When Variety-Seeking Meets Unexpectedness: Incorporating Variety-Seeking Behaviors into Design of Unexpected Recommender Systems

Pan Li[1]& Alexander Tuzhilin[2]

Abstract

Variety seekers are those customers who easily get bored with the products they purchased before, and therefore prefer new and fresh content to expand their horizons. Despite its prevalence, variety-seeking behavior has hardly been studied in recommendation applications due to various limitations in existing variety-seeking measures. To fill the research gap, we present a variety-seeking framework in this paper to measure the level of variety-seeking behavior of customers in recommendations based on their consumption records. We validate the effectiveness of our framework through user questionnaire studies conducted at Alibaba, where our variety-seeking measures match well with consumers' self-reported levels of their variety-seeking behaviors. Furthermore, we present a recommendation framework that combines the identified variety-seeking levels with unexpected recommender systems in the data mining literature, to address consumers' heterogenous desire for product variety, where we provide more unexpected product recommendations to variety-seeking consumers, and vice versa. Through offline experiments on three different recommendation scenarios and a large-scale online controlled experiment at a major video-streaming platform, we demonstrate that those models following our recommendation framework significantly increase various business performance metrics and generate tangible economic impact for the company. Our findings lead to important managerial implications to better understand consumers' variety-

[1] Scheller College of Business, Georgia Institute of Technology, 800 W Peachtree Street NW, Atlanta, GA, 30332, Email: pli95@gatech.edu

[2] Stern School of Business, New York University, 44 West 4th Street, New York, NY, 10012, Email: at2@stern.nyu.edu

seeking behaviors and design recommender systems. As a result, the best-performing model in our proposed frameworks has been deployed by the company to serve all consumers on the video streaming platform.

Keywords: Variety-Seeking, Recommender System, Unexpected Recommendations

1 Introduction

Variety-seeking characterizes consumers' motives to explore products they have not thought about before, when they get tired of their customary purchased products (McAlister & Pessemier 1982). For example, thriller-lovers may switch to a romantic comedy after binge-watching multiple thrillers on a Friday night, even though their movie interests are predominately confined to thrillers. It constitutes an important dimension of exploratory consumer behavior, as varied experiences provide stimulation to reduce user boredom (Faison 1977; Bench & Lench 2019), satisfy innate human curiosity (Raju 1980), and improve consumer satisfaction with their purchases (Ratner et al. 1999). Variety seekers also tend to increase their overall consumption quantity (Kahn & Wansink 2004; Read et al. 1995) and are more open to promotions (Ailawadi et al. 2001), therefore constituting an important segment of consumers in marketing applications.

While variety-seeking has been studied extensively in marketing (Zeithammer & Thomadsen 2013; McAlister & Pessemier 1982; Kahn et al. 1986), it has been noticeably underexplored in the field of recommender systems due to the following issues. First, existing variety-seeking measures, such as those presented in (Kim et al. 2002; Trijp et al. 1996), only operate at the category or brand levels using classical feature-based techniques, when computing the differences between consumed products. In contrast, it is crucial for modern recommendation methods (Adomavicius & Tuzhilin 2005; Zhang et al. 2019) to capture the heterogeneity of consumer preferences at the (fine-grained) individual product level, which is typically done in the latent space of product embeddings (Covington et al. 2016). Traditional explicit feature-based

models are also not scalable to most industrial platforms (Hinton & Salakhutdinov 2006), resulting in high latency and performance downgrades (Covington et al. 2016). Second, while the importance of time-varying factors in modeling consumer variety-seeking behaviors has been demonstrated in (Helsen & Schmittlein 1993; Braun & Moe 2013; Alba et all. 1992), existing variety-seeking models do not consider the dwell time information between purchasing actions, leading to significantly less effective measures. Third, existing marketing methods hardly studied long-term variety-seeking properties, such as stationarity, making it difficult to extract and generalize useful behavioral patterns (Gorgoglione et al. 2019).

To address the aforementioned issues and incorporate variety-seeking behaviors into the design of recommender systems, we propose a variety-seeking framework that specifies the class of effective variety-seeking measures of each consumer based on consumption records, without requiring explicit consumer feedback on their desire for product variety, as was typically done in existing methods (Baumgartner & Steenkamp 1996; McAlister & Pessemier 1982; Kahn et al. 1995). This proposed framework consists of three key components: a distance function between consumed products, a time-decay function specifying how quickly past consumption memories fade, and the stationarity property of variety-seeking behaviors. When we make specific assumptions about the exact nature of these components, we obtain a specific variety-seeking measure for our framework that corresponds to these assumptions. We demonstrate through a questionnaire study that they provide significant performance improvements over existing measures (Givon 1984; Gullo et al. 2019) in modeling consumers' variety-seeking behaviors.

Furthermore, in this paper, we connect the desire of consumers to seek product variety with the paradigm of unexpected recommender systems (Adamopoulos & Tuzhilin 2014) that simultaneously

provide novel and satisfying recommendations to them. Note that the concept of unexpectedness comes from the data mining literature (Silberschatz & Tuzhilin 1996; Padmanabhan & Tuzhilin 1998), and measures how distant the recommended product is from consumer expectations in a *product-centric* manner, whereas the concept of variety-seeking comes from the marketing literature, and measures the consumer propensity to seek for significantly different content, especially unexpected products, in a *consumer-centric* manner. Therefore, we hypothesize in this paper that those two concepts are *complementary* to each other and need to be properly combined to achieve optimal recommendation performance. In particular, we demonstrate that those consumers with a high level of variety-seeking behavior prefer more unexpected products, and we need to increase the degree of unexpectedness in recommendations accordingly to accommodate their desire, and vice versa. Therefore, we propose a recommendation framework that automatically adjusts the degree of unexpectedness in recommendations according to the level of variety-seeking of each consumer, which significantly enhances the level of personalization in unexpected recommender systems (Adamopoulos & Tuzhilin 2014). Under the proposed framework, we construct a series of recommendation models with different operationalizations that all lead to significant performance improvements over the existing unexpected recommender systems (Adamopoulos & Tuzhilin 2014; Li & Tuzhilin 2020), which we demonstrate through offline experiments conducted on three datasets from Yelp, MovieLens, and Alibaba respectively. Furthermore, we conduct a large-scale online controlled experiment at a major video streaming platform in China, where we compare the best-performing model in our frameworks with the latest production system. The results demonstrate significant business performance improvements and lead to tangible economic impact for the company, both in the short-term and the long

run. Therefore, we bridge the gap between academic research on variety-seeking behaviors and real-world applications in need of fulfilling consumers' desire for product variety and improving business performance.

In this paper, we make the following research contributions. First, we propose a variety-seeking framework that measures various aspects of the variety-seeking behaviors of consumers in recommender systems. Second, we propose a recommendation framework that combines the concepts of unexpectedness and variety-seeking to address the heterogeneous desires of consumers for product variety in recommendations. Finally, we construct multiple variety-seeking-based recommendation models fitting these frameworks and demonstrate through a mixture of user questionnaire studies, offline experiments, and online controlled experiments that these models achieve significant performance improvements over the state-of-the-art solutions described in the marketing and CS literature, and the latest production system in the company, leading to actionable managerial implications on how to effectively incorporate variety-seeking behaviors into modern recommendation platforms. The significant economic impact of our proposed frameworks has led the company to deploy our best-performing variety-seeking model into production, serving consumers on the entire video streaming platform.

2 Literature Review

2.1 Variety-Seeking Behavior

Variety seeking represents the consumer behavior to select novel and diversified products (Ratner et al. 1999; Read & Loewenstein 1995) to fulfill their curiosity (Fiske & Maddi 1961). These behaviors can be driven by self-motivation for stimulation (McAlister & Pessemier 1982; Steenkamp & Baumgartner 1992) or situational factors (Kahn & Isen 1993; Menon & Kahn 1995), such as social desirability (Ratner and Kahn 2002) and compromised personal privacy (Levav & Zhu 2009). In addition, consumers tend to seek

more variety in their consumption after being exposed to a sequence of novel products (Huang and Wyer 2015; Xu et al. 2014), leading to increased overall consumption quantity (Kahn & Wansink 2004; Read et al. 1995) and openness to promotions (Ailawadi et al. 2001). Researchers have also studied the impact of variety-seeking behaviors in terms of intensity, brand preferences, and consumer surplus (Bawa 1990; Feinberg et al. 1992; Woratschek & Horbel 2006; Seetharaman & Che 2009; Sajeesh & Raju 2010).

Due to all these important benefits, extensive research has stemmed from both marketing and IS communities around the topic of variety-seeking that is related to our work. Specifically, variety-seeking behaviors of blog readers have been identified by (Singh et al. 2014), as they dynamically switch from reading one set of topics to another. Researchers have shown that consumers prefer more diversified mobile apps when making download decisions due to their variety-seeking behaviors (Lee et al. 2020). The positive effects of product tags and socially endorsed information on consumers' perceived serendipity have also been studied (Cheng et al. 2017), which encourages consumers to conduct more serendipitous searches. The downstream effects of variety-seeking on product demand distribution have also been analyzed in (Tan et al. 2017; Fong 2017). In addition, the nonconscious effects of consistency-seeking, the opposite of variety-seeking, have also been explored in sequential consumer decisions (Fishbach et al. 2011).

Meanwhile, variety-seeking remains underexplored in recommender systems, resulting in suboptimal performance and repeated types of product recommendations that do not take into account consumers' desire for product variety. In this paper, we identify several dimensions for measuring the variety-seeking level of each consumer, and incorporate such information into the design of unexpected recommender systems. Our method is motivated by the theoretical model of variety-seeking in (Wayne & Nancy 1984),

where researchers hypothesize purchasing exploration to be an interaction between individual-level characteristics (e.g. exploration motivation) and product-level characteristics (e.g. product content or Hedonic/Utilitarian characteristics (Li et al. 2020a)). We extend the theoretical analysis to the application of recommender systems, where we focus on the interaction between the consumer-centric variety-seeking behavior and the product-centric unexpectedness property. The idea of providing more unexpected recommendations to variety-seekers is also motivated by the existing literature (Menon & Kahn 1995; Maimaran & Wheeler 2008), where researchers have shown through lab experiments that consumers' need for stimulation can be met by providing more variety to them. We extend the results in lab experiments and propose to provide variety in a personalized manner through the concept of unexpectedness in data mining, resulting in significant performance improvements. Our research sheds light on the business impact of addressing consumers' desire for product variety, as we demonstrate through offline and online experiments.

2.2 Unexpectedness and Related Concepts in Recommender System

Recommender systems provide numerous economic benefits across various industries (Hosanagar et al. 2014; Senecal & Nantel 2004; Panniello et al. 2016). However, typical methods recommend only similar products (Adomavicius & Tuzhilin 2005) while ignoring the dispersion of consumer preference (Givon 1984) and raising the problem of over-specialization and user boredom (Adamopoulos & Tuzhilin 2014; Li & Tuzhilin 2020) that negatively affect model performance. To address these issues, data mining researchers identified the concept of unexpectedness (Kaminskas & Bridge 2016) as a powerful tool to tackle the problem of exploration-exploitation (Schwartz et al. 2017; Zhang et al. 2020). Unexpectedness-based methods identify those products that depart from consumers' expectations to meet their satisfaction (Adamopoulos & Tuzhilin 2014; Li & Tuzhilin 2020). Specifically, they deploy a hybrid utility function

consisting of the relevance and unexpectedness objectives, while the degree of unexpectedness is only manually determined as a fixed value for all consumers (Adamopoulos & Tuzhilin 2014; Li & Tuzhilin 2020). Therefore, consumers' heterogeneous propensity towards unexpected products and their variety-seeking levels are not taken into account, resulting in deteriorating consumer experiences (Chen et al. 2019), as some consumers prefer to stay within their "comfort zones" and to receive familiar recommendations.

In this paper, we focus on modeling the concept of variety-seeking in recommender systems in tandem with unexpectedness. Specifically, we propose a recommendation framework that incorporates variety-seeking behavior to determine the degree of unexpectedness in recommendations. By doing so, we significantly improve the level of personalization in unexpected recommendations and address consumers' heterogeneous desire to seek product variety. We will now present our variety-seeking framework.

3 Variety-Seeking Framework

One of the key considerations in measuring the level of variety-seeking behavior of a consumer is his or her propensity to explore new products and to seek significantly different content (Givon 1984; McAlister & Pessemier 1982), where the differences between the currently selected and the previously chosen products are defined in terms of a distance function that can be introduced in various ways, as will be explained below. Another fundamental assumption in our framework is that the difference between two products x and y consumed at the corresponding periods t_x and t_y is becoming less relevant if time interval $|t_x - t_y|$ increases, since consumers "forget" their experiences from the distant past and are less motivated to seek products different from the past as time goes by. For example, if a person consumed a pasta dish in an Italian restaurant a while ago, she/he is more willing to eat another pasta dish, compared to

the case when the person just had it last night. In other words, the second component is the time-decay function that monotonically contracts differences between consumed products as the time interval increases. Finally, to assign a certain level of variety-seeking measure to a particular consumer, we assume that this propensity should be stable in the long run and, therefore, the process of seeking product variety in recommendations should be stationary over time, albeit experiencing small changes in the short term.

Mathematically, these components can be formally captured as follows: if consumer i purchased products $\{i_1, i_2, \dots, i_k\}$ at time $\{t_1, t_2, \dots, t_k\}$, then the variety of product j at time t is defined as:

$$Product\_Variety(i, j, t) = \sum_k \mu(t - t_k) * \rho(i_k, j) \qquad (1)$$

where $\mu(\cdot)$ is a time-decay function, $\rho(\cdot\,,\ \cdot)$ is a distance function measuring the differences between two products, and i_k represents the k-th product consumed at time t_k. As explained before, when we examine how different the previously consumed products are from product j (that is consumed at time t) based on equation (1), these differences should be *stationary* and do not depend on particular product j or time t, since variety-seeking is a property of consumer i and should be stable over time. Although this stationarity assumption can be modeled in several ways, the most natural approach would be to take the average value of product variety levels over all previous products as follows: (n is the number of consumptions)

$$Variety\_Seeking(i) = \frac{1}{n}\sum_{k=1}^{n} Product\_Variety(i, i_k, t_k) \qquad (2)$$

To summarize, our variety-seeking framework consists of the following three components that collectively define the concept of variety-seeking of each consumer in equation (2):

- **Distance function $\boldsymbol{\rho}(.,\ .)$** that measures the differences between recommended and consumed products.
- **Time decay function $\boldsymbol{\mu}(\cdot)$** that specifies the phenomenon that consumers "forget" about similarities

between the previously consumed products over time.

- **Stationarity assumption**, where we assume the process of seeking product variety should be an intrinsic characteristic of each consumer and therefore should be stationary over time.

When we make specific assumptions about the exact nature of these three components, in equation (2), we obtain the specific models of variety-seeking measures corresponding to our variety-seeking framework. We visualize our framework in Figure 1 and will now describe these three specific components in detail.

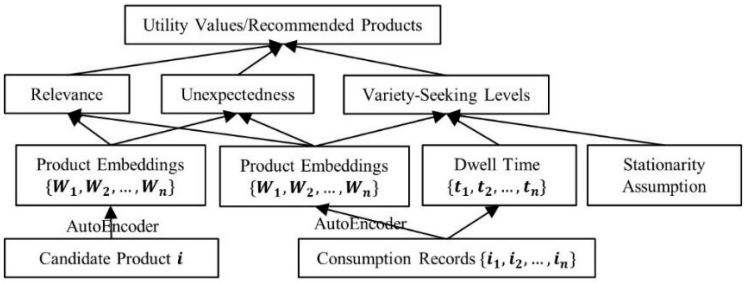


Figure 1: Diagram of Our Variety-Seeking Measurement and Recommendation Frameworks

3.1 Distance Function

The idea of computing distances between various products has been studied in the variety-seeking literature (McAlister & Pessemier 1982), where the level of variety-seeking is computed as a binary value, depending on whether the candidate product appears in the consumption record before or not (Givon 1984; Kahn et al. 1986). However, it does not take into account the degree of differences between various products, as some products can be very similar to each other while others are not. Therefore, researchers proposed to measure the level of variety-seeking as the proportion of explicit features with same values between two products among all feature dimensions (Trijp et al. 1996; Gullo et al. 2019). These feature-based measures managed to achieve significant performance improvements vs. the binary method (Kim et al. 2002).

In this paper, we present another method to measure distances in the *latent* space using the deep-learning-based method (Zhang et al. 2019) to model variety-seeking in a more nuanced way than previous

distances in the feature space. It also effectively captures heterogeneous and complex relationships along different feature dimensions (He et al. 2017; Zhang et al. 2019), and automatically determines the relative importance of them when comparing the differences. These tasks, however, are generally hard for feature-based methods (Boatwright et al. 2008; Sahoo et al. 2012). It also consolidates high-dimensional explicit features into low-dimensional latent embeddings (Hinton & Salakhutdinov 2006), making computations much more efficient while preserving the "essence" of feature information.

While a wide range of latent representation models has been proposed in the CS literature, we focus on the AutoEncoding (AE) (Hinton & Salakhutdinov 2006) model in this paper, as it is the most popular method deployed in industrial platforms (Zhang et al. 2019), such as Alibaba (Li et al. 2020) and Amazon (Hardesty 2019). It is also flexible, scalable, and memory-efficient, making it easier to incorporate into recommendation designs (Zhang et al. 2019), such as variety-seeking in our paper. The AE model learns two separate neural networks simultaneously: the encoder network $F_{encoder}$, which maps explicit features into latent representations; and the decoder network $F_{decoder}$, which reconstructs explicit features from latent representations. These two networks are jointly optimized by minimizing the reconstruction loss for explicit features x: $L_{AE}(x) = F_{decoder}(F_{encoder}(x))$. The product representations are then obtained by applying the encoder network: $W_x = F_{encoder}(x)$. Therefore, to compute the differences between products x and y, we only need to compute the distance between their latent representations $d(W_x, W_y)$, which is formulated as the most popular Euclidean distance, or other alternative distance metrics in the latent space.

To summarize, the distance function between new and previously consumed products is an important component for modeling the level of variety-seeking, which can be defined either as the classical feature-

based distances, or through the Euclidean or other types of distances between latent product representations. As we demonstrate in Section 3.5, the latent Euclidean distance function performs significantly better than other distance metrics, as it fits well with the Euclidean space of product embeddings and is most suitable for recommendation tasks, as shown in the literature (Covington et al. 2016; Zhang et al. 2019).

3.2 Time-Decay Function

We will now introduce another important dimension in the variety-seeking framework - the time decay function $\mu(\cdot)$, which specifies the phenomenon that consumers "forget" about similarities between the previously consumed products over time, as has been the case in other marketing applications, such as advertising (Braun & Moe 2013) and product sales (Helsen & Schmittlein 1993). The most popular time-decay function for consumer modeling is the *exponential decay* function (Helsen & Schmittlein 1993), which follows the Proportional Hazard model (PHM) and the Accelerated Failure Time model (AFT) that apply an exponential penalty for time-related covariates (Chintagunta 1998). Other methods include the Hyperbolic Discounting (Labison 1997; Machado & Sinha 2007) and Additive Risk (Seetharaman 2004) models, which use the hyperbolic function to model time-decay effects. These time-decay functions model dwell time information and obtain a more effective estimation of variety levels of new products as a result.

Meanwhile, time-varying factors are not properly taken into account in prior variety-seeking models (Trijp et al. 1996; Kim et al. 2002), which is unfortunate as they play important roles in shaping consumer online experience, such as accumulating consumer dissatisfaction with repeated choices over time (LaBarbera & Mazursky 1983). In particular, consumer decisions are significantly affected by contextual, time-related factors in recommender systems (Panniello et al. 2016), such as the dwell time between two purchase actions. We demonstrate through the user questionnaire analysis in Section 3.5 that the time-decay

function $\mu(\cdot)$ is an integral component for measuring variety-seeking behavior in recommendations.

3.3 Stationarity of Variety-Seeking Behavior

Finally, we introduce the last dimension in our variety-seeking framework – the stationary property. As discussed earlier in Section 3, it is crucial to guarantee stationarity of the variety-seeking measure over time, since variety-seeking is an intrinsic characteristic of a consumer and, therefore, should be stable. While many plausible variety-seeking statistics can be applied to the set of product variety levels to match with the stable variety-seeking behavior within our framework, the arithmetic mean statistic stands out, and we use it in this paper because of its stationarity property that we empirically validate in the paper and has also been demonstrated in (Gullo et al. 2019). Moreover, several existing marketing studies (Trijp et al. 1996; Kim et al. 2002; Gullo et al. 2019) have also used the arithmetic mean when defining variety-seeking.

Specifically, we formulate the stationarity hypothesis stating that the variety-seeking behavior of a consumer is stable over time in the long run as a stationary time series, assuming that product variety is defined by equation (1). While it is a simplifying hypothesis and not the only plausible way to specify the variety-seeking measure, it constitutes a practical and reasonable assumption, as it matches consumer behavior patterns observed in our studies and manages to provide strong performance results with minimal computational costs, as we empirically validate in this section on three offline datasets and online controlled experiment. We will discuss other methods, such as dynamic variety-seeking measures as our future work.

To start with, we analyzed the dynamic patterns of variety in the video streaming platform studied in our online experiment, where the average mean and variance of product variety levels are 0.2387 and 0.011 respectively, demonstrating little change in variety-seeking behaviors. We also randomly selected 10,000 consumers following the same demographic distribution of the entire consumer population on the platform,

and categorize them into variety-seekers and consistency-seekers, based on whether the variety-seeking level is above average or not. We compute the product variety level of last 10 recently completed videos and first 50 completed videos since they enter the platform, and then plot the average computed values in Figure 2. We observe that there are indeed some fluctuations in the initial part of the viewing history, but not among the later stage or the last 10 completed videos where consumers have watched sufficient amounts of videos and their variety-seeking patterns have converged, making sense for us to determine the variety-seeking level of each consumer through the arithmetic mean of product variety levels in the consumptions.

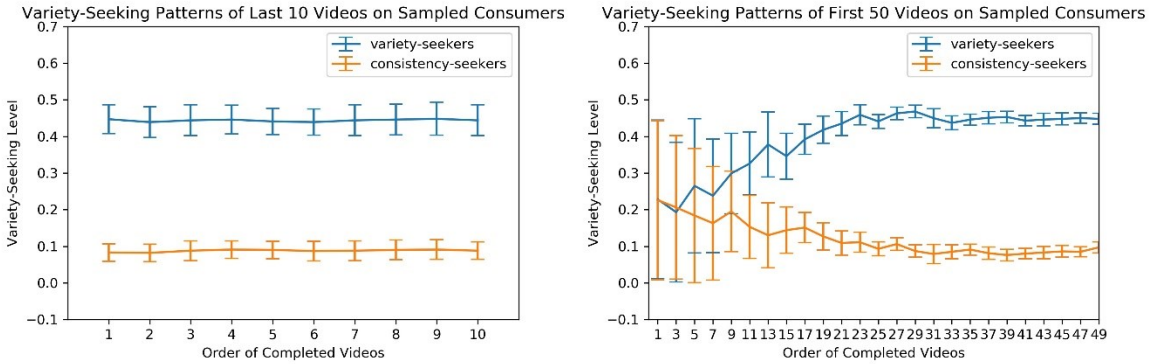


Figure 2: Product variety levels (with std) in first 50 and last 10 completed videos of sampled consumers

In addition to the direct observations, we also conducted the Augmented Dickey–Fuller test (ADF) (Dickey & Fuller 1979) to test for the null hypothesis that a unit root is present, while the alternative hypothesis is that the time series of product variety levels is stationary. We utilize the data collected from three offline experiments (Yelp, MovieLens, Alibaba) and the online experiment (Company A), where the ADF test is conducted separately in the pre-treatment and post-treatment periods to ensure that consumers experience the same recommendation model. As we present in Table 1, for those variety-seeking measures formulated under the variety-seeking framework that we summarize in Section 3.4, the average test statistics are all statistically non-significant at the 95% confidence level. Therefore, the null hypothesis is

rejected, and we verify that the product variety levels are indeed stationary in our offline and online experiments. We also observe from Table 1 that the time-decay function is an integral component of our framework, without which the stationarity property would not hold, as consumers' memory fades over time and they might not remember past experiences with products purchased a long time ago.

Variety-Seeking Framework	Yelp	MovieLens	Alibaba	Online Exp
Binary+Exponential+Mean	-7.62	-17.88	-9.69	-11.03
Binary+Hyperbolic+Mean	-5.44	-13.67	-8.84	-9.74
Binary+No Decay+Mean	-1.12*	-1.55*	-0.96*	-1.68*
Feature+Exponential+Mean	-6.97	-13.66	-8.87	-12.55
Feature+Hyperbolic+Mean	-5.65	-11.79	-8.07	-10.06
Feature+No Decay+Mean	-1.17*	-1.84*	-1.33*	-2.35*
Euclidean+Exponential+Mean	-61.62	-75.33	-77.64	-35.12
Euclidean+Hyperbolic+Mean	-47.35	-71.04	-51.28	-27.68
Euclidean+No Decay+Mean	-2.95*	-3.12*	-2.07*	-2.99*
Cosine+Exponential+Mean	-52.77	-67.94	-73.65	-31.07
Cosine+Hyperbolic+Mean	-43.36	-62.49	-52.97	-26.95
Cosine+No Decay+Mean	-2.99*	-3.03*	-2.06*	-2.94*
Manhattan+Exponential+Mean	-17.61	-57.96	-61.45	-12.65
Manhattan+Hyperbolic+Mean	-10.86	-41.17	-38.86	-7.74
Manhattan+No Decay+Mean	-3.13*	-3.28*	-2.77*	-3.32*
Chybeshev+Exponential+Mean	-15.55	-52.88	-57.95	-10.88
Chybeshev+Hyperbolic+Mean	-8.82	-43.79	-34.41	-6.26
Chybeshev+No Decay+Mean	-3.30*	-3.04*	-2.96*	-3.21*

*Table 1: ADF test statistics of the variety-seeking levels under our proposed framework. *p<0.05 (the critical value used to determine the statistical significance at the 95% confidence level is -3.50)*

To summarize, we empirically demonstrated the stationarity property of our modeling approach to product variety, enabling us to compute the product variety function of a consumer in a practical and useful manner. We would also like to emphasize that this stationarity property does not focus exclusively on the arithmetic mean statistic and may include various alternative statistics to measure variety-seeking, such as weighted mean, geometric mean, harmonic mean, and median, as long as they fit in the stationarity property. However, the arithmetic mean statistic constitutes one simple, reasonable, and effective option that produces better performance over other alternative statistics, as we demonstrate in the Appendix (Part III).

3.4 Summary of the Variety-Seeking Framework

We can now summarize our variety-seeking framework defined by equations (1) and (2) as:

- **the distance function**, which can be selected either as a binary or feature-based distance following existing variety-seeking literature (Trijp et al. 1996; Kim et al. 2002), or as Euclidean or other types of distances in the latent space, as discussed in Section 3.1.
- **the time-decay function**, which can be modeled as Exponential Decay, Hyperbolic Discounting (Helsen & Schmittlein 1993), or other functions discussed in Section 3.2.
- **the stationarity assumption**, which implies that the variety-seeking propensity is a fundamental characteristic of a consumer and, therefore, should be stable over time. Our framework assumes the deployment of any alternative statistic as long as it fits with the stationary property.

These three components of the variety-seeking framework are summarized in Table 2. Each specific option of components in Table 2 leads to a particular variety-seeking measure. For example, we can assume that $\boldsymbol{\rho}$ is the Euclidean distance function in a latent space, the time decay is exponential, and the summary statistic is the arithmetic mean, which leads us to a particular variety-seeking measure of the consumers. Moreover, we demonstrate in Section 3.5 that this particular measure captures consumer desire for variety most accurately and generates the best performance across several alternative models.

Distance Function $\boldsymbol{\rho}(\cdot\,,\,\cdot)$	**Time-Decay Function $\boldsymbol{\mu}(\cdot)$**	**Stationarity Property/Summary Statistics**
•Binary Distance •Feature-based Distance •Distance in Latent Space ▪ Euclidean Distance ▪ Other Distances •Other Distance Functions	•Exponential Decay •Hyperbolic Discounting •Other Time-Decay Functions	•Summary Statistics Satisfying Stationarity Property (e.g. Arithmetic Mean) • Other Summary Statistics

Table 2: Summary of our Variety-Seeking Framework

Finally, our variety-seeking framework stems from existing variety-seeking literature (Trijp et al. 1996; Zeithammer & Thomadsen 2013; Gullo et al. 2019) and significantly advances them from the following three perspectives. First, the framework supports multiple functions to measure distances between the products in the recommendation context, both in the feature and latent spaces, thus leading to a broad set of choices for the distance function. Second, we highlight the importance of adopting the time-decay function to model time-varying factors in the consumer decision-making process, which has been largely ignored in the literature. Finally, we empirically validate the stationarity assumption of variety-seeking behavior in recommender systems, which enables us to determine the variety-seeking level through the arithmetic mean. Under our variety-seeking framework, we will be able to construct several specific variety-seeking models for designing recommender systems, which we will elaborate on in the next section.

3.5 Validation of the Variety-Seeking Framework

To further demonstrate the validity of our variety-seeking framework, we follow the standard marketing practice and conduct the consumer questionnaire analysis (Van & Steenkamp 1992; Wang & Huang 2018) by utilizing a user survey dataset collected by Alibaba (Chen et al. 2019) over three weeks starting from 12/21/2017 to 01/11/2018. Specifically, 2,401 consumers (1,651 females vs. 750 males) on the grocery shopping platform have participated in the survey that evaluates their instant feedback towards the recommended products. Their responses have been carefully checked to make sure there are no invalid records (such as consistently responding with the same answer to all questions or missing answers to some questions). In addition, an analysis of their historical activities over the past three months showed that all of them are well familiar with the online recommendation platform of Alibaba, as they had all clicked at least one recommended product before taking the survey, while 98.5% of them had more than 100 clicks.

More specifically, each participant will first be presented with a recommended product (generated by the company from one of the product domains "clothes", "toys", "home appliances", and "foods") together with its name, image, short description, and price, and then asked to complete a set of questions shown in the Appendix (Part I) based on a five-point Likert scale (i.e., 1 "extremely agree" and 5 "extremely disagree") to assess her/his feedback on this recommendation. One of the questions, "The item recommended to me is different from the types of products I bought before", is directly related to the variety level, as previous marketing literature (Wang & Huang 2018; Yoon & Kim 2018) have adopted similar types of language to evaluate product variety levels. They will then be shown the next recommended product and asked to respond to the same set of questions again until the end of the experiment (they can resume from the point where they left off the previous session), where they will be asked to provide background information and fill out the psychological curiosity quiz of "Ten-item Curiosity and Exploration Inventory-II" (Kashdan et al. 2009) shown in the Appendix (Part I) to determine their curiosity or variety-seeking levels. Finally, all participants were placed in a lottery draw for customized awards as an incentive. The important statistics of consumer responses are reported in Table 3, and other details are reported in the Appendix (Part I).

Survey Question & Response	Mean	Std.	Median	Skewness	Kurtosis
Product Variety Level *"The item recommended to me is different from the types of products I bought before."*	3.39	1.215	4.00	-0.400	-0.813
Variety-Seeking Level *"Ten-item Curiosity and Exploration Inventory-II"*	3.13	0.831	3.10	0.088	-0.402

Table 3: Important Statistics of Consumer Response to the Questionnaire.

For each recommended product, we compare its product variety value reported by the consumer through the questionnaire and the value computed by $Product\_Variety(i,j,t)$ in equation (1). We also compare each consumer's variety-seeking level reported through the curiosity quiz and the value computed

by $Variety\_Seeking(i)$ in equation (2) and two baseline models from the marketing literature: Additive Parametric Function (APF) (Givon 1984; Kim et al. 2002) and Product Assortment Size (PAS) (Trijp et al. 1996; Gullo et al. 2019). Results presented in Table 4 show that those variety-seeking measures constructed under our framework obtain the highest correlation with consumers' self-reported product variety levels and variety-seeking levels, and the improvements over APF and PAS are statistically significant across various configurations. We also identify the best-performing model "Euclidean+Exponential+Mean", which selects the Euclidean distance in the latent space as the distance function, the exponential decay as the time-decay function, and the arithmetic mean statistic to construct a consumer variety-seeking measure. Finally, we observe from Table 4 that if we remove the time-decay function, the resulting variety-seeking measures will not perform well, indicating the importance of time-varying factors in modeling variety-seeking behavior. We also study in the Appendix (Part II) the classification performance of consistency-seeking vs. variety-seeking consumers, where we observe similar levels of performance improvements.

Variety-Seeking Framework	$Product\_Variety(i,j,t)$	$Variety\_Seeking(i)$
Euclidean+Exponential+Mean	**0.775*** **	**0.618*** **
	(0.004)	**(0.003)**
(%Improved)	**+16.77%**	**+8.58%**
Euclidean+Hyperbolic+Mean	0.761***	0.612***
Euclidean+No Decay+Mean	0.658*	0.540
Cosine+Exponential+Mean	0.758***	0.609***
Cosine+Hyperbolic+Mean	0.756***	0.607***
Cosine+No Decay+Mean	0.659*	0.541
Manhattan+Exponential+Mean	0.696***	0.588***
Manhattan+Hyperbolic+Mean	0.685***	0.584***
Manhattan+No Decay+Mean	0.647	0.538
Chebyshev+Exponential+Mean	0.693***	0.589***
Chebyshev+Hyperbolic+Mean	0.686***	0.589***
Chebyshev+No Decay+Mean	0.645	0.537
Feature+Exponential+Mean	0.688***	0.591***
Feature+Hyperbolic+Mean	0.692***	0.589***
Feature+No Decay+Mean	0.645	0.565
Binary+Exponential+Mean	0.667**	0.562
Binary+Hyperbolic+Mean	0.665**	0.558

Binary+No Decay+Mean	0.631	0.532
APF	0.631	0.532
PAS	0.645	0.565

*Table 4: Pearson Correlation Coefficients between Consumers' Self-Reported Variety-Seeking Levels and Our Variety-Seeking Framework. *p<0.1; **p<0.05; ***p<0.01. (compared to APF & PAS)*

To summarize, we demonstrate that our proposed variety-seeking measures defined by equations (1) and (2) correlate well with variety-oriented measures from the survey described in this section. We also demonstrate that each dimension in our framework, namely the distance function, time-decay function, and stationarity assumption, is important when modeling the level of variety-seeking behavior of each consumer.

4 Recommendation Framework

4.1 The Utility Function Design

Based on the variety-seeking framework, we will now focus on building the recommendation framework that incorporates the variety-seeking levels of each consumer in the design of the utility function. In classical recommendation models (Adomavicius & Tuzhilin 2005), the utility value is solely determined by the relevance objective for each product j and consumer i: $Utility(i,j) = Relevance(i,j)$, as the goal is to identify the most relevant types of products for consumers. However, doing so might fail to address consumers' desire for novel content in provided recommendations (Adamopoulos & Tuzhilin 2014), and it is crucial to also take into account the unexpectedness objective to expand consumers' horizons (Chen et al. 2019): $Utility(i,j) = Relevance(i,j) + \alpha \times Unexpectedness(i,j)$, where the value of α controls for the degree of unexpectedness and is typically selected as a fixed value for all consumers on the platform (e.g., see (Li & Tuzhilin 2020)). Note, however, that the variety-seeking levels can vary significantly across different *consumers* based on the discussions in Section 3, as some of them are more adventurers while others have less propensity for desiring new experiences. In contrast, unexpectedness is the property of

individual *products* measuring their deviations from consumer expectations. Therefore, the concepts of unexpectedness and variety-seeking are *complementary* to each other, in the sense that the variety-seeking level of the consumer can be used to determine the degree of unexpectedness in recommendations to that consumer, as we provide more unexpected recommendations to consumers with high variety-seeking levels and are more eager to explore novel content, and vice versa. Specifically, these two concepts are integrated into one unified recommendation framework following the multi-objective optimization paradigm:

$$Utility(i,j) = Relevance(i,j) + f(Variety\_Seeking(i), Unexpectedness(i,j)) \quad (3)$$

where $f(.,.)$ is the aggregation function that models the complementary relationship. Recommendations are subsequently produced by selecting products with the highest utility values. This framework enables us to address consumers' heterogeneous desire for product variety and improve satisfaction as a result. We also do not need to go through the complex process to determine the value of α described previously, since it will be automatically determined by variety-seeking measures, making it more manageable and practical.

$\boldsymbol{Relevance(i,j)}$	$\boldsymbol{Unexpectedness(i,j)}$	**Aggregation Function** $\boldsymbol{f(.,.)}$
•Neural Collaborative Filtering (NCF) •Deep Interest Network (DIN) •Other relevance-focused methods	•Feature-based Methods •Latent Representation Methods •Other unexpectedness methods	•Multiplication Function •Exponential Function •Power Function •Other Aggregation Function

Table 5: The Variety-Seeking Recommendation Framework.

We summarize a series of configuration options in Table 5 along the Relevance, Unexpectedness, and Aggregation Function dimensions. The relevance dimension assumes recommendation methods that focus exclusively on providing relevant content, such as state-of-the-art methods of Neural Collaborative Filtering (NCF) (He et al. 2017) and Deep Interest Network (DIN) (Zhou et al. 2018) that have achieved great success

in practice. The second dimension of the framework focuses on the unexpectedness objective that can be formulated through classical feature-based methods (Adamopoulos & Tuzhilin 2014) or latent modeling (Li & Tuzhilin 2020) deployed by Alibaba (Li et al. 2020). Finally, the aggregation function constitutes the third dimension, and it can be selected as the Multiplication function $Variety\_Seeking(i) \times Unexpectedness(i,j)$, the Exponential function $e^{Variety\_Seeking(i)} \times Unexpectedness(i,j)$ or the Power function $Unexpectedness(i,j)^{Variety\_Seeking(i)}$. We demonstrate in Section 5 that all these options in Table 5 lead to effective unexpected recommender system designs that significantly outperform existing methods, and that the combination of deep interest network for the relevance objective, latent modeling of the unexpectedness objective, and the multiplication function leads to the best performance.

To summarize, our proposed recommendation framework is significantly different from the existing unexpected recommender systems (Adamopoulos & Tuzhilin 2014b; Li et al. 2020), as we determine the degree of unexpectedness in the utility function through the variety-seeking levels identified from the framework that we present in Section 3, resulting in significantly better performance in real business applications vis-à-vis state-of-the-art baselines and the latest production system in Company A.

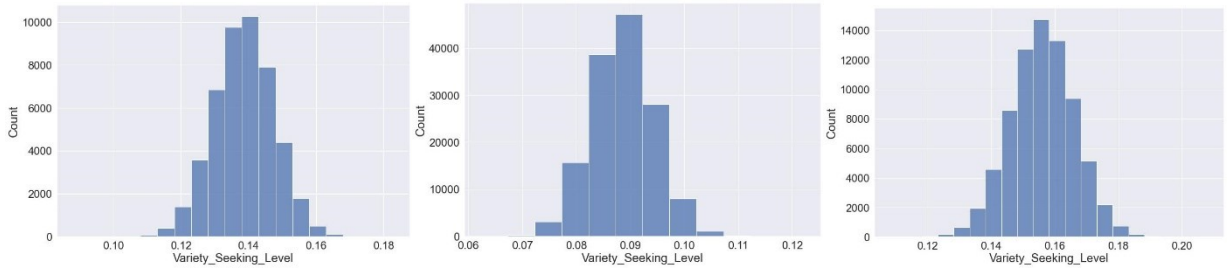
4.2 Validation of the Recommendation Framework – Offline Experiments

In this section, we consider several models fitting the framework and test their performance in the click-through rate prediction task, which is the most correlated with the business revenues generated in the recommendation platform (Zhou et al. 2018). We test on three offline datasets collected from industrial platforms of Yelp, MovieLens, and Alibaba. Each dataset contains the IDs of consumers and products, and the timestamp of purchasing actions. For the Alibaba dataset, we also have binary labels of whether consumers click on the recommended product or not. For the Yelp and MovieLens datasets, however, we

only obtain ratings (scale of 1-5) towards the recommended product, and we simulate consumer response by transforming ratings into binary labels using the threshold of 3.5, following the common practice in the recommender system literature (Zhang et al. 2019; Li et al. 2020). We also test for the alternative thresholds of 2.5 and 3, and obtain the same set of empirical findings. To further test the robustness of our proposed framework, we create three subsets of the Alibaba dataset with different sparsity levels and summarize them in the Appendix (Part VI). Note that the three offline datasets listed in Table 6 represent three vastly different business applications of catering, movie streaming, and short videos, with significant differences in the distribution of variety-seekers demonstrated in Figure 3(a-c), where the x-axis represents the "bins" of variety-seeking levels, and the y-axis represents the number of consumers in each bin. Consequently, results on these datasets significantly enhance the generalizability of our findings across different applications.

Dataset	Yelp	MovieLens	Alibaba
#Consumers	76,564	138,493	46,143
#Products	75,231	15,079	53,657
#Transactions	2,254,589	19,961,113	1,806,157
Sparsity	0.039%	0.956%	0.073%

Table 6: Descriptive Statistics of three Offline Datasets



(a) Yelp Dataset *(b) MovieLens Dataset* *(c) Alibaba Dataset*

Figure 3: Distribution of Variety-Seeking Level in three Offline Datasets

We compare the performance with the following four groups of eight state-of-the-art baselines:

(1) **Relevance-Oriented Recommendation Models**, including DIN (Zhou et al. 2018) and DeepFM (Guo et al. 2017), where we optimize only for the relevance objective: $Utility(i,j) = Relevance(i,j)$.

(2) **Unexpectedness-Oriented Recommendation Models,** including HOM-LIN (Adamopoulos & Tuzhilin 2014) and PURS (Li et al. 2020), where α in the utility function $Utility(i,j) = Relevance(i,j) + \alpha \times Unexpectedness(i,j)$ is determined without taking into account the variety-seeking levels.

(3) **Diversity-Oriented Recommendation Models**, including Re-Ranking (Adomavicius & Kwon 2011) and DPP (Chen et al. 2018) that focus on diversity, rather than unexpectedness in recommendations.

(4) **Bandit-Learning Recommendation Models,** including LinUCB (Li et al. 2010) and COFIBA (Li et al. 2016), which explore consumer preference in recommendation through bandit models.

As we discussed in Section 3, the "Euclidean+Exponential+Mean" method captures variety-seeking behaviors most effectively among all models fitting the variety-seeking framework, and we adopt it to measure $Variety\_Seeking(i)$ in our recommendation framework. To ensure a fair comparison, we use the same set of hyperparameter optimization techniques of Bayesian Hyperparameter Optimization (Feurer et al. 2015) to identify the optimal configurations for our models and all baselines. As a result, the autoencoder in Section 3.1 is constructed using the MLP network with 1 input layer, 3 hidden layers, and 1 output layer, and [256, 128, 16, 128, 256] units in each layer respectively. The neural network parameters are initialized with the Gaussian distribution of mean 0 and standard deviation 0.01, and then optimized by Stochastic Gradient Descent (SGD) with a learning rate 0.001. We normalize $Relevance(i,j)$, $Variety\_Seeking(i)$ and $Unexpectedness(i,j)$ to between -1 and 1 to facilitate the recommender system training process (Covington et al 2016), without distorting differences in the ranges of values or losing information.

4.3 Offline Experiment Results

We evaluate the recommendation performance using the standard ML-based recommendation metrics **AUC** and **Hit Rate@10** (Shani & Gunawardana 2011). The offline experiments are conducted following

time-stratified 5-fold cross-validation, and we report the average performance over multiple runs in Table 7. We observe from the table that all the recommendation models under our recommendation framework significantly outperform all other baselines in terms of evaluation metrics AUC and Hit Rate@10 across all three datasets. On average, we observe an increase of 3.81% for the AUC metric and 3.81% for the HR@10 metric for our proposed models, as compared to the second-best baseline approach. In particular, we identify one specific model "DIN+Latent+Multiply" that works most effectively, where we compute the relevance objective using the DIN model, formulate the unexpectedness objective using latent modeling, and combine variety-seeking with unexpectedness using the multiplication function. These performance improvements are not only statistically significant, but also demonstrate tangible performance gains in terms of the best practices in the recommender system industry (Hardesty 2019). In addition, the results in the Appendix (Part VI) confirm that our proposed framework still performs significantly better across all three Alibaba datasets with different sparsity levels and different consumption quantities for each consumer.

To summarize, as three offline datasets represent vastly different business applications, sparsity levels, and variety-seeker distributions, these results demonstrate the effectiveness, generalizability, and external validity of our recommendation framework. Specifically, we show that combining variety-seeking and unexpectedness is beneficial and works well in practice – by providing more unexpected recommendations to variety-seekers and vice versa, we significantly improve the recommendation performance.

	Yelp		MovieLens		Alibaba	
	AUC	HR@10	AUC	HR@10	AUC	HR@10
DIN+Latent+Multiply	**0.7071*** **	**0.7096*** **	**0.8375*** **	**0.7004*** **	**0.7349*** **	**0.7730*** **
	(0.0071)	**(0.0073)**	**(0.0103)**	**(0.0092)**	**(0.0088)**	**(0.0089)**
(%Improved)	**+5.18%**	**+4.95%**	**+3.52%**	**+3.33%**	**+2.73%**	**+3.15%**
DIN+Latent+Exponential	0.6973***	0.7001***	0.8317***	0.6949***	0.7328***	0.7701***
DIN+Latent+Power	0.6932***	0.6980***	0.8288***	0.6952***	0.7324***	0.7688***
DIN+Feature+Multiply	0.6817**	0.6974***	0.8291***	0.6930***	0.7299***	0.7672***

DIN+Feature+Exponential	0.6810**	0.6982***	0.8269***	0.6937***	0.7303***	0.7654***
DIN+Feature+Power	0.6804**	0.6971***	0.8269***	0.6941***	0.7291***	0.7658***
NCF+Latent+Multiply	0.6776*	0.6950***	0.8208**	0.6873***	0.7266**	0.7649***
NCF+Latent+Exponential	0.6790**	0.6977***	0.8172*	0.6855**	0.7261**	0.7610**
NCF+Latent+Power	0.6794**	0.6956***	0.8189**	0.6852**	0.7249**	0.7587**
NCF+Feature+Multiply	0.6755*	0.6943***	0.8192**	0.6839*	0.7228**	0.7599**
NCF+Feature+Exponential	0.6753*	0.6928***	0.8157*	0.6851**	0.7237**	0.7576**
NCF+Feature+Power	0.6771*	0.6936***	0.8170*	0.6851**	0.7240**	0.7573**
DIN	0.6694	0.6702	0.7021	0.6485	0.6957	0.6972
DeepFM	0.6396	0.6682	0.7056	0.6169	0.5519	0.5164
PURS	0.6723	0.6761	0.8090	0.6778	0.7154	0.7494
HOM-LIN	0.6287	0.6490	0.7177	0.5894	0.5812	0.5493
Re-Ranking	0.6295	0.6502	0.7236	0.6468	0.6025	0.5776
DPP	0.6448	0.6575	0.7490	0.6551	0.6517	0.7026
LinUCB	0.6324	0.6373	0.6883	0.6363	0.6365	0.6525
COFIBA	0.6411	0.6417	0.7162	0.6485	0.6471	0.6913

*Table 7: Offline results on the three industrial datasets. *p<0.1; **p<0.05; ***p<0.01. Improvement percentages were reported over the second-best baseline models (underlined).*

5 Online Controlled Experiments

5.1 Institutional Background

To further demonstrate the economic benefits and practical impact of our proposed framework, we conduct a large-scale online controlled experiment at a major video streaming company in China (denoted as Company A). On the streaming platform of Company A, users cannot follow any specific accounts or other users, search for specific videos, or share content. Videos are uploaded by users and distributed solely by the personalized recommendation service. Users receive video recommendations immediately when they open the App, and they can either click on the recommended videos or seek a new set of recommendations by scrolling down and refreshing the page. In that sense, consumers have little control over their video exposures, which reduces the moderating effect of self-selection bias, and enables us to estimate the average treatment effects through a simple regression in our experiment. While the video streaming platform at Company A is different from other platforms, such as TikTok or YouTube, and we might not be able to directly extend the business impact to the entire video streaming industry, the online controlled experiment

was conducted to demonstrate the advantages of our proposed frameworks and models fitting them for a leading recommendation platform at scale (Company A has over 500 million monthly active users and 800 million daily video views and has developed a powerful recommendation service over the past decade).

The online experiment was conducted over a full month in September 2020, and included 37,965,781 video-watching records by 444,765 users on 8,442,402 videos. The duration of one month is considered to be long-term at Company A where the vast majority of A/B tests are done over only one week, and is also sufficient to demonstrate the treatment effect of recommender system design for Company A which runs thousands of A/B tests per year. We have confirmed with the company that no other A/B test overlapped with our focal experiment. During the experiment, we record consumer and video features that we summarize in Table 8 and in the Appendix (Part IV), which constitute exogenous factors that might affect consumer responses and increase the estimation variance. We will now introduce our identification strategy.

Category	Variable	Description	Format
Consumer Features	Gender	Gender of the consumer	Categorical
	Age	Age of the consumer	Numerical
	Province	The province where the consumer lives in	Categorical
	City	The city where the consumer lives in	Categorical
	Operating System	The operation system on the consumer's device	Categorical
	VIP Status	Subscription to the premium service or not	Categorical
	Active Days	The number of days that the consumer has logged into the platform over the past month	Numerical
	Confidential Features	Confidential features developed by the company to describe consumer behaviors/preferences	Confidential
Video Features	Genre	The genre of the video	Categorical
	View Count	Total view number over the past month	Numerical
	Comment Count	Total comment number over the past month	Numerical
	Release Days	Days since it has been released on the platform	Numerical
	Video Length	The duration of the video	Numerical

	Confidential Features	Confidential features developed by the company to describe the video's content	Confidential

Table 8: Summary of consumer and video features recorded in the online experiment. String variables (e.g., city) will be converted to categorical features as dummy variables

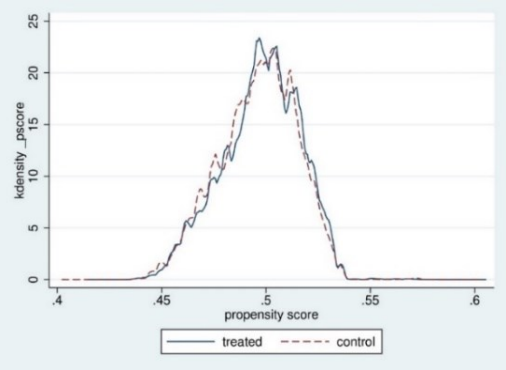


Figure 4: Comparison of propensity score distribution between the treatment group and the control group

5.2 User Splitting and Identification Strategy

We identify the Average Treatment Effects (ATEs) by diverting the video-watching requests from consumers (Kohavi et al. 2009) following binary hashing (Salakhutdinov and Hinton 2009) over user ids in the experiment pool, which is also the standard practice at Company A — if the hash index is 0, the user will be diverted to the control group and receive recommendations from the latest production system in the company described in (Li et al. 2020): $Utility(i,j) = Relevance(i,j) + \alpha \times Unexpectedness(i,j)$; otherwise, the user will be diverted to the treatment group and receive recommendations from the best-performing model "DIN+Latent+Multiply" under our framework: $Utility(i,j) = Relevance(i,j) + Variety\_Seeking(i) \times Unexpectedness(i,j)$, where $Variety\_Seeking(i)$ is computed through the best-performing measure "Euclidean+Exponential+Mean" under our variety-seeking framework. In our experiment, product embeddings and $Variety\_Seeking(i)$ are computed offline, and updated on a daily basis, while $Relevance(i,j)$ and $Unexpectedness(i,j)$ are updated in real-time to reflect dynamic consumer preferences. Similar to the offline experiments, we also normalize the scale of $Relevance(i,j)$, $Variety\_Seeking(i)$ and $Unexpectedness(i,j)$ to between -1 and 1 to facilitate the training process. This

user-splitting strategy keeps the balance between the two groups, as we demonstrate in Figure 4, where the differences between the propensity score distribution are negligible. Users are also unaware of the assignment in our experiment as Graphical User Interface (GUI) remains the same. Therefore, we validate the randomized setting of our experiment, which enables us to assess ATEs directly through OLS regression.

We subsequently conduct the two-sample hypothesis test by comparing users' responses and business performance respectively among the two groups. Based on the practical guidelines of Company A, the following three performance metrics are the most important business revenue indicators of the video streaming services: (a) CTR (Click-Through Rate), the binary variable that indicates whether the user has clicked on the recommended video or not; (b) VV (Video View), the binary variable that indicates whether the user has finished watching the recommended video or not, and (c) TS (Time Spent), the continuous variables that record the dwell time the user has spent on the recommended video, and it will be 0 if the user hasn't clicked. For each user i and video j, we specify the ATEs using the following identification:

$$Metric_{ij} = \alpha_0 + \alpha_1 * Treatment_i + \overrightarrow{\alpha_2} * \overrightarrow{X_i} + \overrightarrow{\alpha_3} * \overrightarrow{Y_j} + D_t + \varepsilon_{ij} \quad (4)$$

where $Metric_{ij} \in \{CTR_{ij}, VV_{ij}, TS_{ij}\}$, $Treatment_i$ is the dummy variable, $\overrightarrow{X_i}$ represents user features, $\overrightarrow{Y_j}$ represents video features, and D_t represents time-fixed effects including dates and hours. Our findings still hold if we adopt various types of alternative identification methods, as we show in Section 5.8.

5.3 Average Treatment Effect

We start with the regression analysis that directly compares video-watching behaviors between two user groups. We observe in Table 9 that consumers who are assigned to be served by our proposed model are 2.29% more likely to click on the recommended videos and 4.56% more likely to finish watching them, compared to the latest production model in Company A. In addition, our model increases the average time

spent on each recommended video by 39.219 seconds. These results indicate that by incorporating consumers' variety-seeking behavior into the unexpected recommender system, our proposed framework significantly increases video consumption in Company A (having $p < 0.01$ in all experimental settings). We have also provided the empirical analysis at the user level and the difference-in-difference analysis in the Appendix (Part V), where we observe significant performance improvements across all the cases.

	CTR_{ij}	VV_{ij}	TS_{ij}
$Treatment_i$	0.0229***	0.0456***	39.219***
	(0.0043)	(0.0028)	(0.9895)
User & Video Features	Yes	Yes	Yes
Time Fixed Effect	Yes	Yes	Yes
R-Squared	0.0081	0.0044	0.2133
Observations	37,965,781	37,965,781	37,965,781

*Table 9: Average Treatment Effect of Variety-Seeking Based Unexpected Recommendations. The table shows a regression with robust standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01*

These improvements, generating a very significant economic impact for the video streaming platform of Company A, are not at all surprising. By taking into account heterogeneous variety-seeking levels, instead of a fixed term α, we provide personalized and properly balanced unexpected recommendations for targeted consumers according to their variety-seeking propensities, which leads to a significant increase in business performance in the treatment group. Our findings demonstrate substantial potential to increase revenues of Company A, which is in fact one of the largest improvements that the engineering team has observed during the entire 2020. In addition, as the video streaming platform of Company A achieved 8,728 million RMB revenue in the fiscal year of 2020, our model would potentially bring an additional 30 million USD revenue to the company, based on the 2.29% CTR improvement that is directly related to the platform profits. These significant economic benefits are achieved with only a 1.3% increase in the serving latency and a 0.5% increase in memory usage compared to the existing model, which is negligible according to the engineering

team. We have also conducted additional product-level analysis, churn rate analysis, and variety-seeking behavior analysis shown in Appendix (Part VIII – Part X) to further demonstrate the robustness and impact of our method. Based on the results of this A/B test and the resulting significant business improvements, Company A has already deployed our model to serve customers on the entire video streaming platform.

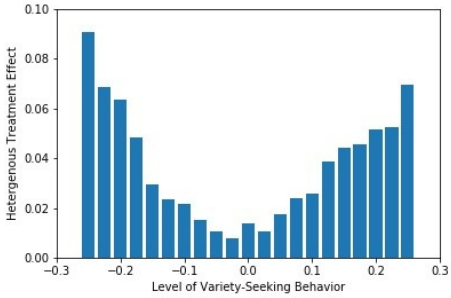


Figure 5: HTEs towards the business metric "CTR" over different levels of variety-seeking behavior. We have witnessed the same phenomena for the other two business metrics as well.

5.4 Heterogeneous Treatment Effects over Variety-Seeking Behavior

In this section, we will show that the performance improvements are not uniform for all consumers, but are rather heterogenous across different consumer groups according to their variety-seeking levels, which we categorize into 20 bins, namely [-0.250, -0.225], [-0.225, -0.200], …, [0.225, 0.250]. We select -0.25 and 0.25 as the boundary threshold for this analysis, since there are only less than 10 consumers in our records, whose level of variety-seeking is outside of this range. We subsequently identify three business metrics within each bin by adding the interaction term $Treatment_i \times Variety\_Seeking_{ij}$ and estimate its coefficients accordingly. The results are shown in Figure 5, where we can make the following observations. First, all consumers in the treatment group, regardless of their variety-seeking levels, enjoy a significant positive effect if served by our variety-seeking recommender system. Second, those consumers who are either truly variety-seeking or are strongly opposed to receiving variety in their recommendations, obtain even greater positive effects from our model. This result is natural, since our model provides more

unexpected recommendations to variety-seekers and less to consistency-seekers, which resonates with both groups, as our study confirms. By addressing the heterogeneous desire for product variety, we avoid the mistakes of providing too similar products for variety-seekers or too irrelevant products for consistency-seekers, thus improving business performance. And finally, those consumers having a medium level of variety-seeking desire still marginally benefit from our model, though the benefits are not as great as for those who strongly prefer or are against variety.

5.5 Parallel Trend Analysis

In this section, we conduct the parallel trend analysis in Figure 6, where we observe that there are no statistical differences between the two groups in the pre-treatment period and that our proposed model consistently and significantly outperforms the latest production system during the post-treatment period. Specifically, our model achieves significant performance improvements during the first week of deployment, partly due to the novelty effect. While the improvements deteriorate slightly in week 2, they still remain significant and no further performance decreases have been observed after week 2 until the end of the experiment. This observation demonstrates the strong performance of our method in the long term, which is not significantly affected by consumer curiosity or short-term factors. We have also conducted additional difference-in-difference analysis in the Appendix (Part V) to further justify our empirical findings.

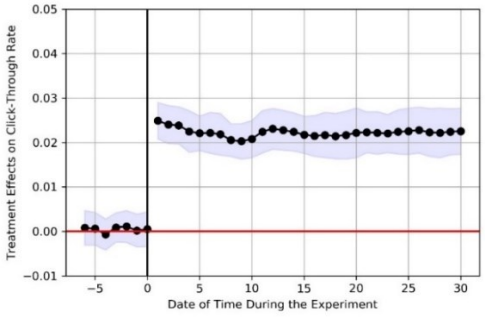


Figure 6: Parallel Trend Analysis of the Treatment Effect on Click-Through Rate (and the

Confidence Interval) in Our Online Experiment on a Daily Basis

5.6 Robustness Check

We also conduct additional experiments to check the robustness of the results, where we replicated our analysis under the following settings: (a) we include different combinations of consumer features, video features, and time-fixed effects in the regression model of equation (4) to evaluate the treatment effects; (b) we use alternative models to specify the binary outcome variables CTR_{ijt} and VV_{ijt}, including the discrete choice models of Logit and Probit; (c) we exclude records of video content uploaded by Company A itself; (d) we drop the records from those users in the regions where the platform was launched recently. The detailed results listed in the Appendix (Part VII) demonstrate that our empirical findings still hold under all these conditions, further illustrating the benefits and robustness of our proposed frameworks in this paper.

6 Conclusions

Variety-seeking plays a significant role in modeling consumers' intentions and understanding their behaviors, and we need to address consumer desire for product variety in recommendations. To this end, we first propose a variety-seeking framework, where we identify three key dimensions to measure consumer variety-seeking levels: the distance function, the time-decay function, and the stationarity assumption that are cohesively combined into the framework. We subsequently propose a recommendation framework where we utilize the identified variety-seeking levels to determine the degree of unexpectedness in the utility function for providing recommendations. By doing so, we can produce more unexpected products for variety-seekers and more familiar types of products for those consumers who prefer to stay within their own “comfort zones”, thus improving consumer satisfaction and business performance significantly.

To demonstrate the validity and effectiveness of our proposed frameworks, we conduct extensive offline

experiments and a large-scale online controlled experiment at a major video streaming platform in China. We demonstrate that by incorporating variety-seeking behavior into the design of unexpected recommender systems, we significantly increase the quantity of video consumption compared to the latest production model deployed at the company. We further demonstrate that the improvements in business performance are not homogenous for all consumers: those consumers who either strongly prefer or dislike product variety would receive the greatest benefits from our model. Nevertheless, our model has a significant impact on all consumers on the platform, as it provides them with additional variety of fresh video content, while still delivering useful recommendations and improving consumer online experience. Due to the strong economic effect demonstrated at Company A, our model has been deployed to serve consumers on the entire platform.

Note that horizontal variety is frequently observed to occur in industries with high rates of consumption, especially entertainment products (Kim et al. 2002), on which we largely focus in this paper. To further understand the consumers' desire for variety-seeking, we plan to also model and incorporate vertically differentiated variety-seeking behavior into the design of recommender systems in future work. In addition, as our analysis focuses primarily on the video streaming platform at Company A, we plan to conduct similar online experiments on other platforms, such as TikTok and YouTube, and also in other business applications, including music, movies, TV shows, and books. Furthermore, we plan to study more complex formulations of the variety-seeking measures, such as dynamic measures, to relax the stationarity assumption. Finally, we plan to utilize the proposed frameworks to shed light on which video genres or categories will be more appealing to the consumers, for example, based on their Hedonic/Utilitarian characteristics (Li et al. 2020a), which would help us to better understand the heterogeneous treatment effect from the product side.

References

Adamopoulos, P., & Tuzhilin, A. 2014. On unexpectedness in recommender systems: Or how to better expect the unexpected. ACM Transactions on Intelligent Systems and Technology (TIST), 5(4), 1-32.

Adomavicius, G. and Kwon, Y., 2011. Improving aggregate recommendation diversity using ranking-based techniques. IEEE Transactions on Knowledge and Data Engineering, 24(5), pp.896-911.

Adomavicius, G., & Tuzhilin, A. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE TKDE, 17(6), 734-749.

Ailawadi, K., Neslin, S. & Gedenk, K., 2001. Pursuing the value-conscious consumer: store brands versus national brand promotions. The Journal of Marketing, 65(1), pp.71–89.

Alba, J.W., Marmorstein, H. and Chattopadhyay, A., 1992. Transitions in preference over time: The effects of memory on message persuasiveness. Journal of Marketing Research, 29(4), pp.406-416.

Baumgartner, H. & Steenkamp, J., 1996. Exploratory consumer buying behavior: Conceptualization and measurement. International Journal of Research in Marketing, 13, pp.121–137.

Bench, S.W. and Lench, H.C., 2019. Boredom as a seeking state: Boredom prompts the pursuit of novel (even negative) experiences. Emotion, 19(2), p.242.

Boatwright, P., Kalra, A., & Zhang, W. 2008. Research Note—Should Consumers Use the Halo to Form Product Evaluations? Management Science, 54(1), 217-223.

Braun, M. and Moe, W.W., 2013. Online display advertising: Modeling the effects of multiple creatives and individual impression histories. Marketing science, 32(5), pp.753-767.

Chen, L., Yang, Y., Wang, N., Yang, K. and Yuan, Q., 2019, May. How serendipity improves user satisfaction with recommendations? A large-scale user evaluation. In WWW Conference (pp. 240-250).

Chen, L., Zhang, G. and Zhou, H., 2018, December. Fast greedy map inference for determinantal point process to improve recommendation diversity. In NeurIPS 2018 (pp. 5627-5638).

Cheng, Y, ZJ Jiang, I Benbasat (2017) "Designing for Diagnosticity and Serendipity: An Investigation of Social Product-Search Mechanisms," Information Systems Research 28(2).

Chintagunta, P.K., 1998. Inertia and variety seeking in a model of brand-purchase timing. Marketing Science, 17(3), pp.253-270.

Covington, P., Adams, J. and Sargin, E., 2016, September. Deep neural networks for youtube recommendations. In Proceedings of 10th ACM conference on recommender systems (pp. 191-198).

Dickey, D.A. and Fuller, W.A., 1979. Distribution of the estimators for autoregressive time series with a unit root. Journal of the American statistical association, 74(366a), pp.427-431.

Faison, E.W., 1977. The neglected variety drive: A useful concept for consumer behavior. Journal of consumer research, pp.172-175.

Feurer, M., Springenberg, J. and Hutter, F., 2015, February. Initializing bayesian hyperparameter optimization via meta-learning. In AAAI Conference on Artificial Intelligence (Vol. 29, No. 1).

Fishbach, A., Ratner, R.K. and Zhang, Y., 2011. Inherently loyal or easily bored?: Nonconscious activation of consistency versus variety-seeking behavior. Journal of Consumer Psychology, 21(1), pp.38-48.

Fong, NM (2017) "How Targeting Affects Customer Search: A Field Experiment," Management Science 63(7).

Givon, M. Variety seeking through brand switching. 1984 Marketing Science, 3(1), 1-22.
Gorgoglione, M., Panniello, U. and Tuzhilin, A., 2019. Recommendation strategies in personalization applications. Information & Management, 56(6), p.103143.
Gullo, K., Berger, J., Etkin, J. and Bollinger, B., 2019. Does time of day affect variety-seeking?. Journal of Consumer Research, 46(1), pp.20-35.
Hardesty, L., 2019. The history of Amazon's recommendation algorithm. Amazon Science, 22.
Helsen, K. and Schmittlein, D.C., 1993. Analyzing duration times in marketing: Evidence for the effectiveness of hazard rate models. Marketing Science, 12(4), pp.395-414.
Hinton, G. E., & Salakhutdinov, R. R. 2006. Reducing the dimensionality of data with neural networks. Science, 313(5786), 504-507.
Hosanagar, K., Fleder, D., Lee, D., & Buja, A. 2014. Will the global village fracture into tribes? Recommender systems and their effects on consumer fragmentation. Management Science, 60(4), 805-823.
Huang, Zhongqiang (Tak), and Robert S. Wyer Jr. 2015, "Diverging Effects of Mortality Salience on Variety Seeking: The Different Roles of Death Anxiety and Semantic Concept Activation," Journal of Experimental Social Psychology, 58 (May), 112–23.
Kahn, B. E., Kalwani, M. U., & Morrison, D. G. 1986. Measuring variety-seeking and reinforcement behaviors using panel data. Journal of Marketing Research, 23(2), 89-100.
Kahn, Barbara E., and Alice M. Isen 1993, "The Influence of Positive Affect on Variety Seeking Among Safe, Enjoyable Products," Journal of Consumer Research, 20 (2), 257–70.
Kahn, B. & Wansink, B., 2004. The influence of assortment structure on perceived variety and consumption quantities. Journal of Consumer Research, 30(4), pp.519–533.
Kaminskas, M. and Bridge, D., 2016. Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. ACM Transactions on Interactive Intelligent Systems (TiiS), 7(1), pp.1-42.
Kashdan, T.B., Gallagher, M.W., Silvia, P.J., Winterstein, B.P., Breen, W.E., Terhar, D. and Steger, M.F., 2009. The curiosity and exploration inventory-II: Development, factor structure, and psychometrics. Journal of research in personality, 43(6), pp.987-998.
Kim, J., Allenby, G.M. and Rossi, P.E., 2002. Modeling consumer demand for variety. Marketing Science, 21(3), pp.229-250.
Kohavi, R., Longbotham, R., Sommerfield, D. and Henne, R.M., 2009. Controlled experiments on the web: survey and practical guide. Data mining and knowledge discovery, 18, pp.140-181.
Laibson, D., 1997. Golden eggs and hyperbolic discounting. The Quarterly Journal of Economics, 112(2), pp.443-478.
LaBarbera, P.A. and Mazursky, D., 1983. A longitudinal assessment of consumer satisfaction/dissatisfaction: the dynamic aspect of the cognitive process. Journal of marketing research, 20(4), pp.393-404.
Lee, GM, S He, J Lee, AB Whinston 2020 "Matching Mobile Applications for Cross-Promotion," Information Systems Research 31(3).
Levav, Jonathan, and Rui (Juliet) Zhu. 2009, Seeking Freedom through Variety, Journal of Consumer Research, 36 (4), 600–610.

Li, L., Chu, W., Langford, J. and Schapire, R.E., 2010, April. A contextual-bandit approach to personalized news article recommendation. In 19th conference on World Wide Web (pp. 661-670).

Li, S., Karatzoglou, A. and Gentile, C., 2016, July. Collaborative filtering bandits. In Proceedings of the 39th International ACM SIGIR Conference (pp. 539-548).

Li, P. and Tuzhilin, A., (2020). Latent Unexpected Recommendations. ACM Transactions on Intelligent Systems and Technology (TIST), 11(6), pp.1-25.

Li, J., Abbasi, A., Cheema, A., & Abraham, L. B. (2020a). Path to purpose? How online customer journeys differ for hedonic versus utilitarian purchases. Journal of Marketing, 84(4), 127-146.

Li, P., Que, M., Jiang, Z., Hu, Y. and Tuzhilin, A., (2020), PURS: Personalized Unexpected Recommender System for Improving User Satisfaction. In Fourteenth ACM Conference on RecSys (pp. 279-288).

Machado, F.S. and Sinha, R.K., 2007. Smoking cessation: A model of planned vs. actual behavior for time-inconsistent consumers. Marketing Science, 26(6), pp.834-850.

Maimaran, Michal, and S. Christian Wheeler (2008), "Circles, Squares, and Choice: The Effect of Shape Arrays on Uniqueness and Variety Seeking," Journal of Marketing Research, 45 (6), 731–40.

Mann, D.H., 1975. Optimal theoretic advertising stock models: A generalization incorporating the effects of delayed response from promotional expenditure. Management Science, 21(7) pp.823-832.

McAlister, L., & Pessemier, E. (1982). Variety seeking behavior: An interdisciplinary review. Journal of Consumer research, 9(3), 311-322.

Mei, Hongyuan, and Jason M. Eisner., 2017. "The Neural Hawkes Process: A Neurally Self-Modulating Multivariate Point Process." Advances in Neural Information Processing Systems 30 (2017).

Menon, Satya, and Barbara E. Kahn (1995), "The Impact of Context on Variety Seeking in Product Choices," Journal of Consumer Research, 22 (3), 285–95.

Padmanabhan, B. and Tuzhilin, A., 1998, August. A Belief-Driven Method for Discovering Unexpected Patterns. In KDD (Vol. 98, pp. 94-100).

Panniello, U., Gorgoglione, M. and Tuzhilin, A., (2016). In CARSs we trust: How context-aware recommendations affect customers' trust and other business performance measures of recommender systems. Information Systems Research, 27(1), pp.182-196.

Raju, P.S., 1980. Optimum stimulation level: Its relationship to personality, demographics, and exploratory behavior. Journal of consumer research, 7(3), pp.272-282.

Ratner, Rebecca K., and Barbara E. Kahn (2002), "The Impact of Private Versus Public Consumption on Variety-Seeking Behavior," Journal of Consumer Research, 29 (2), 246–57.

Ratner, R.K., Kahn, B.E. and Kahneman, D., 1999. Choosing less-preferred experiences for the sake of variety. Journal of consumer research, 26(1), pp.1-15.

Read, Daniel, and George Loewenstein (1995), "Diversification Bias: Explaining the Discrepancy in Variety Seeking Between Combined and Separated Choices," Journal of Experimental Psychology: Applied, 1 (1), 34–9.

Sahoo, N., Krishnan, R., Duncan, G., & Callan, J. (2012). Research note—the halo effect in multicomponent ratings and its implications for recommender systems: The case of yahoo! movies. Information Systems Research, 23(1), 231-246.

Salakhutdinov, R. and Hinton, G., 2009. Semantic hashing. International Journal of Approximate Reasoning, 50(7), pp.969-978.

Schwartz, E.M., Bradlow, E.T. and Fader, P.S., 2017. Customer acquisition via display advertising using multi-armed bandit experiments. Marketing Science, 36(4), pp.500-522.

Seetharaman, P.B., 2004. The additive risk model for purchase timing. Marketing Science, 23(2), pp.234-242.

Senecal, Sylvain, and Jacques Nantel. (2004). "The influence of online product recommendations on consumers' online choices." Journal of retailing 80.2: 159-169.

Shani, G. and Gunawardana, A., (2011). Evaluating recommendation systems. In Recommender systems handbook (pp. 257-297). Springer, Boston, MA.

Silberschatz, A. and Tuzhilin, A., 1996. What makes patterns interesting in knowledge discovery systems. IEEE Transactions on Knowledge and data engineering, 8(6), pp.970-974.

Singh, PV, N Sahoo, T Mukhopadhyay (2014) "How to Attract and Retain Readers in Enterprise Blobbing?" Information Systems Research 25(1).

Steenkamp, Jan-Benedict E.M., and Hans Baumgartner (1992), "The Role of Optimum Stimulation Level in Exploratory Consumer Behavior," Journal of Consumer Research, 19 (3), 434–48.

Tan, TF, S Netessine, L Hitt (2017) "Is Tom Cruise Threatened? An Empirical Study of the Impact of Product Variety on Demand Concentration," Information Systems Research 28(3).

Trijp, H.C.M. Van, Hoyer, W.D. & Inman, J., 1996. Why Switch? Product Category Level Explanations for True Variety-Seeking Behavior. Journal of Marketing Research, 33(3), p.281.

Van Trijp, H.C. and Steenkamp, J.B.E., 1992. Consumers' variety seeking tendency with respect to foods: measurement and managerial implications. European Review of Agricultural Economics, 19(2), pp.181-195.

Wang, C. and Huang, Y., 2018. "I Want to Know the Answer! Give Me Fish'n'Chips!": The Impact of Curiosity on Indulgent Choice. Journal of Consumer Research, 44(5), pp.1052-1067.

Wang, S., Gong, M., Li, H. and Yang, J., 2016. Multi-objective optimization for long tail recommendation. Knowledge-Based Systems, 104, pp.145-155.

Wayne D. Hoyer and Nancy M. Ridgway (1984),Variety Seeking As an Explanation For Exploratory Purchase Behavior: a Theoretical Model, in Advances in Consumer Research Volume 11, 114-119.

Xu, L., Duan, J.A. and Whinston, A., 2014. Path to purchase: A mutually exciting point process model for online advertising and conversion. Management Science, 60(6), pp.1392-1412.

Yin, H., Cui, B., Li, J., Yao, J. and Chen, C., 2012. Challenging the Long Tail Recommendation. Proceedings of the VLDB Endowment, 5(9).

Yoon, S. and Kim, H.C., 2018. Feeling economically stuck: The effect of perceived economic mobility and socioeconomic status on variety seeking. Journal of Consumer Research, 44(5), pp.1141-1156.

Zeithammer, R., & Thomadsen, R. (2013). Vertical differentiation with variety-seeking consumers. Management Science, 59(2), 390-401.

Zhang, S., Yao, L., Sun, A. and Tay, Y., 2019. Deep learning based recommender system: A survey and new perspectives. ACM Computing Surveys (CSUR), 52(1), pp.1-38.

Zhang, J., Adomavicius, G., Gupta, A. and Ketter, W., 2020. Consumption and performance: Understanding longitudinal dynamics of recommender systems via an agent-based simulation framework. Information Systems Research, 31(1), pp.76-101.

When Variety-Seeking Meets Unexpectedness: Incorporating Variety-Seeking Behaviors into Design of Unexpected Recommender Systems

Appendix

Part I: The Curiosity Questionnaire and Statistics of Consumer Response

Questions	Response Scale
Q1: "I actively seek as much information as I can in new situations."	7-point Likert scale 1–strongly disagree 2–moderately disagree 3-disagree a little 4–neither agree nor disagree 5-agree a little 6–moderately agree 7–strongly agree
Q2: "I am the type of person who really enjoys the uncertainty of everyday life."	
Q3: "I am at my best when doing something that is complex or challenging."	
Q4: "Everywhere I go, I am out looking for new things or experiences."	
Q5: "I view challenging situations as an opportunity to grow and learn."	
Q6: "I like to do things that are a little frightening."	
Q7: "I am always looking for experiences that challenge how I think about myself and the world."	
Q8: "I prefer jobs that are excitingly unpredictable."	
Q9: "I frequently seek out opportunities to challenge myself and grow as a person."	
Q10: "I am the kind of person who embraces unfamiliar people, events, and places."	

Table 1: "Ten-item Curiosity and Exploration Inventory-II" Questionnaire (Kashdan et al. 2009)

Survey Question & Response	Mean	Std.	Median	Skewness	Kurtosis
"The item recommended to me matches my interests."	3.32	1.410	4.00	-0.419	-1.192
The item recommended to me is novel."	3.06	1.424	3.00	-0.146	-1.391
"The item recommended to me is different from the types of products I bought before."	3.39	1.215	4.00	-0.400	-0.813
The item recommended to me is similar to the system's prior recommendations."	2.93	1.302	3.00	0.214	-1.109
"The item recommended to me is unexpected."	3.16	1.437	3.00	-0.199	-1.337
"The item recommended to me is a pleasant surprise."	2.73	1.456	2.50	0.195	-1.400
"The item recommended to me is very timely."	3.00	1.484	3.00	-0.074	-1.450
"I am satisfied with this recommendation."	3.21	1.140	3.00	-0.286	-0.466
"I would buy the item recommended, given the opportunity."	2.83	1.456	3.00	0.003	-1.418
"Ten-item Curiosity and Exploration Inventory-II"	3.13	0.831	3.10	0.088	-0.402

Table 2: Statistics of the Consumer Responses to the Questionnaire

Part II: Classification Performance on the Curiosity Questionnaire

Variety-Seeking Framework	Accuracy	F1-Score
Euclidean+Exponential+Mean **(%Improved)**	**0.937***** **(0.017)** **+%**	**0.867***** **(0.012)** **+%**
Euclidean+Hyperbolic+Mean	0.919***	0.839***
Euclidean+No Decay+Mean	0.790***	0.749***
Cosine+Exponential+Mean	0.908***	0.837***
Cosine+Hyperbolic+Mean	0.901***	0.808***
Cosine+No Decay+Mean	0.779**	0.716***
Manhattan+Exponential+Mean	0.873***	0.825***
Manhattan+Hyperbolic+Mean	0.864***	0.801***
Manhattan+No Decay+Mean	0.776**	0.709***
Chebyshev+Exponential+Mean	0.877***	0.784***
Chebyshev+Hyperbolic+Mean	0.865***	0.763***
Chebyshev+No Decay+Mean	0.765	0.682***
Feature+Exponential+Mean	0.832***	0.726***
Feature+Hyperbolic+Mean	0.837***	0.721***
Feature+No Decay+Mean	0.761	0.643
Binary+Exponential+Mean	0.778**	0.668**
Binary+Hyperbolic+Mean	0.771**	0.641
Binary+No Decay+Mean	0.738	0.640
APF	0.730	0.629
PAS	0.763	0.639

*Table 3: Classification Performance of Consistency- v.s. Variety-Seeking Consumers Based on Our Variety-Seeking Framework. The threshold is the average variety-seeking level across all consumers. *p<0.1; **p<0.05; ***p<0.01. (compared to APF & PAS)*

Part III: Analysis of the Statistics to Use for the Stationarity Assumption

Variety-Seeking Framework	$Product\_Variety(i,j,t)$	$Variety\_Seeking(i)$
Euclidean+Exponential+Arithmetic Mean	**0.775**** **(0.004)**	**0.618**** **(0.003)**
Euclidean+Exponential+Weighted Mean	0.773	0.616
Euclidean+Exponential+Geometric Mean	0.771	0.613
Euclidean+Exponential+Harmonic Mean	0.770	0.613
Euclidean+Exponential+Median	0.728	0.589

*Table 4: Pearson Correlation Coefficients between Self-Reported Variety-Seeking Levels and Our Variety-Seeking Framework. *p<0.1; **p<0.05; ***p<0.01. (compared to APF & PAS)*

Part IV: Summary Statistics of Consumer and Video Features

Consumer /Video Features		Mean	25th percentile	Median	75th percentile	Variance
Gender	Treatment	0.50025	0.000	1.000	1.000	0.495
	Control	0.50015	0.000	1.000	1.000	0.495
Age	Treatment	41.119	30.000	40.000	50.000	21.100

	Control	40.076	30.000	40.000	50.000	20.070
VIP Status	Treatment	0.361	0.000	0.000	1.000	0.480
	Control	0.365	0.000	0.000	1.000	0.480
Activity Days	Treatment	20.668	8.000	19.000	27.000	8.520
	Control	20.404	8.000	19.000	27.000	8.450
Genre		4.731	1.000	4.000	6.000	2.150
View Count		9.684	8.000	10.000	12.000	3.128
Comment Count		7.070	0.000	2.000	5.000	4.283
Release Days		6.325	5.000	7.000	8.000	2.304

Table 5: Summary statistics of consumer and video features for the control and treatment groups.

Part V: Additional Results in the Online Experiment: Difference-in-Difference Analysis and User-Level Analysis

To further justify the validity of our empirical findings, we replicate our analysis in Section 5.3 using the Difference-in-Difference method instead, where we specify the ATEs in equation (4) of the paper using the following alternative identification:

$$Metric_{ij} = \alpha_0 + \alpha_1 * Treatment_i + \alpha_2 * Experiment_t + \alpha_3 * Treatment_i * Experiment_t + \overrightarrow{\alpha_4} * \overrightarrow{X_i} + \overrightarrow{\alpha_5} * \overrightarrow{Y_j} + \varepsilon_{ij}$$

where $Metric_{ij} \in \{CTR_{ij}, VV_{ij}, TS_{ij}\}$, $Treatment_i$ is the dummy variable indicating whether consumer i is in the treatment group or not, $Experiment_t$ is the dummy variable indicating the post-treatment period versus the pre-treatment period, $\overrightarrow{X_i}$ represents explicit user features, and $\overrightarrow{Y_j}$ represents explicit video features. We can observe from Table 6 almost identical treatment effects as we report in Section 5.3, as our proposed model achieves significant performance improvements on those consumers in the treatment group. Besides, we observe that $Experiment_t$ does not hold a significant coefficient, suggesting that the parallel trends assumption is fulfilled and that the observed relationship is unlikely to arise as an artifact from events occurred before our treatment.

	CTR_{ij}	VV_{ij}	TS_{ij}
$Treatment_i$	0.0229*** (0.0042)	0.0455*** (0.0028)	39.210*** (0.9891)
User Features Video Features	Yes Yes	Yes Yes	Yes Yes
R-Squared	0.0079	0.0043	0.2126
Observations	46,794,473	46,794,473	46,794,473

*Table 6: Average Treatment Effect Using the Difference-in-Difference Analysis. Robust standard errors are in parentheses. *p<0.1; **p<0.05; ***p<0.01*

Furthermore, in addition to the three business metrics that we study in the paper, we also analyze the treatment effects on the following user-level metrics in this section: (a) CTR_i, the average click-through rate of user i in each session; (b) VV_i, the average finish-watching percentage of user i in each session, and (c) TS (Time Spent), the average time the user has spent in each session. We specify the ATEs using the following identification:

$$Metric_i = \alpha_0 + \alpha_1 * Treatment_i + \overrightarrow{\alpha_2} * \overrightarrow{X_i} + D_t + \varepsilon_{ij}$$

where $Metric_{ij} \in \{CTR_i, VV_i, TS_i\}$, $Treatment_i$ is the dummy variable, $\overrightarrow{X_i}$ represents explicit user features, and D_t represents time-fixed effects including dates and hours. We can observe from Table 7 that consumers who are served by our proposed model achieve significant improvements in all three user-level metrics compared to the control group.

	CTR_i	VV_i	TS_i
$Treatment_i$	0.0227*** (0.0049)	0.0459*** (0.0033)	225.378*** (5.7478)
User Features Time Fixed Effect	Yes Yes	Yes Yes	Yes Yes
R-Squared	0.0075	0.0041	0.2025
Observations	37,965,781	37,965,781	37,965,781

*Table 7: Average Treatment Effect on the User-Level Metrics. The table shows a regression with robust standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01*

Part VI: Statistics and Experiment Results of the Sparsity Analysis

Dataset	Alibaba-1	Alibaba-2	Alibaba-3
#Consumers	46,143	21,152	10,737
#Products	53,657	22,809	13,751
#Transactions	1,806,157	1,061,948	646,933
#Records Per Consumer	39.14	50.21	60.25
Sparsity	0.073%	0.220%	0.438%

Table 8: Descriptive Statistics of the Datasets for Sparsity Analysis

	Alibaba-1		Alibaba-2		Alibaba-3	
	AUC	HR@10	AUC	HR@10	AUC	HR@10
DIN+Latent+Multiply **(%Improved)**	**0.7349***** **(0.0088)** **+2.73%**	**0.7730***** **(0.0089)** **+3.15%**	**0.7511***** **(0.0092)** **+4.62%**	**0.7802***** **(0.0097)** **+3.99%**	**0.7670***** **(0.0094)** **+6.22%**	**0.7885***** **(0.0101)** **+4.85%**
DIN+Latent+Exponential	0.7328***	0.7701***	0.7470***	0.7763***	0.7618***	0.7852***

DIN+Latent+Power	0.7324***	0.7688***	0.7461***	0.7751***	0.7599***	0.7838***
DIN+Feature+Multiply	0.7299***	0.7672***	0.7458***	0.7749***	0.7580***	0.7832***
DIN+Feature+Exponential	0.7303***	0.7654***	0.7427***	0.7738***	0.7564***	0.7815***
DIN+Feature+Power	0.7291***	0.7658***	0.7411***	0.7726***	0.7573***	0.7806***
NCF+Latent+Multiply	0.7266**	0.7649***	0.7409***	0.7730***	0.7562***	0.7802***
NCF+Latent+Exponential	0.7261**	0.7610**	0.7388***	0.7722***	0.7558***	0.7794***
NCF+Latent+Power	0.7249**	0.7587**	0.7395***	0.7719***	0.7549***	0.7799***
NCF+Feature+Multiply	0.7228**	0.7599**	0.7392***	0.7698***	0.7522***	0.7777***
NCF+Feature+Exponential	0.7237**	0.7576**	0.7376***	0.7684***	0.7515***	0.7760***
NCF+Feature+Power	0.7240**	0.7573**	0.7383***	0.7672***	0.7516***	0.7765***
DIN	0.6957	0.6972	0.7026	0.7028	0.7055	0.7047
DeepFM	0.5519	0.5164	0.5917	0.5579	0.6214	0.5892
PURS	0.7154	0.7494	0.7179	0.7503	0.7221	0.7520
HOM-LIN	0.5812	0.5493	0.6032	0.5871	0.6075	0.6061
Re-Ranking	0.6025	0.5776	0.6164	0.6011	0.6215	0.6163
DPP	0.6517	0.7026	0.6662	0.7075	0.6707	0.7124
LinUCB	0.6775	0.6662	0.6825	0.7101	0.7028	0.6925
COFIBA	0.6796	0.6798	0.6906	0.7129	0.7032	0.7016

*Table 9: Offline results on the three Alibaba datasets. *p<0.1; **p<0.05; ***p<0.01. Improvement percentages were reported over the second-best baseline models (underlined).*

Part VII: Additional Results for Robustness Check

(a) We include different combinations of consumer features, video features, and time-fixed effects in the regression model to evaluate the treatment effects of adopting our proposed model. As shown in Table 10, our results will not be affected by the particular model specification of features or fixed effects during the estimation process.

	CTR_{ij}	CTR_{ij}	CTR_{ij}	CTR_{ij}
$Treatment_i$	0.0229*** (0.0043)	0.0233*** (0.0047)	0.0240*** (0.0102)	0.0238*** (0.0129)
User & Video Features Time Fixed Effect	Yes Yes	Yes No	No Yes	No No
R-Squared	0.0081	0.0069	0.0022	0.0011
Observations	37,965,781	37,965,781	37,965,781	37,965,781

*Table 10: Average Treatment Effect with Different Combinations of Features and Fixed Effects. Robust standard errors are in parentheses. *p<0.1; **p<0.05; ***p<0.01*

(b) We use alternative models to specify the binary outcome variables CTR_{ijt} and VV_{ijt}, including the discrete choice models of Logit and Probit. The results in Table 11 show that treatment effects are still positive and statistically significant under different specifications.

	CTR_{ij} (Linear)	CTR_{ij} (Logit)	CTR_{ij} (Probit)

$Treatment_i$	0.0229*** (0.0043)	0.571*** (0.076)	0.403*** (0.059)
User & Video Features Time Fixed Effect	Yes Yes	Yes Yes	Yes Yes
R-Squared	0.0081	0.0084	0.0082
Observations	37,965,781	37,965,781	37,965,781

*Table 11: Average Treatment Effect of Different Specification Models. The table shows a regression with robust standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01*

(c) We also exclude recommendation records of video content uploaded by Company A itself during the experiment, as some consumers might be more willing to click on these video recommendations due to their loyalty to the platform. As shown in Table 12, the estimation results are not significantly different from the original estimation.

	CTR_{ij}	VV_{ij}	TS_{ij}
$Treatment_i$	0.0229*** (0.0044)	0.0456*** (0.0028)	39.206*** (1.0003)
User & Video Features Time Fixed Effect	Yes Yes	Yes Yes	Yes Yes
R-Squared	0.0079	0.0043	0.2126
Observations	37,146,255	37,146,255	37,146,255

*Table 12: Average Treatment Effect after Excluding Self-Uploaded Video Content. The table shows a regression with robust standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01*

(d) Finally, we drop the records from those users in the regions where the platform of Company A was launched recently, as those users might be more willing to click on whatever recommendations they receive due to the novelty effect. The results demonstrated in Table 13 show that our treatment effects remain robust when we exclude those records.

	CTR_{ij}	VV_{ij}	TS_{ij}
$Treatment_i$	0.0229*** (0.0043)	0.0455*** (0.0028)	39.222*** (0.9897)
User & Video Features Time Fixed Effect	Yes Yes	Yes Yes	Yes Yes
R-Squared	0.0080	0.0045	0.2136
Observations	37,894,442	37,894,442	37,894,442

*Table 13: Average Treatment Effect after Excluding Recently Launched Records. The table shows a regression with robust standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01*

Part VIII: Product-Level Analysis

As we have already demonstrated the significant impact of our proposed model on the consumer side, we will study its influence on the product side in this section, particularly on product demand

distribution (Tan et al. 2017; Fong 2017). To do this, we compare the Lorenz curve & Gini Coefficient of the clicked videos between the treatment group and the control group in our experiment. Note that the Gini Coefficients of both groups are the same over the pre-treatment period, as part of the randomized setting. As shown in Figure 1, the inequalities of video distributions in the treatment group (Gini Index=0.44) have significantly decreased, compared to those in the control group (Gini Index=0.59). This is the case, as our proposed model improves consumers' variety-seeking levels in general, resulting in more unexpected video content in recommendations, which are typically novel and have little exposure under classical recommender systems. Therefore, our model has the potential to alleviate the "winners-take-most" and fairness problems in recommendations (Wang et al. 2016) and contribute to a better consumer experience.

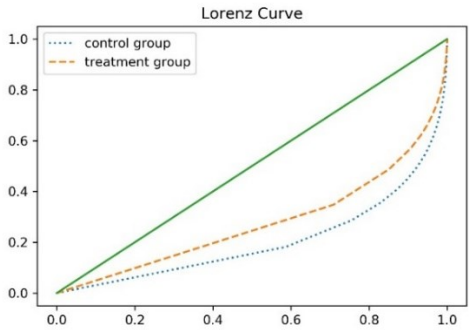


Figure 1: Lorenz Curve of the video view count in the treatment and control group

Part IX: Churn Rate Analysis

To make sure that consumer abandonment behavior (i.e., consumers might reduce their usage or leave the platform if served by our proposed model) and curiosity factors (i.e., consumers might be curious to try out the new recommender system design) do not significantly contribute to the performance improvements, we conduct additional analysis to study the long-term treatment effects and changes of the churn rate, which is measured as a binary variable on a daily basis of whether the consumer watches any video on that day or not. In particular, we specify the churn rate variable of consumer i at day t using the logistic regression:

$$Churn\_Rate_{it} = \frac{1}{1 + e^{-(\alpha_0 + \alpha_1 * Treatment_i + \overrightarrow{\alpha_2} * \overrightarrow{X_i} + D_t + \varepsilon_{it})}} \qquad (5)$$

and we also specify the long-term treatment effects through the OLS regression model:

$$Metric_{ijt} = \alpha_0 + \alpha_1 * Treatment_i * D_t + \overrightarrow{\alpha_2} * \overrightarrow{X_\iota} + \overrightarrow{\alpha_3} * \overrightarrow{Y_j} + \varepsilon_{ijt} \qquad (6)$$

where D_t represents the date in our online experiments. The regression results in Table 14 show that the churn rate will be significantly lower if consumers adopt the treatment of our proposed model, indicating that *more* users choose to stay with the platform compared to the existing model. This is the case, as our proposed model produces more satisfying video recommendations for the consumers to keep them within the platform. Therefore, we demonstrate that those significant improvements achieved by our model do not come from the abandonment effect, as our model reduces the churn rate significantly.

	$Churn\_Rate_{it}$
$Treatment_i$	-0.0611*** (0.0043)
User Features Time Fixed Effect	Yes Yes
R-Squared	0.0081
Observations	37,965,781

*Table 14: Average Treatment Effect of Variety-Seeking Based Unexpected Recommendations on Churn Rate. Robust standard errors are in parentheses. *p<0.1; **p<0.05; ***p<0.01*

Part X: Variety-Seeking Behavior Analysis

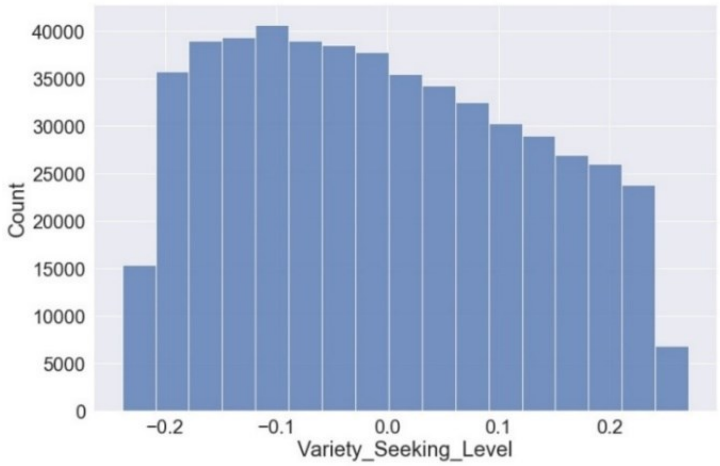
To improve our understanding of the variety-seeking behavior, we plot the pre-experiment distribution of variety-seeking levels in the treatment group in Figure 2(a), which is close to a skewed normal distribution where the majority of consumers have low variety-seeking levels. After being served by our model, their variety-seeking behaviors have significantly changed based on their experience with the new system, where we have the following observations in Figure 2(b):

First, consumers in general would seek more variety in the video recommendations after being served by our proposed method, as the average variety-seeking level $Variety\_Seekiing(i)$ in the treatment group has significantly increased from -0.0053 to 0.0232 after the adoption.

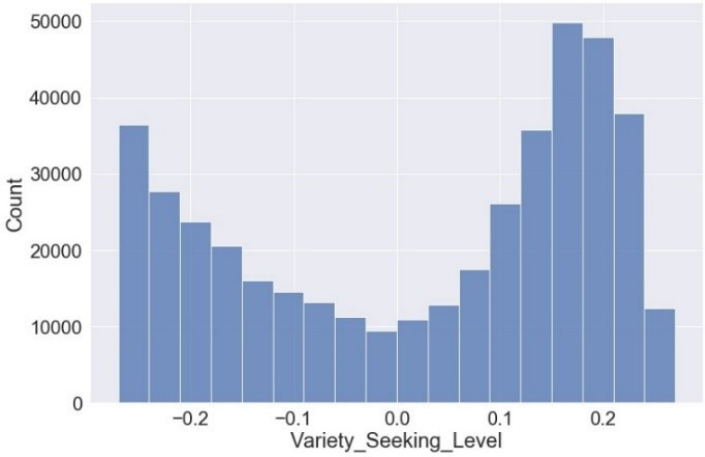
Second, our method reinforces the variety-seeking behavior for those consumers with high-level of variety-seeking behavior ($Variety(i)$>0.1), as we witness a significant increase in the average

variety-seeking level among these consumers from 0.1434 to 0.1781, an improvement of 24.20%. At the same time, our method further reduces the level of variety in the recommended videos for those consumers with low-level of variety-seeking behavior ($Variety(i)$<-0.1), a decrease of 25.53% from -0.1594 to -0.2001.

Third, our stationary assumption still holds for consumers served by our proposed model. In particular, we repeat the ADF test on post-treatment records, and the average statistics in the treatment group is -33.28, more negative than the critical value of −3.5 at the 95% confidence level. This observation further demonstrates the validity of the stationarity assumption that we make in our variety-seeking framework.



(a) Before the Adoption



(b) After the Adoption

Figure 2: Comparisons of the Variety-Seeking Levels of Consumers