# Uncovering the Heterogeneous Effects of Preference Diversity on User Activeness: A Dynamic Mixture Model

Yunfei Lu
Tsinghua University
luyunfeicx@163.com

Peng Cui
Tsinghua University
cuip@tsinghua.edu.cn

Linyun Yu
Bytedance Inc.
yulinyun@bytedance.com

Lei Li
University of California, Santa Barbara
lilei@ucsb.edu

Wenwu Zhu
Tsinghua University
wwzhu@tsinghua.edu.cn

## ABSTRACT

Preference diversity arouses much research attention in recent years, as it is believed to be closely related to many profound problems such as user activeness in social media or recommendation systems. However, due to the lack of large-scale data with comprehensive user behavior log and accurate content labels, the real quantitative effect of preference diversity on user activeness is still largely unknown. This paper studies the heterogeneous effect of preference diversity on user activeness in social media. We examine large-scale real-world datasets collected from two of the most popular video-sharing social platforms in China, including the behavior logs of more than 787 thousand users and 1.95 million videos, with accurate content category information. We investigate the distribution and evolution of preference diversity, and find rich heterogeneity in the effect of preference diversity on the dynamic activeness. Furthermore, we discover the divergence of preference diversity mechanisms for the same user under different usage scenarios, such as active (where users actively seek information) and passive (where users passively receive information) modes. Unlike existing qualitative studies, we propose a universal mixture model with the capability of accurately fitting dynamic activeness curves while reflecting the heterogeneous patterns of preference diversity. To our best knowledge, this is the first quantitative model that incorporates the effect of preference diversity on user activeness. With the modeling parameters, we are able to make accurate churn and activeness predictions and provide decision support for increasing user activity through the intervention of diversity. Our findings and model comprehensively reveal the significance of preference diversity and provide potential implications for the design of future recommendation systems and social media.

## CCS CONCEPTS

• **Human-centered computing** → **Social media**; **Social network analysis**; **Social recommendation**.


*KDD '22, August 14–18, 2022, Washington, DC, USA*

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539033

## KEYWORDS

Preference Diversity; Activeness Model; Heterogeneity; Dynamics

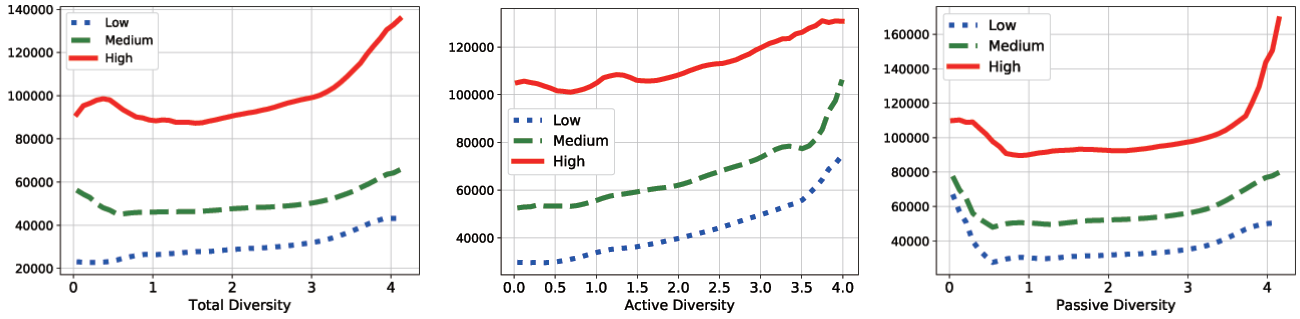
**ACM Reference Format:**
Yunfei Lu, Peng Cui, Linyun Yu, Lei Li, and Wenwu Zhu. 2022. Uncovering the Heterogeneous Effects of Preference Diversity on User Activeness: A Dynamic Mixture Model. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3534678.3539033

## 1 INTRODUCTION

In recent years, diversity plays an increasingly important role in social media for understanding and modeling complex social phenomena and human behavior such as filter bubbles[17], opinion formation[3], and social cooperations[21]. Low diversity is usually regarded as harmful to users (e.g., echo chamber), for which numerous efforts are devoted to expanding diversity, aiming to improve user experience and activeness, especially in recommendation systems[1, 2]. However, much less is known about preference diversity of users, which refers to how diverse one's interests are and its effects on user behaviors and activeness[24].

In literature, earlier works have proven the usefulness of user preference in activeness modeling and prediction[9]. The dynamic nature of preference is found to be a critical issue that affects user activeness[27]. Previous studies also recognize the dynamic predictive effect of preference diversity in human behavior[6]. Besides, there are also qualitative discussions about preference diversity, such as content diversity should match preference diversity to achieve the best user experience[24]. However, none of the previous works uncover the quantitative effect of preference diversity on user activeness, which is still a long-standing open problem.

In this paper, we collect two large-scale and fine-grained datasets from Douyin and Xigua, two popular video-sharing social platforms in China with different product forms. The datasets explicitly record the like behaviors of 787 thousand users over 1.95 million videos in 8 months. For each like behavior, its happening time and scenario (e.g., recommendation or search page) are recorded. Each video is tagged with its content category, enabling the research on user preference diversity. We investigate the relationship between preference diversity and user activeness with other factors that may affect activeness controlled, such as user portrait and registration time. With comprehensive analysis, we find similar rich heterogeneous patterns in the effect of preference diversity




(a) (b) (c)

**Figure 1:** ***We discover that the relationships between preference diversity and user activeness vary across users with different activity levels. The intricate heterogeneous patterns come from the mixture of the exponential mechanism in active diversity and the U-shaped distribution in passive diversity.*** **(a) is the pattern of preference diversity, while (b) is the pattern of active diversity and (c) is the pattern of passive diversity. The Y-axis shows the activeness represented by active time in seconds. We sort the users according to their active time, and the top 10%, 40%-60%, and bottom 10% are divided into groups High, Medium, and Low, respectively, represented by curves with different colors.**

on activeness in both datasets, which are shown in Fig.1(a), and we surprisingly find it presents heterogeneous patterns for users with different levels of activity: For the most inactive users, the activeness exponentially increases with diversity. For moderately active users, activeness showed a U-shaped distribution in diversity. As for the most active users, the distribution presents a mixed bimodal form. We find that the complex pattern results from the mixture of different mechanisms and cannot be well captured by a simple distribution. Therefore, we divide the preference diversity into active modes and passive modes according to the scenarios of behavior. Specifically, if one likes a video during a purposeful process of seeking, such as following or searching for some specified content, this behavior will be classified into active preference. If one likes a video during an aimless process, such as browsing videos, it will be counted into passive preference. The effects of active diversity and passive diversity are illustrated in Fig.1(b) and (c). We discover the divergent patterns and mechanisms for the same user in these two modes. The activeness exponentially increases with diversity in active mode, while it shows a U-shaped distribution on passive mode diversity. The different mechanisms of active and passive modes lead to the heterogeneity of overall preference diversity. We further delve into the distribution patterns and the dynamic nature of preference diversity and reveal how preference diversity interacts with other factors that influence user activeness.

Based on our findings, we propose a mixture model incorporating the heterogeneous effects of preference diversity to capture dynamic user activeness. The effectiveness of our model is validated by fitting the activeness distribution on diversity and capturing the temporal dynamic activeness patterns of each individual on large scale real data. Furthermore, our model accurately predicts user activeness and churn in the future. Through parameter analysis, our model shows usefulness in identifying users' different needs for diversity and assisting decision-making on increasing activeness. Our findings and model reveal the heterogeneous effects of preference diversity and may have potential implications for future recommendation systems and social media design.

In summary, our contributions are highlighted as follows:

- **Novel Findings:** We conduct systematical analysis around preference diversity and uncover its heterogeneous and dynamic effects on user activeness, filling a gap in previous researches.
- **A Novel Activeness Model:** We propose a universal dynamic model to capture the complex patterns of activeness on preference diversity. To our best knowledge, this is the first quantitative activeness model that corporate the effect of preference diversity.
- **Accuracy and Usefulness:** Our model fits and predicts the empirical and future user activeness accurately. By applying our model, we successfully predict the churn event and provide decision support for increasing activeness with preference diversity intervention.

## 2 RELATED WORK

As the investigated problem is closely related to the activeness model, and the main advantage of our work is revealing the significance of user preference diversity, we mainly review the related works in these two fields.

**Activeness Model.** As the basis of social media, user activeness is generally represented by daily online time. Modeling and predicting user activeness are regarded as a crucial problem with highlighted application value[14]. Traditional activeness models have been approached as feature-based problems, focus on seeking relevant features with strong predictive power[22]. Among them, Hu et al. indicate that the adoption of individual features improves prediction accuracy significantly[7]. Tan et al. find that user activeness is highly correlated to structural features[23]. Gonccalves et al. explore the relationship between activeness and stable social relationships[5]. Other methods take note of the dynamic nature of user activity and focus on modeling activeness over time[12]. Zhu et al. develop a personalized and socially regularized time-decay model to predict activeness[30]. Wang et al. propose a framework for predicting user activities based on point processes[25].

However, none has explored the effect of preference diversity in user activity nor its relationship to other factors, for which there is still no comprehensive activeness model that incorporates the effect of preference diversity.

**Preference Diversity.** While a variety of models try to increase content diversity without compromising accuracy[19], it has been found that preference diversity might be the key to strike a balance between accuracy and diversity[24]. Wu et al. discover the relationship between personality and preference in recommendation systems[26]. Lathia et al. indicate that user preferences change over time and present rich temporal patterns[10]. Due to the lack of data with accurate content labels, the study of the significance of preference diversity is limited to specific areas. Park et al. propose a useful metric for the diversity of musical tastes[18]. Liu et al. model the process of how media influences preference diversity for news[13]. McGinty et al. discuss the role of diversity in conversational recommendation systems[15].

However, no previous researches have uncovered the heterogeneous effect of preference diversity nor model its dynamic effect on activeness successfully.

## 3 PREFERENCE DIVERSITY ANALYSIS

In this section, we introduce our findings around preference diversity uncovered through data analysis, which is the basis of our model and experiments.

### 3.1 Metrics

We first introduce how to quantify user activeness and preference diversity. User activeness is measured by daily online time, which is consistent with most relevant researches[30]. Diversity should reflect not only the number of categories of preferences but also the distribution of preferences across those categories. Therefore, we adopt Shannon's entropy to measure preference diversity, which is a standard metric in diversity researches[20]. The following equation determines the information entropy for $u$:

$$E_u = -\sum_i p_u(i) \log_2 p_u(i) \tag{1}$$

where $p_u(i)$ represents the proportion of category $i$ in the preferred content of $u$. Higher entropy means more diversity. In our analysis and experiments, we filter out the inactive users whose number of watched videos is below a threshold to make sure that the remaining users' entropy can reflect preference diversity accurately. For instance, in Douyin dataset, all the sampled users watch more than 100 videos per week[1]. This data selection can also help alleviate the spurious phenomenon that a higher number of watched videos necessarily result in high entropy when the number of watched videos is quite low.

### 3.2 Datasets

We collect a large-scale fine-grained dataset from Douyin[2], one of the most popular social platforms for creating and sharing short-form videos in China. The dataset explicitly records the active time for every user and all the videos she/he liked and the corresponding scenarios during the period of 244 days from April 1, 2019, to November 30, 2019. We sample 620 thousand users and 1.66 million videos they liked in the chosen period, and all the videos are accurately labeled with their content categories according to a unified classification system.

Another supplementary dataset is collected from Xigua[3], one of the most popular video platforms in China. Compared with Douyin which is dominated by short-form videos like TikTok, Xigua focuses on longer videos like YouTube. This dataset includes 167 thousand users with their daily active time in different scenarios, and 290 thousand videos with their accurate content categories, from July 1, 2020, to December 31, 2020. Since Xigua users have limited thumb up, we also regard finishing a video as liking it, which is consistent with the real setting in Xigua platform.

All the data that we could access were anonymized for a strict privacy policy. Although Douyin and Xigua are both video platforms, they are very different in many aspects, such as product form, user base, and interface style. For example, Douyin users are mainly young people, while Xigua users are mainly elderly. Douyin focuses on user-generated content(UGC), while Xigua is positioned as a professional-generated content(PGC) platform. Most videos on Douyin are only 15 seconds long, while those on Xigua are usually longer than five minutes. Different video lengths lead to the difference in popular content on the two platforms. Talent shows, such as singing and dancing, have attracted the most attention on Douyin, while Xigua users are most attracted to popular science lectures or movies.

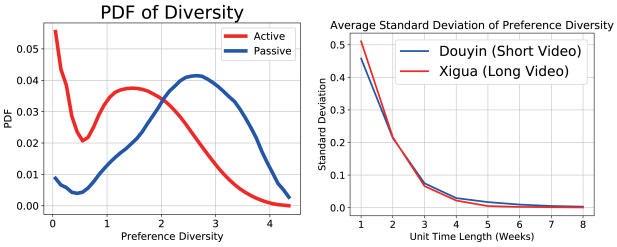
### 3.3 Analysis

Despite the evident distinctions between Douyin and Xigua, we find similar heterogeneous dynamic effects of preference diversity on activeness in these two datasets, which show the generality of our findings. In this section, the figures presented are mainly from the analysis of Douyin since it has a larger scale and a finer granularity, while some figures from Xigua are omiited for brevity.

*3.3.1 Distribution.* To figure out the heterogeneous mechanism of preference diversity, we first compare the distributions of active and passive diversity in Fig.2(a). Both distributions are bimodal with one peak at 0, indicating the users who prefer only one category. Another peak of diversity distribution is 1.55 bits for active diversity and 2.72 bits for passive diversity. The overall entropy of passive diversity is obviously higher, which indicates that users are more tolerant of the content pushed to them. The two different distribution patterns reflect the heterogeneity of preference diversity in different modes. We further find that the active and passive diversity are mutually independent and should be modeled separately. The details are shown in the Appendix.

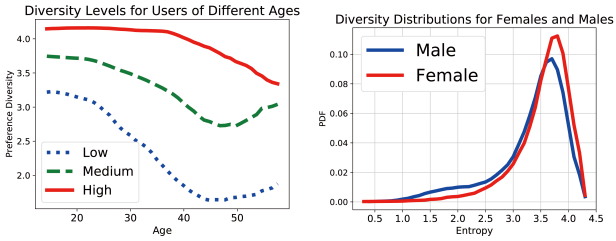
*3.3.2 Evolution.* Previous researches reveal the fact that user preference may change over time[10]. We select the users who are new to the social platform and use the standard deviation of preference diversity over different unit time lengths to measure how much their preference diversity changed over time. The unit time length represents how often we calculate the diversity. For example, when

[1]Since most videos on Douyin are only 15 seconds long, most users (96.1%) satisfy this condition, which means that we only filter out the most inactive users.
[2]https://www.douyin.com/
[3]https://www.ixigua.com/




(a) (b)

**Figure 2:** ***(a) Probability distribution functions of active diversity (red) and passive diversity (blue). (b) The average standard deviation decreases with the increase of unit time length for Douyin (blue) and Xigua (red). The similar curves show that the two datasets share the same pattern.***




(a) (b)

**Figure 3:** ***Preference diversity vs. individual factors.*** **(a) The relationship between preference diversity and age. The curves in different colors represent users with different levels of diversity. (b) The distribution of stable preference diversity for women and men.**

the unit time length is two weeks, it means that we calculate the each user's preference diversity every two weeks according to the videos liked by this user in these two weeks, and get the curves of how diversity evolves during the whole observation for each user. The evolving curves for different users are shown in the Appendix. We calculate each evolving curve's standard deviation, then plot its average value and the corresponding unit time length in Fig.2(b). It indicates that preference diversity is relatively volatile in the short term but quite stable in the long term (more than four weeks) for both datasets, which is similar to user activeness[29]. A similar dynamic nature enables us to model user activeness with preference diversity over time. It is also worth noting that a stable passive preference diversity means that user interests are not gradually narrowed by the recommendation system, which is consistent with some previous studies on filter bubbles in recommendation-based platforms[16]. From the evolution patterns of diversity for different users shown in the Appendix, we also find that the stable entropy values in the long term vary from person to person, reflecting the heterogeneity of preference diversity of different users.

*3.3.3 Relations to Other Factors.* Since user activeness is affected by various factors, it is necessary to figure out how the effects of preference diversity relate to the effects of other factors, which mainly include individual[7] and structural factors[23]. By analyzing various individual factors, we find that preference diversity is mainly negatively correlated with age, but for those users with fewer interests, after their mid-40s, the older they get, the more diverse their preferences are. We also find that the preference diversity of women is significantly higher than men, as shown in Fig.3. We do not find significant correlations for other factors, including the location and economy of the user's area.

As for the structural factors, we find that network structure features and diversity influence activeness independently, for which it is necessary for activeness models to incorporate preference diversity directly. The details are shown in the Appendix.

# 4 PROPOSED METHOD

This section presents our proposed model in detail and analyzes it, introducing how to capture the heterogeneous effect of preference diversity on dynamic user activeness.

## 4.1 Model intuition

To depict the activeness over time for each user, we set up our dynamic model at the individual level and denote the activeness of user $u$ at time $t$ as $Act_u(t)$. Our primary goal is to figure out the dynamic relationship between user activeness and diversity. However, given the existence of the recommendation system, which is usually a black box but has a huge impact on user activity, it is impossible to derive models from the bottom up through microscopic mechanisms. Inspired by Zang et al.[28], we try to fit empirical data distributions with parametric models directly and then interpret them in a generative way to incorporate the effect of the latent factors in social platforms, such as the recommendation system.

## 4.2 Our Activeness Model

Since the activeness from active and passive online scenarios should be modeled separately, we split $Act_u(t)$ into the corresponding two parts, $Act_{u,a}(t)$ and $Act_{u,p}(t)$, which are the active time on active and passive modes respectively:

$$Act_u(t) = w_a Act_{u,a}(t) + w_p Act_{u,p}(t) \tag{2}$$

How time is allocated between active and passive scenarios is part of user habit and usually stable, for which we denote the propensity of users $u$ to spend time in active and passive scenarios by static parameters $w_a$ and $w_p$. Based on the previous researches and our findings, the main factors that affect active time are structure, diversity, and individual factors. From the former analysis we know the effects of the structural factors are independent with preference diversity, which enables us to decompose $Act_{u,a}(t)$ and $Act_{u,p}(t)$ into the following formation:

$$\begin{aligned} Act_{u,a}(t) &= F_u(S_u(t))G_{u,a}(d_{u,a}(t)) + I_{u,a} \\ Act_{u,p}(t) &= F_u(S_u(t))G_{u,p}(d_{u,p}(t)) + I_{u,p} \end{aligned} \tag{3}$$

In the equation above, $F_u(S_u(t))$ describes the effect of structural factors $S_u(t)$, while $G_u(d_u(t))$ represents the impact of preference diversity $d_u(t)$. $I_u$ captures the effect of individual factors. Since individual factors such as gender and personality hardly change over time, $I_u$ is constant during the observation of several months.

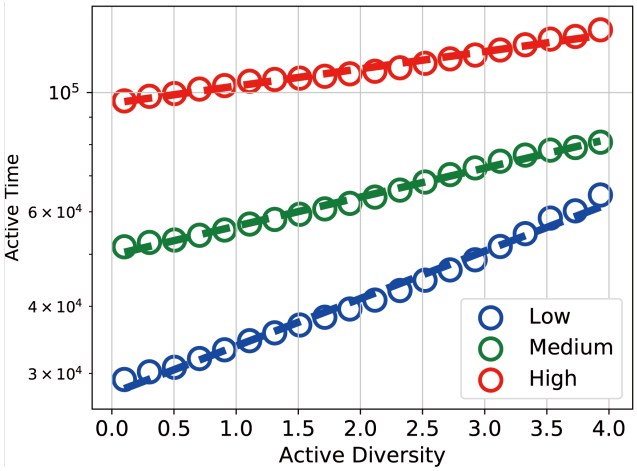



**Figure 4:** $log(Act_{u,a}(t))$ ***increases linearly with active diversity, indicating the exponential growth mechanism.*** **The curves in different colors represent users with different activity levels, which share the same pattern.**

We first focus on the active part $Act_{u,a}(t)$ of the model. We collect a set of users with a similar number of links, namely, controlling the structural factors to make $F_{u,a}(S_u(t))$ constant for different users, and display the relationship between activeness and active preference diversity in Fig.4. For all the users, the curves show linear patterns with activeness at a log scale. With $k$ and $b$ representing the slope and intercept of a linear function, we have

$$\begin{aligned} \log Act_{u,a}(t) &= kd_{u,a}(t) + b \\ Act_{u,a}(t) &= e^{kd_{u,a}(t)} \cdot e^{b} \end{aligned} \tag{4}$$

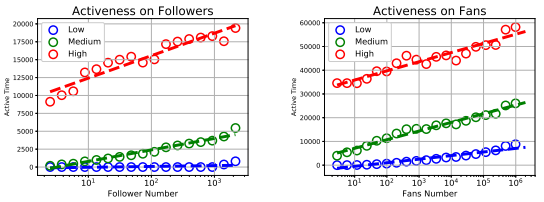
Combine the equation 3 with the equation 4 and consider $F_{u,a}(S_u(t))$ as a constant $F_a$, we get

$$\begin{aligned} Act_{u,a}(t) &= F_a G_{u,a}(d_{u,a}(t)) + I_{u,a} = e^{kd_{u,a}(t)} \cdot e^{b} \\ G_{u,a}(d_{u,a}(t)) &\sim e^{kd_{u,a}(t)} \end{aligned} \tag{5}$$

Then we decompose $Act_{u,p}(t)$ in the same way and observe the relationship between activeness and passive diversity with structure factors controlled. From Fig.1(c) we can see, activeness is U-shaped distributed on passive diversity for users with various activity levels. Since Beta distribution is one of the simplest distributions that can generate U-shaped curves with only two parameters[8], we adopt it as the basis for the passive part of our model.

$$\begin{aligned} r_{u,p}(F_p G_{u,p}(d_{u,p}(t)) + I_{u,p}) &= k\frac{d_{u,p}(t)^{\alpha-1}(1-d_{u,p}(t))^{\beta-1}}{\int u^{\alpha-1}(1-u)^{\beta-1}du} + b \\ G_{u,p}(d_{u,p}(t)) &\sim \frac{d_{u,p}(t)^{\alpha-1}(1-d_{u,p}(t))^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du} \end{aligned} \tag{6}$$

where $\frac{d_{u,p}(t)^{\alpha-1}(1-d_{u,p}(t))^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du}$ is the value of the Beta distribution on $d_{u,p}(t)$, with $\alpha$ and $\beta$ as parameters. Since the domain of definition for Beta distribution is $[0,1]$, $d_{u,p}(t)$ must be normalized according to the maximal diversity before used for computation.




(a) Activeness vs. Outward links (b) Activeness vs. Inward links

**Figure 5:** ***The relationships between user activeness and structural factors on a semilog scale are linear.*** **The curves in different colors represent users with different activity levels, which share the same pattern.**

The last step is to figure out the function form of the structural factors $F_u(S_u(t))$. Former researches demonstrate that the effects of the links with different direction are mutually independent[14], for which $F_u(S_u(t))$ can be rewritten as $F_u(n_{u,out}(t))F_u(n_{u,in}(t))$. We control the effect from other factors that may affect activeness as a constant $c$ and plot the relationships between activeness and the numbers of outward and inward links. As shown in Fig.5, all the curves share the same pattern, increasing linearly with the number of links at a log scale. Taking the effect of the number of outward links on activeness at active modes as an instance, we get

$$\begin{aligned} Act_{u,a}(t) &= k\log n_{u,out}(t) + b = cF_u(n_{u,out}(t)) + I_{u,a}) \\ F_u(n_{u,out}(t)) &\sim \log n_{u,out}(t) \\ F_u(S_u(t)) &\sim \log n_{u,out}(t)\log n_{u,in}(t) \end{aligned} \tag{7}$$

Substitute the results of equation 5, 6 and 7 into equation 3 and 2 and merge the constant terms into $I$, then we have the final form of our model

$$\begin{aligned} Act_u(t) = &\log n_{u,out}(t)\log n_{u,in}(t)(w_a e^{k_u d_{u,a}(t)} + \\ &w_p\frac{d_{u,p}(t)^{\alpha-1}(1-d_{u,p}(t))^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du} + b) + I \end{aligned} \tag{8}$$

**Justification of the model:**

**Active preference diversity.** $k$ captures the exponential growth mechanism of active preference diversity on activeness. Higher diversity for actively seeking content on social media is always beneficial to improve user activity, and the effect is especially significant when $k > 1$.

**Passive preference diversity.** The user activity level first decreases and then increases with passive preference diversity controlled by a Beta distribution with parameters $\alpha$ and $\beta$. According to Beta distribution properties, $\alpha$ and $\beta$ controls the low and high diversity part respectively and act symmetrically. With the decrease of diversity in the low diversity part, activeness decreases rapidly for $\alpha > 1$ and increases rapidly for $0 < \alpha \leq 1$. With the increase of diversity in the high diversity part, activeness decreases rapidly for $\beta > 1$ and increases rapidly for $0 < \beta \leq 1$.

**Structural effect.** $s$ describes the amplification effect of the number of different kinds of links. More links are always helpful to achieve higher activity levels. However, since the activity increases

with the logarithm of the number of links, this effect is feeble for users who already established lots of links of the same type.

**Personality factors.** $I$ captures the effect of personality factors that hardly change over time, such as gender and educational background. $w_a$ and $w_p$ reflect the browsing habits of different users on social media.

### 4.3 Parameter Learning

Our model consists of seven parameters: $\theta = \{w_a, w_p, k, \alpha, \beta, b, I\}$. Given a real time sequence $A_u(t)$, which represents the activeness sequences of $u$ from the beginning of the observation period $t_0$ to the end $t_e$, we use Levenberg-Marquardt (LM)[11] to minimize the sum of squared errors and learn the parameters of our model.

$$\underset{\theta}{Min}\, D(A_u, \theta) = \sum_{t=t_0}^{t_e} (A_u(t) - Act_u(t))^2 \tag{9}$$

To find a good region of parameter space and get a faster learning process, we set the initial value of parameters using some prior knowledge from empirical data. The setting methods for the initial values are shown in the Appendix.

## 5 EXPERIMENTS

In this section, we evaluate the effectiveness of our model on real data and give insights on improving user activity by parameter analysis. We first introduce the baselines used in our experiments, then show the accuracy of our model on fitting the heterogeneous and dynamic effect on user activeness. We further demonstrate the predicting power of our model on churn and activeness in the future, analyze the distribution of parameters and give insight into how to improve user activity. The datasets used in our experiments are the same as Sec.3.2.

### 5.1 Baselines for experiments

Since previous activeness models did not take the effect of preference diversity into consideration, we adopt some representative and universal activeness models as baselines to exemplify the performance of our model, described as follows:

*1)* Linear regression (LR): LR is one of the most frequently used methods, which shows the effect of regarding preference diversity as a regular feature without capturing the mechanism.

*2)* Support vector regression (SVR): SVR can capture the nonlinear correlations between features and target values. We try the frequently used kernels and regularization parameters for the optimal combination.

*3)* Our model without preference diversity (Null): We remove the preference diversity part from our model to reveal the significance of the effect of preference diversity.

*4)* Our model without passive diversity (Active): We only use the exponential growth mechanism of the active part for capturing the effect of overall diversity.

*5)* Our model without active diversity (Passive): We only use the U-shaped distribution of the passive part for capturing the effect of overall diversity.

### 5.2 Accuracy

We validate the accuracy of our model by answering whether it can capture the heterogeneous effect of preference diversity on user activeness in reality and fit the dynamic activeness of individuals. With the parameters learned through our model and baselines, we generate the distributions of activeness on preference diversity, compare them with empirical data, and try to fit the dynamic activeness for each user, which is measured by daily online time.

As shown in the upper part of Fig.6, our model successfully captures the heterogeneity of dynamic user activeness for both datasets from different platforms, superior to all the baselines. Our model regenerates the overall distribution of activeness on preference diversity, which is consistent with reality. Although individual activeness has quite multifarious temporal patterns, our model shows strong unification power and is accurate enough to capture dynamic activeness fluctuation. In contrast, the regression-based baselines can only capture the trend.

We adopt the frequently-used Kolmogorov-Smirnov statistics(KS-stat) to evaluate the performance of fitting the preference diversity effects quantitatively. It is a frequently-used method for comparing two subsets by quantifying a distance between the cumulative distribution functions of two samples. If the KS-stat is small, the hypothesis that the distributions of the two subsets are the same cannot be rejected. KS-stat is calculated by

$$ksstat = sup_x |F_1(x) - F_2(x)| \tag{10}$$

As for the metrics for fitting accuracy on dynamic individual activeness, we calculate the mean absolute percentage error (MAPE) and root mean square percentage error (RMSPE) as follows:

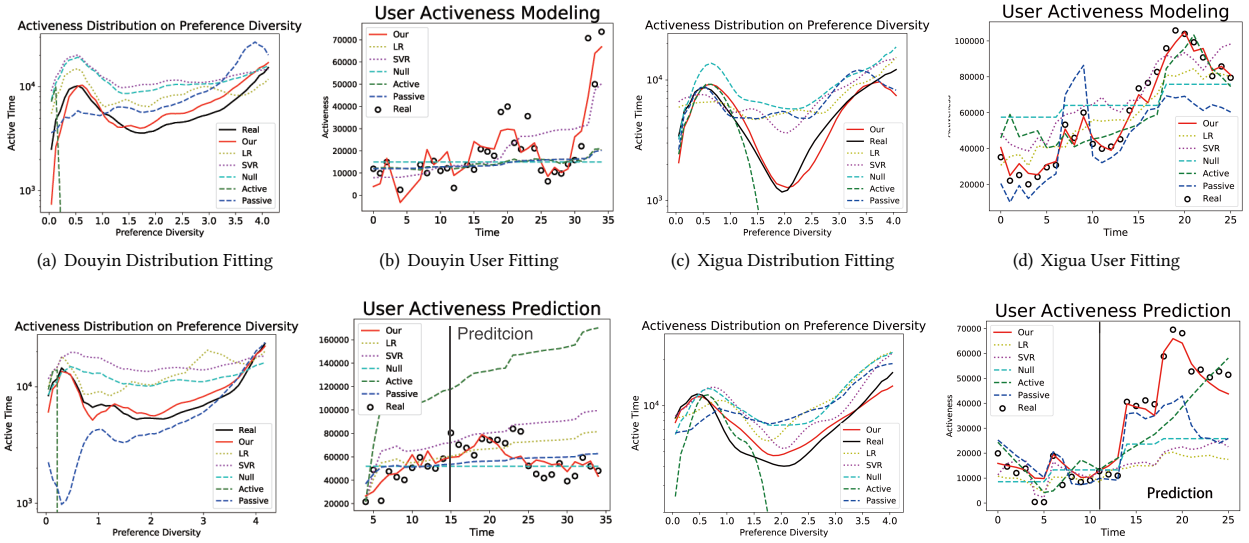
$$\begin{aligned} MAPE(u) &= \frac{1}{N}\sum_u \left(\frac{1}{t_e - t_0 + 1}\sum_{t=t_0}^{t_e} \left|\frac{A_u(t) - \widehat{A_u}(t)}{A_u(t)}\right|\right) \\ RMSPE(u) &= \frac{1}{N}\sum_u \sqrt{\frac{1}{t_e - t_0 + 1}\sum_{t=t_0}^{t_e} \left(\frac{A_u(t) - \widehat{A_u}(t)}{A_u(t)}\right)^2} \end{aligned} \tag{11}$$

where we denote the true activeness of $u$ at $t$ as $A_u(t)$ and the corresponding fitting results from models as $\widehat{A_u}(t)$. The fitting results from all the models are listed in Table 1.

**Table 1: Fitting results on empirical data from Douyin(D) and Xigua(X). Winner in bold.**

| Metrics | Our | LR | SVR | Null | Active | Passive |
|---|---|---|---|---|---|---|
| KS-stat(D) | **0.14** | 0.38 | 0.36 | 0.42 | 0.82 | 0.30 |
| KS-stat(X) | **0.17** | 0.34 | 0.32 | 0.49 | 0.66 | 0.32 |
| P-Value(D) | **0.68** | $9.7\times10^{-4}$ | $2.1\times10^{-3}$ | $1.8\times10^{-4}$ | $7.2\times10^{-16}$ | $1.7\times10^{-2}$ |
| P-Value(X) | **0.55** | 0.01 | 0.02 | $6.2\times10^{-5}$ | $1.2\times10^{-8}$ | 0.02 |
| MAPE(D) | **0.151** | 0.256 | 0.249 | 0.486 | 0.641 | 0.306 |
| MAPE(X) | **0.177** | 0.242 | 0.235 | 0.499 | 0.589 | 0.374 |
| RMSPE(D) | **0.237** | 0.331 | 0.378 | 0.623 | 0.766 | 0.417 |
| RMSPE(X) | **0.286** | 0.321 | 0.343 | 0.657 | 0.689 | 0.450 |

In KS-stat, only our model passes the test at the default 5% significance level. As for MAPE and RMSPE, the accuracy of our model




(a) Douyin Distribution Fitting (b) Douyin User Fitting (c) Xigua Distribution Fitting (d) Xigua User Fitting

(e) Douyin Distribution Prediction (f) Douyin User Prediction (g) Xigua Distribution Prediction (h) Xigua User Prediction

**Figure 6: *Our model fits and predicts the heterogeneous effects of preference diversity on user activeness pretty well for both datasets.* The upper are fitting results, and the below are prediction results. We compare the distribution of activeness generated from different models with real data on preference diversity in (a)(c) for fitting and (e)(g) for prediction. Our model can also handle the intricate dynamic patterns of individual activeness in (b)(d) for fitting and (f)(h) for prediction. The vertical bars in (f)(h) are the boundaries of training and prediction.**

has obvious advantages over the baselines that simply treat preference diversity as another feature, which demonstrates that uncovering the mechanism of preference diversity is critical for activeness modeling. The bad performance of Active is because more users spend most of their time on recommendations, which belongs to the passive mode. Although Passive performs well in fitting overall distribution, it fails to capture individual activeness patterns. Incorporating the heterogeneous and dynamic effects from both active and passive modes, our model fits various patterns accurately.

## 5.3 Prediction

As a dynamic model, our model can predict user activeness over time in the future measured by daily online time. We designed two practical tasks, churn prediction and activity prediction, to measure the predicting power of our model.

*5.3.1 Churn Prediction.* Since the retention of users is the basis of other behavior in social media, user churn is regarded as one of the most important indicators for user activeness. We define churn for a user if she/he does not log in for over 4 consecutive weeks, which is consistent with the real setting in Douyin and Xigua platform. 10937 users (8531 for Douyin and 2406 for Xigua) in our datasets churn at the end of the observation. We randomly sample an equal number of users from those who remain until the last and mix them with the churn users to make the data subset balanced.

We use the parameters of our model and the baseline models as features where the distribution for each feature was standardized. Then we conduct the Logistics Regression and SVM classifier for the classification problem of churn or not. To verify the effect of prediction, we also conduct the same classifiers on a feature set as another baseline, including age, gender, time, followers number, fans number, preference diversity, active diversity, and passive diversity. It is evident that the dimensions of this feature set are more than the parameters of our model.

**Table 2: Accuracy for churn prediction on Douyin(D) and Xigua(X) dataset. Winner in bold.**

| Methods | Our | Null | Active | Passive | Features |
|---|---|---|---|---|---|
| Logistic Regression(D) | **0.808** | 0.597 | 0.620 | 0.708 | 0.729 |
| Logistic Regression(X) | **0.741** | 0.556 | 0.572 | 0.709 | 0.723 |
| SVM Classifier(D) | **0.825** | 0.681 | 0.639 | 0.717 | 0.747 |
| SVM Classifier(X) | **0.783** | 0.605 | 0.626 | 0.748 | 0.766 |

We use 5-fold cross-validation and repeat each experiment for 5 times, and the average performances are reported in Table.2. We can see that our model outperforms all the baselines with both classifiers and achieve up to 82.5% accuracy. Although the feature dimensionality of our method is much smaller than the feature-based baseline, our method can still achieve much higher performances, demonstrating that the parameters of our dynamic model can capture more intrinsic and predictive factors of the dynamic process than manually defined features.

*5.3.2 Activity Prediction.* A more challenging and practical task to evaluate the prediction power of a model is the activity prediction problem: with the parameters learned from early-stage information, can we generate the active time of the user in the future? For Douyin(Xigua) dataset, we use the data of the first 15(11) weeks for

training, with the purpose of predicting the active time of different users in the future 20(15) weeks.

As shown in the lower row of Fig.6, our model outperforms all the baselines significantly in various user activeness patterns. Our model not only precisely forecasts the active time in the future, but also discovers the patterns that are latent in training data. For example, in Fig.6(f), this user shows a rising trend in the training data and deceives the baselines into overestimating the future activity, but our model remains accurate. As for Fig.6(h), the user never shows a tendency to be more active in the first 11 weeks, for which the regression-based baselines think this user will keep the current level of activity. However, our model makes the correct prediction based on diversity. We also generate the distribution of activeness on preference diversity in the future from our model and baselines and compare them with empirical data in Fig.6(e)(g). Only the results of our model are consistent with reality for both datasets from different applications.

**Table 3: Prediction results on empirical data from Douyin(D) and Xigua(X). Winner in bold.**

| Metrics | Our | LR | SVR | Null | Active | Passive |
|---|---|---|---|---|---|---|
| KS-stat(D) | **0.22** | 0.52 | 0.58 | 0.54 | 0.88 | 0.62 |
| KS-stat(X) | **0.24** | 0.44 | 0.40 | 0.51 | 0.68 | 0.44 |
| P-Value(D) | **0.15** | $1.2\times10^{-6}$ | $3.8\times10^{-8}$ | $4.0\times10^{-7}$ | $3.3\times10^{-18}$ | $2.9\times10^{-9}$ |
| P-Value(X) | **0.15** | $4.4\times10^{-4}$ | $2.6\times10^{-3}$ | $2.1\times10^{-5}$ | $2.8\times10^{-9}$ | $4.4\times10^{-4}$ |
| MAPE(D) | **0.204** | 0.439 | 0.425 | 0.542 | 0.710 | 0.332 |
| MAPE(X) | **0.248** | 0.367 | 0.326 | 0.505 | 0.627 | 0.401 |
| RMSPE(D) | **0.355** | 0.557 | 0.587 | 0.828 | 0.895 | 0.475 |
| RMSPE(X) | **0.396** | 0.481 | 0.448 | 0.773 | 0.729 | 0.506 |

Table 3 shows the prediction results from our model and the baselines compared with reality. Our model beats all the baselines in various metrics on distribution generation and activeness prediction, showing the impressive benefit of successfully capturing the heterogeneous and dynamic effect of preference diversity.

## 5.4 Parameter Analysis

A specific advantage of our model is that all of the parameters have clear physical meanings. Since the main novelty of our model lies in capturing the effect of preference diversity, we analyze the relevant parameters to seek insights into improving user activeness.

The probability distribution functions (PDF) of the related parameters or metrics are given in Fig.7. We first focus on the active mode. From equation 5 we know that user activeness increases exponentially with $k$ times the active diversity. From the PDF of $k$, we can see that $k > 0$ for all the users, which means that higher active diversity never hurts activeness. However, the distribution is bimodal, and the smaller peak is very close to 0, which means that the activeness of corresponding users is entirely insensitive to active diversity. Another peak position is bigger than 10, indicating the existence of users who are very sensitive to active diversity.

As for the passive diversity, which is closely related to the recommendation system, the corresponding parameters are $\alpha$ and $\beta$. According to Beta distribution properties, the PDF shows that $\alpha < 1$ for all the users, which means that low diversity always leads to higher activeness. It may be explained by the fact that recommending similar content liked by users in succession can keep them from

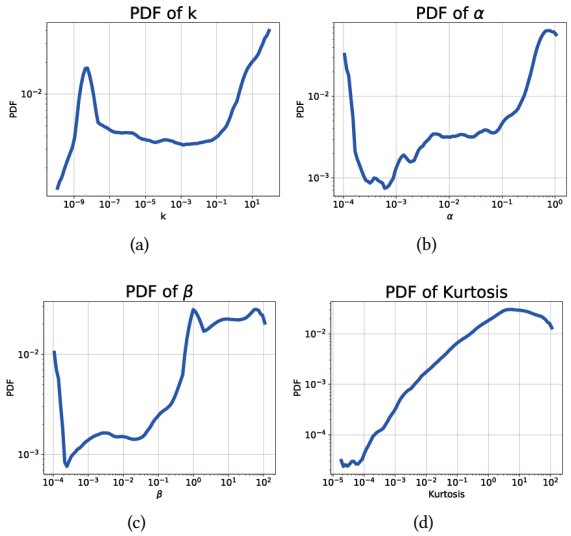



**Figure 7: *The parameter distributions of (a)k, (b)$\alpha$, (c)$\beta$ and (d)kurtosis on a log-log scale.***

leaving in the short term, which already becomes a common strategy in recommendation systems[4]. There are two peaks in the PDF of $\beta$, where one peak is around 0 while another peak is bigger than 1. It indicates that high passive diversity has completely different effects on these two types of users, leading to extraordinarily active or extremely inactive, which results from the various preference diversity of different users. This phenomenon may result from the different performances on the accuracy of the recommendation algorithm.

Since the effect of passive preference diversity is U-shaped on activeness, we further calculate its kurtosis $K$ as follow and draw the distribution of kurtosis in Fig.7(d).

$$K = \frac{6[(\alpha-\beta)^2(\alpha+\beta+1) - \alpha\beta(\alpha+\beta+2)]}{\alpha\beta(\alpha+\beta+2)(\alpha+\beta+3)} \tag{12}$$

Kurtosis measures the gap between the top and the bottom of the distribution and reflects the extent to which different preference diversity can affect activeness. As shown in Fig.7(d), there is a mass of users with kurtosis bigger than 1, which means a significant effect from activeness. With our model, the effect of preference diversity changes on activeness can be quantified by calculating the partial derivatives of $Act_u$ for $d_{u,a}$ and $d_{u,p}$

$$\begin{aligned}
\frac{\partial Act_u}{\partial d_{u,a}} &= \log n_{u,out} \log n_{u,in} w_{u,a} k_u e^{k_u d_{u,a}(t)} \\
\frac{\partial Act_u}{\partial d_{u,p}} &= \log n_{u,out} \log n_{u,in} w_{u,p} \frac{(\alpha-1)d_{u,p}^{\alpha-2} - (\alpha+\beta-2)d_{u,p}^{\alpha+\beta-3}}{\log n_{u,out} \log n_{u,in} w_{u,a}}
\end{aligned} \tag{13}$$

Positive derivatives demonstrate that higher diversity can improve activeness, while negative derivatives indicate that lower diversity leads to better user experiences. We can obtain the best strategy to improve activeness by intervening in diversity according to the values of the derivatives above.

## 6 DISCUSSIONS & CONCLUSIONS

This paper studies the effect of preference diversity on user activeness with two large-scale real-world datasets. Through comprehensive analysis with other factors controlled, we reveal the distribution and evolution patterns of preference diversity and find rich heterogeneity in its effect on user activeness, which results from the mixture of the divergent mechanisms in active and passive modes. We propose a universal mixture model to incorporate the dynamic effect of preference diversity on user activeness. To our best knowledge, this is the first quantitative model that reveals the relationship between activeness and preference diversity. Our activeness model achieves high accuracy on fitting the heterogeneity of the preference diversity effect and the dynamic activeness patterns of different users. The application value of our model lies in the strong power of churn prediction and activeness prediction, as well as the decision support for improving user activity through the intervention of preference diversity. Our findings and model fill a gap in previous researches around preference diversity, and provide potential implications for the design of future recommendation systems and social media.

To make the proposed model and the conclusions general and credible, we pay much effort to avoid bias from the recommendation system and search engine. First, we try to fit empirical data distributions with parametric models directly and then interpret them in a generative way to incorporate the effect of the latent factors in social platforms, including recommendation systems and search engines. Second, we try to minimize the bias in sampling strategy of analysis and modeling. For example, we select users new to the platforms to ensure that the recommendation systems treat them equally. Third, we conduct the analysis and experiments on datasets from two different platforms, where the traffic ratios from the recommendation system and search engine are significantly distinct. However, we find similar patterns and mechanisms, and the proposed model performs well on both, which means the recommendation system and search engine do not affect the conclusions of this paper decisively, and our model remains stable.

## Acknowledgement

This work was supported in part by National Key RD Program of China (No. 2018AAA0102004, No. 2020AAA0106300), National Natural Science Foundation of China (No. U1936219, No. 62141607), and Beijing Academy of Artificial Intelligence (BAAI). The authors thank anonymous reviewers for many useful discussions and insightful suggestions.

## REFERENCES

[1] Gediminas Adomavicius and YoungOk Kwon. 2011. Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Transactions on Knowledge and Data Engineering* 24, 5 (2011), 896–911.

[2] Laming Chen, Guoxin Zhang, and Eric Zhou. 2018. Fast greedy map inference for determinantal point process to improve recommendation diversity. In *Advances in Neural Information Processing Systems*. 5622–5633.

[3] Michael D Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. In *Fifth international AAAI conference on weblogs and social media*.

[4] Munmun De Choudhury, Scott Counts, and Mary Czerwinski. 2011. Find me the right content! Diversity-based sampling of social media spaces for topic-centric search. In *Fifth International AAAI Conference on Weblogs and Social Media*.

[5] Bruno Gonçalves, Nicola Perra, and Alessandro Vespignani. 2011. Modeling users' activity on twitter networks: Validation of dunbar's number. *PloS one* 6, 8 (2011).

[6] Qiang Guo, Lei Ji, Jian-Guo Liu, and Jingti Han. 2017. Evolution properties of online user preference diversity. *Physica A: Statistical Mechanics and its Applications* 468 (2017), 698–713.

[7] Meiqun Hu, Ee-Peng Lim, and Ramayya Krishnan. 2009. Predicting outcome for collaborative featured article nomination in wikipedia. In *Third International AAAI Conference on Weblogs and Social Media*.

[8] David Johnson. 1997. The triangular distribution as a proxy for the beta distribution in risk analysis. *Journal of the Royal Statistical Society: Series D (The Statistician)* 46, 3 (1997), 387–398.

[9] Kin-Che Lam, AL Brown, Lawal Marafa, and Kwai-Cheong Chau. 2010. Human preference for countryside soundscapes. *Acta Acustica united with Acustica* 96, 3 (2010), 463–471.

[10] Neal Lathia, Stephen Hailes, Licia Capra, and Xavier Amatriain. 2010. Temporal diversity in recommender systems. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 210–217.

[11] Kenneth Levenberg. 1944. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics* 2, 2 (1944), 164–168.

[12] Shuyang Lin, Xiangnan Kong, and Philip S Yu. 2013. Predicting trends in social networks via dynamic activeness model. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 1661–1666.

[13] Frank Liu and Paul E Johnson. 2011. News media environment, selective perception, and the survival of preference diversity within communication networks. (2011).

[14] Yunfei Lu, Linyun Yu, Peng Cui, Chengxi Zang, Renzhe Xu, Yihao Liu, Lei Li, and Wenwu Zhu. 2019. Uncovering the Co-driven Mechanism of Social and Content Links in User Churn Phenomena. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 3093–3101.

[15] Lorraine McGinty and Barry Smyth. 2003. On the role of diversity in conversational recommender systems. In *International Conference on Case-Based Reasoning*. Springer, 276–290.

[16] Judith Möller, Damian Trilling, Natali Helberger, and Bram van Es. 2018. Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society* 21, 7 (2018), 959–977.

[17] Tien T Nguyen, Pik-Mai Hui, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. 2014. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web*. 677–686.

[18] Minsu Park, Ingmar Weber, Mor Naaman, and Sarah Vieweg. 2015. Understanding musical diversity via online social media. In *Ninth International AAAI Conference on Web and Social Media*.

[19] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2011. Introduction to recommender systems handbook. In *Recommender systems handbook*. Springer, 1–35.

[20] Carlo Ricotta and Laszlo Szeidl. 2006. Towards a unifying approach to diversity measures: bridging the gap between the Shannon entropy and Rao's quadratic index. *Theoretical population biology* 70, 3 (2006), 237–243.

[21] Francisco C Santos, Marta D Santos, and Jorge M Pacheco. 2008. Social diversity promotes the emergence of cooperation in public goods games. *Nature* 454, 7201 (2008), 213–216.

[22] Vibha Singhal Sinha, Senthil Mani, and Monika Gupta. 2013. Exploring activeness of users in QA forums. In *2013 10th Working Conference on Mining Software Repositories (MSR)*. IEEE, 77–80.

[23] Chenhao Tan and Lillian Lee. 2015. All who wander: On the prevalence and characteristics of multi-community engagement. In *Proceedings of the 24th International Conference on World Wide Web*. 1056–1066.

[24] Muh-Chyun Tang. 2014. Exploring the Impact of Users' Preference Diversity on Recommender System Performance. In *International Conference on HCI in Business*. Springer, 681–689.

[25] Yichen Wang, Xiaojing Ye, Hongyuan Zha, and Le Song. 2017. Predicting user activity level in point processes with mass transport equation. In *Advances in Neural Information Processing Systems*. 1645–1655.

[26] Wen Wu, Li Chen, and Liang He. 2013. Using personality to adjust diversity in recommender systems. In *Proceedings of the 24th ACM conference on hypertext and social media*. 225–229.

[27] Bin Yin, Yujiu Yang, and Wenhuang Liu. 2014. Exploring social activeness and dynamic interest in community-based recommender system. In *Proceedings of the 23rd International Conference on World Wide Web*. 771–776.

[28] Chengxi Zang, Peng Cui, and Wenwu Zhu. 2018. Learning and Interpreting Complex Distributions in Empirical Data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*. ACM, 2682–2691.

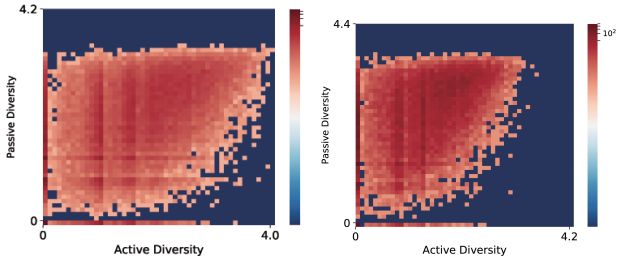
[29] Chengxi Zang, Cui Peng, Christos Faloutsos, and Wenwu Zhu. 2017. Long Short Memory Process: Modeling Growth Dynamics of Microscopic Social Connectivity. In *Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*.

[30] Yin Zhu, Erheng Zhong, Sinno Jialin Pan, Xiao Wang, Minzhe Zhou, and Qiang Yang. 2013. Predicting user activity level in social networks. In *Acm International Conference on Information and Knowledge Management*.

# A APPENDIX

## A.1 More Analysis on Preference Diversity

*A.1.1 Distribution.* We explore the correlation between active and passive diversity by plotting the empirical joint distribution in Fig.8. Compared with Fig.2, we find that the joint distribution is very close to the product of the distributions of these two kinds of diversity respectively, which indicates that active and passive diversity are mutually independent and should be modeled separately.




(a) Douyin (b) Xigua

**Figure 8: *The empirical joint distribution of active and passive diversity is very close to the product of the two marginal distributions of them.* A brighter color indicates a higher density of users.**

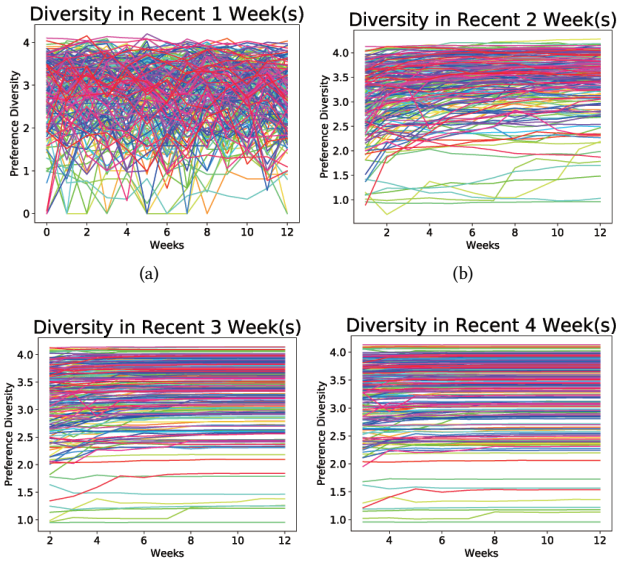
*A.1.2 Evolution.* We randomly select 200 users who are new to the social platform and plot their preference diversity with time for 3 months in Fig.9. For most users, the preference diversity in the short term (e.g., one week) is volatile, but the long-term diversity in more than three weeks is quite stable. The same conclusion applies to both active and passive diversity, and we omit the corresponding curves for short.

*A.1.3 Relations to Structural Factors.* In Fig.10, we illustrate the heat map of average activeness on the joint distribution of diversity and the number of outward links (e.g., followers) or inward links (e.g., fans). We can see the shape of the distribution is very close to the product of the effects from links and preference diversity (see section 4.2), which means the effect of both kinds of links are mutually independent with both kinds of diversity, for which the factors from structure and diversity also influence activeness independently.

## A.2 Parameter Learning

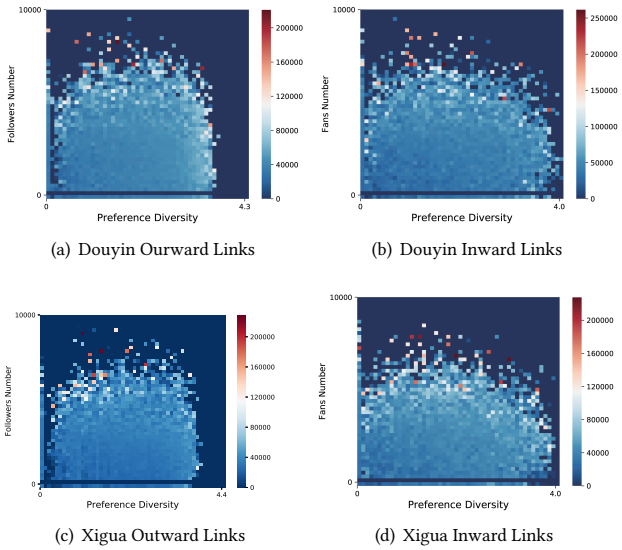
Our model consists of seven parameters: $\theta = \{w_a, w_p, k, \alpha, \beta, b, I\}$. For each parameter the setting method for its initial value is as follows:

- $w_a$ and $w_p$ are set according to the proportion of time the user spends watching videos in active modes and passive modes respectively.
- $k$ is set as 0 or 10, which are the two peaks of the distribution of $k$ according to Fig.7. For users who spend more time in active modes, we choose 10 as the initial value, otherwise, we choose 0.
- $\alpha$ and $\beta$ are set according to the preference diversity of the user. The higher the diversity, the lower the initial value of $\alpha$, and the greater the initial value of $\beta$.




(a) (b) (c) (d)

**Figure 9: *The evolution curves of the preference diversity of 200 users, where each curve represents a user. The x-axis is time, and the y-axis is the preference diversity in recent one week (a), two weeks (b), three weeks (c), and four weeks (d).***




(a) Douyin Ourward Links (b) Douyin Inward Links

(c) Xigua Outward Links (d) Xigua Inward Links

**Figure 10: *The heat maps of average activeness on the joint distribution of preference diversity and the number of links.* (a)(c) Preference diversity vs. Outward links. (b)(d) Preference diversity vs. Inward links. Brighter colors indicate higher average activeness.**

- $b$ and $I$ are set according to the learning results of our null model, which only contain these two parameters. See the experiment section.