Finding Nemo (in a sea of options): An analysis of catalog size and movie concentration on a Video-on-Demand platform

Chiara Thérèse Dalle Donne

Dissertation written under the supervision of professor
Miguel Godinho de Matos

Dissertation submitted in partial fulfillment of requirements for the MSc in International Management, at the Universidade Católica Portuguesa, June 1, 2023.

Finding Nemo (in a sea of options): An analysis of catalog size and movie concentration on a Video-on-Demand platform

Chiara Thérèse Dalle Donne

June 1, 2023

Abstract

Recently, the media industry has undergone a significant transformation, expanding its offering to include blockbusters and niche titles. Consequently, catalog managers face the challenge of deciding which products to include on their platforms and striking the right balance in terms of catalog size. This study aims to examine the influence of catalog size on consumption concentration. To investigate this relationship, we analyze secondary data from a two-month period provided by a Video-on-Demand platform, during which the catalog size varied. While previous studies supported that catalog size significantly impacts the concentration of views, our findings reveal that the relationship between catalog size and view concentration is not statistically significant. Instead, we find that factors related to the nature of titles, such as genre and position on the platform, have more substantial explanatory power in understanding viewership concentration. These results emphasize that there are other critical factors to consider when examining the concentration of views, which outweigh the significance of catalog size.

Keywords: Catalog size, Concentration, Video-on-Demand, Variety, Assortment

Finding Nemo (in a sea of options): An analysis of catalog size and movie concentration on a Video-on-Demand platform

Chiara Thérèse Dalle Donne

June 1, 2023

Abstract

Recentemente, a indústria dos media sofreu uma transformação significativa, expandindo a sua oferta para incluir êxitos de bilheteira e títulos de nicho. Consequentemente, os gestores de catálogos enfrentam o desafio de decidir quais os produtos a incluir nas suas plataformas e de encontrar o equilíbrio certo em termos da dimensão do catálogo. Este estudo tem como objectivo analisar a influência da dimensão do catálogo na concentração do consumo. Para investigar esta relação, analisamos dados secundários de um período de dois meses fornecidos por uma plataforma de vídeo a pedido, durante o qual a dimensão do catálogo variou. Embora estudos anteriores tenham sustentado que a dimensão do catálogo tem um impacto significativo na concentração de visualizações, os nossos resultados revelam que a relação entre a dimensão do catálogo e a concentração de visualizações não é estatisticamente significativa. Em vez disso, descobrimos que os factores relacionados com a natureza dos títulos, como o género e a posição na plataforma, têm um poder explicativo mais substancial para compreender a concentração. Esses resultados enfatizam que há outros fatores críticos a serem considerados ao examinar a concentração de visualizações, que superam a importância do tamanho do catálogo.

Keywords: Catalog size, Concentration, Video-on-Demand, Variety, Assortment

Acknowledgements

I would like to express my gratitude to my supervisor Miguel Godinho de Matos, for his support and guidance, which made this work possible. The project has been partially supported by the Portuguese Foundation for Science and Technology through research grant PTDC/EGE-OGE/27968/2017. We thank the industry partner for their support.

Additionally, I would like to express my gratitude towards my supportive friends who accompanied me on this journey.

Lastly, I extend my sincere appreciation to my family for their unwavering support and encouragement, which has been invaluable to my academic and personal development.

Contents

List of Figures

List of Tables

1 Introduction

The Media industry, formerly known as the blockbuster industry (Anderson, 2006), has undergone significant changes in recent years. For decades, booksellers and movie sellers focused on selling hits rather than catering to niche markets (Brynjolfsson et al., 2011). However, with the rise of the internet, there has been a shift from brick-and-mortar stores to online stores. With the new channel, niche products emerged on the virtual shelves. Research has identified two key factors contributing to this trend of niche products in the market. On the supply side, the lower costs and ease of logistics of online stores have made it more feasible for sellers to include niche titles in their catalogs (Brynjolfsson et al., 2003). On the demand side, consumers now have greater access to information and resources, allowing them to explore and discover niche products more easily (Brynjolfsson et al., 2003). This landscape presents a new challenge for catalog managers, who must carefully evaluate the potential benefits and drawbacks of promoting hits versus niche products in order to meet customers' needs and maximize sales and engagement. As a result, catalog managers must review their strategies carefully, evaluating which movies to add, remove or change in order to meet their goals. One identified factor affecting whether consumption concentrates around hits or niches is the size of the catalog. Several authors showed that as the size of catalogs increased over time, consumer behavior has undergone significant changes.

Our analysis focuses on the movie industry and analyzes secondary data of a Video-on-Demand provider offering movies and series on a subscription basis. We contrast data from two months in different years, characterized by different catalog sizes. At the core of our analysis lies the impact of catalog size on the concentration of views.

This study aims to replicate the analysis of Brynjolfsson et al. (2011) in the context of movies and TV shows. In contrast to the prevailing literature on the concentration of consumption in the movie industry, which broadly aligns with the "Long Tail" Theory (Anderson, 2006), our findings suggest that there is limited evidence supporting the existence of a significant effect of catalog size on concentration. When controlling for the genre of a title, our results indicate a small effect. However, it is essential to note that this effect is primarily attributed to changes in the nature of the catalog rather than its size alone. Therefore, it is more accurate to conclude that the observed effect is a consequence of the catalog's evolving composition rather than its overall magnitude. We further strengthen our findings by examining other factors significantly impacting viewer behavior. Given these findings, it becomes imperative to highlight the dynamic nature of catalogs and the profound impact of genre, positioning, and format on shaping user concentration.

This thesis is structured as follows: Section 2 provides a literature review on choices, catalog sizes, and concentration. Section 3 introduces the underlying data, provides descriptive statistics, and outlines the empirical strategy. The regression analyses and their outcomes are presented in Section 4, along with additional analysis to explore findings further. A critical

evaluation of the distributional assumptions is provided in Section 5 to ensure the robustness of our results. The results are discussed in section 6, the limitations outlined in section 7, and the main findings concluded in section 8.

2 Literature Review

In the domain of catalog management, managers face decisions regarding the composition of their catalogs, such as determining which products to add or remove to achieve their goals, which may involve maximizing sales and views. Along with the selection of products also goes the size of the catalog, which has been a topic of ongoing debate in the literature. While there is no definite answer, we first discuss the advantages and disadvantages of large catalogs.

Gourville and Soman (2005) argue that large assortments are necessary and advantageous due to the heterogeneity in customer needs and tastes. The presence of greater variety increases the likelihood of customers finding options that align with their specific needs (Gourville and Soman, 2005) and preferences (Baumol and Ide, 1956). This phenomenon extends to various product characteristics such as sizes (Gourville and Soman, 2005) and variations (Baumol and Ide, 1956). Furthermore, the concept of "option value" suggests that offering more choices can enhance individuals' preference for a product or service by enabling them to select what aligns with their preferences and explore alternative possibilities (Reibstein et al., 1975). Consistent with these findings, Kahn (1998) finds that the introduction of new products can lead to increased market share. Moreover, Kahn and Lehmann (1991) explore the behavior of consumers who are still in the process of determining their preferences. The study reveals that having various desirable options allows customers to be flexible and adaptable in their choices, enabling them to experiment with different alternatives and switch between options if their preferences change over time (Kahn and Lehmann, 1991).

Along with improving the match of consumers' preferences, the research found that introducing new products generates additional consumer surplus. Brynjolfsson et al. (2003) found that in the field of bookstores, in 2000, the consumer surplus increased from $731 million to $1.03 billion by introducing new products and corresponding growth in assortment size. The authors indicate that this welfare gain is "at least five times as large as the consumer welfare gain from increased competition and lower prices in this market" (Brynjolfsson et al., 2003, p. 1).

Moreover, Tversky and Shafir (1992) address whether larger assortments can create difficulties for consumers in decision-making. The authors conclude that as long as a set of options includes a clear distinction between good and bad choices, there are no issues with decision-making (Tversky and Shafir, 1992). These arguments explain why a consistent expansion in catalog size is often observed across catalogs over time, including the dataset we are analyzing.

On the contrary, extensive research indicates obstacles with large assortments and variety, coined as the idea of "choice overload" by Toffler (1970). The phenomenon was further popu-

larized by psychologist Schwartz (2004) as the "Paradox of Choice", which shows that having several options can lead to decision paralysis and decreased satisfaction with the final choice. Diehl and Poynor (2010) provide supporting evidence, showing that when consumers are presented with large assortments, their expectations regarding the extent to which their preferences can be met are heightened. If the catalog does not meet the consumers' expectations, this can translate into a reduction in choice satisfaction. Thus, high hopes can result in more significant disappointment when browsing the catalog (Diehl and Poynor, 2010). Furthermore, Tversky and Shafir (1992) suggest that when faced with a large set of options, consumers are more likely to delay making a choice. This tendency is especially true when each option has significant advantages and disadvantages (Tversky and Shafir, 1992). This notion is further supported by the findings of Dhar (1997), which suggest that having multiple options can even result in consumers not making a choice at all.

Moreover, the research found that introducing new products does not necessarily result in more sales (Dhar, 1997), nor does it always increase market share (Gourville and Soman, 2005). In order to address this question, Gourville and Soman (2005) differentiate between alignable and non-alignable assortments. Alignable choices involve options that can be directly compared, such as the origin of a food product (e.g., corn chips from mid-western or south-western corn). On the other hand, non-alignable choices involve options that cannot be easily compared, like choosing between a car with a sunroof or leather seats (Gourville and Soman, 2005). According to Gourville and Soman (2005), an increase in variety leads to an increase in needed cognitive effort, thereby decreasing market share for non-alignable assortments. Similarly, research on the grocery industry found that brand image is not necessarily decreased when assortment is reduced (Broniarczyk et al., 1998). One explanation for this phenomenon is that consumers do not necessarily have a direct perception of the size of an assortment, which could only be achieved in a grocery store by counting the available products (Broniarczyk et al., 1998). In addition, change in assortment size is only detected by customers over a certain threshold (Marks, 2014). These points, therefore, underscore the need to exercise caution when continuously adding new products, or movies, to a platform. Careful consideration must be given to the factors mentioned above in order to optimize the platform's performance.

While the discussion on the advantages and disadvantages of large catalogs provides valuable insights into the complex nature of catalog management, understanding the effects of changes in catalog size on the concentration of views can provide additional insights into the relationship between catalog size and viewer behavior.

Many researchers have found that consumption pattern changes occur alongside catalog size variations. Tan et al. (2017) discovered that an increase in assortment size can lead to an increase in concentration. Increased product concentration implies that a small proportion of products are responsible for most sales. This phenomenon is known as the "Pareto principle", which states that approximately 80% of the outcomes can be attributed to just 20% of the underlying causes (Pareto, 1964). Tan et al. (2017) observe this phenomenon in the movie industry and

show that demand for individual movies drops with larger catalog sizes. The authors find that this effect is even more prominent for niche products than hits, resulting in a higher demand for hits and overall higher concentration (Tan et al., 2017). In addition, the authors provide evidence that the increase in concentration is not a result of adding mainly niche titles to the platform.

With increased catalog size, recommender systems become increasingly important on platforms. These recommenders aim to "help people move from the world they know [“blockbusters”] to the world they don’t [“niches”]" (Anderson, 2006, p. 109). However, Fleder and Hosanagar (2007) found that these recommenders can further increase the concentration as movies with a successful history are more likely to be recommended. Furthermore, the authors describe the phenomenon as a "self-reinforcing cycle of popularity" (Fleder and Hosanagar, 2007, p. 198), where regularly purchased items are recommended more, and items recommended more are more likely to be purchased. This is one reason why an increase in catalog size may further increase the concentration of views.

Conversely, increasing catalog size can lead to a less concentrated distribution, particularly in the context of the internet. Zentner et al. (2013) found that when consumers move from brick-and-mortar shops to the internet, viewership concentration decreases, meaning that consumption goes towards niche products instead of the hits (Zentner et al., 2013). This phenomenon coined “Long Tail” (Anderson, 2004) means that sales of niche products can grow to become a large share of sales. In other words, the tail of the distribution expands, highlighting the increasing importance and popularity of niche offerings alongside mainstream products.

Brynjolfsson et al. (2003) provide two reasons for the emergence of “Long Tails” in content industries. On the supply side, the internet channel can offer a significantly larger range of products as they do not face the restrictions of space, logistics, and holding costs (Brynjolfsson et al., 2003). The cost of stocking is therefore much lower in the internet channel. In addition, as brick-and-mortar stores sell only locally, mainstream consumers are served first (Brynjolfsson et al., 2006). In the internet channel, also niche customers can be addressed. On the demand side, the search costs for consumers drastically decreased. The introduction of the internet enables consumers to easily gather information more conveniently and at reduced expenses through internet search tools and recommendation engines (Brynjolfsson et al., 2011). The authors prove these findings hold, even when offering the same products and prices (Brynjolfsson et al., 2011). Anderson (2004) concludes that as consumers are now able to discover and purchase products more closely aligned with their tastes, it is likely that they will become less interested in popular products.

Our analysis aims to investigate whether the expansion of catalog size within a specific VoD platform over time leads to changes in the concentration of views. Contrary to the broad findings indicating a long-tail effect, our study does not find sufficient evidence for a significant impact of catalog size on concentration. The discrepancy may arise from the fact that most authors who support the “Long Tail” theory examine the context of catalog size changes occur-

ring during the transition from brick-and-mortar stores to the online environment (Brynjolfsson et al., 2003). In contrast, our analysis focuses on a specific platform with similar settings over time, providing insights into the concentration dynamics within this particular context. On a methodological level, our approach aims to replicate the study conducted by Brynjolfsson et al. (2011). However, we find that our data does not fit the used model very well. Therefore, an alternative model is applied.

In addition, our analysis investigates further factors influencing concentration, such as the potential impact of sort order and the nature of the title, including the genre and whether it is a TV show or movie, on the consumption of titles on our platform. By examining these additional factors, we aim to gain a comprehensive understanding of the dynamics that contribute to the behavior of views.

3 Data and Empirical Methods

3.1 Data

We use secondary data from a telecommunications company, which we refer to as TELCO. TELCO provides services to over a million households in the studied country. In addition, to pay TV, internet, and phone service, TELCO also provides a Video-on-Demand (VoD) system, which will be the focus of this analysis.

Our dataset includes information on the number of views and clients for each movie title in two different months, February 2014 and February 2016. In addition, we received catalog data indicating the period in which the title was listed in the VoD catalog and title level information such as the genre, duration, episode, and season (if applicable) and whether the title is available in high definition. We aggregated title daily views on a monthly level to mitigate the challenges associated with calendar seasonality across different years. We focus our analysis on movies and TV show episodes. We exclude trailers, previews, and TV show marathon titles from the dataset. We thoroughly explain our criteria for excluding specific titles in Appendix A.1.

In addition, we have available information on the daily position of each title within the catalog. The platform has a variety of menus and submenus. Under each of those, titles are sorted in a particular order. To reach the sub-menus, users must click several times. We have information on the depth and position of the titles in the catalog. In any sub-menu, the first six titles are displayed to the users. To view the following items, the user must click again. A title may simultaneously display in multiple sub-menus and change positions within each sub-menu. We use a combination of both level and sort order as a proxy for the number of clicks needed to reach a title. To match the daily catalog position data with the monthly title visualization data, we calculated minimum values for the number of clicks monthly.

Finally, we received a dataset that provides daily views for individual users (client IDs). While most analyses run at the title level, we use this individual level dataset in Section 4.4 to

conduct a robustness check to test the sensitivity of our main findings.

Table 1 presents the key variables used in our analysis to answer our research question.

Table 1: Variables

Variable	**Definition**	**Min.**	**Max.**	**Mean**	**St. Dev.**
monthly views	Number of views of the corresponding movie on a monthly basis	0	2,223	23.70	68.32
proportional monthly views	Number of views as a proportion of total views of corresponding month	0	0.03	0.00031	0.00085
cumulative proportional monthly views	Cumulative number of proportional monthly views	0	1	0.13	0.19
large catalog	Dummy variable, which takes the value 1 for data from the large catalog (February 2016) and value 0 for the small catalog (February 2014)	0	1	0.71	0.45
rank	Ranks the movies by monthly views in descending order	1	265	195.46	79.57
series	Dummy variable, which takes value 1 if the title is a TV show episode, and value 0 if it is a movie	0	1	0.82	0.38
genre	Titles belong to 23 different main genres	N/A	N/A	N/A	N/A
age	Number of days since the title is available on the platform	0	5,630	121.84	120.96
catalog days	Number of days a title was available on the platform in the corresponding month	1	29	24.82	9.64
total length	Length of title in minutes	4.02	209.98	42.04	30.41
min. sort order	Proxy for the number of clicks needed to reach a title. We take the minimum on a monthly basis	6	122	22.33	15.99

3.2 Descriptive Statistics

Every observation in our dataset represents a movie title or TV show episode in a month. TV shows and movies show significant differences in duration, 33.2 minutes on average for episodes of TV shows and 95.12 minutes for movies.

Our dataset contains data for February 2014 and February 2016. Looking at the availability of titles on the platform during these periods, we see a substantial change (Figure 1, left). In February 2014, the platform only offered 1,873 titles, which jumped to 4,559 titles in February 2016. Along with the catalog size, the total *monthly views* show significant changes displayed on the right of Figure 1. In February 2014, the platform reported 20,612 total views, which rose to 131,795 views in February 2016.

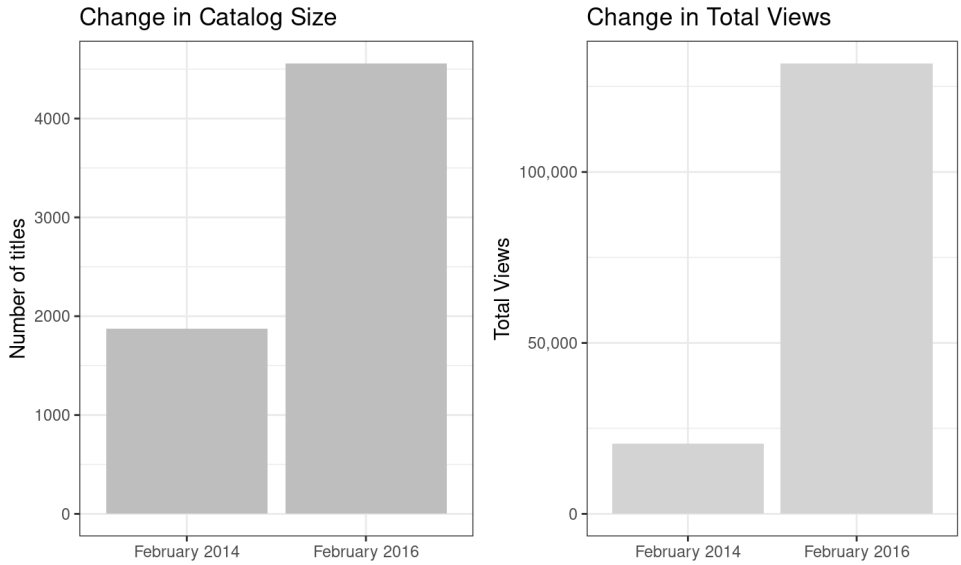

Figure 1: Changes in catalog size and total views over time

The catalog data provides information on every title's *genre*. There are 23 main genres. The majority of titles belong to the "Animation" genre, followed by the "Comedy", "Drama", "Crime" and "Action" genres. Figure 2 extracts the five genres, emphasizing the shift of listed titles between February 2014 and 2016 regarding genres. The plot on the left highlights the changes in the absolute number of provided titles. We see a notable increase in the availability of titles in the "Animation" genre from four titles in February 2014 to 2,290 titles in February 2016. Also, the "Action" genre increased from 183 titles to 344 titles. On the other hand, the "Crime" and "Comedy" genres experienced a decrease in the number of listed titles from February 2014 to 2016. The number of titles in the genre "Drama" remained almost unchanged. Given the overall change in catalog sizes, we additionally provide a comprehensive overview of the proportional number of titles on the right side of Figure 2. We can see that the proportion of titles from the "Action" genre actually decreased. Conversely, there is a significant increase in the proportion of "Animation", making up nearly 60% titles in February 2016. The

proportional overview highlights a significant decrease in the proportional number of titles in the "Comedy" and "Crime" genres. Furthermore, analyzing the proportional number of titles from the "Drama" genre reveals that what initially appeared unchanged in absolute numbers represents a substantial proportional decrease.

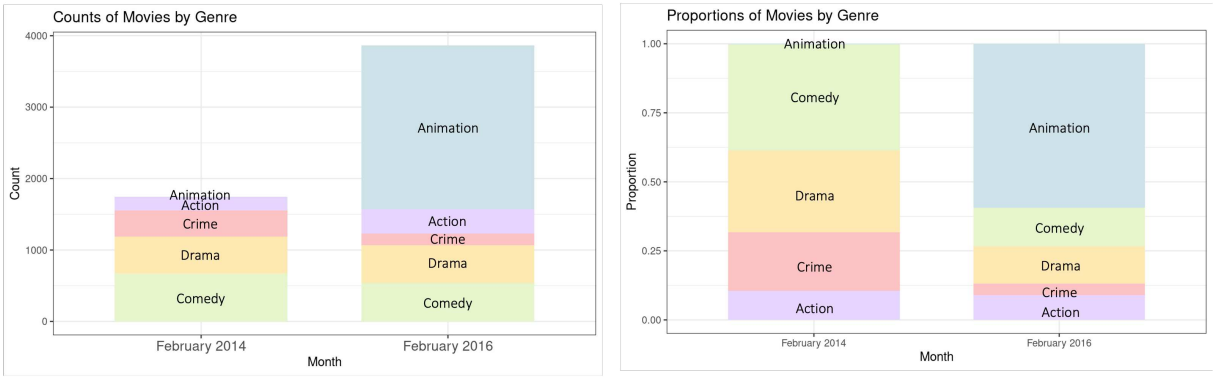

Figure 2: Count and proportion of available titles by *genre* per month

We have data on the *monthly views* for each movie as well as the number of clients who watched the titles. However, the values of both variables are very similar, which can be explained by the short period we are analyzing. Since we are examining only two individual months, it is unlikely that many individuals watched the same title more than once. Therefore, our analysis focuses on the *monthly views* variable.

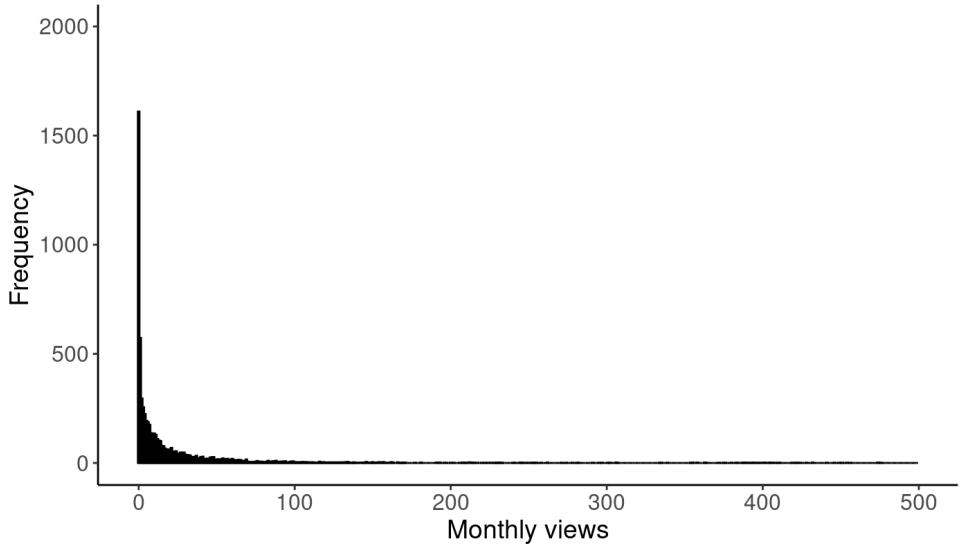

Figure 3: Distribution of variable *monthly views*

Figure 3 provides a zoomed-in view of the distribution of the *monthly views* variable, which is skewed to the right. This indicates that most of the observations are clustered around the lower end of the scale. However, there is a long tail on the right side of the distribution, suggesting some observations with very high values. The title with the highest value for *monthly*

views received 2,223 views. To account for the right skewness of the variable, it is common to log-transform it before conducting regression analysis (Brynjolfsson et al., 2011). Hence, we will log-transform the variable in our regressions. However, for taking the log of a variable, we have to exclude observations with zero values in our regression, as the logarithm function is undefined for zero. In section 5, we run a zero-inflated model with a negative binomial distribution to account for the skewness of the *monthly views* variable while still including the observations with zero views. In order to have a log-log relation, which fits the relationship between sales and rank according to the research of Brynjolfsson et al. (2003) well, the variable *rank* is also log-transformed in our regressions.

3.3 Empirical Strategy

To answer the research question regarding the impact of catalog size changes on the concentration of views on a Video-on-Demand platform, we compare the data from two specific months, showing a significant increase in catalog size. As we received data for the same calendar months of two years, this controls for the effects of holidays or special events. The two months differ significantly in their number of offered titles and will be referred to as "small catalog" and "large catalog" hereafter.

To examine changes in the concentration of views, we plot the cumulative number of titles and the cumulative number of proportional views, also known as the Lorenz curve (Brynjolfsson et al., 2011). Figure 4 shows the Lorenz Curve of February 2014 (small catalog) and February 2016 (large catalog). The curves show a rather concentrated distribution for both catalog sizes. The dotted diagonal line illustrates the opposite extreme, being a completely uniform distribution, where every title would account for the same number of views.

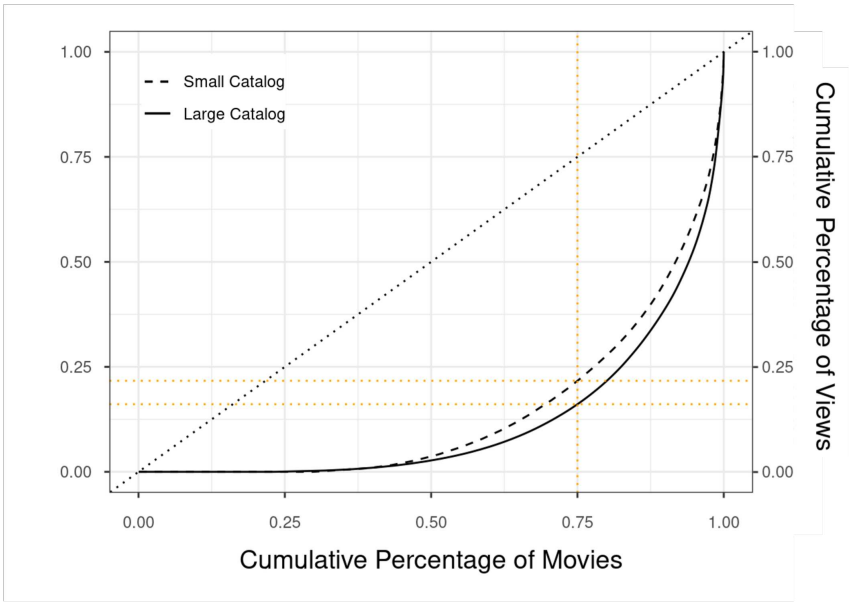

Figure 4: Lorenz curves for both catalog sizes

We observe that for the *small catalog*, 75% of movies were responsible for 22% of proportional views (represented by the dotted orange lines). This implies that, on the other hand, only 25% of movies were responsible for the remaining 78% of views. When we consider the curve for the *large catalog*, we notice that the concentration of views increased further, with 25% of movies accounting for 84% of views.

Our findings suggest a change in the concentration of views among different catalog sizes. In order to examine if there is a statistically significant difference, we run a linear regression with the log-linear of *monthly views* as a dependent variable and the log-linear of *views rank* as an independent variable, as already conducted by Brynjolfsson et al. (2011). The *views rank* variable ranks the movies in descending order according to their *monthly views*, assigning the lowest rank to the title with the highest *monthly views*. The described relationship and corresponding curve are commonly known as the "Pareto Curve" and match the relationship well according to previous research (Brynjolfsson et al., 2003; Pareto, 1964). In this specific case, the Pareto curve shows how the number of *monthly views* changes as the *views rank* of titles increases. We conduct this regression analysis for both discussed catalog sizes separately:

$$\log(\textit{monthly views}_j) = \beta_0 + \beta_1 \log(\textit{views rank}_j) + \varepsilon_j \tag{1}$$

If positive, the coefficient of *log(views rank)* would indicate that an increase of 1% in *views rank* would lead to an increase of β_1% in *monthly views*. If negative, the increase would lead to a decrease in β_1% *monthly views*.

To analyze if the influence of the *views rank* variable is different for different sizes of catalogs, we run a regression including the data of both months. We add a dummy variable representing the catalog sizes.

$$\log(\textit{monthly views}_j) = \beta_0 + \beta_1 \log(\textit{views rank}_j) + \beta_2\, \textit{large catalog}_j + \varepsilon_j \tag{2}$$

The variable *large catalog* represents the two catalog sizes. A positive prefix of the coefficient would indicate that, on average, *monthly views* are higher for the large catalog than for the small catalog. On the other hand, a negative prefix would suggest that *monthly views* are higher for the small catalog.

Afterward, we add an interaction term between the *large catalog* variable and the *log(views rank)* to our regression analysis. This allows us to test whether the relationship between the *views rank* of titles and the *monthly views* varies across different catalog sizes:

$$\log(\textit{monthly views}_j) = \beta_0 + \beta_1 \log(\textit{views rank}_j) + \beta_2\, \textit{large catalog}_j + \beta_3\, \textit{large catalog}_j \times \log(\textit{views rank}_j) + \varepsilon_j \tag{3}$$

The prefix of the interaction term demonstrates if the impact of the variable *views rank* on *monthly views* is higher or lower for the large catalog than for the small catalog.

4 Results

4.1 Main Regressions

Table 2 presents the results of our first regression models with the logarithm of *monthly views* as the dependent variable for all four regressions. The first two columns include only the data for the respective catalogs. The coefficient of the variable *log(views rank)* is negative and statistically significant for both regressions. This indicates that an increase in a title's *views rank* by 1%, on average, results in a decrease in *monthly views* by 2.2% for the small catalog and by 2.4% for the large catalog, c.p.. However, the coefficients of two different regressions cannot be directly compared. The moderate values for R^2 indicate that the independent variable in the model can explain a large proportion of the variation in the dependent variable.

Table 2: Main linear regressions

	Dependent variable:			
	log(monthly views)			
	(1)	(2)	(3)	(4)
	Small catalog	Large catalog	Whole sample	Whole sample
log(views rank)	-2.221^{***}	-2.403^{***}	-2.359^{***}	-2.221^{***}
	(0.066)	(0.051)	(0.041)	(0.083)
large catalog			2.983^{***}	3.824^{***}
			(0.056)	(0.444)
log(views rank)× large catalog				-0.182^{*}
				(0.095)
constant	11.708^{***}	15.532^{***}	12.307^{***}	11.708^{***}
	(0.287)	(0.274)	(0.182)	(0.362)
Observations	1,317	3,505	4,822	4,822
R^2	0.466	0.393	0.420	0.421
Adjusted R^2	0.466	0.393	0.420	0.420

Note: *p<0.1; **p<0.05; ***p<0.01

[*] We do not cluster on the movie level due to the limited time horizon of two months and the infrequent occurrence of titles appearing more than once. Additionally, genre-level clustering would result in a loss of statistical significance.

Column 3 contains the aggregated data for both catalogs and the dummy variable *large catalog*. The coefficient of the *log(views rank)* variable is still significantly different from zero. It has a negative prefix, implying that (on average) an increase in 1% of *views rank* is associated with a decrease of *monthly views* by 2.4%, c.p.. Furthermore, the positive coefficient of the dummy variable demonstrates that, on average, *monthly views* are significantly higher for the large catalog than for the small catalog. However, we have to be careful when interpreting this result, as with the larger catalog also more customers were subscribing to the platform.

To further assess if the *views rank* variable significantly differs for both months, we add an interaction term of *log(views rank)* and the dummy of the *large catalog* in the regression. The regression results are displayed in column 4 of Table 2. The negative prefix of the interaction term indicates that the ß of *log(views rank)* is more negative for the large catalog than for the small catalog. This means that the effect of *views rank* on *monthly views* is more negative for the large catalog. This indicates that there is less concentration and a longer tail when the catalog size is smaller and, on the contrary, a higher concentration and shorter tail when the catalog size is larger. The analysis suggests that catalog size significantly impacts the concentration of views, with larger catalogs leading to more concentrated views.

4.2 Types of Titles

Brynjolfsson et al. (2011) highlight that an increase in the concentration of views can also be explained by introducing a large number of products with minimal sales. Sales may appear more concentrated than before, even if they did not change for existing products. This is explained in Figure 5, inspired by a figure of Brynjolfsson et al. (2011). The authors present a scenario where 50% of 100 available products account for 75% of sales (left plot). However, when additional 100 niche titles with small sales are introduced, the distribution shifts so that 50% of products now account for 95% of sales. The shift happens without any sales changes of the existing products and is solely due to adding new niche products.

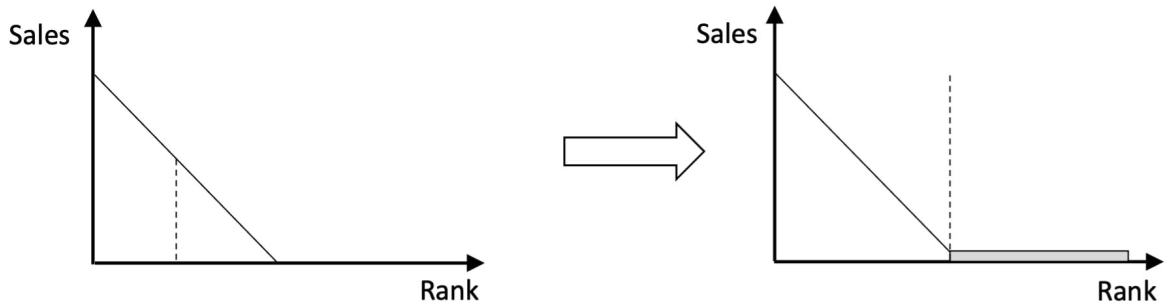

Figure 5: Impact of niche title introduction on concentration[1]

[1] Visualization adapted from Brynjolfsson et al. (2011)

To analyze whether our finding of increased concentration results from an introduction of mainly niche titles, we analyze the fraction of observations with zero views for both catalog sizes. For the smaller catalog, 29.69% of titles report zero views. However, for the larger catalog, the proportion of observations with zero views decreases to 23.12%. This indicates that the increase in concentration is not solely due to an introduction of low-performing titles.

4.3 Extended Analysis

To enhance the robustness of our analysis, we consider additional variables that might influence *monthly views* of our VoD platform.

Column 1 of Table 3 represents our benchmark regression, including only our main variables of interest and their interaction term. In column 2, we include the dummy variable *series*, indicating if a title is a TV show or movie. We expect this variable to influence *monthly views* due to potential differences in consumer preferences regarding TV shows and movies. Moreover, TV shows and movies are associated with different consumer behavior. Movies are typically standalone content watched in one sitting. TV shows, on the other hand, consist of multiple episodes, and consumers might engage with them differently (Matrix, 2014). Also, the catalog managers of TELCO might have had a release strategy linked to the differentiation between movies and TV shows. As expected, the regression shows that variable *series* significantly differs from zero. The coefficient has a negative prefix, indicating that TV show episodes, on average, have fewer *monthly views* than movies, c.p.. However, the interaction term of *log(views rank)* and *large catalog* remains significantly negative, implying that even when controlling for *series*, the impact of *views rank* on *monthly views* under consideration of catalog size holds. Compared with our baseline regression (Table 3, column 1), we observe a slight increase in the R^2 value, indicating that the addition of variable *series* improves the explanatory power of the model.

In addition, we add variable *age* to our regression (Table 3, column 3). The variable explains how many days have passed since the title was introduced to the platform. We expect the variable to impact *monthly views*, as titles might attract an audience right after being introduced on the platform, and others might need some time to be discovered. The importance of optimal timing was already discussed by Frank (1994). However, our platform had a relatively short history when our first exemplary month (February 2014) started. Therefore, most titles were available only for a very short period, leading to limited variation in the variable. This is also reflected in the regression, where the coefficient is not significantly different from zero. Therefore we cannot make any statements regarding the impact of *age* on our *monthly views* variable. However, we see that when controlling for *age*, the significantly negative coefficient of our interaction term *log(views rank)* $\times$ *large catalog* holds. This shows that even when controlling for the age of a title, the impact of catalog size on concentration holds.

Furthermore, we include the number of days a movie was listed in the catalog in the cor-

responding month (Table 3, column 4). We expect this variable to affect *monthly views*, as it directly influences the number of days customers can discover and watch the title. The regression shows that, as expected, the variable *catalog days* is significant and has a positive prefix. This suggests that, on average, an increase in days during which the title is available in the catalog is associated with an increase in *monthly views*, c.p..

We further analyze if the increase in concentration can be explained by an increase in difficulty in finding titles. Therefore we include the variable *min. sort order* in the regression (Table 3, column 4). The variable is a proxy for the minimum number of clicks needed to reach a title in the respective month. The results show that *min. sort order* has a significant negative impact on monthly views, meaning that an increase in *sort order* results in a decrease in *monthly views*. This is reasonable as an increase in sort order means that a title is ranked further down in the catalog and needs more clicks to be discovered. However, our interaction term of interest remains unchanged. This indicates that even when controlling for the number of days a title was available and the number of clicks needed to reach the title, the impact of *views rank* on *monthly views* under consideration of catalog size holds.

Another essential factor to consider is the impact of *genre* on *monthly views*. Due to consumer preferences and the availability of titles, it is expected that the genre has an impact on the number of views. Therefore, we want to control for the variable in our regression. However, as there are 23 different genres, we include dummies for the five most popular genres (Table 3, column 5). The results show that the significance of *log(views rank)* $\times$ *large catalog* disappears for the first time. The coefficient is no longer statistically significantly different from zero. This indicates that when controlling for the *genre* of a title, there is no significant impact of the size of a catalog on view concentration. Moreover, we see a substantial increase in R^2, implying that including the *genre* in our regression significantly improves its explanatory power.

Based on the results obtained, we conclude that the impact of catalog size on concentration holds when controlling for the title type (TV show vs. movie), the title's age, and the number of days the title was available in the respective month. However, catalog size does not significantly affect the concentration of views when additionally controlling for the title *genre*. Our findings suggest that its nature undergoes a transformation as the catalog size increases. The fact that controlling for genre effectively eliminates the effect indicates that it is not just the availability of more options that influences consumer behavior but rather the specific characteristics and nature of the new titles being added. This evidence indicates that the changing consumer behavior can be attributed to the nature of the additional titles rather than simply the presence of more choices. We can conclude that a title's *genre* plays a crucial role in explaining the concentration of views, overriding the impact of catalog size.

Table 3: Extended linear regressions

	Dependent variable:				
	log(monthly views)				
	(1)	(2)	(3)	(4)	(5)
log(views rank)	-2.221^{***}	-1.965^{***}	-1.967^{***}	-1.722^{***}	-1.943^{***}
	(0.083)	(0.082)	(0.083)	(0.079)	(0.073)
log(views rank) ×	-0.182^{*}	-0.309^{***}	-0.308^{***}	-0.341^{***}	0.084
large catalog	(0.095)	(0.093)	(0.093)	(0.090)	(0.083)
large catalog	3.824^{***}	4.195^{***}	4.189^{***}	4.099^{***}	2.562^{***}
	(0.444)	(0.434)	(0.434)	(0.414)	(0.382)
series		-0.602^{***}	-0.601^{***}	-0.714^{***}	-0.370^{***}
		(0.039)	(0.039)	(0.038)	(0.038)
age			0.00005	-0.001^{***}	-0.0002
			(0.0002)	(0.0002)	(0.0001)
catalog days				0.055^{***}	0.047^{***}
				(0.004)	(0.003)
min. sort order				-0.014^{***}	-0.010^{***}
				(0.001)	(0.001)
genre comedy					1.121^{***}
					(0.045)
genre drama					1.205^{***}
					(0.047)
genre crome					1.168^{***}
					(0.056)
genre action					1.007^{***}
					(0.063)
genre adventure					1.085^{***}
					(0.080)
genre other					0.819^{***}
					(0.052)
constant	11.708^{***}	11.093^{***}	11.096^{***}	9.038^{***}	8.678^{***}
	(0.362)	(0.355)	(0.356)	(0.367)	(0.337)
Observations	4,822	4,822	4,822	4,622	4,622
R^2	0.421	0.448	0.448	0.504	0.589
Adjusted R^2	0.420	0.448	0.448	0.504	0.588
Residual Std. Error	1.108	1.081	1.081	1.020	0.928

Note: *p<0.1; **p<0.05; ***p<0.01

* We do not cluster on movie level due to the limited time horizon of two months and the infrequent occurrence of titles appearing more than once. Additionally, genre-level clustering would result in a loss of statistical significance.

4.4 Changes in customers and preferences

In our primary analyzes (Section 4.1 and 4.3), we included views data of customers that either watched in February 2014 only, in February 2016 only, or in both months. Therefore, the shift in concentration could be attributed to the change in consumers and corresponding preferences from February 2014 to 2016. To account for this possibility, we are conducting a robustness check with the same set of customers. We extract the views data of customers present in both months and replicate the analysis. The resulting Lorenz curve is displayed in Figure 6 and indicates a rise in concentration from 2014 to 2016 when the catalog size grew. This trend is similar to when observing the whole dataset (see Figure 4) and gives a first indication that controlling for shifts in customers and preferences yields the same result.

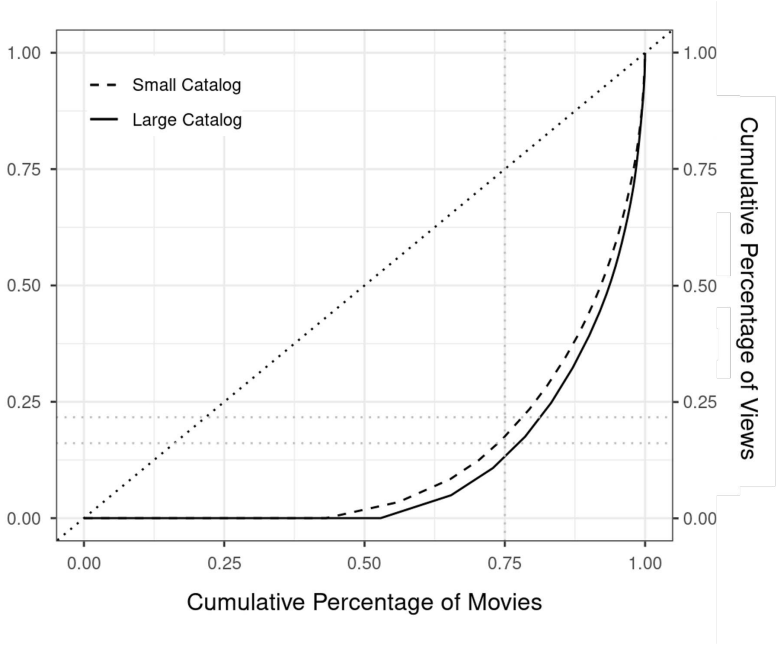

Figure 6: Lorenz curve for both catalog sizes, containing data on a subset of customers

In order to analyze if the statistical significance holds while analyzing the same customers, we run a linear regression. We replicate the earlier regression analysis and present our results in Table 4. The Table compares the baseline regression of the whole data set (column 1) and the same regression run on the subset of the same customers (column 2). The results show that the independent variable's prefixes remain consistent between the whole dataset's and the subset's regression. Additionally, all variables remain statistically significantly different from zero. In particular, the p-value of the interaction term *log(views rank)* $\times$ *large catalog* increased from the 10% significance level (whole dataset) to the 1% significance level (subset of same customers). The smaller p-value indicates more substantial evidence against the null hypothesis, being that there is no relationship between the variables. This finding supports the presence of a significant relationship between catalog size and concentration. It suggests that the effect of an increase in concentration as catalog size increases remains consistent when observing the same customers

over time.

Furthermore, we replicate the extended analysis from Table 3, including additional variables and in particular variable *genre*, for our subset. Column 3 shows the extended analysis based on the whole dataset, while column 4 only considers the subset of the same customers. Section 4.3 revealed that controlling for the *genre* of a title cancels the significance of the effect of catalog size on concentration (Table 4, column 3). However, when analyzing only the same set of customers, we see that controlling for the *genre* decreases the statistical significance but does not cancel it entirely (Table 4, column 4). This implies that focussing on the same customers reduces the potential effect of *genre* on *monthly views*.

Furthermore, we include variable *min. sort order*, indicating how many clicks are needed to reach a title in columns 5 and 6. We see that the interaction term of *log(views rank)* $\times$ *large catalog* now loses its statistical significance even for the subset of the same customers. This indicates that controlling for preferences and changes in customers can reduce the effect of *genre* on *monthly views*, but not the effect of *min. sort order*.

This shows that the relationship between catalog size and concentration, under the influence of *genre*, is more robust when analyzing the same customers. This is reasonable as we examine a more homogeneous group regarding preferences and behaviors. By focusing on the same customers, we reduce the potential influence of individual variations and external factors. This allows us to isolate the impact of catalog size on concentration more accurately. However, when also controlling for the number of clicks needed to reach a title, analyzing the same set of customers does not prevent the effect of catalog size on concentration from disappearing. This is an indication that the power of a title's position in the catalog has a higher influence on explaining the view concentration than its size.

We can conclude that by accounting for customers' preferences and changes, we can diminish the impact of *genre* on *monthly views*. However, even when analyzing only the subset of customers, the effect of *min. sortorder* remains significant in explaining the concentration of views, overriding the impact of catalog size.

Table 4: Regressions controlling for the same customers

	Dependent variable:					
	log(monthly views)					
	Whole Dataset	Same customers	Whole Dataset	Same customers	Whole Dataset	Same customers
	(1)	(2)	(3)	(4)	(5)	(6)
log(views rank)	-2.221^{***}	-2.421^{***}	-1.924^{***}	-2.334^{***}	-1.943^{***}	-2.316^{***}
	(0.083)	(0.070)	(0.074)	(0.069)	(0.073)	(0.069)
log(views rank) ×	-0.182^{*}	-0.261^{***}	-0.025	-0.149^{*}	0.084	-0.115
large catalog	(0.095)	(0.091)	(0.083)	(0.088)	(0.083)	(0.089)
large catalog	3.824^{***}	0.943^{***}	3.048^{***}	0.608^{*}	2.562^{***}	0.504
	(0.444)	(0.342)	(0.385)	(0.328)	(0.382)	(0.329)
series			-0.368^{***}	0.010	-0.370^{***}	-0.001
			(0.037)	(0.028)	(0.038)	(0.029)
age			-0.0002	-0.0003^{**}	-0.0002	-0.0002^{**}
			(0.0001)	(0.0001)	(0.0001)	(0.0001)
catalog days			0.055^{***}	0.023^{***}	0.047^{***}	0.024^{***}
			(0.002)	(0.002)	(0.003)	(0.003)
genre comedy			1.104^{***}	0.545^{***}	1.121^{***}	0.576^{***}
			(0.045)	(0.039)	(0.045)	(0.039)
genre drama			1.200^{***}	0.545^{***}	1.205^{***}	0.530^{***}
			(0.045)	(0.039)	(0.047)	(0.040)
genre crime			1.219^{***}	0.393^{***}	1.168^{***}	0.361^{***}
			(0.056)	(0.046)	(0.056)	(0.047)
genre action			1.020^{***}	0.535^{***}	1.007^{***}	0.554^{***}
			(0.059)	(0.051)	(0.063)	(0.052)
genre adventure			1.114^{***}	0.703^{***}	1.085^{***}	0.712^{***}
			(0.081)	(0.062)	(0.080)	(0.062)
genre other			0.839^{***}	0.405^{***}	0.819^{***}	0.414^{***}
			(0.052)	(0.045)	(0.052)	(0.045)
min. sort order					-0.010^{***}	-0.004^{***}
					(0.001)	(0.001)
constant	11.708^{***}	10.294^{***}	8.219^{***}	8.867^{***}	8.678^{***}	8.856^{***}
	(0.362)	(0.262)	(0.329)	(0.267)	(0.337)	(0.277)
Observations	4,822	3,213	4,822	3,213	4,622	3,117
R^2	0.421	0.510	0.580	0.563	0.589	0.566
Adjusted R^2	0.420	0.509	0.579	0.561	0.588	0.564
Res. Std. Error	1.108	0.684	0.944	0.647	0.928	0.647

Note: *p<0.1; **p<0.05; ***p<0.01

[*] We do not cluster on the movie level due to the limited time horizon of two months and the infrequent occurrence of titles appearing more than once. Additionally, clustering on the genre level would result in a loss of statistical significance.

5 Change in distributional assumptions

In this section, we aim to examine critically and evaluate the robustness of our findings. We will reconsider some crucial assumptions as we look at a more complex composition than the underlying dataset of Brynjolfsson et al.'s paper, which we are replicating.

Like the original paper, we initially assume that the Pareto curve captures the relationship between rank and sales well. However, when we fit the regression model to our dataset, we observe that the linear regression with log-log transformation does not provide a good fit (see Figure 7). The red and blue lines in the graph represent the empirical values obtained from our dataset. It is evident from the curves that the relationship between *log(monthly views)* and *log(views rank)* is non-linear. The two black lines represent the values estimated by fitting our linear function to the data. However, the fit between the predicted (black lines) and actual (red and blue lines) data is not ideal. The discrepancy arises because the assumed linear relationship in our regression models does not accurately capture the underlying non-linear nature of the data. As a result, the linear prediction lines fail to align closely with the observed data, leading to a poor fit. Consequently, we seek to identify a more suitable model that could capture the relationship between *monthly views* and *rank* more accurately.

In addition, as the paper we are replicating is applying a log-log relationship, they exclude zero sales from their analysis. However, since we have a significant proportion of observations with zero views, we intend to include them in our robustness analysis.

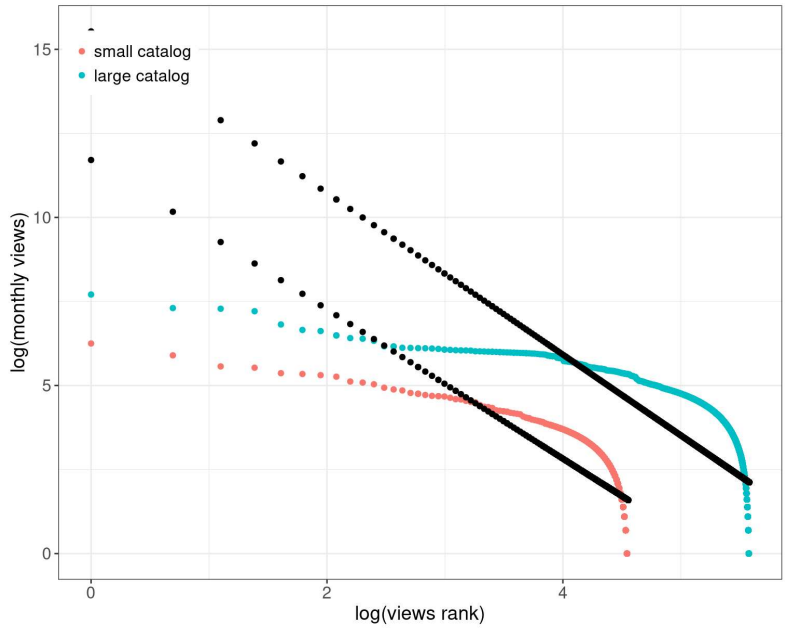

Figure 7: Linear regression model fit

We consider various count models, including count models with zero inflation, to understand their fit to our data better. Specifically, we examine zero-inflated Poisson regression and

zero-inflated negative binomial regression models. To compare the performances of these models, we incorporate the models as a rootogram in Figure 8. The rootogram serves as a visual representation that allows us to assess the goodness of fit between the observed data and the predicted values from the regression model. The rootogram displays the discrepancies between the observed and predicted values by plotting the square root of the absolute difference or residuals against specific intervals. By examining the rootogram, we can evaluate the overall quality of the model's fit and assess any areas of over- or under-prediction. The left rootogramm of Figure 8 displays the fit of the zero-inflated Poisson regression model. The observed fit exhibits substantial irregularities, indicating a poor match to the data. Specifically, we observe significant underprediction for observations with low *monthly views*. This suggests that the model does not adequately capture the underlying factors influencing the occurrence of instances with low *monthly views*. The model exhibits high overprediction for increased *monthly views*, indicating a mismatch between the predicted values and the actual data points in this interval. Furthermore, the model again underpredicts instances for high *monthly views*. These observations collectively show the inadequacy of the zero-inflated Poisson regression model, as it fails to provide a reliable and consistent fit across the entire range of *monthly views*.

The right rootogramm in Figure 8 visualizes the fit obtained from the zero-inflated negative binomial regression model. Examining this fit reveals certain discrepancies between the predicted values and the actual data. Specifically, there is a tendency for over-prediction in the range of low *monthly views*. This indicates that the model tends to estimate higher values for instances falling within this range than what is observed in the data. Conversely, the model exhibits under-prediction in the area of very low and high *monthly views*. However, considering the overall fit, we can conclude that the zero-inflated negative binomial regression model fits the data significantly better than the Poisson model.

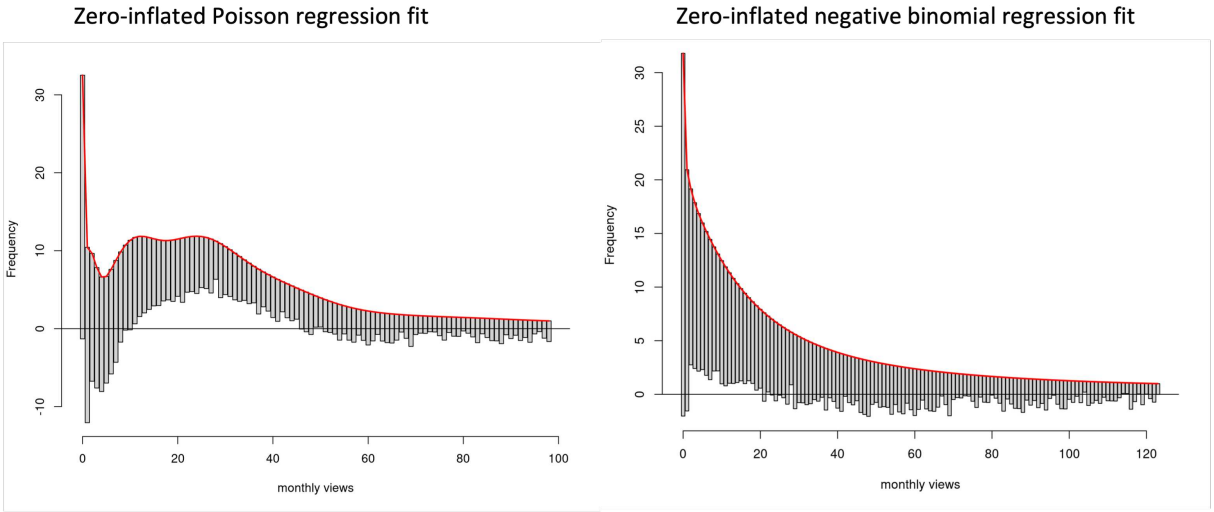

Figure 8: Comparison of model fits

This is reasonable, as the Poisson regression assumes that the mean (= 23.70) and variance (= 4667.24) of the dependent are equal, which is not the case for our dataset. On the other

hand, the zero-inflated negative binomial model does not make this assumption, allowing for more flexibility in capturing the inherent variability present in our data. By accommodating overdispersion, where the variance exceeds the mean, the zero-inflated negative binomial model provides a more suitable framework for accurately modeling count data.

Furthermore, the zero inflation of the model aligns with our data, where a substantial proportion of observations consists of zero values. The zero-inflated negative binomial model addresses the zero values by incorporating a separate process that accounts for these values, improving the model's fit and accuracy.

We replicate the previous regressions using the new zero-inflated negative binomial model, including zero values for *monthly views*, in Table 5.

The results show that by utilizing the negative binomial zero-inflated model and incorporating zero values for *monthly views* in our analysis, the significance of the effect of catalog size on concentration diminishes. Furthermore, when controlling for factors like whether the title is a movie or TV show, the age, number of days the title was listed, minimum sort order, and genres, the impact of catalog size on concentration is not significantly different from zero. However, it is noteworthy that all other factors exhibit statistical significance when explaining the number of *monthly views*.

The results show the importance of viewer preferences and highlight the need to consider various factors beyond catalog size when investigating the determinants of view concentration.

Table 5: Zero-inflated negative binomial regressions

	Dependent variable:				
	monthly views				
	(1)	(2)	(3)	(4)	(5)
log(views rank)	-4.353^{***}	-4.134^{***}	-4.101^{***}	-2.967^{***}	-3.135^{***}
	(0.456)	(0.455)	(0.458)	(0.336)	(0.366)
log(views rank) ×	0.572	0.446	0.420	−0.531	0.244
large catalog	(0.479)	(0.476)	(0.478)	(0.397)	(0.387)
large catalog	2.379	2.792	2.913	6.658^{***}	3.015^{*}
	(2.222)	(2.207)	(2.221)	(1.898)	(1.771)
series		-0.362^{***}	-0.371^{***}	-0.512^{***}	-0.352^{***}
		(0.022)	(0.022)	(0.025)	(0.030)
age			-0.0003^{***}	-0.001^{***}	-0.0002^{***}
			(0.00003)	(0.0001)	(0.0001)
catalog days				0.074^{***}	0.070^{***}
				(0.004)	(0.004)
min. sort order				-0.008^{***}	-0.009^{***}
				(0.001)	(0.001)
genre comedy					0.852^{***}
					(0.042)
genre drama					0.987^{***}
					(0.044)
genre crime					1.028^{***}
					(0.051)
genre action					0.892^{***}
					(0.055)
genre adventure					0.820^{***}
					(0.073)
genre other					0.803^{***}
					(0.049)
constant	21.045^{***}	20.389^{***}	20.278^{***}	13.722^{***}	13.477^{***}
	(1.997)	(1.993)	(2.005)	(1.518)	(1.642)
Observations	5,766	5,766	5,766	5,766	5,766
Log Likelihood	−20,210.170	−20,155.100	−20,152.200	−19,854.340	−19,501.840

Note: *p<0.1; **p<0.05; ***p<0.01

[*] We do not cluster on the movie level due to the limited time horizon of two months and the infrequent occurrence of titles appearing more than once. Additionally, genre-level clustering would result in a loss of statistical significance.

6 Discussion

The findings of our analysis contribute to the ongoing discourse surrounding catalog size and its impact on concentration. While the relationship between catalog size and concentration has been a subject of debate, our study indicates that catalog size alone does not provide robust

evidence of its influence on concentration. Instead, we observe that various factors, including genre, content type (movie or TV show), and position in a catalog, play significant roles in shaping consumer behavior.

Our findings present a contrasting perspective to Anderson's "Long Tail" theory, which suggests that as variety increases due to the adoption of information technology, consumption tends to disperse and shift toward niche products. Amongst other things, Brynjolfsson et al. (2011) explain the "Long Tail" phenomenon with increased variety enabling to serve also niche customers. However, our analysis challenges this finding, as we do not find evidence that increased catalog size automatically leads to increased consumption of niche products.

A possible explanation for the disparities in findings could be that most authors who support the "Long Tail" theory examine the context of catalog size changes occurring during the transition from brick-and-mortar stores to the online environment (Brynjolfsson et al., 2003). This transition introduces significant differences in the dynamics of consumption and product availability (Brynjolfsson et al., 2003). In contrast, our analysis focuses on a specific digital platform operating within the same channel and with a similar user base over time.

Furthermore, our findings diverge from those of Tan et al. (2017), who examine the effect of product variety on demand concentration in one channel. In their study, they observe an increase in concentration as catalog size increases, which aligns with the "Pareto principle" that suggests that a significant proportion of outcomes can be attributed to a small proportion of cases (Pareto, 1964). A potential reason for the disparities in results could be the differences in underlying datasets. Whereas we focus on VoD data from a specific platform from 2014 and 2016, Tan et al. (2017) analyze DVD rentals from January 2001 to July 2005. The variations in findings could be attributed to several factors, including changes in consumer behavior, technology advancements, and market dynamics over time. The shift from physical DVD rentals to digital streaming platforms introduces distinct consumption patterns and preferences among users (Zentner et al., 2013). Furthermore, the platforms' characteristics, such as unique features, user interfaces, and recommendation algorithms, can significantly influence consumer behavior, potentially leading to different outcomes compared to traditional DVD rentals.

Overall, our findings challenge the prevailing theories that posit a direct relationship between changes in catalog size and shifts in the concentration of views. Instead, our research highlights the complex and context-dependent nature of concentration dynamics. The lack of a direct relationship between catalog size and changes in the concentration of views emphasizes the need for a deeper understanding of consumer behavior in the digital era.

Moreover, our analysis reveals that other factors, like the genre of a title, play a crucial role in explaining viewer behavior, surpassing the impact of catalog size alone. We find that different genres result in diverse consumption patterns and preferences among users. Additionally, we observe that the distinction between movies and TV shows plays a significant role in explaining views. The format of the content, whether it is a movie or a TV series, influences how viewers engage with and consume the content. Furthermore, we find that the number of clicks required

to discover a title has a notable effect on the number of views it receives. Titles that need fewer clicks to be discovered tend to attract more views compared to those that require more effort to find. Another factor influencing the number of views is the time the title has been listed on the platform, as well as the days it is available in the corresponding month. Titles that have been available for a longer period of time tend to accumulate higher viewership. These findings highlight the complex nature of viewer behavior and emphasize the importance for digital platforms to consider various factors beyond catalog size. By implementing genre-specific strategies, strategically positioning content on a platform, and recognizing the influence of variables such as content type (movie or TV show) and duration of availability, platforms can better understand and anticipate viewer preferences. By leveraging these insights, platforms can make informed choices regarding content selection, recommendation algorithms, and user engagement strategies, ultimately enhancing the overall viewing experience for users.

With a substantial fraction of listed titles in our dataset generating very few views, the question arises whether retaining all these titles in the catalog is reasonable. Beyond the additional costs of holding them in the catalog, there is also the risk of overwhelming customers with too many choices, a phenomenon commonly referred to as "choice overload" (Dhar, 1997; Toffler, 1970). Broniarczyk et al. (1998) suggest that consumers do not even notice a decrease in a variety up to a certain threshold. Furthermore, our analysis reveals that a change in catalog size does not necessarily have a significant impact on the concentration of views, nor does it result in the dispersal of consumption towards the tail. In light of these findings, which indicate that a change in catalog size does not have a robust impact on viewer behavior and niche titles often generate costs without attracting significant views, it raises the question of whether these titles should be retained in the catalog.

On the other hand, Tversky and Shafir (1992) argue that providing options with a clear distinction between good and bad options eases the choice for customers. Furthermore, Kahn (1998) directly links introducing additional products to an increase in market share. These findings indicate that retaining unpopular titles in the catalog may have its benefits.

Given these inconsistent findings, conducting an extensive analysis of the profitability and performance of long-tail titles becomes crucial. Furthermore, exploring the extent to which the size of the assortment influences customers' decision-making when selecting a VoD provider would provide valuable insights into the role of catalog size in shaping consumer behavior.

7 Limitations

The dataset of our analysis only consists of two months. This sample may not represent the overall trend of catalog sizes for the remaining months of the VoD platform. Therefore it is essential to consider this when concluding long-term patterns based on this restricted time frame. A more extensive dataset would provide a more robust understanding of catalog size changes over a longer time period. In general, we do not have any information on the strategies fol-

lowed by TELCO's catalog managers. We do not know if there was a rationale behind adding or removing specific titles in the respective months. Without insights into the decision-making process of catalog management, it becomes challenging to determine the underlying reasons for the catalog sizes. Different goals, such as targeting specific customers, optimizing revenue, and clearing out inventory, could result in varying catalogs. Thus we cannot determine if the change in catalog size was intentional or accidental. We also have to consider that external factors like market conditions and industry trends may have influenced the catalog size of our company. In addition, we do not have any information regarding promotional activities or experiments conducted during the period, which could have influenced the consumption of titles. Furthermore, we have no information on potential release strategies regarding episodes of TV shows, which could have impacted the views of specific episodes.

Moreover, it is essential to note that our analysis is based on data of a specific geographical scope. Local factors may have influenced our findings and may differ in other geographical scopes, preventing making them generalizable. Thus it is essential to acknowledge that the conclusions drawn from our analysis may not apply to other time periods or contexts.

From a managerial standpoint, the focus should extend beyond the effect of adding more movies to the catalog on *monthly views*. Instead, the overall impact on revenues resulting from a larger catalog becomes essential. We could run a randomized experiment where the treatment group would receive an expanded catalog, and the control group would keep the smaller catalog. Observing the two groups over time, we could determine whether offering a larger catalog leads to more concentrated consumption and its effects on revenues. Since the necessary data is unavailable, our analysis is limited to examining and comparing the two months only. As we discovered that catalog size does not significantly impact the concentration of views in our dataset, and considering the limited availability of additional factors, it would be intriguing to explore other variables and factors. Given the indication of the importance of sort order, it would be interesting to explore the effects of sort order and the platform's recommenders on viewer behavior in more depth.

8 Conclusion

This dissertation analyzes the effect of change in catalog size on the concentration of views in the media industry. It adds managerial value to the industry of Video-on-Demand, where catalog managers are faced with strategic decisions regarding adding and removing specific titles to the platform. While most previous studies in the media industry observe a trend toward the "Long Tail" theory, our research does not support this notion. We analyze the catalog data of a VoD provider for two different months, characterized by big differences in catalog size.

Our first finding, when relying on the assumptions of our reference paper (Brynjolfsson et al., 2011), is an increase in concentration as catalog size increases. This suggests that when adding more titles to a VoD catalog, consumption goes towards hits rather than niche titles. We

provide evidence that introducing mainly niche titles does not cause this trend. In addition, our results hold even when controlling for additional factors like if the title is a TV show or movie, how long the movie has been listed in the catalog, and how many days it was available in the respective month. However, when controlling for the genre of a title, the effect of catalog size on concentration loses its significance. This implies that the genre of a title overrides the impact of catalog size when explaining the concentration of views. Moreover, we find that when analyzing the same set of customers across both months, our results hold even when controlling for the genre of a title. Analyzing a more homogeneous group with consistent preferences and behaviors reduces the potential influence of individual variations, allowing us to isolate the impact of catalog size on concentration accurately. However, we find that when controlling for the number of clicks needed to reach a title, the significance of the effect of catalog size disappears. This indicates that the power of a title's position in the catalog has a greater influence on explaining the view concentration than its size.

Furthermore, we critically examine the underlying distributional model derived from our reference paper by Brynjolfsson et al. (2011) and find that the suggested linear regression model does not fit our data well. We explore alternative distributional models and find that the zero-inflated negative binomial regression, which addresses explicitly zero values through a separate process, more effectively captures the characteristics of our data. Applying this model, we find that catalog size does not statistically significantly impact the concentration of views. This finding contradicts the previous results obtained by applying the log-log relationship and linear regression. In conclusion, our study differs from the paper we replicated as we operate within a more complex context with different products and a higher proportion of zero values for the dependent variable. Our findings indicate that the replicated model is not applicable in this complex environment. By applying a more suitable model, we find that catalog size does not significantly impact the concentration of views in our context. However, we identify other significant factors, such as title position, genre, and whether it is a movie or TV show that shape viewership behavior. These results emphasize the importance of considering the unique characteristics of the dataset and the specific context for a more accurate understanding of viewer behavior in the media industry.

References

Anderson, C. (2004). The long tail. *Wired Magazine*, 12:170–177.

Anderson, C. (2006). The long tail. *New York: Hyperion*.

Baumol, W. J. and Ide, E. A. (1956). Variety in retailing. *Management Science*, 3(1):93–101.

Broniarczyk, S. M., Hoyer, W. D., and McAlister, L. (1998). Consumers' perceptions of the assortment offered in a grocery category: The impact of item reduction. *Journal of marketing research*, 35(2):166–176.

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

Brynjolfsson, E., Hu, Y., and Smith, M. D. (2003). Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. *Management science*, 49(11):1580–1596.

Brynjolfsson, E., Hu, Y. J., and Smith, M. D. (2006). From niches to riches: Anatomy of the long tail. *Sloan management review*, 47(4):67–71.

Dhar, R. (1997). Consumer preference for a no-choice option. *Journal of consumer research*, 24(2):215–231.

Diehl, K. and Poynor, C. (2010). Great expectations?! Assortment size, expectations, and satisfaction. *Journal of marketing research*, 47(2):312–322.

Fleder, D. M. and Hosanagar, K. (2007). Recommender systems and their impact on sales diversity. In *Proceedings of the 8th ACM conference on Electronic commerce*, pages 192–199.

Frank, B. (1994). Optimal timing of movie releases in ancillary markets: The case of video releases. *Journal of Cultural Economics*, 18:125–133.

Gourville, J. T. and Soman, D. (2005). Overchoice and assortment type: When and why variety backfires. *Marketing science*, 24(3):382–395.

Kahn, B. E. (1998). Dynamic relationships with customers: High-variety strategies. *Journal of the Academy of Marketing Science*, 26(1):45–53.

Kahn, B. E. and Lehmann, D. R. (1991). Modeling choice among assortments. *Journal of Retailing*, 67(3):274–299.

Marks, L. (2014). *Sensory processes: The new psychophysics*. Elsevier.

Matrix, S. (2014). The netflix effect: Teens, binge watching, and on-demand digital media trends. *Jeunesse: young people, texts, cultures*, 6(1):119–138.

Pareto, V. (1964). *Cours d'économie politique*, volume 1. Librairie Droz.

Reibstein, D. J., Youngblood, S. A., and Fromkin, H. L. (1975). Number of choices and perceived decision freedom as a determinant of satisfaction and consumer behavior. *Journal of Applied Psychology*, 60(4):434.

Schwartz, B. (2004). *The Paradox of Choice: Why More Is Less*. Ecco, New York.

Tan, T. F., Netessine, S., and Hitt, L. (2017). Is Tom Cruise threatened? An empirical study of the impact of product variety on demand concentration. *Information Systems Research*, 28(3):643–660.

Toffler, A. (1970). Future shock, 1970. *Sydney. Pan*.

Tversky, A. and Shafir, E. (1992). Choice under conflict: The dynamics of deferred decision. *Psychological science*, 3(6):358–361.

Zentner, A., Smith, M., and Kaya, C. (2013). How video rental patterns change as consumers move online. *Management Science*, 59(11):2622–2634.

A Appendix

A.1 Identification of Trailers

To identify trailers and previews in the dataset, we assume they should have a significantly shorter duration than the other titles. However, we noticed significant differences in the duration of titles across different genres. Figure 9 shows the distribution of lengths (in minutes) for six exemplary genres. For instance, titles of the *genre* "Family" tend to have shorter durations than other genres, which could result in them being wrongly identified as outliers (trailers) in the overall dataset. Figure 9 shows how the distribution and potential outliers in terms of durations differ between the six genres. Therefore, we analyze the duration of titles for each *genre* separately to ensure that outliers are only removed where appropriate. In addition, some observations in our dataset contained keywords such as "preview" in their title, which allowed us to identify them. After excluding observations that qualified as trailers or previews due to their short duration or title, we analyze observations that stand out due to their long durations. We find observations with durations of several hours, representing whole marathons of series. Moreover, some marathons could be identified because they included the flag "marathon" in their title.

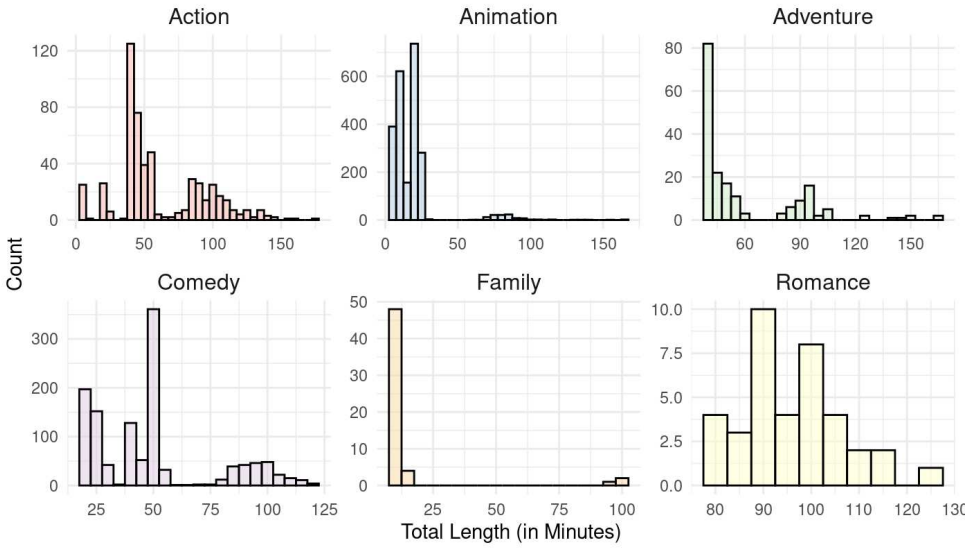

Figure 9: Distribution of TV show and movie lengths by genre