Exploration of Popularity Effects in Music

Jose Eduardo De Moraes Lopez, Camila Navarro Llaven
Web Science, University of Twente
Enschede, Netherlands
j.e.demoraeslopez@student.utwente.nl
c.navarrollaven@student.utwente.nl

I. Introduction

A. Dataset selection

Several options of datasets and retrieval methods are available for analysing network effects in music. The focus on this work will be on internet streaming platforms, specifically Youtube and Spotify. Information on streaming numbers can be either accessed through these platforms' open APIs, web scraping, or third party sources.

When deciding the method to select two main features should be taken into account. As in the attributes that should be retrieved for the analysis, view counts (streams for Spotify) are probably the number one indicator for popularity. Other metrics may include the like and dislike counts, number of shares and number of comments. The sample to be used should also be taken into account. As these are very large platforms, a small sample size isn't likely to represent their whole music collection, so it should be as large as possible. One should also take into account the limitations of having a representative sample on a geographical and music genre basis.

Let us now discuss the advantages and disadvantages respecting the available methods. API-based search offers direct access to YouTube's comprehensive metrics applying custom filters (region, keywords, etc.). It should be noted that in the case of Spotify, though its API provides access to many important features, it doesn't report the number of streams. Another option is web scraping. Though more labor-intensive, it captures data that may not be accessible through APIs but may face limitations due to website policies or technical constraints. Surveys and user studies also provide first-hand information, as they offer valuable insights into viewer behavior and preferences. However, collecting this information on a large scale, especially regarding Popularity Effects and Power Laws, can be challenging depending on the researchers' reach. Finally, publicly available datasets may make data retrieval an overaly easier process.

In this project, a combination of data retrieval methods was employed, with a primary focus on utilizing the YouTube API to gather essential information. Specifically, an API request was made to obtain view count data for the most relevant songs on YouTube ([1]).

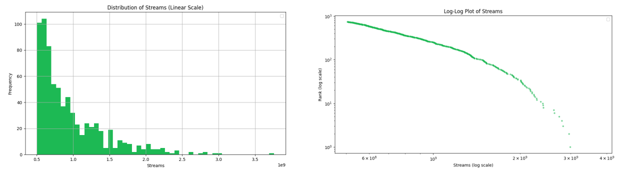
The study involved the curation of three key datasets from YouTube. The initial and main dataset was created through a YouTube-API search query, encompassing crucial details such as video IDs, titles, view counts, as well as like and dislike counts. This dataset was centered around the currently most-viewed 777 VEVO videos within a dedicated playlist. Notably, the primary parameter of interest in this analysis was the view count. It's worth noting that changes in YouTube's policies have led to the unavailability of dislike counts, rendering the difference between likes and dislikes irrelevant for this particular study.

The second dataset was also established using YouTube-API search functionality and had the specific objective of assessing the prominence of the reggaeton music genre. Similar to the initial dataset, it included video IDs, titles, view counts, and the counts of likes and dislikes. This dataset served a dual purpose: as a valuable point of reference and as a foundational experience that contributed to the refinement of the dataset collection methodology. Additionally, it shed light on certain inherent limitations and mistakes.

Lastly, the project incorporated a pre-existing dataset obtained in a prior stage. This dataset contained information on the top 100 YouTube videos and featured daily updates with a range of parameters. However, for the scope of this study, our exclusive focus was directed towards two critical variables: the date and the view count. These meticulously curated datasets collectively form the foundation for a comprehensive examination of YouTube video trends and dynamics.

In the case of Spotify, as the streaming counts of songs are not available in their public-free API, a third party source was used. providing you with the list of most streamed tracks on Spotify. ChartMasters ([2]) is a public webpage that lists the most popular Spotify songs and aims to list all tracks with over 500 million streams. It is worth noting that it lists the number of streams per song as of today. This data was retrieved using webscraping to return the songs available for years 2010 to 2017.

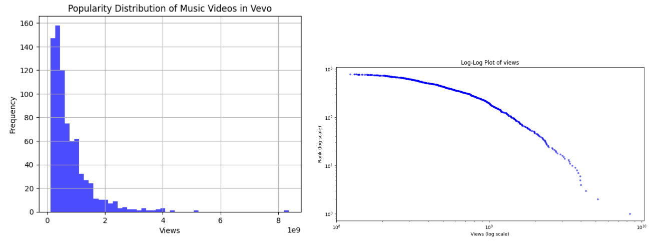
II. Popularity Distribution



(a) Linear distribution

(b) Log-log distribution

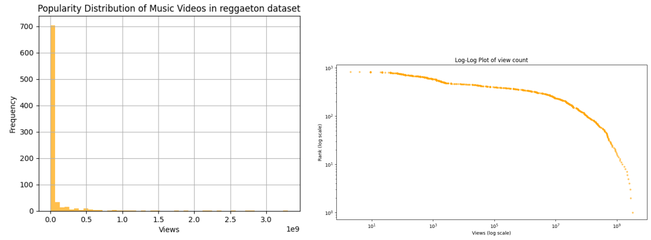
Fig. 1: Popularity distribution of Spotify streams.



(a) Linear distribution

(b) Log-log distribution

Fig. 2: Popularity distribution of Youtube VEVO.



(a) Linear distribution

(b) Log-log distribution

Fig. 3: Popularity distribution of Youtube Reggaeton.

To test if a distribution follows a power law, both visual and statistical methods can be followed.

For the Spotify top songs data, we first test using a visual method by comparing how much the log-log plot resembles a straight line. Plotting the distribution as $log(y)$ against $log(x)$ 4 we see that the Spotify distribution resembles a straight line but not as closely that the results can be conclusive.

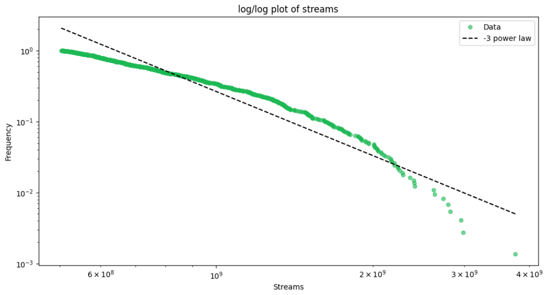


Fig. 4: Power law visual test Spotify.

To confirm if the data follows a power law or an exponential function statistical tests are needed. The first applied was the Log-likelihood ratio test with the Power Law library in Python ([3]). The Likelihood Ratio (R) is a statistical measure used to compare the goodness of fit of two competing models. It is calculated by taking the logarithm of the ratio of the maximum likelihoods of the two models:

$$R = ln(\frac{L_1}{L_0})$$

Where L_1 is the maximum likelihood of the more complex model (in this case, power law), and L_0 is the maximum likelihood of the simpler model (in this case, exponential). The log-likelihood ratio gives us a measure of how much more likely the data is under one model than the other. If R is positive, it indicates that the data is more likely under the more complex model. If R is negative, the data is more likely under the simpler model.

The Powerlaw python library contains a built-in function to apply the Log-likelihood ratio test, which uses the fit of each distribution to the data to calculate the maximum likelihoods and returns R and a p-value for statistical significance, with the null hypothesis $H_0 : p > \alpha$ meaning that there's a statistically significant difference and $H_1 : p < \alpha$ that the R is statistically significant.

The results in the Spotify total views data are: $R = -4.002$ and $p_{value} = 6.275e^-5$. This means that the distribution is a better fit to an exponential model than to a power law.

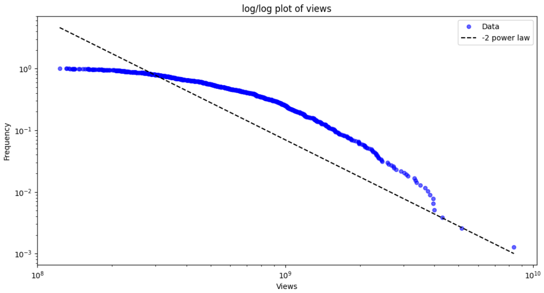


Fig. 5: Power law visual test Youtube Vevo.

First by visual method in the Youtube VEVO data it seems that the log-log plot doesn't fully reassemble a straight line.

The results in the Youtube VEVO total views data are: $R = 0.1439$ and $p_{value} = 0.9595$. This means that the distribution may be a better fit to a power law but the p-value is not statistically significant enough.

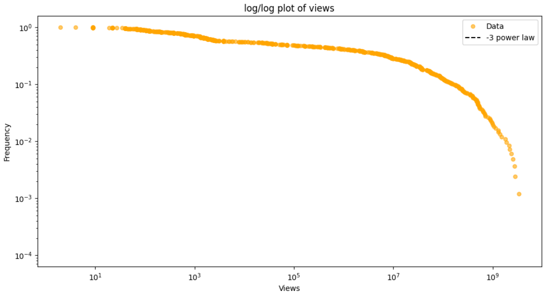


Fig. 6: Power law visual test Youtube Reggaeton.

First by visual method in the Youtube reggaeton data it seems that the log-log plot doesn't follow a straight line in any way showing the flaws in the data collection as several low outlier values are present in the dataset as the query wasn't as specific as desired.

The results in the Youtube Reggaeton total views data are: $R = 0.3783$ and $p_{value} = 0.7051$. This means that the distribution may be a better fit to a power law than to an exponential model but the p-value is not statistically significant enough.

Now the Akaike Information Criterion (AIC), is utilized as a second methodology for discerning whether a dataset align more closely with a power law or exponential distribution, one must account for both the quality of the model fit and its simplicity. The AIC is calculated using the formula $AIC = 2k - 2\ln(\hat{L})$, where k represents the number of parameters in the model and $ln(\hat{L})$ denotes the natural logarithm of the model's maximum likelihood, values of which are derived from the same likelihood estimations used in the Log-likelihood ratio test. The lower AIC score of the two models suggests a more parsimonious model, meaning it provides an adequate fit using fewer parameters. Thus, by comparing the AIC scores of the power law and exponential models, one can determine which model is more likely to accurately reflect the underlying distribution of the dataset in question.

First the results for spotify were a Power-law AIC: 31291 and an Exponential AIC: 30827, proving furthermore like in the Log-likelihood ratio test, that in this case the exponential model is more efficient at explaining the variability in the data without overfitting better fit to an exponential model.

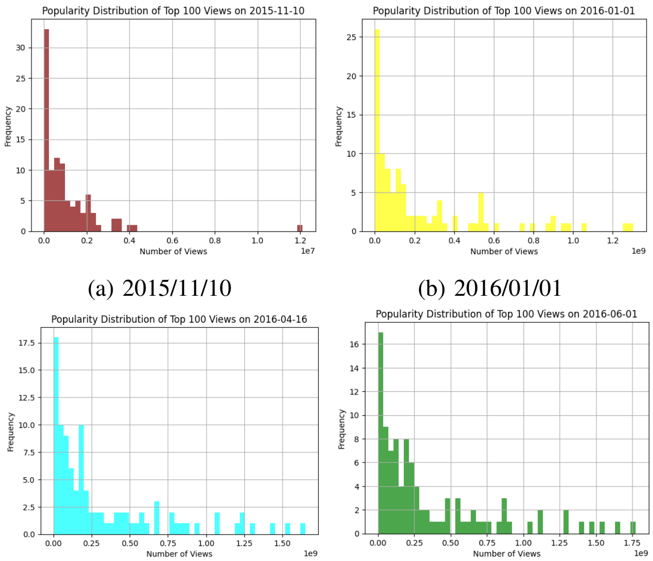
Then the results for Youtube Vevo show the Power-law AIC: 34023 and the Exponential AIC: 33126, this suggest that the exponential model is a better fit, the clear discrepancy here could be due to the fact that the AIC accounts for the number of parameters in the model, whereas the Log-likelihood ratio test focuses solely on the fit. When the AIC results differ from the Log-likelihood ratio test, it indicates that the goodness of fit needs to be considered alongside the model complexity. The lower AIC for the exponential model suggests that it is the more parsimonious model for the data, despite the Log-likelihood ratio test potentially favoring the power-law model.

Finally for the results for Youtube reggaeton the results are Power-law AIC: 23198 Exponential AIC: 31813, this does match the conclusion Log-likelihood ratio test result and the Power-law AIC is much lower than the Exponential AIC, suggesting the power-law model as the better fit. The significant difference in AIC values here indicates a strong preference for the power-law model. Even though the p-value from the Log-likelihood ratio test was not statistically significant, the AIC provides a separate validation for preferring the power-law model by penalizing the exponential model's fit relative to its complexity.

A. *Total Views vs. Unique Viewers vs. Daily Views*

Analyzing the total number of views over an extended period is suitable for investigating power law popularity distribution. This method provides a comprehensive view of a video's overall popularity and accounts for cumulative popularity over time. It's suitable to analyse long-term trends and provides a comprehensive analysis that is not affected by daily fluctuations or short-term spikes in popularity. In contrast, data of daily views can be useful for assessing the immediate impact of a song or to see if it has virality effects. The two answer different research questions, but to assess the overall popularity of songs, cumulative popularity may be more suitable.

Using the class provided data of Youtube songs from 2015 and 2016 dataset that includes the top 100 songs per day accroding to Spotify's ranking, three dates were selected, the initial date (2015/11/10), a random date (2016/01/01) and the end date (2016/04/16) resulting on the following graphs.



(a) 2015/11/10

(b) 2016/01/01

(c) 2016/04/16

(d) 2016/06/01

Fig. 7: Popularity distribution of daily views in Youtube.

The metrics resulting of the statistical tests are:

	LLR		AIC	
	R	p-value	Power law	Exponential
2015/11/10	+1	0.31	2979	2974
2016/01/01	-1.32	0.18	3982	4054
2016/04/16	-1.84	0.06	3985	4043
2016/06/01	-1.61	0.1	4026	4082

TABLE I: Comparative table of metrics

Some hypotheses may explain the results. Power laws are often observed when analyzing cumulative data over time because a few videos accumulate a disproportionately large number of views. These few highly popular videos skew the distribution, resulting in a long tail. Daily views may be exponential as they are more influenced by short-term factors like marketing, trends, or social media sharing. Each day, different videos may capture viewer attention, leading to fluctuations in daily views that don't necessarily conform to a power law distribution.

III. Views Growth

Prompt*: In the weekly project, two assumptions were stated with respect to the number views of a music video: 1. Without network effects, users visit a website with a certain frequency and view the music video if it matches their taste; 2. The number of views of a music video in the next day is proportional to the number of views up to the day before, resulting in an exponential growth of views in time.*

The assumptions that users visit a website at a certain frequency and watch music videos according only to their taste captures the core concept of user engagement, but oversimplifies the complexity of a user's decision to watch a video. In real life users may not visit the website with a fixed frequency. It's very uncommon in fact for a person to follow a fixed Youtube usage pattern. Other real-world behaviors that aren't taken into account are the user recommendations, trending content, word of mouth recommendations and random exploration. Therefore, this assumption is somewhat simplistic.

Regarding the assumption of exponential growth, in reality, the growth of views is influenced by various factors, such as initial promotion, social sharing, trends, viral events, interaction with other social media platforms and connectedness between artists. The exponential growth assumption doesn't account for these external influences, and it may not hold true for all videos.

A more realistic yet still relatively simple assumption that can explain the number of views over time could be a modified version of the "rich-get-richer" model that acknowledges the influence of network effects. In Rich-Get-Richer the number of views of a music video on a given day is influenced proportionally by its previous popularity, but it is also affected by external factors such as trending algorithms, social media sharing, and collaborations with popular content creators or artists. Though these external factors cannot be as easily modelled, we can add this to the previous assumptions:

1) Without network effects, users visit a website with a certain frequency and view the a video if it matches their taste and/or has good initial exposure (very famous artists, external media coverage, marketing and viral trends).
2) The number of views of a music video with good initial exposure in the next day is proportional to the number of views up to the day before, resulting in an exponential growth of views in time.

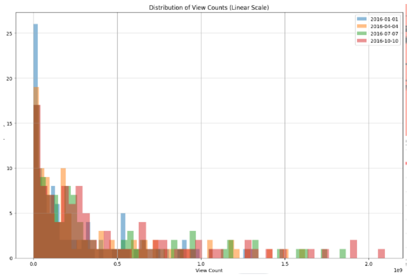
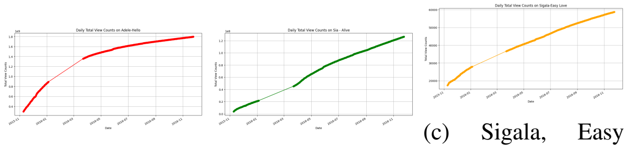


Fig. 8: Distribution of Youtube views on different days. (Class given dataset)

Using the class given datasset we can see how our assumptions play into the distribution. In the early phase of a video's life, if it matches the users' taste without the influence of network effects, it will garner views at a rate dependent on the frequency of interested user visits, and this initial number of interested users is dependent on the fame of the artist and previous marketing among other things discussed. This could explain why there's a large number of videos with a lower view count: they have been seen by users who happened upon them because they matched their taste. Videos that manage to get past a certain threshold of views might start to experience the early stages of network effects or might be featured more prominently by the platform's recommendation algorithms, hence the faster growth in views. This can be seen on individual video basis.



(a) Adele, Hello

(b) Sia, Alive

(c) Sigala, Easy Love

Fig. 9: Popularity distribution of individual songs over time.

The view count plots for the individual songs "Hello" by Adele, for example, shows a typical pattern of popularity growth for successful music videos on YouTube. When the song was first released, it likely received a surge of views due to initial interest from fans, media coverage, and promotional efforts. This phase represents the immediate impact of the release where the song matches the taste of a large number of users who are actively searching for the new content or are exposed to it through various media channels. Even without considering network effects, the songs likely matched the tastes of a broad audience, leading to a rapid increase in view counts. This is evident in the steep slope of the graph, which indicates a high rate of growth. On the other hand for a song like Sia's Alive, there are less initial views as she was a less stablished artist at the time, but the graph shows a surge in the middle, which could indicate an unexpected trend.

References

[1] Google Developers, *YouTube Data API v3*, Google, 2023. [Online]. Available: https://developers.google.com/youtube/v3

[2] "Chartmasters: Music industry - one step further," https://chartmasters.org/, accessed: 2024-01-26.

[3] J. Alstott, E. Bullmore, and D. Plenz, "powerlaw: a python package for analysis of heavy-tailed distributions," 2014, accessed: 2024-01-26. [Online]. Available: https://pythonhosted.org/powerlaw/#powerlaw.Fit.distribution_compare