Unsupervised methods to group users' consumption behaviour for service degradation policies personalisation.

In Alphabet order:

Roberto Casaluce [a, b] - Maciej Zuziak [a,c]

[a] Department of Computer Science, University of Pisa, Pisa, Italy;

[b] Institute of Economics and EMbeDS, Sant'Anna School of Adv. Studies, Pisa, Italy;

[c] KDD Lab, National Research Council of Italy (CNR), Pisa, Italy.

ABSTRACT:

Over-the-top services (OTTs) are the backbone of the modern digital entertainment industry and shape the predominant part of the consumers' digital behaviour. Due to the steadily increasing demand for such content, it is vital to study the users' behavioural patterns to better personalise the products as well as design infrastructural constraints (degradation policies) in a less encumbering manner. In this report, we use an open-source dataset to perform a user classification based solely on the time of consumption. What makes the task even more challenging is that the dataset is devoid of any demographical or geographical labels due to privacy constraints. In this report, we present that grouping the users in just a few clusters is possible based on their time spent on these online platforms. Indeed, despite the challenges posed by employing very sparse data, we have found that the number of clusters that more correctly group the users is three, and those clusters can describe consumers' behaviour in an insightful way.

1. Introduction

Over-the-top services gained much popularity over the recent fifteen years, drawing the attention of consumers, researchers and regulators all over the globe. Taking into account that the pandemic has only enhanced that trend [1], it would be quite reasonable to expect that those services will be taking the lion's share in the global multimedia market in the upcoming ten years, at least in regions that are more prone to digitalisation [2]. In connection with the growing popularity of the OTT-based business, many researchers try to tackle the relationship between the various emerging consumption petters, provide a potential classification of the group of consumers or explain a relationship between existing groups of consumers and their engagement time. In this paper, we will try to approach the issue of classifying consumers based on their consumption behaviour. Such classification can be used, among other things, to better understand emerging consumer groups and their needs or to study service degradation policies' relevance.

> ***Over-the-top services (OTT)*** *are commonly understood as a media service offered directly to the viewers via the Internet, without using traditional means such as cable television, broadcasting television systems or satellite television. For a wider audience, they are also known as "subscription-based video-on-demand (SVOD)[1]", as this second name is perhaps much more popular among consumers [3], [4].*

> ***Service Degradation Policy*** *is a measure to throttle the use of a network by users generating heavy traffic in order to protect the capacity of such network and preserve it for other users (possibly generating much less severe strain on the infrastructure. In* [5], *it was proposed to personalise such policies to individualise usage caps placed on particular types of users.*

Dataset used in this project comes from the research described in [5] and published on the Kaggle platform [6]. Concerning the topic of the research, we pose several research questions:

[1] Also sometimes as: transactional video on demand (TVOD), or advertising-based video on demand (AVOD).

- Is it possible to identify common groups of consumers based on time spent in particular applications?
    - (If yes) then how we can possibly explain or describe the characteristics of those groups?
- What are the relationships between users of different (constituting possible adversaries in terms of the competition) platforms? Can one platform influence (positively or negatively) the time spent on another platform?

The underlying research is primarily exploratory in its nature – as the majority of users' data belongs to the companies hosting the services provided to those users, we are primarily interested in the preliminary, unsupervised study – to assert the existence of any emerging patterns and verify our assumptions about the ecosystem.

2. Methods

Exploratory Data analysis

The original dataset contains 1.249 variables (statistical units, entries) and 114 features. The features can be divided into three categories:

- IP_ADRESSES (contained only in the first column);
- SERVICE_TIME_OCCUPATION (expressed in seconds, one feature represents one service);
- SERIVCE_DATA_CONSUMPTION (expressed in bytes, one feature represents one service).

Amazon	AmazonVideo	Apple	AppleiCloud
AppleiTunes	AppleStore	DataSaver	Dropbox
eBay	Facebook	GMail	Google
GoogleDocs	GoogleDrive	GoogleHangoutDuo	GoogleMaps
GooglePlus	GoogleServices	HTTP	HTTP_Proxy
IMO	Instagram	LinkedIn	LotusNotes
Messenger	MS_OneDrive	MSN	NetFlix
Playstation	PlayStore	PS_VUE	Sina_Weibo
Skype	SkypeCall	Snapchat	SoundCloud

Spotify	**Steam**	**TeamViewer**	**Telegram**
TikTok	**Twitch**	**Twitter**	**Waze**
WeChat	**WhatsApp**	**WhatsAppCall**	**Wikipedia**
Xbox	**Yahoo**	**YouTube**	**Zoom**

Table 1: Matrix containing all the features (services) included in the dataset.

The first natural question regarding the features concerned the relationship between time consumption and service data consumption. A study of cross-correlations indicated a high correlation between those two types of features. The phenomenon is relatively easy to explain, as all OTT services rely on data consumption. Apart from the heavy correlation between those two types of features, a few more questions regarding data usage may arise. For example, it would be quite natural that some services would require heavy data usage (one example of that could consist of video streaming services), while others do not impose such strains on the network, even if the user is spending a lot of time browsing particular application. Some may also point out that it is unclear how we could measure the overall data consumption – whether we should consider only download or download and upload. We have decided to remove this category of features and rely solely on time-related features for all of those reasons.

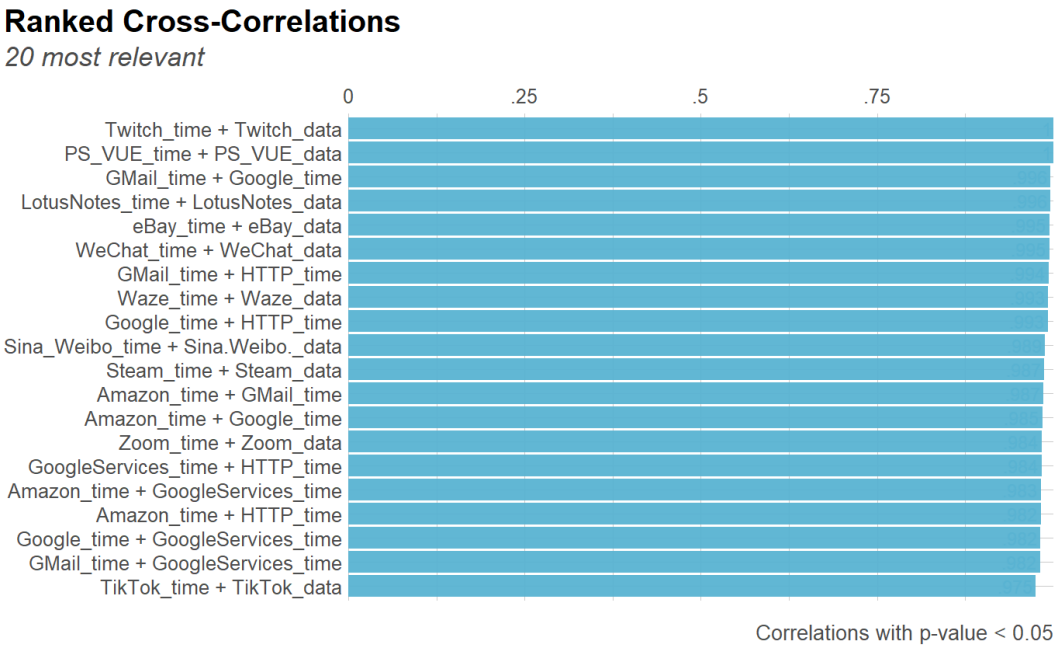

Diagram I: *Ranked overview of the 20 most relevant cross-correlations among the features contained in the dataset. On the left: selected pair of two features. On the right, pairwise correlation value.*

Because each feature describes a different service (and related time/data consumption), it is natural to assume that while the most popular services are represented in most statistical units, some of the less popular will not be represented even in the quarter of the variables. In the most deprived case, we will have to deal with a null vector, although, more naturally, it would be to assume that the vector will contain one or two non-zero entries. Such features theoretically increase the dataset dimensionality, but they do not possess any informative value in our scenario. Thus, we have decided to drop features with less than two entries.

To gain a better understanding of the distribution underlying the dataset, we have decided to sum up all the time-related features for individual users [$y = \sum_{i=0}^{j} x_i$, where y represents individual users and x_i, x_i +1,…, x_j represent subsequent time entries for different services] and all the usage time for individual services [users [$y = \sum_{i=0}^{j} x_i$, where y represents individual service and x_i,x_i+1,…,x_j represents time-usage of subsequent users for the same service] and to create a histogram of the total time consumption for individual users and services.

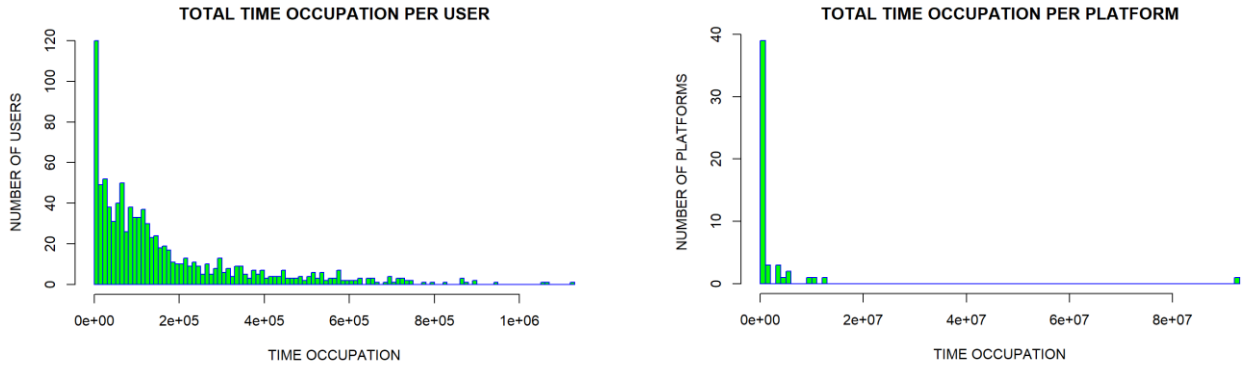

Diagram II: *Histograms presenting Total Time Occupation per User [top] and Total Time Occupation per Platform [bottom]. As seen above, the majority of users occupy the left side of the histogram – with a minimal registered usage time or with none at all. The same applies to the second histogram presenting total time occupation per platform – the majority of platforms are minimally used according to the sampling data, while a small fraction of them sustains a heavy traffic flow (rectangle on the right side of the histogram).*

Due to the heavily skewed distribution of the data, we have decided to perform a transformative operation – we add a small pseudo-random number between 0 and 1 to each

entry and then perform log transformation on the whole dataset [more explanation on that operation is provided in the subsequent the Log Transformation Section]

LOG Transformation

In order to render the dataset more operational, we perform a transformative operation – firstly by adding a pseudo-random generated number to every element of the dataset[2], then by performing a logarithmic transformation of every element in the matrix. By doing that, we expect that we could shift the distribution towards canter and normalise it without adding additional bias.

$$\forall ij\ a_{ij} + \lambda \text{ where } \sim N1,0$$

and a_{ij} is the i-th element of the j-th column of the dataset

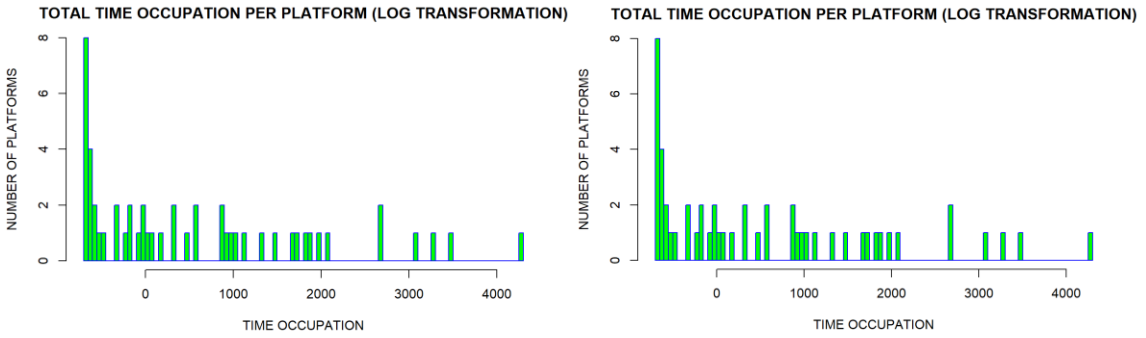

Diagram III: *Histograms presenting Total Time Occupation per User [top] and Total Time Occupation per Platform [bottom] after performing the aforementioned transformative operation.*

We were cautious in interpreting the results during applying the log transformation correctly. Such practice is common in the community [7]. However, it can cause issues with further interpretation of the data, especially where the data is further used in, among other things, linear or logistic regression [8].

Data processing summary and Further remarks

During the data processing stage, we had:

- Removed features (services) that have less than two active users;

[2] The added number is the same for every element in the dataset.

- Removed every feature that was describing different quantities than total time spent on a selected platform;
- Removed duplicated statistical units (registered users);
- The performed transformation is described in the Log Transformation Section.

By performing all the actions described above, we have reduced the size of the dataset from 1.249 statistical units and 114 features to 973 statistical units and 52 variables, respectively. Each variable now corresponds to the time spent on the platform by particular users (identified by their IP address) expressed as an argument of the logarithm with base 10.

PCA for Data Visualization

We have employed Principal Component Analysis (PCA) to display in a lower dimension our data. PCA helps us discover patterns in the data that are difficult to learn only by plotting a summary of the data. As explained above, we still have some highly correlated variables even after removing all the highly correlated features by dropping the features representing the data consumption ones. For example, in Diagram I, the features representing the time spent on Gmail and Google Services are, as expected, highly correlated. In this regard, PCA would help us summarise the data in a smaller set of representative variables by keeping most of the information in the original features set [9]. This new set of variables, called principal components, are uncorrelated and represent the maximum variance of the original data. **Unscaled data**: In most situations, it is essential to scale the data before using PCA when the features are measured with different scales. Those with the highest variance can produce very large loadings that contribute to the principal components. Therefore, loadings calculated on unscaled data with features measured with different scales can create a distorted magnitude of the relationship between the variables with the highest variance and the principal components. We have left our data unscaled since all the features are measured with the same units, i.e., in seconds; therefore, we have employed PCA with only log-transformed data. **Method to flag outliners:** We have applied the log transformation to our data to reduce the skewness of the distribution, but this does not guarantee that the data is approximated to a normal distribution and has fixed the outliner problem [10].[3] Besides using a classic

[3] The data used in the present work is a very sparse dataset since around 40% of its entrances are zeros which corresponds to users that did not use some platforms at all (see the plot in Diagram II). This high number of zeros is why the distribution of our dataset is very skewed.

version of PCA to visualise the data in a lower dimension, we also use a robust version of it, better suited to cope with skewed data to flag outliners. More specifically, we use the ROBust method for Principal Components Analysis (ROBPCA), which calculates the loadings using a projection-pursuit technique, and the Minimum Covariance Determinant (MCD) method is less sensitive to outliers [11]. This robust version of the PCA helps categorise the different types of outliners. Indeed, it helps to distinguish outliers points that can be either leverage points or orthogonal outliers. In brief, the latter are data points with higher orthogonal distance (OD) values than regular points. This distance measures the deviation of a data point from the k dimensional PCA subspace. Conversely, the leverage points have higher score distance values than the regular points, which measures the distance between observation of a k dimensional PCA subspace and its origin. In the next section, when we discuss the results of our analysis, we examine the type of outliners found in our data.

Clustering technique

The lack of a target variable and any features representing the users' socio-demographic features have prevented us from applying any supervised techniques that could help us classify the users into different classes. Hence, we had to rely on unsupervised techniques to group in clusters the users based on their engaging time on the online platforms. We have run many prelaminar experiments with different clustering techniques. However, the results were not encouraging since it was challenging to identify the exact number of clusters obtained in our data.[4] This problem of defining the number of clusters was even more marked with the hierarchical clustering (HC) techniques. Therefore, at first, we have run the gap static method [12], which tries to find the best number of clusters by comparing 'the total within intra-cluster variation for different values of k with their expected values under null reference distribution of the data' [13]. The gap static method has found the number of clusters in our dataset that was not consistent with a previous work that used the same data to cluster users based on their engaging time [5]. In [5], they have found 4 clusters dividing the users into low, medium, high, and very high based on their time spent and the data used on the online platforms using the K-means clustering methods. **Choosing the best cluster solution:** To choose the best cluster solution, we have run the hartigans's method [14] to identify the best

[4] In the Appendix, we have added a few plots of the dendrograms obtained during our prelaminar experiments.

number of clusters and other validation methods [15] to check the internal and stability measures over three different clustering algorithms, namely HC, K-means, and Partition Around Medoids (PAM) algorithms. Finally, as the last evaluation method to assess the best cluster solution, we have run the Hubert index algorithm that assesses the best number of clusters by majority rule over different evaluation methods and indexes. **Clustering algorithm**: The clustering algorithm that we have chosen is the K-means clustering method. As discussed in the section dedicated to the results, the evaluation methods explained above have indicated K-means clustering is the best solution to divide the individuals of our data into groups.

3. Results

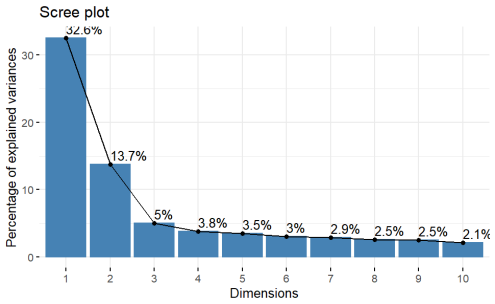

Figure 1 *Scree Plot*

We have run the Principal Component Analysis on our log-transformed data without scaling further the data. In Fig. 1, we can see the Scree plot depicting the proportion of explained variance in each dimension. For the first two principal components, the percentage of variances explained is only around 46%, and we can observe an inflation point in the third dimension. The inflation point is where there is a drop-off, an elbow, after the third dimension. This inflation point should be used to select the number of components that should preserve most of the information in the data. In our data, we have 52 features that correspond to 52 dimensions, and the first three components explain around 51% of the variance.

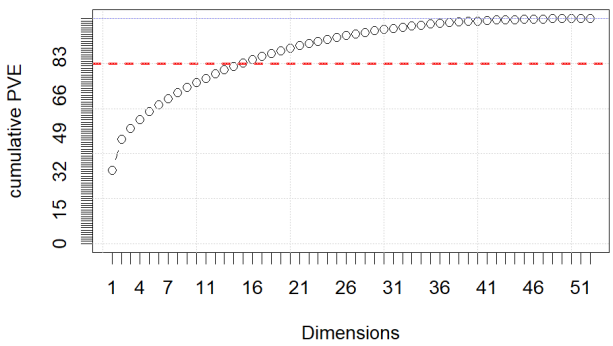

Figure 2 *Cumulative Percentage of variance explained*

In contrast, all the remaining 49 components explain less variance in the data than the first three. Fig. 2 depicts the Cumulative Percentage of Variance Explained, in which we can see that each of the components after the first three ones does not add a lot of variances explained. We have added the dashed red line indicating the number of components that we should

retain to reach a threshold of 80% of the variance explained. To reach this threshold, we should need around 15 components.

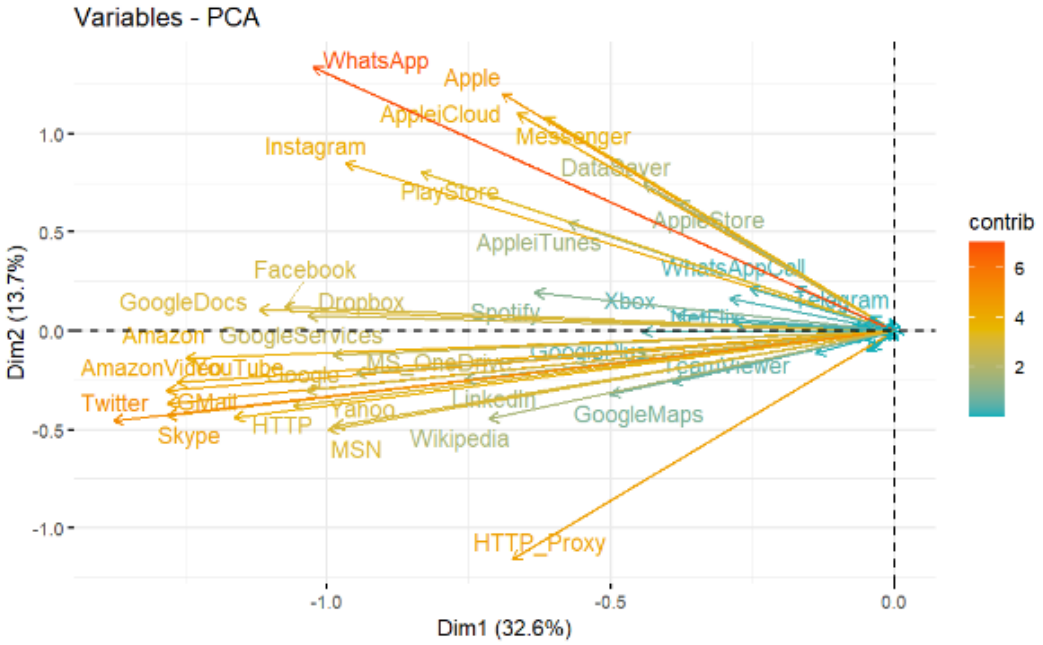

Figure 3 *Loadings in the first two dimensions*

We have also plotted the loadings in a 2D plot representing the first two dimensions, Fig. 3. The loadings display how much each original variable contributes to the corresponding principal component. The loadings are the correlations between the variables and the components. Correlated variables have a slight angle and point in the same direction. We can see that Twitter and WhatsApp have the two longest arrows, and both contribute more, respectively, to the first and second components. To better display the contribution of the first 20 features in each of the two dimensions, we have plotted them in Fig. 4. The most

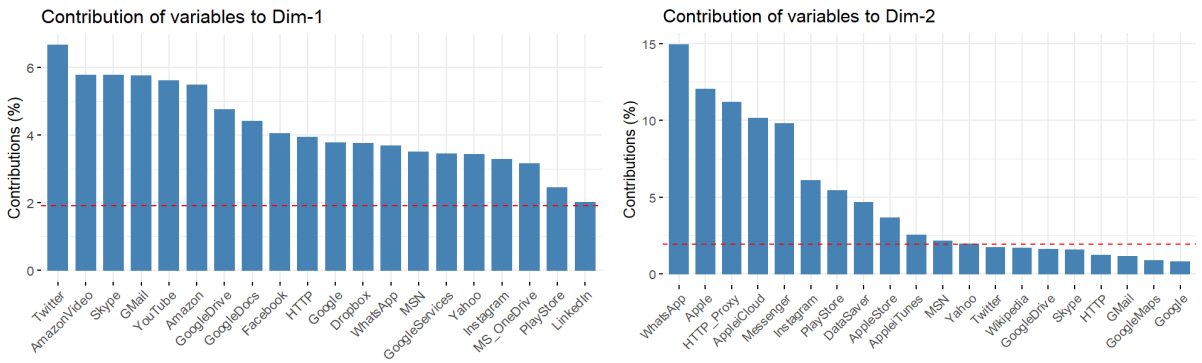

Figure 4 *Contribution of the variables for the First Dimension (a) and Second Dimension (b)*

interesting information provided by these two plots is that in the second dimension(Fig. 4 (b))

, four features correspond to Apple platforms that are absent in the first dimension. The direction of the loadings representing Apple services and the small angles between them confirm that users who utilise one of those services are most likely to use others in the same brand. Fig. 5 (a) depicts both loadings and individuals, i.e., columns and rows of our log-transformed data, in a Biplot to better appreciate the magnitude of the loadings in the first two dimensions. We have added a second Biplot, Fig. 5 (b), for both loadings and individuals obtained after applying PCA on the scaled version of our data without applying the log transformation. As we can see, most individuals in Fig. 5 (b) are more concentrated in a small area but with a large spread of some individuals located far from that area. In contrast, in Fig. 5 (a), most individuals are not concentrated in a small area but more equally spread around the origin with less extreme values thanks to the log transformation. The two figures, Fig. 5 (a) and (b), demonstrate how a log transformation creates a more symmetric distribution and helps to smooth the values collocated to the extreme of the original distribution.

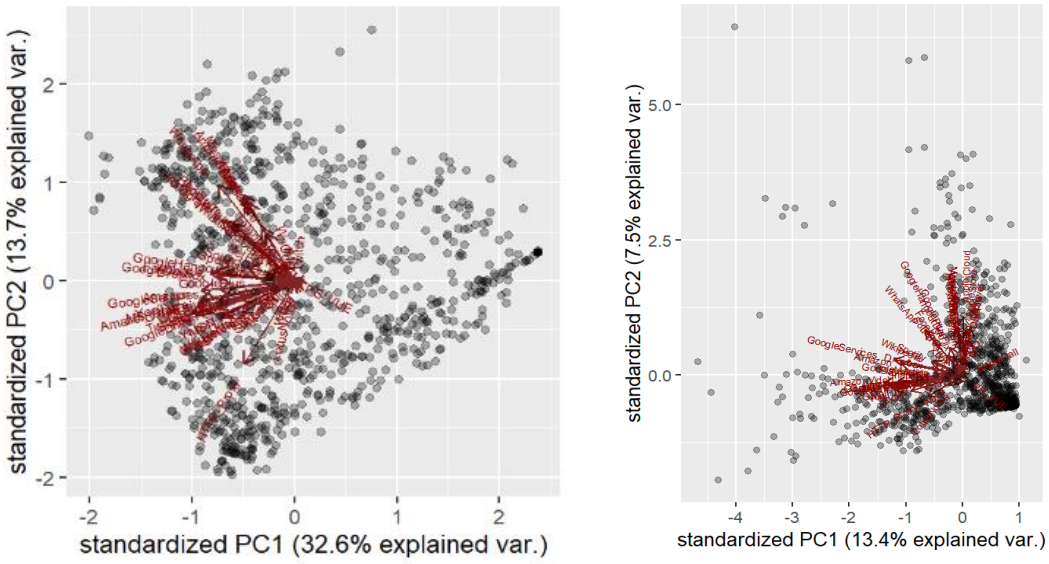

Figure 5 *Biplot for the Log Transformed data (a) and for the Scaled data (b)*

In the Section 2, we have explained that the log transformation can help create a more symmetric distribution but that it cannot completely fix the outliners problem. In Fig. 6, we have plotted the ROBPCA results that can help us to flag the outliners. We can see that the data still have some outliners, and many of them are orthogonal outliners, such as the points 621 and 286, and others are good leverage points, such as the point 492. The latter points can improve the accuracy of the fitted PCA subspace [16] since they are close to the PCA space but far from the regular observations. On the other hand, there are no bad leverage points

with a large orthogonal distance to the PCA space, and their projection on the PCA space is distant from most of the projected data [17].

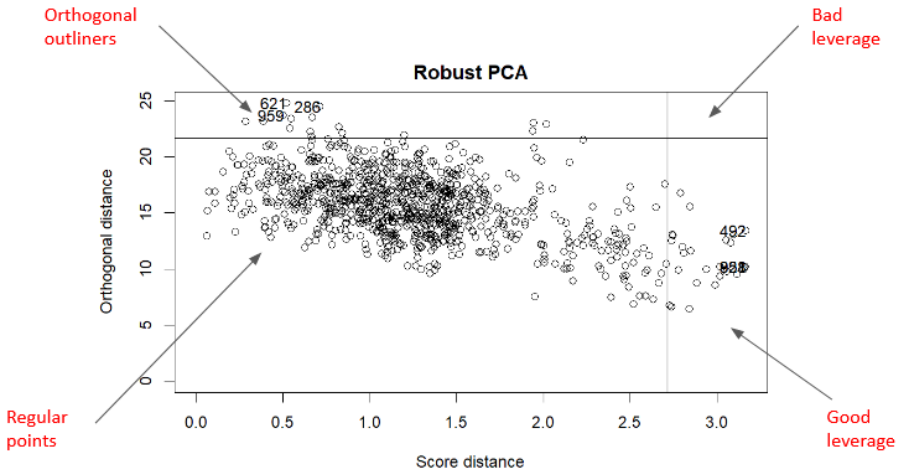

Figure 6 *Robust PCA: outlier map: plot of orthogonal distances versus score distances*

About the actual clustering method, in Fig. 7 are plotted the results of Hartigan's method over two clustering algorithms, namely Agglomerative HC (a) and K-means clustering (b) methods, used to find the best number of clusters for these two clustering methods.[5]

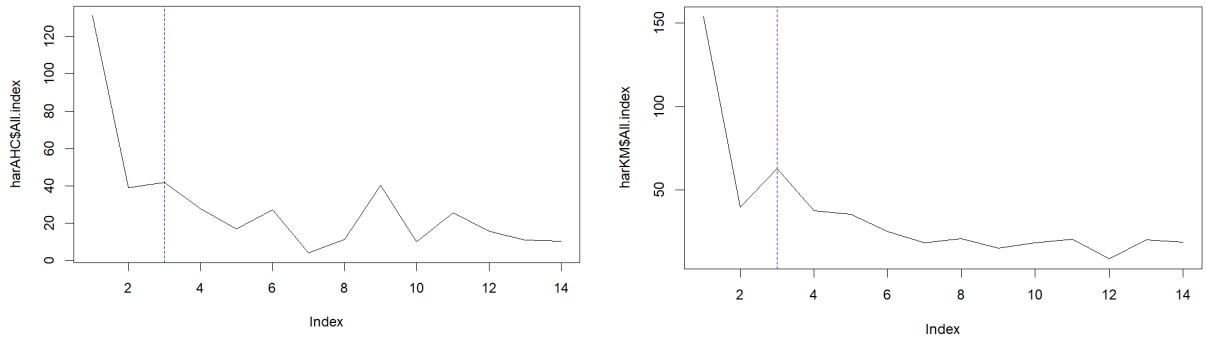

Figure 7 *Hartigan Indexes: Agglomerative HC (a) and K-means (b)*

For both clustering methods, the Hartigan index returned three as the best number of clusters to group the users in our data. After that, we evaluated the best cluster solutions for internal and stability validation measures [15]. The stability measures evaluate the stability of a

[5] We have also run the gap static method that estimates the number of clusters, and for this method, the number of possible groups in our data is 10 clusters using the k-means clustering method. This high number of clusters completely contradicts what [5] found about the number of groups in our data. We have added the plot obtained in the Appendix.

clustering result by comparing it with the clusters obtained by removing one column at a time. The results are listed in Table 1. The last column indicates if the values of each measure should be either minimised or maximised.

	Measure		**Clustering method**	**# of Clusters**	**Range**	
Stability measures	APN	0.009	K-means	3	From 0 to 1	Minimised
	AD	10.2	K-means	4	From 0 to infinity	Minimised
	ADM	0.095	K-means	3	From 0 to 1	Minimised
	FOM	0.93	K-means	4	From 0 to 1	Minimised
Internal measures	Connectivity	139.75	Hierarchical	3	From 0 to infinity	Minimised
	Dunn	0.26	Hierarchical	3	From 0 to 1	maximized
	Silhouette	0.20	K-means	3	From -1 to +1	maximized

APN - average proportion of non-overlap, AD - average distance, ADM - average distance between means, FOM - figure of merit.

Table 1 *The results of the internal and stability measures*

Most measures point toward the K-means clustering algorithm as the best method to divide the users into clusters. Finally, to verify that the K-means clustering method is the best choice, we have decided to run the Hubert index algorithm that assesses the best number of clusters by a majority rule by going over different evaluation methods and indexes. Also, according to the majority rule, the best number of clusters for the Hubert index is 3. Therefore, we have run the K-means clustering algorithm with k = 3, where k is the number of clusters with the Euclidian metric used for calculating dissimilarities between observations.

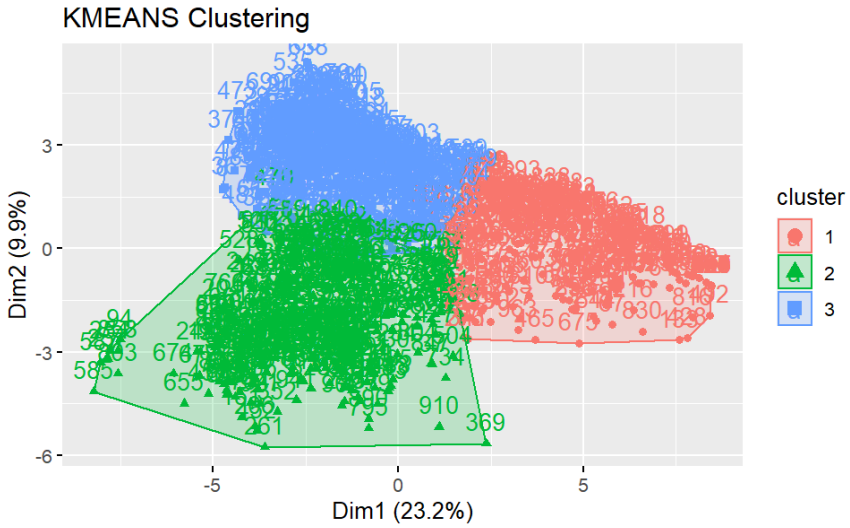

Figure 8 *2D plot of three clusters*

The three clusters found correspond respectively to low, medium, and high levels of the users' time spent on the online platforms, with a size of each cluster respectively of 266, 343, and 364. Fig. 8 depicts the 2D plot of the clusters. We can see that the clustering algorithm was able to cluster in the three groups the users, and only some of them located on the bounders of each cluster were not correctly grouped in their respective clusters.

To verify that the actual clusters found to correspond to low, medium, and high levels of the engaging time of users on the online platforms, we have calculated the mean value of each group.

Clusters	**1 - Low**	**2 - Medium**	**3 - High**
Size of the clusters	266	343	364
Mean time spend on the platforms	359.42	2472.27	5920.06

Table 2 *Sizes and mean values of the three clusters identified*

The last row of Table 2 lists the mean values of the time spent on the online platforms in each group. Those mean values confirm that the first group users have spent less time on an average than the users in the other two groups, and the users in the second group have spent more time online than users in the first group but less time than the users in the last group.

4. Discussion

In the present work, we have employed data representing the users' engagement time on some of the most famous online platforms providing the so-called Over-the-top services. We have shown how to group the users in just a few clusters. Indeed, despite the challenges posed by employing very sparse data, we have found that three are the number of clusters that more correctly represents the use of those platforms based on the time spent on them. In addition, during the EDA phase, we have identified a strong correlation between using a platform and others within the same services. For instance, users that utilise Gmail could also use other Google services, or an Apple user most likely utilises many other services of Apple's brand. Nevertheless, the lack of any of the users' socio-demographic features has prevented us from answering our first sub-question, 'how we can explain or describe the characteristics of those groups?' since, besides an apparent correlation of the use of some online services within the same brand, we cannot describe other characteristics of those users. It would be interesting to describe the users' engaging time by their age and gender. This could help those online platforms better improve the availability of their services to those groups.

By employing the Principal Component Analysis and visualising the loadings [Figure 3], we have also reached a particular conclusion regarding the relationships between users of different (constituting possible adversaries in terms of the competition) platforms. As unexpected as it may be, it seems that there is no strong negative correlation between the time usage of different OTC services included in the dataset. In another way, users that spend much time on platform X may also spend much time on platform Y despite the fact that X and Y are both in-market competitors. If this observation is correct, it could lead to some valuable inferences about the digital consumer population. For example – the individual preferences do not matter as much as the class to which the users belong – whether they are low, medium or heavy-consumption users. Those who spend a lot of time using OTC services are likely to use more than one channel, possibly simultaneously paying for services that are labelled as “market competitors”. On the other hand, those who do not use OTC or digital services are at all are quite indifferent to all the offers that are out there – preferably staying out of the digital loop at all. However, we must all emphasise that this observation is quite dubious – as our dataset does not contain enough competing services. It would be fascinating to compare

different OTC services that are in direct competition by offering the very same content with slightly different terms & conditions of supply (we could provide an example of Netflix, HBO GO, Amazon Prime Video, Disney Plus and Hulu) but this stays outside the scope of this report and unfortunately – for now, we do not possess sufficient data. Nevertheless, it may be a quite valuable research path to follow in the upcoming years.

Appendix

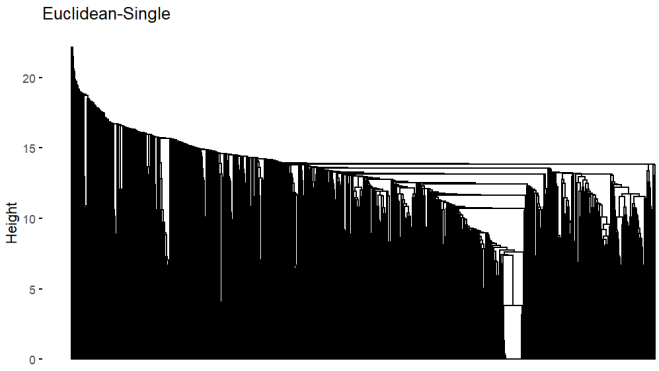

Figure 9 *Preliminary experiments: HC single Linkage*

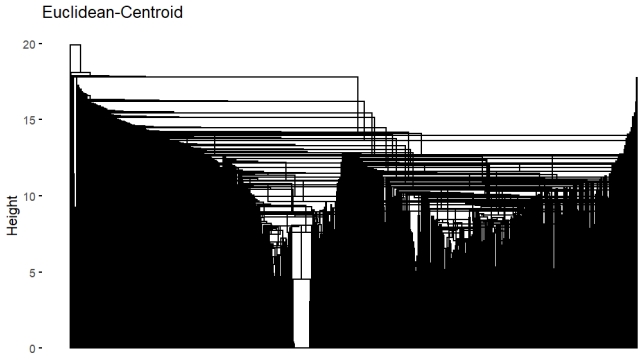

Figure 10 *Preliminary experiments: HC centroid Linkage*

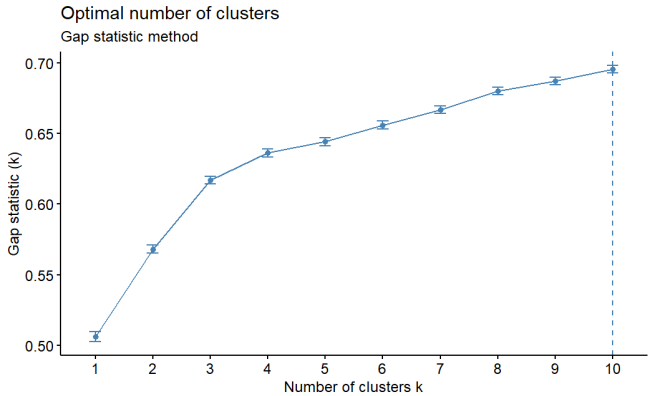

Figure 11 *The number of clusters found by the gap static method for the K-means clustering algorithm*

References

[1] R. T. Watson, 'World-Wide Streaming Subscriptions Pass One Billion During Pandemic', *Wall Street Journal*, Mar. 18, 2021. Accessed: May 30, 2022. [Online]. Available: https://www.wsj.com/articles/worldwide-streaming-subscriptions-pass-one-billion-during-pandemic-11616079600

[2] E. Sundaravel and N. Elangovan, 'Emergence and future of Over-the-top (OTT) video services in India: an analytical research', 2020, doi: 10.18801/ijbmsr.080220.50.

[3] 'Over-the-top media service', *Wikipedia*. May 22, 2022. Accessed: May 30, 2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Over-the-top_media_service&oldid=1089134030

[4] 'What are SVOD, TVOD, AVOD?', *Imagen*. https://imagen.io/blog/what-are-svod-tvod-avod/ (accessed May 30, 2022).

[5] J. Rojas Meléndez, A. Pekar, A. Rendón, and J. Corrales, 'Smart User Consumption Profiling: Incremental Learning-Based OTT Service Degradation', *IEEE Access*, vol. 8, Nov. 2020, doi: 10.1109/ACCESS.2020.3037971.

[6] 'User OTT Consumption Profile - 2019'. https://www.kaggle.com/jsrojas/user-ott-consumption-profile-2019 (accessed May 30, 2022).

[7] C. FENG *et al.*, 'Log-transformation and its implications for data analysis', *Shanghai Arch. Psychiatry*, vol. 26, no. 2, pp. 105–109, Apr. 2014, doi: 10.3969/j.issn.1002-0829.2014.02.009.

[8] C. Feng, H. Wang, N. Lu, and X. M. Tu, 'Log transformation: application and interpretation in biomedical research', *Stat. Med.*, vol. 32, no. 2, pp. 230–239, Jan. 2013, doi: 10.1002/sim.5486.

[9] G. James, D. Witten, T. Hastie, R. Tibshirani, An introduction to statistical learning, 112, Springer, 2013.

[10] C. FENG et al., 'Log-transformation and its implications for data analysis', Shanghai Arch Psychiatry, vol. 26, no. 2, pp. 105–109, Apr. 2014, doi: 10.3969/j.issn.1002-0829.2014.02.009.

[11] M. Hubert, P. Rousseeuw, and T. Verdonck, 'Robust PCA for skewed data and its outlier map', Computational Statistics & Data Analysis, vol. 53, no. 6, pp. 2264–2274, Apr. 2009

[12] R. Tibshirani, G. Walther, and T. Hastie, 'Estimating the Number of Clusters in a Data Set via the Gap Statistic', Journal of the Royal Statistical Society. Series B (Statistical Methodology), vol. 63, no. 2, pp. 411–423, 2001.

[13] 'Determining The Optimal Number Of Clusters: 3 Must Know Methods', Datanovia. https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/ (accessed Jun. 04, 2022).

[14] J. A. Hartigan, Clustering algorithms. New York : Wiley, 1975. Accessed: Jun. 04, 2022. [Online]. Available: http://archive.org/details/clusteringalgori0000hart

[15] 'clValid: An R Package for Cluster Validation by Guy Brock, Vasyl Pihur, Susmita Datta, Somnath Datta', Mar. 2008, Accessed: Jun. 04, 2022. [Online]. Available: https://www.jstatsoft.org/article/view/v025i04

[16] P. Rousseeuw and M. Hubert, 'Robust statistics for outlier detection', Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery, vol. 1, pp. 73–79, Jan. 2011, doi: 10.1002/widm.2.

[17] X. Chen, B. Zhang, T. Wang, A. Bonni, and G. Zhao, 'Robust principal component analysis for accurate outlier sample detection in RNA-Seq data', BMC Bioinformatics, vol. 21, no. 1, p. 269, Jun. 2020, doi: 10.1186/s12859-020-03608-0.