The Value of Personalized Recommendations: Evidence from Netflix*

Kevin Zielnicki[1], Guy Aridor[2], Aurélien Bibaut[1], Allen Tran[1], Winston Chou[1], and Nathan Kallus[1,3]

[1]Netflix
[2]Northwestern University, Kellogg School of Management
[3]Cornell University and Cornell Tech

November 12, 2025

Abstract

Personalized recommendation systems shape much of user choice online, yet their targeted nature makes separating out the value of recommendation and the underlying goods challenging. We build a discrete choice model that embeds recommendation-induced utility, low-rank heterogeneity, and flexible state dependence and apply the model to viewership data at Netflix. We exploit idiosyncratic variation introduced by the recommendation algorithm to identify and separately value these components as well as to recover model-free diversion ratios that we can use to validate our structural model. We use the model to evaluate counterfactuals that quantify the incremental engagement generated by personalized recommendations. First, we show that replacing the current recommender system with a matrix factorization or popularity-based algorithm would lead to 4% and 12% reduction in engagement, respectively, and decreased consumption diversity. Second, most of the consumption increase from recommendations comes from effective targeting, not mechanical exposure, with the largest gains for mid-popularity goods (as opposed to broadly appealing or very niche goods).

*Corresponding authors: kzielnicki@netflix.com, guy.aridor@kellogg.northwestern.edu, and abibaut@netflix.com. Kevin, Aurélien, Allen, Winston, and Nathan were all employed and hold a financial stake in Netflix. Guy was a paid contractor at Netflix, but currently holds no financial stake in Netflix. We thank Rafael Jiménez-Durán, Brett Gordon, Malika Korganbekova, and Andrey Simonov as well as conference participants at EARIE for helpful comments. We thank Meghana Bhatt, Patric Glynn, Apoorva Lal, Adam Lindsay, Danielle Rich, Max Schmeiser, Qi Wang, and Patrick Zheng for internal assistance at Netflix with various aspects of this project. All errors are our own.

1 Introduction

Online platforms increasingly mediate user choice through personalized recommendations that alleviate the informational and search frictions of selecting high quality goods from large catalogs. These systems are deployed across a wide range of platforms in the digital economy, such as e-commerce and media streaming platforms. Their large role raises a fundamental question: how much of user demand reflects the inherent demand for goods versus the effects of the recommendations? This distinction is important for quantifying the impact that the recommendation system (RecSys) has on platform engagement as well as quantifying the underlying incremental engagement generated by goods, which can be crucial for catalog optimization. However, this problem is challenging as recommendations are personalized to users' tastes, making it unclear whether a user consumed the good due to the recommendation or their underlying preferences.

In this paper, we estimate a discrete-choice model of user choice that incorporates the effects of recommendations on choice on a large video streaming platform with a prominent RecSys—Netflix. With this model in hand, we are able to both separately quantify the value of recommendations for engagement and provide a credible method for estimating the incremental demand for the underlying goods. A key aspect of our approach, both in the model itself and in its validation, is to exploit exogenous variation in the recommendations generated by the RecSys to separate the effects of recommendations from "intrinsic" preferences for goods and to estimate model-free diversion ratios that we use to validate our model.

Modeling user choices in this setting faces several challenges that are present on many other online platforms. First, conditional on subscribing to the platform, users costlessly consume goods, meaning that there is no price variation that we can exploit to help us measure user valuation of goods or ascertain the role of recommendations. Second, estimating movie demand has historically been challenging as traditional good characteristics, such as cast and genre, have done a relatively poor job of predicting user demand (Einav, 2007; McKenzie, 2023). Finally, there is potential serial dependence in choices (i.e., the good a user consumes today influences their choice tomorrow), and both the number of goods and users is incredibly large. We overcome each of these challenges and provide a model that we can apply to estimate the value of goods and recommendations.

The structure of our model follows a low-rank discrete choice model (Kallus and Udell, 2016), which assumes that the good $\times$ user matrix of user preferences can be represented by a low-rank representation of good and user embeddings. We extend this baseline model in several ways to solve the aforementioned challenges. For the first challenge, we model the effects of recommendation via an additive recommendation "bonus" that provides a utility boost for goods that are recommended. This is a reduced-form characterization of the effects of the recommendation as reducing information frictions for these goods (Aridor et al., 2023). For the second challenge,

we endogenously learn the good-level "characteristics" directly from user choices, circumventing the need to define a reasonable set of characteristics that can rationalize choices. For the third challenge, we adapt a popular approach in the machine learning literature to our setting and consider a transformer-style architecture (Vaswani et al., 2017) that extends standard state-dependent demand methods to allow for user preferences to flexibly adapt across time and rely on a user's consumption history.[1]

We estimate our model using 2 million Netflix users and the approximately 7,000 goods available to them in the United States. To feasibly estimate user and good representations at this scale, we adopt a modeling approach that avoids the need to estimate a user embedding for each user. Instead, we represent each user dynamically as a function of their past watch history using a sequence model operating over good embeddings. In this architecture, user preferences are encoded by the parameters of the sequence model and the user's observed history, rather than by a user-specific embedding vector. As a result, the number of parameters to estimate grows only with the number of goods and the complexity of the sequence model, rather than with the number of users. This dramatically reduces the memory and computational requirements for model estimation, enabling us to scale to large populations while capturing heterogeneous and time-varying user preferences.

A key aspect of our approach and its validation is to rely on idiosyncratic randomization that the recommender system (RecSys) induces to learn user preferences, which allows us to separate out the effects of recommendation from the "intrinsic" preferences of users. This randomization occasionally induces variation in the selection and ordering of goods in the recommendations. To validate our model, we run an experiment that exogenously "nudges" the salience of a random set of goods, allowing us to generate model-free diversion ratio estimates for this set of goods (Conlon and Mortimer, 2021). We generate the same diversion ratio estimates using our model and find that, out of sample, the model's predicted diversion has a 0.86 correlation and 0.73 R^2 when compared to the model-free diversion ratios. This approach builds upon a recent line of work in industrial organization that relies on exogenous non-price variation, such as good availability experiments, in order to estimate diversion ratios in settings without price variation (Conlon and Mortimer, 2013; Conlon et al., 2023; Aridor, 2025).

We use our model to better understand to what extent recommendations influence user choice. There are two distinct perspectives on how the RecSys affects choices. First, the RecSys can reduce the informational and search costs associated with choosing satisfying items from the catalog, thus increasing *overall* consumption. Second, the personalization capabilities of the RecSys can especially help to enable consumption of niche goods, and thus have heterogeneous impacts across

[1]The central concept of the transformer architecture is "attention", which is a context-dependent weight computed for each vector in a sequence, enabling a weighted-average summarization of an arbitrary-length sequence into a single vector. In this work, we use a simpler form of attention weighting described in Section 3.

different types of goods. Motivated by this, we use our model to quantify how much engagement and consumption diversity is driven by the current RecSys compared to reasonable algorithmic benchmarks and quantify the role of effective targeting in driving consumption across the distribution of goods.

In order to quantify the overall value of the RecSys, we evaluate counterfactuals where the set of recommendations is generated by (i) randomized, (ii) popularity-based, and (iii) traditional matrix factorization (Koren et al., 2009) recommendation systems.[2] Our main finding is that the current RecSys balances high engagement with high consumption diversity. Moving to an entirely random recommendation policy decreases engagement by 16%, with minimal improvements to overall consumption diversity. Moving to a popularity-based policy or a traditional matrix factorization approach decreases engagement by 12% and 4%, respectively, while having large negative effects on consumption diversity. Internal estimates of the value of engagement suggest that these are economically large changes (cf. Gomez-Uribe and Hunt, 2015). Furthermore, the current RecSys dramatically improves consumption diversity relative to the latter benchmarks, which is important for long-run user satisfaction (Anderson et al., 2020) and a healthy two-sided market that distributes viewership more evenly across good producers.

How does the value of the recommendation vary across goods and how important is targeting in driving this? We consider the difference in the model-implied probability of consuming a good for a targeted user who is recommended a good and an average user who is counterfactually not. The difference can be driven by three components: baseline propensity to consume (selection), the base effect of recommending a good to any user (exposure), and the larger incremental change in consumption probability on users who receive the recommendations (targeting). We find that, for the observation-weighted average across goods, the exposure effect is relatively large—exposing a good to the average user triples its consumption probability. However, in our decomposition, exposure only accounts for 6.8% of the difference, with selection accounting for 51.3% and targeting for 41.9%. The magnitude of this effect is dramatically large, with targeting being 7 times larger than the mechanical exposure effect in explaining the consumption differences, whereas under the counterfactual matrix factorization algorithm both of these channels are nearly equally as important. We explore heterogeneity across goods and find that targeting most benefits mid-sized goods that appeal to a sizeable but specific subset of users relative to broadly appealing or very niche goods.

Finally, apart from the value of recommendations, we discuss how the model can be used to measure the incremental engagement generated by particular goods. We define the *incremental* engagement for a particular good as the difference between the engagement with the good avail-

[2] As discussed in Gomez-Uribe and Hunt (2015), a matrix factorization based RecSys was the algorithm used at Netflix 10 years prior to this work.

able on the platform compared to it being removed. We can credibly simulate this counterfactual since we can compute choice probabilities and recommendations under different counterfactual catalogs. One key limitation of our baseline model is that, since the characteristics of goods are learned endogenously based on choices, we cannot compute these counterfactuals for new goods that are not available on the platform. Thus, we provide an extension of the model that relies on high-dimensional, pre-consumption good embeddings that are derived from baseline good characteristics and human labels about each of the goods. Put together, these different components enable the estimation of the incremental demand associated with both new and existing goods that is not biased by the influence of recommendations on consumption choices.

Related Work. Our paper provides both substantive and methodological contributions. Substantively, we contribute to the literature on the impact of online recommendations on user choices. Methodologically, we provide innovations in using good recommendations to estimate diversion ratios and include a more flexible characterization of state-dependent demand estimation models.

The Impact of Online Recommendations. We contribute to a burgeoning literature that quantifies the effects of recommendations and rankings on choices. Senecal and Nantel (2004), Das et al. (2007), Freyne et al. (2009), Zhou et al. (2010), Peukert et al. (2024), Holtz et al. (2020), Aridor et al. (2023), and Donnelly et al. (2024) show that, compared to non-personalized benchmarks, recommendation systems increased consumption in hypothetical choices in a lab experiment, Google News, a social network, a news website, YouTube, Spotify, MovieLens, and Wayfair, respectively. These papers collectively provide evidence that personalized recommendations, relative to a non-personalized benchmark, meaningfully increase consumption. There has subsequently been a significant amount of work measuring how personalization impacts both the user-level and aggregate-level diversity of the consumption distribution (Fleder and Hosanagar, 2009; Nguyen et al., 2014; Brynjolfsson et al., 2011; Hosanagar et al., 2013; Lee and Hosanagar, 2019; Holtz et al., 2020; Korganbekova and Zuber, 2023; Calvano et al., 2022).

Our work contributes to the literature in several ways. First, we quantify the incremental value of recommendation on Netflix, a platform with a highly influential and economically significant RecSys (Bennett and Lanning, 2007). By comparing not only to non-personalized benchmarks, but also a previous state of the art algorithm (e.g., matrix factorization (Koren et al., 2009)), we quantify the value of advances in personalization technology in the past decade (e.g., relative to those discussed in Gomez-Uribe and Hunt (2015)). We showcase how this leads to differences in overall engagement as well as its influence on consumption diversity, which is especially important in two-sided markets. Second, by systematically quantifying the value of recommendation across the good distribution and isolating how much of this comes from effective targeting, we more comprehensively document the heterogeneity in the value of recommendations across different goods—highlighting that mid-sized goods benefit the most and not goods with broad appeal or very

niche goods (in contrast to the hypothesis of Anderson et al. (2006)). Third, similar to the insights of Morozov et al. (2021) in other markets with search and information frictions, we highlight the importance of accounting for the influence of recommendations when estimating user preferences.

Measuring Substitution Patterns. Methodologically, our paper contributes to the literature on measuring user substitution patterns.

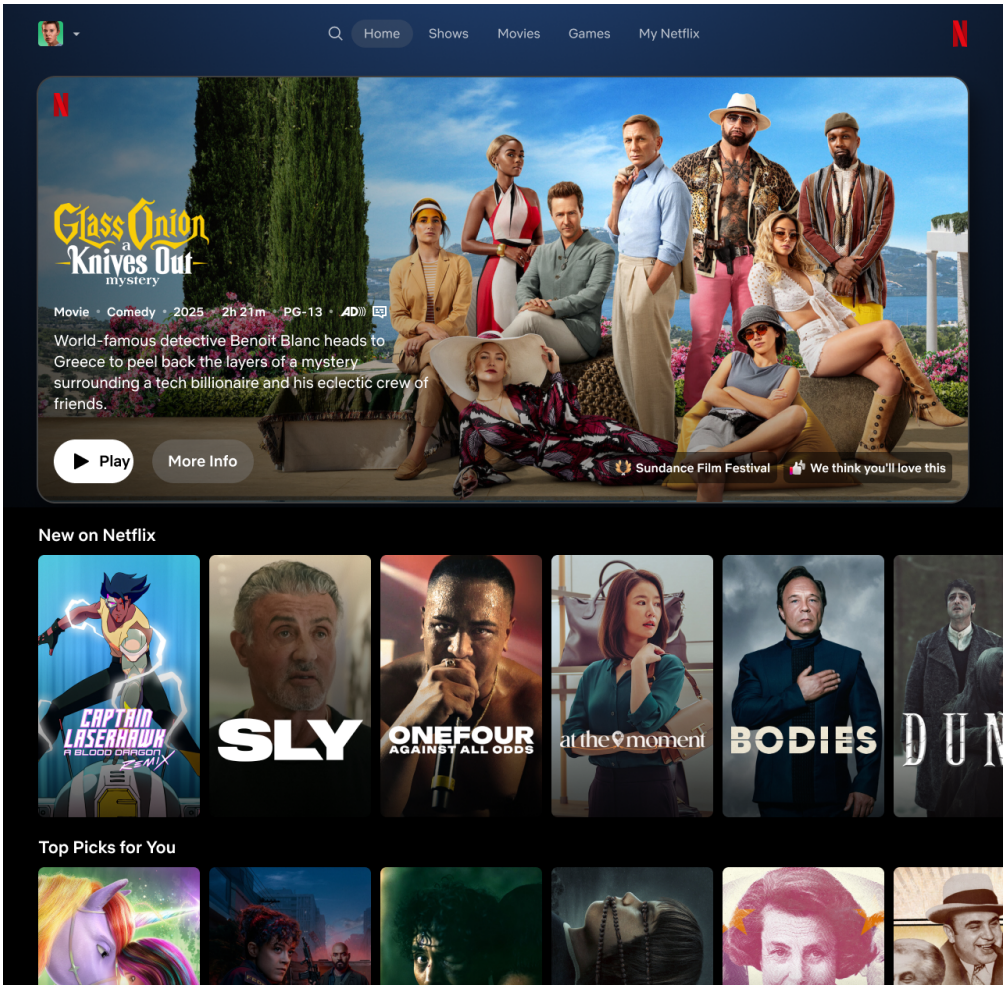
First, we contribute to the growing literature on measuring substitution patterns in markets with limited price variation. While these methods are broadly applicable, they are particularly relevant in the "attention economy," where the scarce resource is user attention and goods are typically costless to consume (Brynjolfsson et al., 2019; Calvano and Polo, 2021; Yuan, 2025). Existing work relies on generated or exogenous good unavailability variation (Conlon and Mortimer, 2013; Raval et al., 2022; Aridor, 2025) or using survey-based methods to elicit choices under hypothetical scenarios (Dertwinkel-Kalt et al., 2024; Bursztyn et al., 2025) to estimate second-choice diversion ratios. Facing a similar challenge, we highlight that platforms can use exploration experiments in the RecSys as plausibly exogenous variation to estimate both choice models and model-free diversion ratios (Conlon and Mortimer, 2021).

Second, we contribute to the literature on state-dependent demand estimation, which provides a class of models to rationalize persistence in user choices. The empirical challenge in this literature, dating back to Heckman (1981, 1991), is to disentangle whether past choices causally affect future choices or if choices are spuriously serially correlated due to unobserved preference heterogeneity. The canonical models of state-dependent demand (Dubé et al., 2010; Simonov et al., 2020) provide a Markovian formulation with rich preference heterogeneity to solve this problem. We formulate our problem in a similar manner, except in our context we are more interested in how to model the evolution of preferences over time as opposed to decoupling these two channels. In order to do so, adapt methods from transformer-style architectures that are popular in machine learning (Vaswani et al., 2017) to characterize state-dependence based on the full consumption history. This allows users' tastes to flexibly change over time and computationally simplifies our model by requiring us to only keep track of different consumption histories, instead of the full matrix of user embeddings. The latter enables us to estimate this model at the scale of millions of users with minimal compute requirements. Topically, Zeller (2022); Lee et al. (2025) apply models with state dependence in media streaming contexts, albeit with substantively different models than we consider here.

2 Institutional Details

Our empirical context is the video-streaming platform Netflix. Users pay a subscription fee to access Netflix and can then costlessly consume any good they wish on the platform. The set of goods consists of both television shows (which have multiple episodes and seasons) and movies.

Figure 1: Netflix Homepage



The Netflix catalog is vast, spanning thousands of goods, and we focus our empirical analysis on the United States, which has approximately 7,000 goods available as of October 2025.

To help users find satisfying goods, Netflix's homepage is personalized. There has been considerable development on the recommendation system, dating back to the public Netflix prize (Bennett and Lanning, 2007) and disclosed in Gomez-Uribe and Hunt (2015). The implementation discussed in Gomez-Uribe and Hunt (2015) relies on matrix factorization approaches, but since then the RecSys has evolved to incorporate newer algorithmic advances in generating recommendations (e.g., see Steck et al. (2021), Joshi et al. (2024)).

Figure 1 provides an example of the homepage for one particular user. There are several distinct components of the homepage, which are important for us to consider when modeling it. First, there is one good that is highlighted for each user that spans across the width of the screen, which we will refer to as the billboard. Second, the example in Figure 1 shows that the recommendations are arranged in different rows that have a similar good style grouping, such as New on Netflix and

Top Picks for You in this example. Based on this, the top-left 5×5 block of titles (i.e., the first 5 rows and the first 5 goods in each row) is where most users engage with the recommendations and we refer to these as the "top 25" recommendations. Third, expanding to include the top-left 10×10 block of titles (i.e., the first 10 rows and the first 10 goods in each row), is where the vast majority of engagement from recommendations occurs and we refer to these as the "top 100" recommendations. In our empirical model, we consider each of these canvases separately. Another important aspect of the homepage is that it includes a "continue watching" row for users to continue consumption of a previously consumed good (e.g., an unfinished movie or a continuing TV show), which we do not consider a recommendation and is unimpacted by our interventions and counterfactuals.

3 Demand Model

We consider a user i who chooses to consume a good from $\mathcal{M} \cup \{0\}$ where $\mathcal{M}$ denotes the set of J goods that are present on the platform and 0 denotes the outside option (not consuming any good). In principle, within a given time period (for instance, a day), users can choose several goods and allocate continuous amounts of time to each of them. However, to capture the consumption of the modal user, we cast this problem as a discrete choice problem where a user chooses to consume a single good from $\mathcal{M}$ or take up the outside option at the time granularity of one day.[3]

Model Formulation. Given this formulation, we consider the following utility specification for user i and good j at time t:

$$u_{ijt} = A_{it}B_j^\top + \sum_{r \in \{\text{billboard, top 25, top 100}\}} \beta_{jr}\mathcal{C}_{irt} + \varepsilon_{ijt} \tag{1}$$

and we normalize the utility of the outside option $u_{i0t} = \varepsilon_{i0t}$. Again, the outside option here denotes not consuming any good on Netflix on a given day.

This utility formulation is intrinsically in good space as it considers that our goal is to learn the matrix of user $\times$ good utilities. We follow a similar formulation as Kallus and Udell (2016) by assuming that this can be decomposed into low-rank good and user matrices. We exogenously impose the rank of these matrices, such that B is $\mathbb{R}^{J \times d}$ and A_t is $\mathbb{R}^{I \times d}$. Beyond these two terms, which represent the "intrinsic" preferences that users have for the goods, we explicitly model the role of recommendations where $\mathcal{C}_{it} \equiv \cup_{r \in \{\text{billboard, top 25, top 100}\}} \mathcal{C}_{irt}$ denotes the set of recommendations that user i observes at time t. We handle the granularity of the different types of recommendations (as

[3] We consider consumption of a good as whether a user watches the good for a pre-specified number of minutes. In the cases when a user watches multiple times in a given day, we randomly choose among those goods. We make this assumption since it simplifies the problem dramatically and empirically captures most viewership.

described in Section 2) by having separate recommendation utility booster terms for the billboard, top 25, and top 100 recommendations (denoted by β_{jr}). Finally, ε_{ijt} denotes the standard Type-1 Extreme Value error. Under the standard assumption that ε_{ijt} are independent and identically distributed, this induces the probability that good j will be chosen by user i in time period t:

$$\Pr(\text{i chooses j at time t}) = \frac{\exp u_{ijt}}{\sum_{k \in \mathcal{M} \cup \{0\}} \exp u_{ikt}}$$

A particularly important quantity for our counterfactuals and applications of the model is whether a user chooses any good. Thus, we define whether a user i *engages* in a time period t as follows:

$$\text{ENGAGEMENT}_{it}(\mathcal{M}) = \sum_{m \in \mathcal{M}} \Pr(\text{i chooses m at time t}) = 1 - \Pr(\text{i chooses } 0 \text{ at time t}) \quad (2)$$

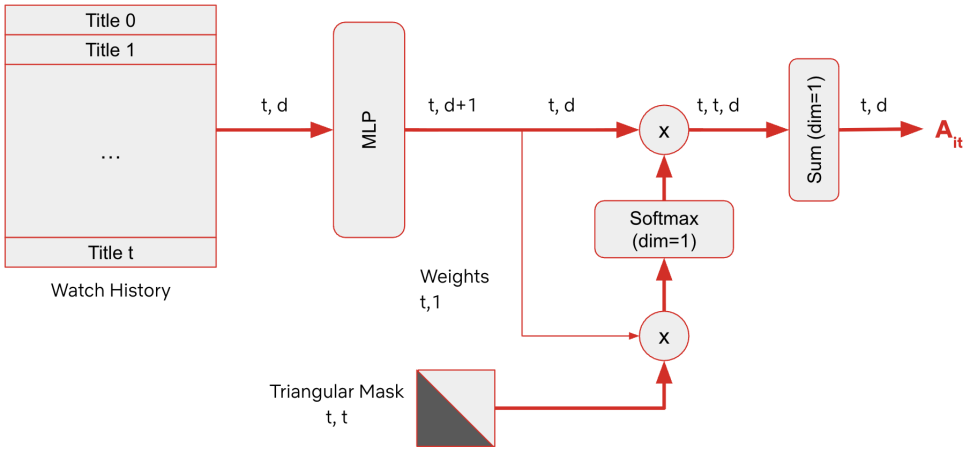
Model Challenges and Assumptions. There are three key challenges for credibly modeling user choices that we tackle with the structure of the choice model.

The first challenge is that, in our empirical context, recommendations play a large role in driving user choices. In principle, recommendations can influence user choices through two mechanisms: making users aware of goods they did not know existed and alleviating information or search frictions conditional on awareness. We follow Aridor et al. (2023) in assuming that the latter is the dominant force and, as such, model the effect of recommendation as providing an additive utility bonus for the set of goods that are recommended, which provides a reduced-form formulation of the (positive) informational role that recommendations play in helping guide user choices.

The second challenge is that modeling demand for movies and TV shows has traditionally been difficult since traditional good characteristics (e.g., genre, cast, etc.) are typically not sufficient to explain user choices (Basuroy et al., 2003; Einav, 2007; McKenzie, 2023). Instead of relying on traditional characteristics, we endogenously learn the good embeddings via user choices in a similar spirit as traditional matrix factorization approaches used in recommender systems (Mnih and Salakhutdinov, 2007; Koren et al., 2009). As in Kallus and Udell (2016), the key difference is that we embed this within a discrete-choice model that allows us to flexibly learn these good embeddings. While this is our preferred specification, we nonetheless show that the model performs reasonably well if we use good-level embeddings that are learned on pre-consumption good characteristics and tags in Appendix B.

The third challenge is that there may be serial dependence in choices, for instance due to users being more likely to watch more sports-related show after watching a sports-related movie. We let the preference weights, A_{it}, vary over time and encapsulate everything the user i has watched up

Figure 2: Preference Weight (A_{it}) Sequence Model



to period $t-1$, based on a sequence model of prior watch history as diagrammed in Figure 2. At a given time t, we encode the good embeddings B_j in an ordered sequence of when the user watched them. We pass that vector through a two-layer fully-connected MLP that produces, for each time period t, a set of attention weights for each previous time period and a transformed, user-specific, set of characteristics. Taking the weighted sum of these and passing it through a two-layer fully connected MLP again provides our vector of estimates A_{it}. Intuitively, the formulation is similar to a traditional sequence model in machine learning where the future goods that a user consumes depend on their previous sequence and the model endogenously learns position-specific attention weights. This aspect of the formulation is a modified variant of state-dependent demand models (Dubé et al., 2010) that combines both time-variation in preferences and "intrinsic" preferences. In addition to addressing serial dependence, this architecture avoids the need to learn user embeddings directly, which becomes highly computationally challenging at the scale of millions of users.

Identification. There are several identification challenges that we face. The first is that recommendations are endogenously targeted towards users because the recommender system predicts that they have high predicted utility for that user. We rely on the intrinsic randomization introduced into the recommender system for the purposes of exploration as exogenous variation in the set of recommendations that breaks this endogenous targeting. This exogenous variation also allows us to identify the effect of state-dependence since users are more likely to consume goods that are present in the recommendations and thus the exogenous variation in choice probabilities provides us with randomization in the initial set of consumed goods. Additionally, the same identification argument as Zeller (2022) applies in our context since goods have a finite amount of consumption time.

Computational Details for Estimation. To estimate our model at scale, we employ stochastic gra-

dient descent (SGD) to maximize the likelihood function, using mini-batch training implemented in PyTorch. Given the scale – millions of users and thousands of goods – mini-batch SGD is essential for both computational efficiency and tractability. We train our model on GPUs using AWS cloud infrastructure, which provides significant speedups over CPU-based computation. To manage the scale of the data, we preprocess user and viewing records using Spark. During training, batches of data are dynamically sampled and pre-fetched into memory, ensuring high GPU utilization and efficient learning.

4 Model Estimates and Validation

We estimate our model using the viewership data for approximately 2 million active Netflix U.S. users for a period of 35 days (February 18, 2025 to March 25, 2025). We first provide the model estimates and then discuss our experimental validation.

4.1 Model Estimates

Since the model endogenously learns the characteristic space, we show in Figure 3 that the model produces reasonable good-level similarities. For instance, it considers "Love Is Blind" most similar to other dating shows like "Married at First Sight" but dissimilar to action films like "Faster", and "Despicable Me 4" most similar to other kids movies like "Sonic the Hedgehog 2" but dissimilar to adult dramas like "A Man in Full".

In Figure 4 we show how a projection of the good embeddings via UMAP (McInnes et al., 2018) corresponds to characteristic good attributes. Figure 4a shows that goods that are distinctly different from others (e.g., kids and documentaries content) are clustered together, while the remaining genres are similarly clustered together though there is considerable overlap with other genres.[4] Additionally, Figure 4b shows a distinct separation between non-English and English goods and Figure 4c highlights that, within clusters, there is a clear separation between movies and television shows. Overall, these figures highlight that the endogenously learned characteristic space is reasonable and accords with intuition, but also subtly captures more complex similarity between goods than can be discerned from standard observable characteristics.

4.2 Model Validation

We conduct several exercises to validate that our model accurately captures substitution patterns and user choices.

[4]This highlights the challenges with using observable characteristics, such as cast and genre, alone.

Figure 3: Most and Least Similar Goods for Selected Goods

Good	Comparison Goods	Similarity
Love Is Blind	Love Is Blind: UK	0.502
	Married at First Sight	0.476
	The Millionaire Matchmaker	0.470
	Red Notice	-0.310
	Faster	-0.356
Zero Day	To Catch a Killer	0.690
	The Sum of All Fears	0.559
	Trial by Fire	0.499
	Total Drama	-0.433
	That Girl Lay Lay	-0.448
American Murder: Gabby Petito	The Search For Instagram's Worst Con Artist	0.559
	Trial by Fire	0.507
	American Murder: Laci Peterson	0.487
	Robin Hood	-0.359
	Wolf King	-0.420
Despicable Me 4	Sonic the Hedgehog 2	0.726
	Plankton: The Movie	0.709
	Minions	0.598
	Gangs of London	-0.448
	A Man in Full	-0.525

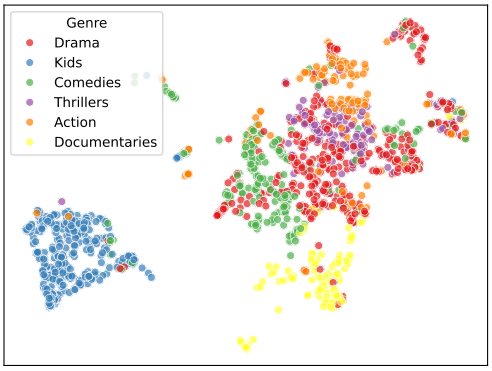
NOTES: Similarity is cosine similarity between learned good embeddings B_j from the baseline model estimated on U.S. viewing. Positive values indicate higher embedding similarity and negative values indicate relative dissimilarity. We show the three most and two least similar goods. The set of comparison goods is chosen for illustration and is not exhaustive.

One key challenge in these environments is that the consumption distribution follows a heavy-tail distribution, with some goods having an outsized share of consumption (McKenzie, 2023), and the challenge with traditional characteristics space models is that the good characteristics such as cast and genre cannot rationalize these patterns. Thus, our first simple validation exercise, presented in Figure 5, shows that our model (in-sample) recovers good-level shares reasonably well.

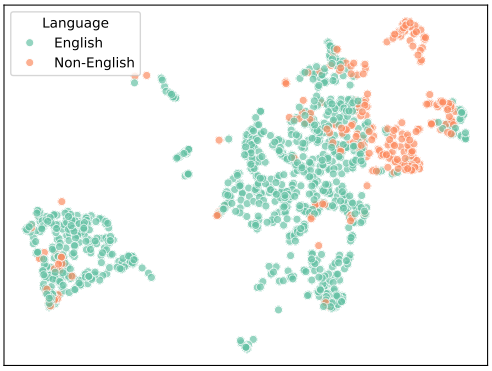
Our next set of validation exercises rely on experimental modifications of the salience of particular goods in the recommendations in order to estimate model-free diversion ratios that we can benchmark against our model.

Figure 4: UMAP-projected Good Embeddings

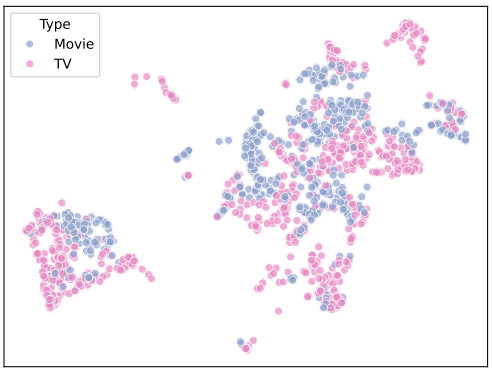
(a) Genre



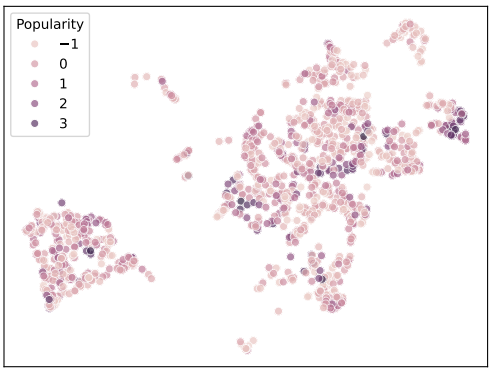
(b) Language



(c) Movie vs. TV

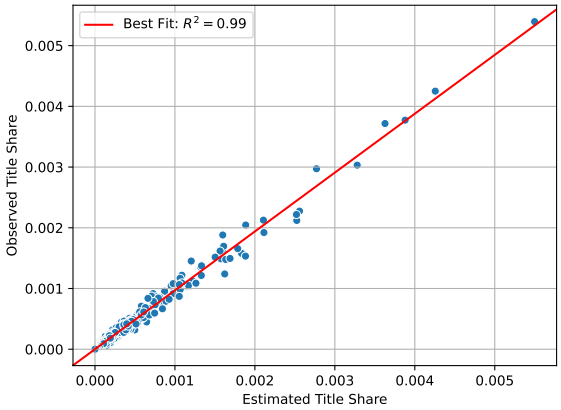


(d) Popularity



NOTES: These figures present two-dimensional UMAP projections of learned good embeddings B_j. Panels color points by (a) genre, (b) language, (c) movie vs. TV, and (d) popularity (share). Projections are for visualization only; all estimation is done in the original embedding space.

Figure 5: Observed Good Shares vs In-Sample Model Estimates



NOTES: This figure presents the comparison of the estimated and actual viewership shares for different goods. Each point is a good; x-axis shows observed share over the estimation window, y-axis shows model-implied in-sample share holding realized recommendations fixed.

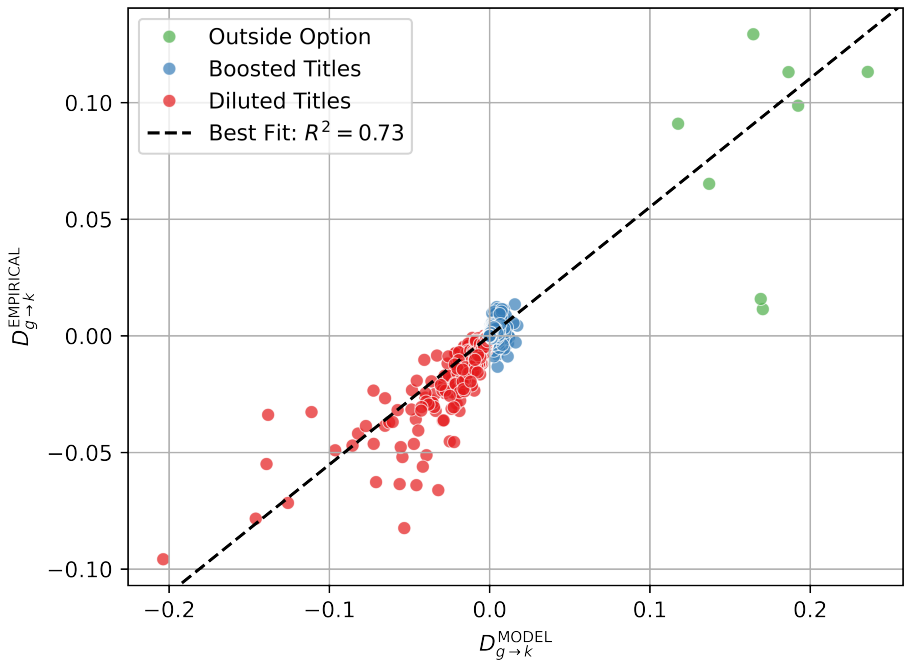
4.2.1 Good Salience Experiments

As discussed in Section 3, we rely on the exploration experiments that the RecSys does for model identification. Re-purposing this randomization, we run a separate experiment that allows us to provide additional validation of the model. In this experiment, we randomly allocate users into eight treatment arms corresponding to a focal good category (e.g., scripted vs. unscripted goods, language, film vs. television series). We exogenously "boost" the salience of the focal category in each arm by increasing the frequency with which goods in that category appear in recommendations. Note that, because the total supply of recommendations is finite, boosting goods in the focal category implies that some goods not in the focal category appear with reduced frequency.

Thus, in each treatment arm, we observe a set of focal goods whose salience is boosted (i.e., show up more in the recommendations) and a set of nonfocal goods whose salience is diluted (i.e., show up less in the recommendations). The resulting fluctuations in good salience are comparable with those in routine experiments that are run on the RecSys (e.g., to test algorithmic improvements). Our experiment coordinates this variation across goods and users to generate more precise and interpretable model-free diversion ratios. We have eight different experimental arms and a control arm. The experiment is run for 5 weeks with each experimental arm including approximately 1 million users.

4.2.2 Comparison to Experimental Estimates

Figure 6: Modeled ($D^{\text{MODEL}}_{g\to k}$) vs. Empirical ($D^{\text{EMPIRICAL}}_{g\to k}$) Good Diversion



NOTES: This figure presents comparison of diversion $D_{g\to k}$ computed from (i) experimental salience interventions (empirical) and (ii) model simulations that emulate the same recommendation restrictions (modeled). For the model-based diversion estimates, we estimate our model on the control arm and simulate the experimental intervention by imputing recommendations using the recommendation-model described in Appendix A. The blue and red points denote goods whose recommendations were boosted and diluted, respectively. The green points denote diversion to the outside for each of the eight separate interventions.

We denote each of the treatment arms as $g \in \mathcal{G}$ and denote $P_k(g)$ as the probability that good k gets chosen under the recommendation restriction g. Therefore, following Conlon and Mortimer (2021) we can use the Wald Estimator to obtain the main diversion measure of interest:

$$D_{g\to k} = \frac{P_k(g) - P_k(\emptyset)}{-\sum_{j\in g}\left(P_j(g) - P_j(\emptyset)\right)} \tag{3}$$

This provides the diversion to a specific good k that results from experimental restriction g. We can therefore compare the performance of our model on these experimental measures of diversion in order to benchmark its performance in learning substitution patterns. We can easily estimate this quantity using the experimental data since $P_k(\emptyset)$ and $P_j(\emptyset)$ represent the empirical probability of choice for k and j, respectively, in the control arm and $P_k(g)$ and $P_j(g)$ represent the empirical

probability of choice, respectively, for k and j in treatment group g. We note that for $k = 0$ this captures the overall change in ENGAGEMENT as a result of the intervention. We denote the experimental estimates as $D^{\text{EMPIRICAL}}_{g \to k}$.

In order to compare the performance of our model to these empirical quantities, we first estimate our model on the control arm. Then, we simulate each experimental group $g \in \mathcal{G}$ separately and produce the model estimates of $D^{\text{MODEL}}_{g \to k}$. Our main model validation test will then be to compare $D^{\text{MODEL}}_{g \to k}$ to $D^{\text{EMPIRICAL}}_{g \to k}$. In order to perform this validation exercise, we need to emulate the good salience intervention on the control arm. This requires us to define a procedure that replaces goods that appear in the recommendations with the goods that the RecSys would replace them with. In Appendix A, we provide two imputation procedures and validate them on the treatment arms of the good salience experiments. For the validation exercise, we use the recommendation model described in Appendix A, which leads to an $R^2 = 0.69$ when benchmarked against the actual recommendations generated during the salience experiment.

In Figure 6, we compare the estimated diversion ratios between the model ($D^{\text{MODEL}}_{g \to k}$) and the experimental estimates ($D^{\text{EMPIRICAL}}_{g \to k}$). The figure shows the diversion to goods that were "boosted" in blue, to goods that were "diluted" in the treatment groups $\mathcal{G}$ in red, and to the outside option in green. We find that the association between the counterfactual diversion ratios is strong with a correlation of 0.86 and an R^2 of 0.73. These results show that our model is able to reproduce out-of-sample estimates of both good-level diversion as well as diversion to the outside option, indicating its ability to capture substitution patterns.

5 Estimating the Value of Recommendations and Goods

In this section, we describe two applications of our model: first, we quantify the value of recommendations in terms of the incremental engagement they generate; second, we illustrate how to use the model to quantify the incremental demand for goods.

5.1 Value of Recommendation and Targeting

We quantify the overall value of the current RecSys on overall engagement and then investigate how this varies across goods. First, we compare the engagement under different reasonable recommendation algorithms to the currently deployed one. This allows us to quantify the overall gains in platform-wide engagement as a result of the current RecSys. Second, we compare the value of recommendations for specific goods and quantify the incremental value of targeting for each of these goods under the current RecSys. This provides us with a precise measure for the value of targeting and how this varies across good-level characteristics.

5.1.1 Value of the Current RecSys

Since we have separately isolated the effect of recommendations from the baseline preferences of users, we can credibly simulate the impact of different recommender systems on choices. We consider the impact of these different recommendation policies on two key variables: (i) overall engagement – how much incremental consumption is generated – and (ii) the diversity of goods that get consumed. Although overall engagement is the most relevant measure for the platform as it provides a measure for how much the recommender system is helping users find relevant goods, diversity is also important. This is because consumption diversity is beneficial in the long-run: it has been shown to correlate with higher long-run satisfaction for users (Anderson et al., 2020; Chen et al., 2024) and it helps to sustain a healthier two-sided market by distributing viewership more evenly across good producers.

We consider two measures of good-level diversity in consumption, where s_x denotes consumption share of good x:

$$\text{HHI} = \sum_{i=1}^{N} s_i^2, \quad G = \frac{1}{2N} \sum_{i=1}^{N} \sum_{j=1}^{N} |s_i - s_j| \tag{4}$$

The Herfindahl–Hirschman Index (HHI) captures absolute concentration, as it places greater weight on goods with large exposure. It effectively measures how many goods receive meaningful viewership: as the consumption share concentrates on fewer goods, HHI rises toward one. The Gini coefficient, in contrast, captures relative inequality in exposure. A higher Gini indicates that exposure is more unevenly distributed, even if many goods are still being viewed. Conceptually, HHI summarizes how broad the set of consumed goods is, while Gini describes how balanced exposure is within that set.

We compute ENGAGEMENT and consumption diversity under the current RecSys compared to when $\mathcal{C}_{it}$ is generated using the following alternative algorithms:

1. **Random Recommendations** - $\mathcal{C}_{it}^{\text{RANDOM}}$: Every user gets a set of recommendations chosen uniformly at random from $\mathcal{M}$.

2. **Popularity Recommendations** - $\mathcal{C}_{it}^{\text{POPULAR}}$: Every user gets the same set of recommendations that are the top N goods in terms of market share in $\mathcal{M}$.

3. **Matrix Factorization Recommendations** - $\mathcal{C}_{it}^{\text{FACTOR}}$: The recommendations are computed using a matrix factorization algorithm (Koren et al., 2009). This captures the "state of the art" approach to computing recommendations from 2015 as discussed in Gomez-Uribe and Hunt (2015).[5]

[5]We employ a vanilla SVD matrix factorization approach, without tuning or hyperparameter optimization. This

To perform these counterfactual estimates, we simulate each alternative recommendation system by replacing the set of recommended goods according to the relevant algorithm, keeping the number of recommendations each user receives constant. Our implementation takes into account several complexities of the actual good interface and business logic. Specifically, we model the effects of recommendations via a separate parameter for each of four different locations on the page the recommendations are present. This accounts for the fact that recommendations in certain locations (e.g., as described in Section 2) are more salient and influential. When constructing counterfactual recommendations, we also disallow recommending goods that a user has already watched, in line with the platform's discovery-oriented business rules. Changes to these modeling choices can significantly affect the resulting engagement and diversity metrics for a counterfactual recommendation algorithm, but not the relative ordering.

The results of the comparison across algorithms is summarized in Table 1, reported in terms of percent deviations from the current RecSys.

Table 1: Engagement and Diversity under Counterfactual RecSys Algorithms

Model	**Δ ENGAGEMENT (%)**	**Δ Gini (%)**	**Δ HHI (%)**
Current RecSys	-	-	-
Random	-16	-0.7	-2.5
Popularity	-12	-0.1	$+42.5$
Matrix Factorization	-4	$+1.1$	$+37.5$

NOTES: This table reports percent deviations from the currently deployed recommender system ("Current RecSys") for overall engagement and two consumption-diversity metrics. ENGAGEMENT is the probability of any play in a period as defined in Equation (2). The Gini and HHI values are computed using Equation (4).

First, these results imply that the current algorithm leads to large improvements in engagement relative to each of the benchmark algorithms. Reverting to random, popularity, or matrix factorization algorithms would reduce engagement by 16%, 12%, and 4%, respectively. These gains are quantitatively large and economically meaningful—e.g., internal estimates would imply that reverting back to matrix factorization recommendations would lead to similar monetary losses as the estimates from Gomez-Uribe and Hunt (2015) indicate reverting back to popularity from matrix factorization would induce in 2015.[6] This improvement highlights that, despite recent concerns in the RecSys community that modern algorithmic improvements are providing little gain relative to traditional approaches such as matrix factorization (Ferrari Dacrema et al., 2019), the latest algo-

is intended to provide a reasonable baseline for personalized recommendations, but does not represent a specific recommendation system previously employed by Netflix.

[6]These estimates are likely lower-bounds, given that our implementation of the recommendation bonus implicitly incorporates the trust that the users have with the RecSys, which would likely be lower under these counterfactual policies.

rithmic improvements applied at one of the most prominent deployments of RecSys are leading to meaningful gains.

Second, we report both the HHI and Gini Index deviations from moving from the current RecSys to the alternative recommendation policies. We expect that random recommendation should lead to higher diversity, since by design it provides exposure to a wider set of goods. In line with this, we find that random recommendations lead to higher good diversity across both measures than each of the proposed recommendation policies. However, the difference is not dramatic—moving from the current RecSys to random recommendations leads to only a 0.7% and 2.5% reduction in these metrics, but with a dramatic increase in engagement. In contrast, both matrix factorization and popularity-based recommendation policies would lead to dramatic increases in concentration – 37.5% and 42.5%, respectively – as well as a +1.1% and -0.1% change in the Gini coefficient. These patterns indicate that both algorithms would substantially narrow the set of goods receiving meaningful viewership. The popularity-based policy concentrates exposure on a small group of equally dominant goods, whereas matrix factorization further amplifies inequality within this group.

Overall, these results show that the current RecSys meaningfully improves engagement while maintaining relatively high consumption diversity. In contrast, traditional matrix-factorization and popularity-based algorithms would lead to large reductions in engagement and concentrate viewership on a much narrower set of goods. This underscores that, despite the concerns in Ferrari Dacrema et al. (2019), modern improvements to recommendation algorithms are both helping match users with their preferred goods and resulting in improved consumption diversity which is important for sustaining a broader good ecosystem.

5.1.2 Good-Level Value of Recommendation

We now turn to understanding the value of recommendation for *particular* goods by characterizing the incremental consumption for a good that is generated by the recommendation as well as the value of targeting in the effectiveness of recommendations. To begin, we define the potential outcomes for a particular user i and good j as follows:

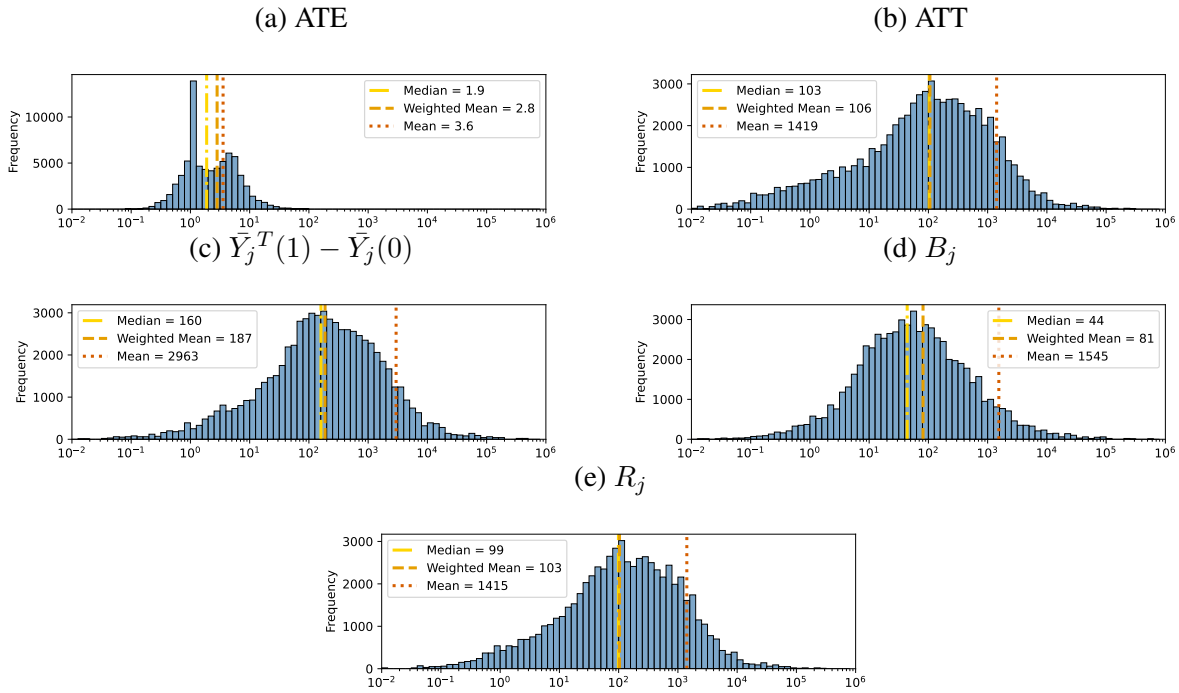
$$Y_{ij}(1): \text{ consumption probability if good } j \text{ is recommended to user } i$$
$$Y_{ij}(0): \text{ consumption probability if good } j \text{ is } \textit{not} \text{ recommended to user } i$$

Thus, using our model we can simulate both potential outcomes. To estimate $Y_{ij}(1)$ for each user i, we force good j to be in each user's recommended set[7] and to estimate $Y_{ij}(0)$ we remove

[7]If $j \in \mathcal{C}_{it}$ originally, then we leave the rest of $\mathcal{C}_{it}$ same. If $j \notin \mathcal{C}_{it}$, then we take the average consumption probability of N random goods to be replaced with good j in $\mathcal{C}_{it}$.

good j from $\mathcal{C}_{it}$. Thus, given this the causal effect of recommending good j to user i is $\tau_{ij} \equiv Y_{ij}(1) - Y_{ij}(0)$.

Figure 7: Good-level Recommendation Outcomes



NOTES: We plot the distributions of each of the terms in (5). To ease interpretation, we divide each through by $\bar{Y}_j(0)$. The different panels represent the following: (a) ATE_j (average treatment effect of recommending good j to a random user), (b) ATT_j (average treatment effect on the treated, i.e., users who were actually targeted with j), (c) $\bar{Y}_j^T(1) - \bar{Y}_j(0)$ (difference between targeted-with-recommendation and population-without-recommendation probabilities), (d) B_j (selection term), and (e) R_j (targeting term), as defined in (5). For each term, we report the median, baseline observation-weighted average, and mean across all goods.

We can then estimate the average treatment effect (ATE) of recommending good j by

$$\text{ATE}_j = \frac{1}{I}\sum_{i=1}^{I} \tau_{ij}$$

which captures the incremental consumption that would result if j were recommended to a random user and the distribution of ATE_j is presented in Figure 7a. However, the ATE does not fully measure the value of the recommendation for specific goods as they are targeted by design towards users for whom the good would be a good fit. Thus, it is possible that some goods that have a low ATE of recommendations have a large effect for the users that they are actually targeted to. To characterize this, we consider the average treatment effect on the treated (ATT), using the realized

recommendations as an (endogenous) treatment status. We characterize this as

$$\text{ATT}_j = \frac{\sum_{i:\, j \in \mathcal{C}_{it}} \tau_{ij}}{\sum_i \mathbb{1}\{j \in \mathcal{C}_{it}\}}$$

and we plot the distribution of this across goods in Figure 7b.

The substantially larger value of ATT compared to ATE highlights the importance of targeting in driving the effectiveness of recommendations. Intuitively, recommended goods are more likely to be consumed since (i) they have higher baseline consumption probability for targeted users (*selection*), (ii) there is a base increase from having any good show up in recommendations (*exposure*), and (iii) targeted users have a larger differential responsiveness than an average user (*targeting*). We decompose the relative importance of each of these components in driving consumption for recommended goods.

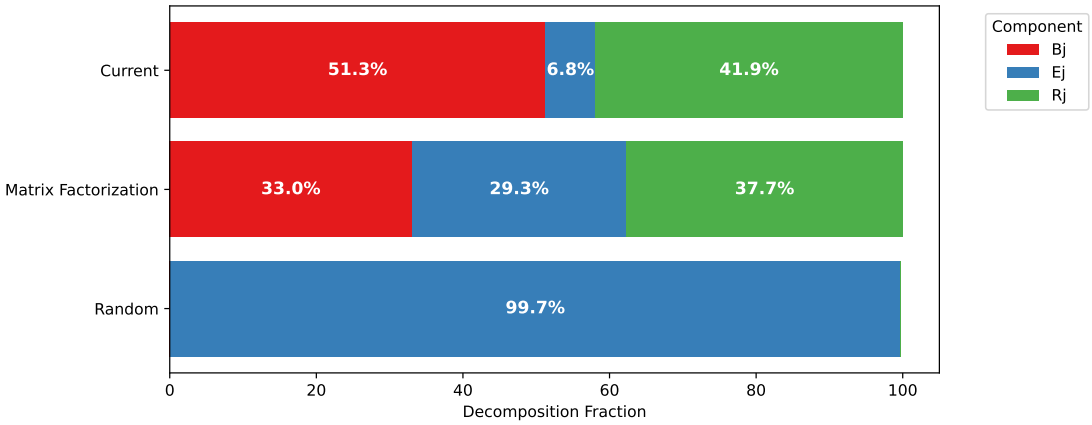
In order to characterize this, we define several additional terms. We define $\bar{Y}_j(d)$ as the average consumption probability in the population with $(d = 1)$ and without $(d = 0)$ recommendation and $\bar{Y}_j^T(d)$ as the average consumption probability in the set of targeted users (i.e., $j \in \mathcal{C}_{jt}$) with $(d = 1)$ and without $(d = 0)$ recommendation. This allows us to express the difference between the consumption probability of a recommended good in the targeted set of users and the consumption probability of a counterfactually not recommended good for an average user as follows:

$$\bar{Y}_j^T(1) - \bar{Y}_j(0) = \underbrace{\bar{Y}_j^T(0) - \bar{Y}_j(0)}_{B_j \text{ (selection)}} + \underbrace{\bar{Y}_j(1) - \bar{Y}_j(0)}_{E_j \equiv \text{ATE}_j \text{ (exposure)}} + \underbrace{\left[(\bar{Y}_j^T(1) - \bar{Y}_j^T(0)) - (\bar{Y}_j(1) - \bar{Y}_j(0))\right]}_{R_j \equiv \text{ATT}_j - \text{ATE}_j \text{ (targeting)}} \quad (5)$$

We compute these terms for each $j \in \mathcal{M}$ and report the distribution as well as the median, mean, and observation-weighted[8] average across goods for each term in Figure 7. We first note that exposure alone has a large effect, leading to a nearly three times increase in consumption probability relative to the baseline consumption probability $\bar{Y}_j(0)$. However, we decompose the relative importance of each term in driving the consumption difference, finding that it is driven by 51.3%, 6.8%, and 41.9% from selection, exposure, and targeting, respectively. This highlights that while recommending a good does meaningfully impact its consumption probability, the impact of targeting is dramatically larger (nearly 7 times larger) than the exposure effect alone. Combined, 93% of the consumption differences comes from the RecSys identifying users with higher baseline consumption probability and with higher incremental value from targeting.

[8]The observation-weighted average weights each good by the total number of times it is observed across user recommendation sets, which corresponds to a typical recommended goods, and mitigates bias from infrequently-recommended long-tail content.

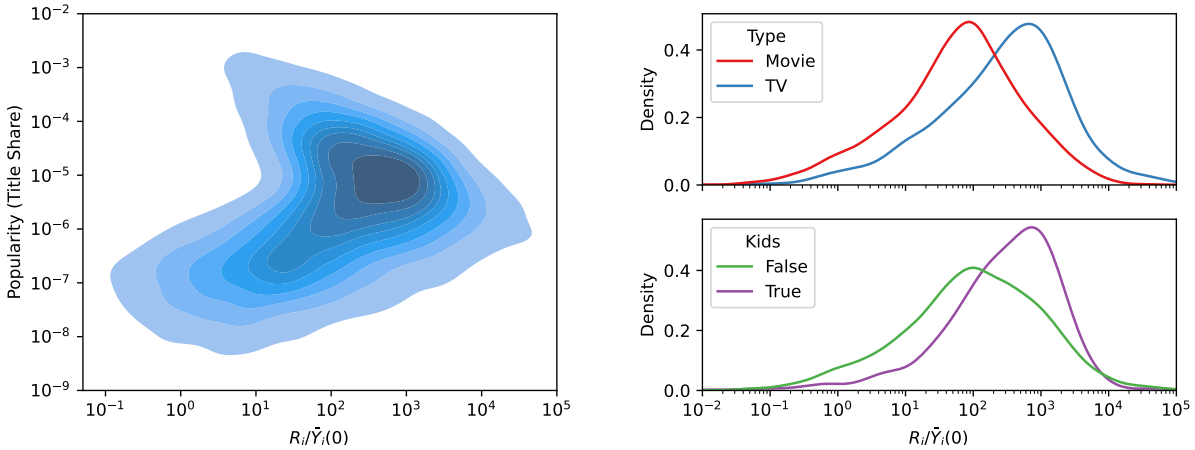
Figure 8: Consumption Decomposition Across Different RecSys Algorithms



NOTES: This figure presents the relative importance of each of the terms in (5) – selection (B_j), exposure (E_j), and targeting (R_j). The top, middle, and bottom rows present the relative importance for the current RecSys, the matrix factorization algorithm described in Section 5.1.1, and the random recommendation algorithm described in Section 5.1.1, respectively.

Figure 9: Targeting Responsiveness Heterogeneity

(a) Density across Popularity vs R_j (b) Density across Good Category vs R_j



NOTES: We plot the kernel densities of R_j (targeting) by (a) baseline popularity ($\bar{Y}_j(0)$) and (b) good categories (TV vs. Movie; Kids vs. Non-Kids). Higher R_j indicates larger incremental responsiveness among targeted users beyond the average exposure effect. Category assignments are derived from internal metadata.

To contextualize these numbers, we provide two benchmarks that are displayed in Figure 8. The top row shows the decomposition for the current RecSys and the bottom two rows show the same decomposition for the matrix factorization and random algorithms discussed in Section 5.1.1. The figure shows that random recommendations are dominated by the exposure effect which is due

to the fact that, by construction, the random recommendation algorithm has no targeting. However, the decomposition for matrix factorization suggests that there is relative parity between each of the different components with exposure playing nearly as large of a role as both selection and targeting. This suggests that the results of increased engagement under the current RecSys from Section 5.1.1 are likely due to effective targeting and that the relative magnitude of targeting would not necessarily be as large even under other popular RecSys algorithms such as matrix factorization.

Finally, in Figure 9 we explore heterogeneity in targeting (R_j) under the current RecSys—the most direct measure of targeting effectiveness for incremental consumption from recommendations—across observables based on baseline popularity ($\bar{Y}_j(0)$) and good category. This comparison allows us to ascertain whether niche goods gain the most from recommendations as well as which categories it is most important for. Figure 9a shows that targeting is most impactful for moderately popular goods. The most popular goods benefit less from targeting because of their extremely broad appeal, while the smallest goods may be too niche to target effectively, and mid-size goods represent a sweet spot of appealing to a sizeable but specific subset of users who may not discover the goods without recommendation. Additionally, we explore variation along other variables where we may expect heterogeneity: whether goods are TV or movies and whether a good is labeled as Kids or non-Kids. Figure 9b shows that TV benefits slightly more from targeting than movies, and Kids goods benefits more from targeting than non-Kids goods.

5.2 The Incremental Demand for Goods

The next application of the model that we consider is to estimate the incremental demand for particular types of goods, including new goods. Estimating this quantity is important for various platform applications, such as catalog optimization or scheduling of releases.

We showcase how one can use the model to estimate these quantities, but due to confidentiality of these numbers do not report the actual estimates. For existing goods, we denote $\mathcal{M}_c$ as the set of existing goods we want to evaluate and define the incrementality of this good as follows:

$$\text{INCREMENTALITY}^{EXISTING}(\mathcal{M}_c) = \sum_i \widehat{\text{ENGAGEMENT}}_{it}(\mathcal{M}) - \widehat{\text{ENGAGEMENT}}_{it}((\mathcal{M} \setminus \mathcal{M}_c))$$

Intuitively, this definition allows us to express how much additional viewership this set of goods brings, accounting for substitution to other goods. In addition to measuring the incremental demand of existing goods, we use a similar definition for the incremental value of a new good ($\mathcal{M}_n$):

$$\text{INCREMENTALITY}^{NEW}(\mathcal{M}_n) = \sum_i \widehat{\text{ENGAGEMENT}}_{it}(\mathcal{M}_n \cup \mathcal{M}) - \widehat{\text{ENGAGEMENT}}_{it}(\mathcal{M})$$

There are two key challenges with estimating these quantities. First, we need to simulate recommendations for goods that aren't currently available on the platform. Second, our model in Section 3 uses endogenously learned good characteristics, which means that it cannot be used for new goods.

For the first challenge, we can rely on the utility-based approach for imputing counterfactual recommendations – described in Appendix A – that uses draws from the predicted user-specific utilities.[9] For the second challenge, we provide an extension of the model that relies on an "exogenous" characteristic space that allows us to simulate counterfactual engagement for new goods.[10]

Model Extension: Exogenous Good Embeddings. We therefore consider an extension of our model that replaces the endogenously learned good-level embeddings with an exogenous set of good characteristics. This allows us to credibly estimate counterfactual engagement for new goods. The empirical challenge is in constructing these good characteristics as this historically has prevented reasonable models of movie demand. We rely on good-level embeddings developed internally at Netflix that provide a high-dimensional representation of a good and capture many hard to directly measure characteristics.[11] Importantly, these embeddings are trained based on pre-consumption characteristics and are primarily driven by human tags associated with each of the goods that are combined with observable characteristics, such as cast and genre, to produce a high-dimensional representation of both existing goods and new goods that are not available on the platform.

We provide estimates of this model and its validation in Appendix B. We report the same set of validation exercises as we conduct for our baseline model in Section 4.2.2 and find that the model performs reasonably well. In particular, we show that the characteristic space is rich enough to predict the empirical good viewership shares with $R^2 = 0.96$ and that it similarly has a strong $R^2 = 0.65$ when comparing the empirical and model-based diversion estimates. This shows that we can use this model to credibly measure the incremental value of good for a wide class of goods.

6 Conclusion

In this paper we studied the problem of separately measuring the value of goods and recommendations on a large streaming platform, Netflix. A key aspect of our approach is to exploit exogenous

[9] We rely on the utility-based estimate procedure since changing the catalog would mean that the recommendation model is based on recommendations under a different catalog which would likely produce incorrect recommendations, especially for new goods.

[10] This has historically been one of the most prominent advantages of characteristic-space approaches to demand (Petrin, 2002).

[11] Recent examples in economics that rely on embeddings in lieu of traditional characteristics spaces in demand models include Quan and Williams (2019); Toubia et al. (2019); Chen et al. (2020); Bach et al. (2024); Magnolfi et al. (2025); Compiani et al. (2025).

variation in the RecSys to both separately identify the effects of recommendations and "intrinsic" preferences on choices in a discrete-choice model as well as to provide model-free diversion ratios that we use to benchmark the performance of our model. We showcase how to apply our model both to quantify the effect of recommendation on engagement as well as to ascertain the incremental value of goods.

Our paper highlights the central role of recommendation in driving user choices in these markets and the large value of the current RecSys on Netflix in several ways. First, exploration experiments in the RecSys can be a powerful lever to learn substitution patterns across goods. Second, we show that compared to both non-personalized and previous state-of-the-art methods for recommendations the current RecSys drives a meaningfully large amount of additional engagement. Third, we decompose the effects of recommendation on individual goods into selection into who gets targeted, the exposure effect of a recommendation, and differential targeting of the recommendations. We show that while exposure meaningfully increases consumption, effective targeting plays a quantitatively larger role and that this effect is most pronounced for mid-sized goods, as opposed to very niche or broadly appealing goods. Overall, our paper provides rich insights into how to measure the value of recommendations for the platform and their impact on consumer behavior.

References

Anderson, A., L. Maystre, I. Anderson, R. Mehrotra, and M. Lalmas (2020). Algorithmic effects on the diversity of consumption on spotify. In *Proceedings of the Web Conference 2020*, pp. 2155–2165.

Anderson, C., C. Nissley, and C. Anderson (2006). The long tail.

Aridor, G. (2025). Measuring substitution patterns in the attention economy: An experimental approach. *The RAND Journal of Economics 56*(3), 302–324.

Aridor, G., D. Goncalves, D. Kluver, R. Kong, and J. Konstan (2023). The economics of recommender systems: Evidence from a field experiment on movielens. In *24th ACM Conference on Economics and Computation, EC 2023*, pp. 117. Association for Computing Machinery, Inc.

Bach, P., V. Chernozhukov, S. Klaassen, M. Spindler, J. Teichert-Kluge, and S. Vijaykumar (2024). Adventures in demand analysis using ai. *arXiv preprint arXiv:2501.00382*.

Basuroy, S., S. Chatterjee, and S. A. Ravid (2003). How critical are critical reviews? the box office effects of film critics, star power, and budgets. *Journal of Marketing 67*(4), 103–117.

Bennett, J. and S. Lanning (2007). The netflix prize.

Brynjolfsson, E., A. Collis, and F. Eggers (2019). Using massive online choice experiments to measure changes in well-being. *Proceedings of the National Academy of Sciences 116*(15), 7250–7255.

Brynjolfsson, E., Y. Hu, and D. Simester (2011). Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales. *Management Science 57*(8), 1373–1386.

Bursztyn, L., M. Gentzkow, R. Jiménez-Durán, A. Leonard, F. Milojević, and C. Roth (2025). Measuring markets for network goods. Technical report, National Bureau of Economic Research.

Calvano, E., G. Calzolari, V. Denicolò, and S. Pastorello (2022). Artificial intelligence, product recommendations and concentration.

Calvano, E. and M. Polo (2021). Market power, competition and innovation in digital markets: A survey. *Information Economics and Policy 54*, 100853.

Chen, F., X. Liu, D. Proserpio, I. Troncoso, and F. Xiong (2020). Studying product competition using representation learning. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1261–1268.

Chen, G., T. Chan, D. Zhang, S. Liu, and Y. Wu (2024). The impact of a more diversified recommender system on digital platforms: Evidence from a large-scale field experiment.

Compiani, G., I. Morozov, and S. Seiler (2025). Demand estimation with text and image data. *arXiv preprint arXiv:2503.20711*.

Conlon, C., J. Mortimer, and P. Sarkis (2023). Estimating preferences and substitution patterns from second choice data alone.

Conlon, C. and J. H. Mortimer (2021). Empirical properties of diversion ratios. *The RAND Journal of Economics 52*(4), 693–726.

Conlon, C. T. and J. H. Mortimer (2013). An experimental approach to merger evaluation. Technical report, National Bureau of Economic Research.

Das, A. S., M. Datar, A. Garg, and S. Rajaram (2007). Google news personalization: scalable online collaborative filtering. *Proceedings of the 16th International Conference on World Wide Web*, 271–280.

Dertwinkel-Kalt, M., V. Eulenberg, and C. Wey (2024). Defining what the relevant market is: A new method for consumer research and antitrust.

Donnelly, R., A. Kanodia, and I. Morozov (2024). Welfare effects of personalized rankings. *Marketing Science 43*(1), 92–113.

Dubé, J.-P., G. J. Hitsch, and P. E. Rossi (2010). State dependence and alternative explanations for consumer inertia. *The RAND Journal of Economics 41*(3), 417–445.

Einav, L. (2007). Seasonality in the us motion picture industry. *The RAND Journal of Economics 38*(1), 127–145.

Ferrari Dacrema, M., P. Cremonesi, and D. Jannach (2019). Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pp. 101–109.

Fleder, D. and K. Hosanagar (2009). Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management Science 55*(5), 697–712.

Freyne, J., M. Jacovi, I. Guy, and W. Geyer (2009). Increasing engagement through early recommender intervention. *Proceedings of the 3rd ACM Conference on Recommender systems*, 85–92.

Gomez-Uribe, C. A. and N. Hunt (2015). The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS) 6*(4), 1–19.

Heckman, J. J. (1981). Heterogeneity and state dependence. In *Studies in Labor Markets*, pp. 91–140. University of Chicago Press.

Heckman, J. J. (1991). Identifying the hand of past: Distinguishing state dependence from heterogeneity. *The American Economic Review 81*(2), 75–79.

Holtz, D., B. Carterette, P. Chandar, Z. Nazari, H. Cramer, and S. Aral (2020). The engagement-diversity connection: Evidence from a field experiment on spotify. pp. 1–57.

Hosanagar, K., D. Fleder, D. Lee, and A. Buja (2013). Will the global village fracture into tribes? recommender systems and their effects on consumer fragmentation. *Management Science 60*(4), 805–823.

Joshi, S., Y. Feng, K.-J. Hsiao, Z. Zhang, and S. Lamkhede (2024). Sliding window training - utilizing historical recommender systems data for foundation models. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys '24, New York, NY, USA, pp. 835–837. Association for Computing Machinery.

Kallus, N. and M. Udell (2016). Revealed preference at scale: Learning personalized preferences from assortment choices. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pp. 821–837.

Koren, Y., R. Bell, and C. Volinsky (2009). Matrix factorization techniques for recommender systems. *Computer 42*(8), 30–37.

Korganbekova, M. and C. Zuber (2023). Balancing user privacy and personalization.

Lee, D. and K. Hosanagar (2019). How do recommender systems affect sales diversity? a cross-category investigation via randomized field experiment. *Information Systems Research 30*(1), 239–259.

Lee, P. S., V. Kumar, and K. Sudhir (2025). Pay now or wait to unlock? cliffhangers and monetization of serialized media.

Magnolfi, L., J. McClure, and A. Sorensen (2025). Triplet embeddings for demand estimation. *American Economic Journal: Microeconomics 17*(1), 282–307.

McInnes, L., J. Healy, N. Saul, and L. Grossberger (2018). Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software 3*(29), 861.

McKenzie, J. (2023). The economics of movies (revisited): A survey of recent literature. *Journal of Economic Surveys 37*(2), 480–525.

Mnih, A. and R. R. Salakhutdinov (2007). Probabilistic matrix factorization. *Advances in Neural Information Processing Systems 20*.

Morozov, I., S. Seiler, X. Dong, and L. Hou (2021). Estimation of preference heterogeneity in markets with costly search. *Marketing Science 40*(5), 871–899.

Nguyen, T. T., P.-M. Hui, F. M. Harper, L. Terveen, and J. A. Konstan (2014). Exploring the filter bubble: the effect of using recommender systems on content diversity. *Proceedings of the 23rd International Conference on World Wide Web*, 677–686.

Petrin, A. (2002). Quantifying the benefits of new products: The case of the minivan. *Journal of Political Economy 110*(4), 705–729.

Peukert, C., A. Sen, and J. Claussen (2024). The editor and the algorithm: Recommendation technology in online news. *Management Science 70*(9), 5816–5831.

Quan, T. W. and K. R. Williams (2019). Extracting characteristics from product images and its application to demand estimation.

Raval, D., T. Rosenbaum, and N. E. Wilson (2022). Using disaster-induced closures to evaluate discrete choice models of hospital demand. *The RAND Journal of Economics 53*(3), 561–589.

Senecal, S. and J. Nantel (2004). The influence of online product recommendations on consumers' online choices. *Journal of Retailing 80*(2), 159–169.

Simonov, A., J.-P. Dubé, G. Hitsch, and P. Rossi (2020). State-dependent demand estimation with initial conditions correction. *Journal of Marketing Research 57*(5), 789–809.

Steck, H., L. Baltrunas, E. Elahi, D. Liang, Y. Raimond, and J. Basilico (2021). Deep learning for recommender systems: A netflix case study. *AI Magazine 42*(3), 7–18.

Toubia, O., G. Iyengar, R. Bunnell, and A. Lemaire (2019). Extracting features of entertainment products: A guided latent dirichlet allocation approach informed by the psychology of media consumption. *Journal of Marketing Research 56*(1), 18–36.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017). Attention is all you need. *Advances in Neural Information Processing Systems 30*.

Yuan, H. (2025). Competing for time: A study of mobile applications. Technical report.

Zeller, J. (2022). Drip or drop: Consumption constraints & habit formation in tv streaming.

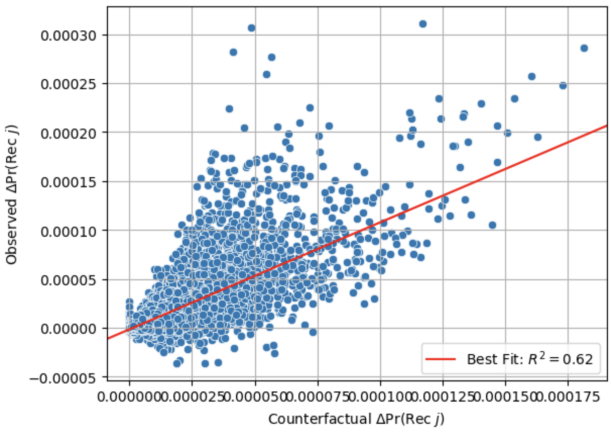
Zhou, R., S. Khemmarat, and L. Gao (2010). The impact of youtube recommendation system on video views. *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, 404–410.

Appendix

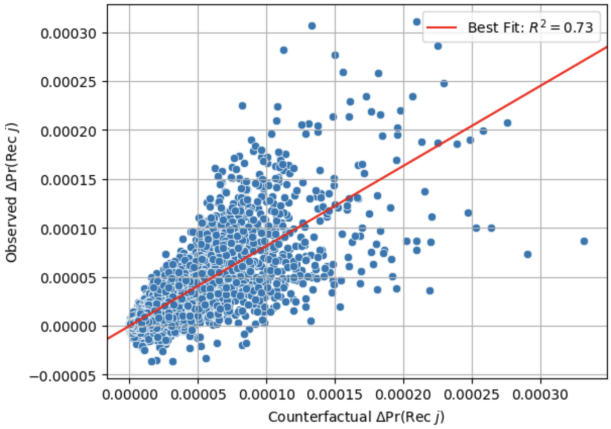
A Imputing Counterfactual Recommendations

Figure A.1: Counterfactual vs. Observed Change in Recommendation Rate for Boosted Goods

(a) Utility-based estimate



(b) Recommendation model



NOTES: For boosted goods in the salience experiments, we plot the change in recommendation rate predicted by (Panel a) the utility-based replacement rule and (Panel b) the recommendation-model replacement against the realized change. Each point aggregates over users within an experimental arm and represents a single good. These procedures are used only to emulate recommendation replacement; demand estimation remains as specified in Section 3.

In this section we describe two methods for determining how to replace goods that appear in the recommendations. Since it is not possible for us to directly query the RecSys, we provide two procedures to do this imputation: a *recommendation model* that estimates a separate model to learn the substitution patterns of the recommender system and a *utility-based* estimate that relies only on the outputs from the demand model itself. Each of these procedures is useful for different applications – for instance, the *utility-based* procedure is important when we are considering counterfactual catalogs and the *recommendation* model is important for ensuring we have an accurate as possible imputation of the recommendations during the intervention that we need for model validation.

The recommendation model approach involves training a model to predict which recommendations a user with a given history will receive. The structure of this model uses the same approach as our demand model for learning good embeddings B_j and user embeddings A_{it} as a function of watch history. However, rather than estimating a probability across a choice set, we estimate the

probability of each good being recommended as:

$$\Pr(\text{Recommended } j \text{ to user } i \text{ at time } t) = \sigma(A_{it} B_j^\top)$$

The utility-based procedure relies on the intuition that goods that appear in recommendations are those that have high predicted utility for user i. Thus, if a good is not allowed to appear in the recommendations, the next most substitutable good will also be one that has high predicted utility. While, of course, there is much more nuance in practice to the recommendation system, our procedure follows this intuition by sampling goods using a softmax rule based on their predicted utility. Specifically, we consider the following replacement rule for $j \notin \mathcal{C}_{it}$:

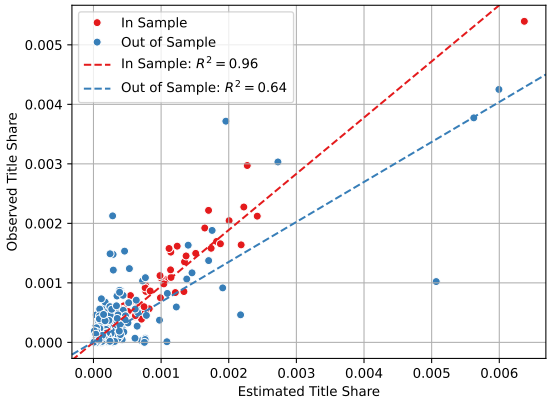
$$\Pr(\text{Recommend } j \text{ to user } i) = \frac{\exp(u_{ij})}{\sum_{k \in \mathcal{M}, k \notin \mathcal{C}_{it}} \exp(u_{ik})}$$

We empirically validate both of these procedures by estimating the counterfactual recommendation rate in each treatment arm using our replacement rule. Then, we can subtract the baseline recommendation rate in the control group to obtain the increase in recommendations for boosted items. Finally, we can compare this counterfactual estimate to the observed change in recommendations for these items, as shown in Figure A.1. This validation demonstrates that we get reasonable estimates for the replacement rate for the boosted recommendations which are noisy but still sufficiently well-correlated for our purposes ($R^2 = 0.62$ under the utility-based estimate and $R^2 = 0.73$ under the recommendation model).

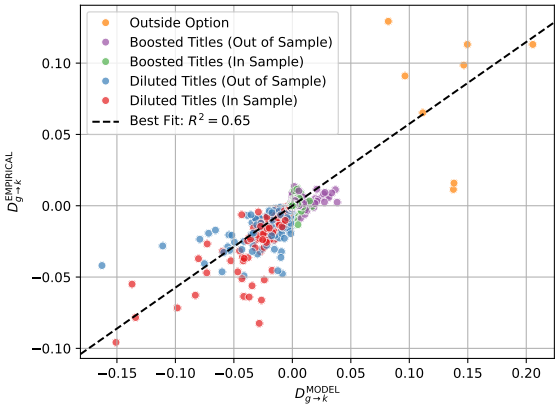
B Exogenous Embeddings Model Validation

Figure B.1: Exogenous Embedding Cross-Validation

(a) Good Share Cross-Validation



(b) Good Diversion Cross-Validation



NOTES: This figure presents the validation of the exogenous embeddings demand model. Panel A presents the comparison between the model-based and empirical good shares. Panel B presents the comparison between the model-based and empirical good diversion. "Out of sample" vs. "in sample" is defined as follows. We randomly split the goods into two groups during model training. For the first group, we use the exogenous embeddings, and for the second group we learn endogenous embeddings according to the model in Section 3. Then for evaluation, we replace all endogenous embeddings with exogenous embeddings, and label these goods as "out of sample".

In this section we describe and validate an alternative version of the model from Section 3 where instead of endogenously learning the good embeddings, we rely on an exogenous characteristic space to represent goods. As discussed in the main text, a key challenge with historical demand models for movies is that choices can typically not be rationalized by standard characteristics such as cast and genre. As such, we rely on high dimensional representations of goods that are developed internally at Netflix and are constructed using pre-consumption human tags for each good.

We incorporate this into the model by imposing that B_j is fixed through the estimation procedure according to this representation of goods. One challenge is that the dimensionality of the embeddings is still too high (256) for us to feasibly estimate in our model and so we include a two-layer, fully-connected multi-layer perceptron within the model to reduce the dimension from 256 dimensional to d dimensional.

We validate this model using the same exact procedure as described in Section 4.2 for the baseline model. Additionally, a concern is that, if the dimensionality reduction is built into the model training and rich enough, the model will effectively convert the exogenous embeddings into the endogenous embeddings, making it unusable for extrapolation to unseen goods (which is the primary purpose for the exogenous embeddings model). Thus, in addition to the exercises

done in Section 4.2, we randomly split the goods into two groups during model training. For the first group, we use the exogenous embeddings, and for the second group we learn endogenous embeddings according to the model in Section 3. Then for evaluation, we replace all endogenous embeddings with exogenous embeddings, and label these goods as "out of sample".

The validation results are presented in Figure B.1. First, Figure B.1a shows that the in-sample fit for the market shares is fairly strong with $R^2 = 0.96$ and with the model able to rationalize the most watched titles. Additionally, it shows that the "out of sample" $R^2 = 0.64$, indicating that the model does reasonably well at extrapolating to unseen goods. Second, Figure B.1b shows that the model also does well at recovering the model-free diversion ratios with an $R^2 = 0.65$. Overall, we conclude that the model with the exogenous embeddings is able to sufficiently capture consumer choices for our purposes.