the level of impact each of these factors has in terms of driving the extent of content personalization for a user. We include violin plots to show the distributions of each factor in Figure 4, and also run the Student’s  $t$ -test for statistical significance for each factor. We assign the level of impact to be “high,” “medium,” or “low” as characterized by the difference in distributions of the factor between the two user groups (a lower  $p$ -value corresponds to a higher impact level).

**Video watch percentage.** We observe a moderate difference between the BQ and TQ user groups in terms of their mean video watch percentage across all videos in these users’ timelines. The TQ user group has a mean watch percentage of 93%, whereas the BQ group has a mean watch percentage

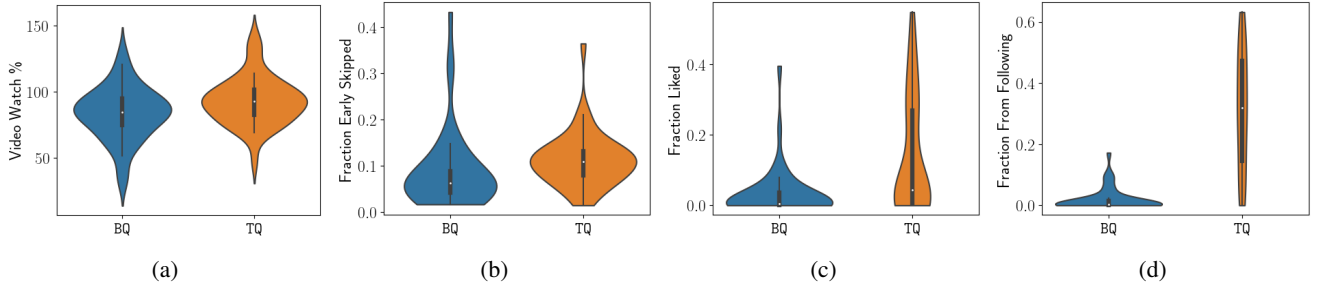


Figure 4: Violin plots showing the distributions of personalization factors between the TQ and BQ user groups.

of 84%. The  $t$ -test yielded a  $p$ -value of 0.03 indicating that there is a significant difference between the two groups (at the 0.95 confidence level). Additionally, we visually observe that the tail of the BQ distribution goes well below 50%, whereas the tail of the TQ distribution does not.

**Early skip rate.** We observe that there is only a marginal difference between BQ and TQ user groups in terms of the fraction of video skipped over quickly, across all videos in users’ timelines. The TQ has a mean early skip rate of 0.11, whereas BQ has a mean skip rate of 0.09. We observe outliers in both distributions, and  $t$ -test yielded a  $p$ -value of 0.14 indicating a less significant difference in the two distributions.

**Fraction liked.** We observe a significant difference between BQ and TQ user groups in terms of the fraction of videos liked. The TQ group liked 14% of videos on average, whereas the BQ group liked 3% of videos on average. Comparing the distributions from the violin plots in figure 4c we observe that the TQ group has a significantly higher fraction of liked videos above 0.2, whereas the BQ group (aside from a couple of outliers) lies entirely below this threshold. The  $t$ -test yielded a  $p$ -value of  $10^{-5}$  indicating a very significant difference in distributions

**Fraction from following.** We observe a significant difference between the BQ and TQ user groups in terms of the videos in a user’s timeline that were uploaded by a creator that the user was following. Visually, the distributions in figure 4d differ greatly. We observed a  $p$ -value of  $10^{-14}$  from the  $t$ -test indicating a very significant difference in distributions between the two groups.

## 6 Discussion and Conclusion

In this work, we proposed a framework that models and assesses personalization on social media feeds. The framework takes as an input a set of recommendations made to users on social media, performs an exhaustive feature selection, and creates links between the recommendations and user behavior. Then, the framework labels each recommendation as exploit (i.e., the result of user personalization) or explore (i.e., it is not the result of user personalization, and the algorithm explores user interests). We applied our framework to real data obtained from TikTok users and data obtained using automated accounts on TikTok, demonstrating our framework’s applicability and validity. Our analysis shows that on TikTok when considering the first one thousand videos in each user’s timeline, the recommendation algorithm ex-

ploits the user interests in 30%-50% of the recommendation videos; this finding indicates that the TikTok algorithm opts to recommend a large number of explore videos in an attempt to either infer better the user interests or maximize user retention by recommending many videos that are outside of the user’s (known) interests. Also, our analysis of the personalization factors finds that the most important aspects that affect the degree of personalization are following other TikTok accounts and liking videos. Our results (based on real TikTok user data) are in-line and confirm the results from Boeker and Urman (2022), which made similar observations based on automated accounts. Our framework can assist various interested stakeholders in further understanding personalization on the Web. We elaborate on these use cases and the implications of our work below.

**For Platforms: Aiding Transparency Efforts.** Online platforms increasingly rely on recommendation algorithms. Motivated by this, policymakers are currently demanding online platforms to provide end-users with explanations on why they are getting recommended specific content and giving them control over their recommendations (e.g., see the DSA regulation (European Commission 2023)). We argue that our framework can be used to generate fine-grained explanations that can be used to inform the users why they are getting recommended specific content. Our framework constructs links between the user’s behavior and the recommendation content; these links can be used to generate explanations. For instance, assuming that the user liked a lot of videos with the #sports, a possible explanation that can be given to a user that gets recommended a sports-related video will be “In the past 50 videos, you liked 20 videos with the #sports, so we inferred you liked sports content.” We believe that our framework is a substantial step toward providing tools and techniques that can be used by online platforms to generate informative and precise explanations to end-users, hence being compliant with emerging regulations like the DSA (European Commission 2023).

**For Users: Insights into User Algorithmic Personalization.** Our framework can act as the backbone for future systems that end-users can leverage to extract insights into how personalized their social media feeds are. We envision that it is possible to implement an easy-to-use system where end-users can request their data from online platforms using the right of access by the data subject as described in the EU’s General Data Protection Regulation (European Commission

2016). Then, they can input their data into this system, which will leverage our framework to assess user personalization and then visualize to the end users the extent of their personalization, which recommendation items are the result of personalization and which are not, as well as extract insights into what the online platform knows about them, through the lens of the content recommendations.

**For Policymakers and Researchers: Auditing Platforms and Algorithms.** We argue that our framework can assist policymakers and researchers that aim to audit online platforms and the effects of AI-based recommendation algorithms. Our flexible framework can be adapted to study other online platforms based on users' social media feeds. Given the increasing popularity of platforms that leverage recommendation systems to deliver content to the users, we believe that our framework is an important leap towards understanding the extent of user personalization on other emerging platforms like YouTube shorts and Facebook/Instagram Reels. Such audits are of paramount importance, and policymakers can use the audit results to assess how compliant online platforms are with emerging regulations and act accordingly.

**Limitations & Future Work.** We conclude with the limitations of our work and our future work. First, the sample of real traces from TikTok users is not necessarily representative; this is an inherent challenge that exists when undertaking studies that aim to recruit a small number of users from huge online platforms like TikTok. Nevertheless, despite this, we argue that the proposed framework and insights have merit and demonstrate that the degree of personalization varies across TikTok users. The fact that our findings match those of prior work that studied bot-based traces on TikTok (Boeker and Urman 2022) also increases our confidence. Another limitation is that our analysis and results are agnostic to changes in the recommendation algorithm done by the platform over time. In this work, we audited the personalization of users irrespectively of when they started using TikTok. In the future, we plan to perform longitudinal audits to understand how the algorithm changes over time. Finally, our work lacks ground truth data on recommendation content that results from user personalization, another inherent challenge when performing such audits. To overcome this challenge, we use the notion of randomization and assume that there should be little personalization in randomized traces, hence evaluating our framework in this way. In the future, we aim to investigate ways to generate ground truth datasets related to user personalization.

## References

- Bandy, J.; and Diakopoulos, N. 2020. #TulsaFlop: A Case Study of Algorithmically-Influenced Collective Action on TikTok. *arXiv preprint arXiv:2012.07716*.
- Boeker, M.; and Urman, A. 2022. An Empirical Investigation of Personalization Factors on TikTok. *arXiv preprint arXiv:2201.12271*.
- Ernala, S. K.; Burke, M.; Leavitt, A.; and Ellison, N. B. 2020. How well do people report time spent on Facebook? An evaluation of established survey questions with recommendations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.
- European Commission. 2016. General Data Protection Regulation. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>.
- European Commission. 2023. The Digital Services Act package. <https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package>.
- Klug, D.; Qin, Y.; Evans, M.; and Kaufman, G. 2021. Trick and please. A mixed-method study on user assumptions about the TikTok algorithm. In *13th ACM Web Science Conference 2021*, 84–92.
- Landsberg, N. 2022. The Ultimate TikTok Video Size Guide for 2023. <https://influencermarketinghub.com/tiktok-video-size/>.
- Ledwich, M.; and Zaitsev, A. 2019. Algorithmic extremism: Examining YouTube's rabbit hole of radicalization. *arXiv preprint arXiv:1912.11211*.
- Lee, A. Y.; Mieczkowski, H.; Ellison, N. B.; and Hancock, J. T. 2022. The algorithmic crystal: Conceptualizing the self through algorithmic personalization on TikTok. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–22.
- Macgowan, A. 2020. The TikTok Algorithm Knew My Sexuality Better Than I Did. <https://repeller.com/tiktok-algorithm-bisexual/>.
- Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.
- Parry, D. A.; Davidson, B. I.; Sewall, C. J.; Fisher, J. T.; Mieczkowski, H.; and Quintana, D. S. 2021. A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. *Nature Human Behaviour*, 5(11): 1535–1547.
- Playwright. 2023. Playwright enables reliable end-to-end testing for modern web apps. <https://playwright.dev/>.
- Ribeiro, M. H.; Ottoni, R.; West, R.; Almeida, V. A.; and Meira Jr, W. 2020. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 131–141.
- Rivers, C. M.; and Lewis, B. L. 2014. Ethical research standards in a world of big data. *F1000Research*, 3.
- Simpson, E.; and Semaan, B. 2021. For You, or For" You"? Everyday LGBTQ+ Encounters with TikTok. *Proceedings of the ACM on human-computer interaction*, 4(CSCW3): 1–34.
- Smith, B. 2021. How TikTok Reads Your Mind. <https://www.nytimes.com/2021/12/05/business/media/tiktok-algorithm.html>.
- Staff, W. 2021. Inside TikTok's Algorithm: A WSJ Video Investigation. <https://www.wsj.com/articles/tiktok-algorithm-video-investigation-11626877477>.
- Statista. 2023. Leading mobile apps worldwide in 2022, by downloads. <https://www.statista.com/statistics/1285960/top-downloaded-mobile-apps-worldwide/>.
- Teather, D. 2022. Unofficial TikTok API in Python. <https://github.com/davidteather/TikTok-API>.
- TikTok. 2020. How TikTok recommends videos #ForYou. <https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you>.
- Tufekci, Z. 2018. YouTube, the great radicalizer. *The New York Times*, 10(3): 2018.
- Verbeij, T.; Pouwels, J. L.; Beyens, I.; and Valkenburg, P. M. 2021. The accuracy and validity of self-reported social media use measures among adolescents. *Computers in Human Behavior Reports*, 3: 100090.
- Zannettou, S.; Nemeth, O.-N.; Ayalon, O.; Goetzen, A.; Gummadi, K. P.; Redmiles, E. M.; and Roesner, F. 2023. Leveraging Rights of Data Subjects for Social Media Analysis: Studying TikTok via Data Donations. *arXiv preprint arXiv:2301.04945*.