

A Real-world Dataset of Netflix Videos and User Watch-Behavior: Analysis and Insights

Shruti Lall and Raghupathy Sivakumar
Georgia Institute of Technology, Atlanta, Georgia
Email: {slall,siva}@ece.gatech.edu

Abstract—Netflix is the most popular video streaming site contributing to nearly a quarter of global video traffic. Given the dominance of Netflix on Internet traffic, understanding how individual users consume content on Netflix is of interest to not only the research community, but to network operators, content creators and providers, users and advertisers. In this context, we collect Netflix viewing activity from 1060 users spanning a 1 year period, and consisting of over 1.7 million episodes and movies. We group the users based on their activity level, and provide key insights pertaining to the user's watch patterns, watch-session length, user preferences, predictability and watch-behavior continuation tendencies. We also implement and evaluate classifiers which are used to predict the user's engagement in a series based on their past behavioral patterns.

I. INTRODUCTION

Video streaming accounts for over 60% of downstream Internet traffic, and is expected to grow to 82% by 2022 [1]. Netflix, the world's most popular video streaming service, is alone responsible for nearly a quarter of the world's video traffic [2]. Given the dominance of Netflix on Internet resources, it is valuable to derive insights on Netflix usage which can be useful to not only the research community, but to network operators, content providers, marketing agencies, content creators as well as users themselves. This serves as the primary motivation for our work in which we conduct a meaningful analysis and provide key insights using a real-world dataset of users' Netflix behavior.

To this end, we use Amazon's Mechanical Turk (mTurk) platform to collect a dataset for Netflix usage from 1060 users. The collected dataset contains 1-year worth of viewing activity for each user, which amounts to over 1.7 million episodes and movies collectively watched. Beyond high-level statistics published by Netflix [3], there has been little work done towards collecting and deriving insights using real world usage data spanning a significant period of time.

Equipped with this dataset, we provide an in-depth analysis on user's watching behavior for movies and series content. Movies and series vary vastly in their form with movies being 3 times longer than episodes from series, and also are non-episodic. It follows that the way a user watches movies will be different to how they consume series content. We thus separate the analysis of user's watching behavior for movies and series. We derive key insights for individual user behavior related to their watch patterns, watch-session length, preferences, predictability of their future viewing, and their series continuation tendencies. Furthermore, we implement and evaluate classification models to predict the user's engagement

in a series, and the likelihood of them continuing to watch a series. We present our results by grouping our users into 3 categories based on the amount of content they consume: *low* active user, *moderately* active user, and *high* active users. We believe that the real value of the dataset lies in researchers using it for their respective problems. A core contribution of this work includes presenting results in the context of problems in the domain of networking and communications. Specifically, we consider the following sets of research questions (RQs):

- 1) Do users have a preferred day of viewing for movies/series? What is the number of days between subsequent watches? How does this differ for varying user activity levels?
- 2) How many episodes do users watch each day? Do active users tend to consume the same amount of content each day?
- 3) Do users watch the same genre(s) content regularly? Are users inclined to binge watch certain genres over others? Do active users prefer more popular and higher rated content? How much content is related to content the user has previously seen?
- 4) How much of a user's future watches are predictable? Is it easier to predict for less active users? Are certain genres easier to predict than others?
- 5) How much of a series does a user watch to its entirety? At what point does a user stop watching a series if they don't complete it? Can we predict when this point will arise? Which classification model is the most well suited for this prediction?

The rest of the paper is organized as follows: In Section II, we discuss the data collection methodology, and present a baseline analysis. In section III, we present results for user's movie watching behavior. In Section IV, we present the analysis pertaining to the aforementioned sets of questions for series; and in Section V we present conclusions of the paper.

II. DATA COLLECTION

A. Methodology

To collect our dataset, we rely on Amazon Mechanical Turk (mTurk) to gather anonymized Netflix viewing history from 1060 users for a 1-year period [4]. The mTurk platform allows a task to be posted for a fee, which in turn can be completed by users known as mTurkers. Studies have shown that mTurk samples can be accurate when studying technology use in the broader population [5]. The mTurkers were required to

navigate to their *viewing activity* page associated with their profile, and download their Netflix viewing activity as a csv file; the file was then anonymously uploaded via a dropbox link¹. The viewing activity file uploaded by the mTurker contains 2 fields: *title* and *date*. The *title* field consists of the name of the feature film or TV series/documentary, as well as the season and episode name where applicable, separated by colons. The *date* field consists of the most recent date that the title was viewed (there is no time of day given).

We then use *The Movie Database* API [6] (TMDB) to obtain the following metadata for each title watched by a user: the release date of the title, the IMDB rating, the number of IMDB votes for the title, the run-time in minutes, the genre(s), director(s), writers, actors, the language of the title, country of production, and related titles. For series, we obtain the number of seasons, and number of episodes each season has through appropriate API calls. A Postgres SQL database is used to store the user's viewing history and metadata.

B. Baseline Characteristics

A high-level overview of the collected dataset is presented in table I. We show the total number of movies and episodes watched by all the users in the dataset, as well as the total number of seasons and series watched by all our users. We also show the average number of hours a user spends watching series during each watch session (WS). We define a (WS) as a day on which at least one episode of some series is watched by the user.

TABLE I: Dataset Overview

Description	Value
Users	1060
Movies	63,296
Episodes	1,632,980
Seasons	121,101
Series	30,224
Hours per WS	1.8

III. ANALYSIS AND KEY INSIGHTS- MOVIES

In this section, we perform an analysis on the user's movie watching behavior. We group the users into 3 categories based on the number of movies they have watched in their submitted 1-year history. Users in *low* active category have watched less than 20 movies (11% of users), users in *moderately* (mod) active category have watched between 21 to 100 movies (81% of users), and *high* active users have watched more than 100 movies (9% of users).

RQ1: How often do users watch movies? An important question to consider for load estimation and content delivery systems is how much and how often the user consumes content. The typical user in our dataset watched 56 movies in 1 year, this equates to approximately 1.1 movies per week. For the users 1-year viewing history, we show the number of days between subsequent movie watches in Fig. 1 (outliers were removed). We find that 75% of the low, mod, and high

¹We were advised by the IRB that IRB approval was not required as no private or personally identifiable information was collected.

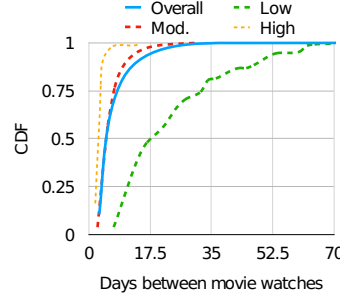


Fig. 1: Days between subsequent movie watches

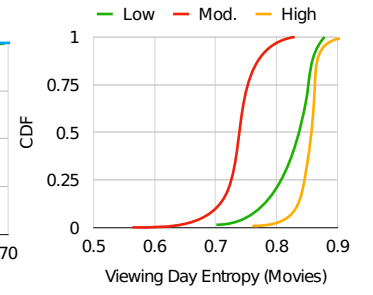


Fig. 2: Viewing day entropy for movies

active users watch movies every 33.2 days, 6.6 days and 3.5 days respectively.

RQ2: Do users watch movies on the same day(s)/week?

To quantify whether users tend to watch movies on the same day(s) each week, we define the *Viewing Day Entropy* (VDE) as given in Eq. 7.

$$VDE = \frac{-\sum_{d \in D} p_d \times \log(p_d)}{\log(N)} \quad (1)$$

where $p_d = \frac{\text{Number of movies watched on day } d}{\text{Total number of episodes watched by user}}$ (2)

and N is the total number of days in a week ($N=7$). The VDE is a value between 0 and 1, where a VDE closer to 0 indicates that the user has a more regular request pattern, and a value close to 1 indicates that the user uniformly watches content across the week. Interestingly, we find that less active and highly active users have a higher VDE than moderately active users. This implies that moderately active users tend to watch movies around the same day of the week as compared to other users.

RQ3: Are movies more often re-watched by active users?

Local caching attempts to speed the access to data by storing data that has recently been accessed by the client. A prerequisite for successful caching is the presence of redundancy in a user's behaviour. Here we analyze if and how often a user re-watches, either parts or the entire, movie. For every user, we compute the fraction of movies that appear in the user's viewing activity more than once (i.e. it was watched on more than 1 day). We find that for low active, moderately active, and high active users, approximately only 3.2%, 7.4% and 8.7% of movies, respectively, are watched more than once. Thus, we conclude that active users tend to re-watch more movies than less active users.

IV. ANALYSIS AND KEY INSIGHTS- SERIES

With 96% of the user's Netflix titles being episodes of series, we perform a larger and more in-depth analysis of users' series watching behavior. In the following subsections, we answer questions categorized into 5 groups to gain insights regarding the users' Netflix series viewing behavior. The groups are related to the user's watch patterns, watch-session length, user's preferences, the predictability of Netflix series videos and the continuation of watching series. In order to

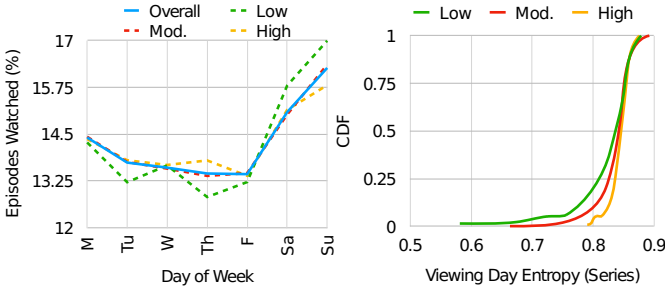


Fig. 3: Distribution of episodes watched per day

Fig. 4: Viewing day entropy across 1 year history

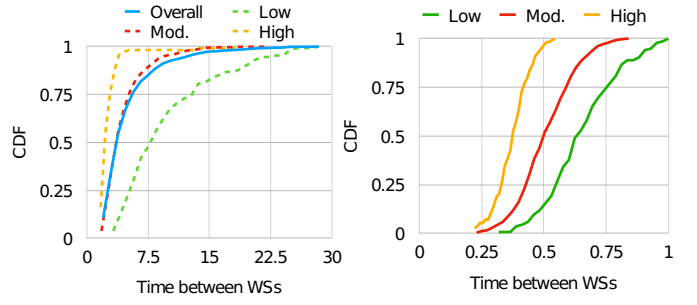


Fig. 5: CDF of time between WSs days

Fig. 6: TBWS entropy across 1 year history

analyze the user's behavior, and how their levels of activity impact our derived insights, we group our users into 3 categories based on the number of episodes they have consumed in their 1-year viewing history, namely, *low* active users that have watched less than 100 episodes, *moderate* (mod) active users who have watched between 101 and 800 episodes, and *high* active users that have watched more than 800 episodes. Approximately 13% of the users are in the *low* active category, 17% in the *high* active category, and the remaining 70% of the users are in the *moderate* active category.

A. User Watch Patterns

RQ1: Do users have a preferred day of viewing?

We explore whether users have a regular schedule in terms of when they view content; knowing what day a user is likely to access content is particularly helpful for load estimation and caching systems, and consequently can improve the user's Quality of Experience (QoE). We first show the distribution of content watched across the day of the week; this can be seen in Fig. 3. In general, the highest % of episodes watched occurred on a Sunday (16.3%), and the lowest on Friday (13.4%). In contrast, low active users watch their least content on Thursdays (12.8%). To quantify whether users tend to watch series content on the same day(s) each week, we compute the VDE as given in Eq. 7, where

$$p_d = \frac{\text{Number of episodes watched on day } d}{\text{Total number of episodes watched by user}} \quad (3)$$

The CDF of the VDE across users is shown in Fig. 4. We find that low active users are slightly more regular in terms of their day of viewing than high active users; however, 50% of all users, regardless of their activity level, have a VDE between 0.6 and 0.83, implying that in general users do not have a regular schedule in terms of their watch pattern.

RQ2: What is the number of days between subsequent watches?

A further important insight related to a user's watch patterns, is what the number of days between subsequent WSs, termed as *time between watch sessions* (TBWS), is. Fig. 5 shows the CDF of the TBWS days across the 1 year viewing history for all the users. We find that the TBWS days for 75% of the users is less than 6 days i.e. a typical user watches Netflix at least every 6 days. The TBWS days is nearly 3 days for 75% of the highly active users, and 13 days for low active users.

Similar to computing the VDE to see if a user's watch pattern follows a regular schedule, we also compute the entropy for the TBWS days. That is, we compute if the user tends to leave the same number of days between watching Netflix series content, regularly. The *TBWS entropy* (TBWSE) is computed as

$$TBWSE = \frac{-\sum_{i \in I} p_t \times \log(p_t)}{\log(N)} \quad (4)$$

where

$$p_t = \frac{\text{No. of instances when TBWS was } t}{\text{Total number of WSs} - 1} \quad (5)$$

and N is the total number of possible TBWS days ($N = 28$, as the maximum number of days between any 2 WSs was 28 days across in our dataset). Essentially, p_t is the probability that the days between 2 WSs for a specific user is t days. Fig. 6 shows the CDF for the TBWSE. We see that highly active users are more regular in terms of the days between subsequent WSs (a smaller TBWSE means a more regular behavior), whereas for the least active users, the TBWSE is closer to 1, implying that the user's watch pattern is sporadic. This is line with the findings from Fig. 4 where the highly active users have a larger VDE, indicating a smaller and more regular TBWS.

B. User Watch-Session Length

RQ3: How many episodes do users watch per day?

In effort of understanding user's viewing behavior as well as for the design of content delivery, caching and load estimation systems, it is crucial to know about how much content is consumed by a user. We show the CDF of the number of episodes watched in each WS across all our users' viewing history in Fig. 7. For 75% of the users in our dataset, at most 4.5 episodes are watched per day. For highly active users, 75% of the users watches at most 6 episodes per day and for the least active users, it is 3.5 episodes per day. The typical user in our dataset watches 2.7 episodes each day- using the runtime associated with watched episodes, this is equivalent to spending approximately 1.5 hours during each WS. This corresponds to 4.5 GB of a user's data when streaming in HD [7]. Furthermore, for a highly active user, the average user watch 5.3 episodes per day, spends 2.9 hours on Netflix series, and uses 9 GB (streaming at HD) of data each day.

RQ4: Do active users watch the same no. of episodes daily?

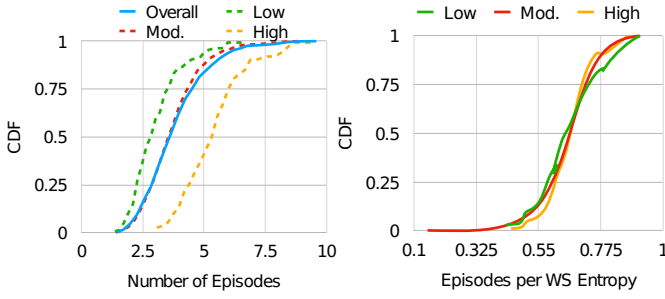


Fig. 7: CDF of No. of episodes watched per WS

Fig. 8: Episodes consumption entropy across 1 year history

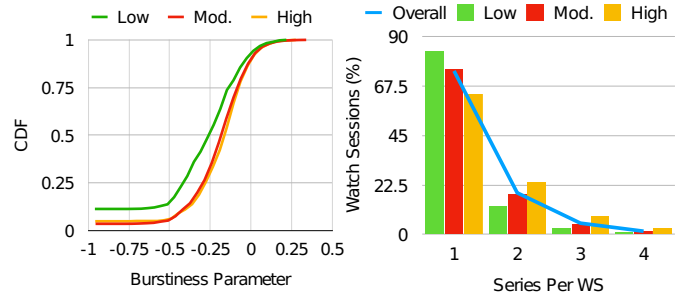


Fig. 9: CDF of burstiness on a per month basis

Fig. 10: Distribution of No. of series watcher per day

An important consideration for prefetching and caching systems, is being able to effectively predict how much content a user will see, usually based on their past behavior. It follows that users with uniform behaviour will be easier to predict for than users with inconsistent behavior. We observed that some users drastically increase or decrease the number of episodes they watch in 1 WS as compared to previous WSs. To quantify if the user tends to watch the same number of episodes during each WS, we compute the *episode consumption entropy* (ECE) as given in Eq. 6.

$$ECE = \frac{-\sum_{w \in W} p_e \times \log(p_e)}{\log(N)} \quad (6)$$

where $p_e = \frac{\text{No. of WSs when episodes watched was } e}{\text{Total number of WSs}}$ (7)

and N is the total number of possible episodes that the user can watch in a WS ($N=9$ as the maximum number of episodes watched by a user during a single WS). The CDF of the entropy is shown Fig. 8; here we find the entropy is very similar across users with different activity levels. We find that 75% of all the users in our dataset have a ECE of more than 0.5 which indicates that the users do not have a regular pattern in terms of the number of episodes consumed during each WS; we observe that users have a large variance in the number of episodes they watch in consecutive WSs.

This insight leads us to investigate the “burstiness” of the amount of content consumed during WSs; this parameter is computed as in Goh and Barabasi [8]. The Burstiness parameter is defined in equation 8 as,

$$B = \frac{\sigma_t - m_t}{\sigma_t + m_t} \quad (8)$$

where σ_t is the standard deviation and m_t is the mean of the user’s episodes per WS, over a period of t days. The parameter is a value between -1 and 1, where a value closer to 1 means that the standard deviation is larger than the mean, implying that the user’s behavior is bursty with regard to the number of episodes they consume in consecutive WSs. A value closer to -1 indicates the user watches almost the same number of episodes each WS. As an example, if user A watches the following number of episodes from Monday to Friday: [M=2, Tu=3, W=2, Th=2, F=3], then the burstiness parameter is -0.6; whereas if user B watches episodes as follows: [M=0, Tu=5, W=0, Th=10, F=0], then the burstiness parameter is 0.2. Fig.

9 shows the average monthly burstiness parameter ($t=30$) for the entire viewing history for the users. We find that the more active the user is, the more bursty their behavior is i.e. there is more variance in the number of episodes consumed per WS for active users.

RQ5: How many series does a user watch in a single day?

It can be argued that the number of episodes a user watches from a particular series will vary depending on what else the user is watching at that time. We explore this by determining the number of different series the user watches episodes from in a single sitting. Fig. 10 shows the distribution of the number of series watched across all WSs of the users. We find that, most of the time (nearly 68% of WSs), a user watches episodes from a single series in one sitting. In general, we see that less active users have a more concentrated viewing experience in that they only watch episodes from a single series in nearly 88% of their WSs, whereas, for highly active users, they only watch content from a single series for 65% of their WSs, and 23% of the time, they watch episodes from 2 series during the same sitting.

C. User Preferences

RQ6: Do active users watch the same genre(s) regularly?

This is an important question for recommendation engines and proactive caching systems, where a prediction of what to cache is made based on the user’s preferences. Understanding users’ preferences would also be useful for targeted advertising. There are 27 genres of Netflix series that are watched by the users in our dataset, and a series can be assigned multiple genres. A distribution of the episodes watched by all the users in our dataset, and the genres of the associated series, is shown in Fig. 11. We have shown the % of episodes watched belonging to the top 12 genres that make up 98% of all series’ genres consumed; the remaining 15 genres are included in “other”. As seen in the figure, the largest % of episodes watched (nearly 27%) are of the “drama” genre; this is the largest for all levels of user activity. Furthermore, we see that regardless of user activity level, the distribution of episode genres is very similar.

This, however, does not tell us if users in different activity levels have a concentrated preference in terms of the genre of content (i.e. they tend to watch content only from 1 or 2 genres) or a more diverse genre preference (i.e. they watch

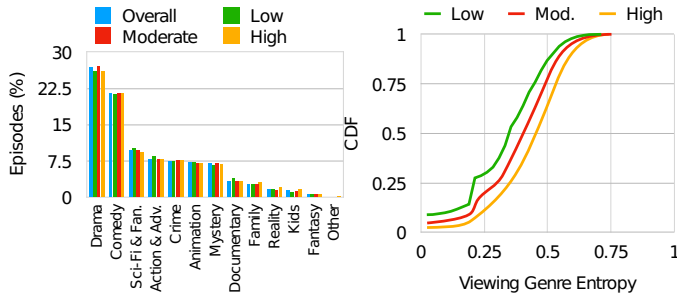


Fig. 11: Episodes distribution per genre Fig. 12: Monthly genre entropy for all users

content from multiple genres). To quantify this, we compute the user's *viewing genre entropy* (VGE) as given in Eq. 9.

$$VGE = \frac{-\sum_{g \in G} p_g \times \log(p_g)}{\log(N)} \quad (9)$$

$$p_g = \frac{\text{Number of episodes in genre } g}{\text{Total number of episodes watched by user}} \quad (10)$$

and N is the total number of genres ($N=27$). The VGE is a number between 0 and 1; a value closer to 0 means that the user has more stability in terms of their preference (they prefer content from a few genres only), whereas a larger VGE means that the user watches content from various genres. We computed the VGE for each month of the user's viewing history, and obtained the average across all the months; the results are shown in Fig. 12. We find that the more active the user is, the higher the VGE and thus, the more diverse the preferred genres (75% of the users in low, moderate and high activity levels have a VGE of less than 0.45, 0.49, and 0.54 respectively).

RQ7: Do active users prefer more popular and higher rated content?

Gaining insight into how the popularity and ratings of content affect the consumption for different activity levels, is helpful for caching and content delivery. For each series watched by users in our dataset, we obtained the number of IMDB votes the series had at the time of retrieval. IMDB is an extensive online database of information related to movies, TV series and streaming content- including rating and reviews that are given by registered IMDB users. A rating that a series has received by IMDB registered users is counted as a vote; thus the number of votes a series received can serve as an indication of how popular that series is. Fig. 13 shows the distribution of the votes that users' watched series have; we find that 33% of series that highly active users watch has between 20,000 and 30,000 votes, whereas 29% of low active users' series fall in this range. We see that for votes higher than 30,000; users in low active categories watch the largest % of series (39%) as compared to moderately and highly active users (32% for both). The average number of votes for series watched by users in low, moderate and high categories are 30786, 29985, 28734 respectively. This implies that less active users tend to watch slightly more popular content than more active users.

Fig. 14 shows the distribution of the ratings (a score out of 10 given by registered IMDB users) of series watched by

users. Here we see that less active users prefer content with higher ratings than more active users; 35% of series watched by low active users have a rating of above 8, whereas 30% of highly active user have a rating of above 8. In conclusion, we find that less active users, even though watch less content, prefer more popular and higher rated content than more active users.

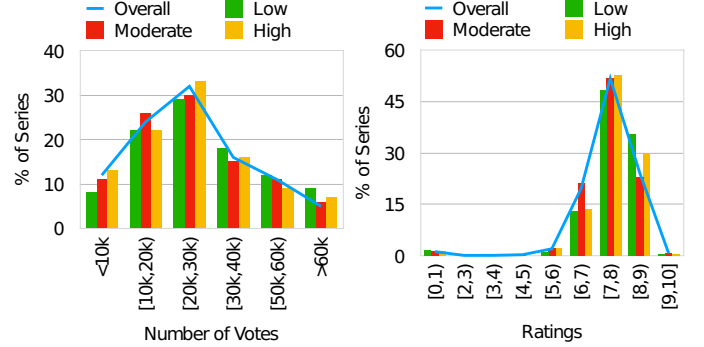


Fig. 13: Distribution of number of votes received

Fig. 14: Distribution of ratings received

RQ8: How much of user's watched series are related to series they have seen in the past?

Recommendation engines predominantly recommend content that is related to what the user has watched in the past. Although we are unable to retrieve the content that is recommended to the user when they are watching content on Netflix, we obtain an approximation of the effectiveness of the engine by computing the fraction of series watches that are related to series that the user has previously seen. This analysis can further aid in the prediction of what content the user will watch. During the meta-data retrieval process, we obtained the 12 related series as listed by IMDB; using this information, for every series that a user has watched, we see if this series is related to any series watched previous to this. We find that approximately 42% of a series watched by a user, was related to a previously watched series. Furthermore, we find that this percentage is similar for users across activity levels; with low, moderate and high active users, watching 41.4%, 42.3% and 40.1% of series that was related to a series they had seen before.

D. Predictability

RQ9: How much of user's future watches are predictable?

Predicting what, and how much, a user will watch next, is crucial for prefetching and caching strategies. These strategies anticipate the content a user is likely to consume, downloads the content ahead of time, and makes the content available at the time of consumption. To see whether we can predict what the user will watch next based on what they have consumed in the past WSs, we do the following: for every WS that appeared in a user's viewing history, and for each episode watched in that session, we check if its preceding episode was watched within a certain number of previous WSs. For example, if episode 20 of series A was watched today, we check if and

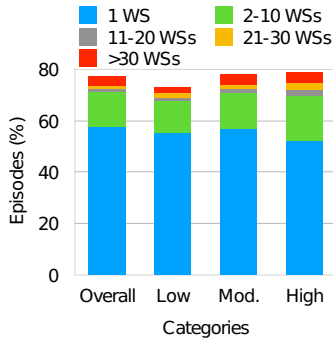


Fig. 15: Predictability from past WSs

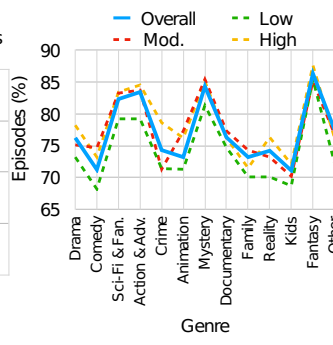


Fig. 16: Predictability across different genres

how many WSs prior, episode 19 was watched. We compute this for all users in our dataset across their entire history, the average is shown in Fig. 15 for various WS intervals.

For the average user in our dataset, we see that nearly 58% of episodes proceed an episode that was watched in the previous WS (1 WS), a further 13% of episodes proceeded an episode that was watched between the previous 2 and 10 WSs, 1% between 11 and 20 WSs ago, 1% between 21 and 30 WSs prior, and 3% was watched more than 30 WSs prior. In general, we find that approximately 77% of the user's episode watches follows an episode that the user has seen in the past. We also find that as the activity level of the user increases, the larger the predictable % of episodes. Thus, we conclude that nearly 77% of a user's future episode watches can be predicted as it proceeds a previously watched episode from the series.

RQ10: Are certain genres easier to predict for than others?

Fig. 16 shows the % of episodes that is predictable for different genres. We consider all WSs in the user's viewing activity for this analysis. We find that nearly 85% of episodes from "Fantasy" series follows a previously watched episode; this is the highest for any genre. We find that the "comedy" genre and "kids" genre has the least % of episodes that are predictable (71.4% and 71.1% respectively), this is the same for low and high active users as well. We speculate that these differences arise due to the episodic (such as for "Fantasy" series) vs non-episodic (such as "kid" shows) nature of series. This insight can further aid prediction and prefetching systems to determine if and how many episodes from a particular series, the user will watch in the near-future.

E. Continuity of User Watch-Behavior

RQ11: How many seasons does a user watch to its entirety?

An effective way of gauging a user's interest and engagement in a particular series, which will be helpful for content creators, marketing agencies and content providers, is to see if they watch a series season to its entirety. Fig. 17 shows that % of seasons users watch to its completion across various series genres. Overall, nearly 55% of series seasons are watched entirely, with series seasons in the "Animation" genre watched to its entirety the most as compared to other genres (60%). We find this to be similar across low and high active users.

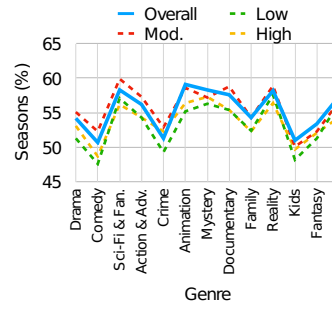


Fig. 17: Seasons watched to its entirety for different genres

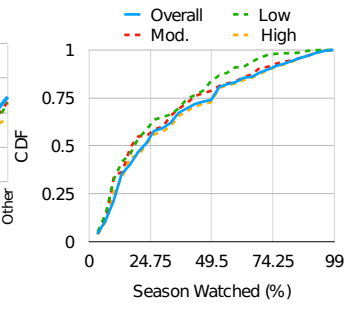


Fig. 18: Point of departure of seasons not watched to its entirety

RQ12: At what point does a user stop watching a series season if they don't complete it?

Interestingly, we found that a large percentage of seasons, nearly 45%, are abandoned at some point, and not watched to completion. For the series seasons that are not watched to its entirety, we explore the point at which a user stops watching a season (we only consider seasons of episodes that are watched contiguously). Fig. 18 shows the CDF of how much a season a user has watched before abandoning it- we term this as the "point of departure". We see that 50% of seasons are abandoned when less than 25% of the season is watched; this is consistent across users of all activity levels. The remaining 50% of the seasons has a point of departure from 25% to 99%, and this is nearly uniformly distributed.

RQ13: Can we predict when a user will abandon a series?

Given that nearly 45% of series seasons are not watched to completion, this leads us to investigate if we can predict the time at which the user will stop watching a series- this could be due to a variety of reasons, but particularly a waning interest in continuing the season. To this end, we employ 4 popular machine learning classification models to answer the following question: *For the latest episode of a series watched in a particular WS, will the user watch proceeding episodes in subsequent WSs?* The models we employ are as follows: Binary Logistic Regression (LR), Support Vector Machine (SVM), Naive Bayes model (NB) and Random Forest (RF). The models use the following features for prediction: % season watched, number of votes the season's series has, the IMDB rating the season's series has, series genre, episode runtime, year of release and number of seasons. In essence, for every series watched in a particular user's WS, we obtain the latest episode watched from that series, extract the appropriate features of the episode's series, feed this into the trained classification model and obtain one of two possible outputs: 1) "continue"- the model predicts that the user will continue watching the seasons, 2) "abandon"- the model predicts that the user will stop watching the series.

Table II shows the results of the classification model using the following performance metrics: the accuracy, the precision, the recall, the and the AUC value. The descriptions of the classification metrics can be found in [9]. We train the models on the first 9 months of the user's data, and perform the testing

TABLE II: Classifier Comparison

Method	Accuracy	Precision	Recall	AUC
LR	66.2	0.57	0.69	0.61
SVM	59.2	0.53	0.59	0.62
NB	65.1	0.56	0.67	0.63
RR	68.1	0.61	0.73	0.69

on the remaining 3 months. To ensure a balanced dataset i.e. approximately the same number of “abandon” instances as there are “continue” instances, we perform under-sampling of the “continue” class during training. We find that we are able to achieve the highest prediction accuracy with the RF model—we are able to correctly predict 68% of the instances of when the user either abandons or continues watching a seasons. In general, we find that the classifiers perform similarly in terms of their classification.

V. RELATED WORKS

There have been several measurement studies performed for understanding video traffic. With YouTube contributing to 15% of the world’s internet traffic [10], there has been a plethora of measurement studies in which YouTube video popularity and YouTube video request patterns were investigated [11]–[13]. Given the short-form nature of YouTube videos, the findings cannot be effectively extrapolated to long-form videos, like Netflix episodes. There has been work towards understand the characteristics of Netflix traffic as performed by Rao *et al.* [14], in which the strategies that Netflix employs to stream their video traffic is performed, and by Adhikari *et al.* [15], in which the authors perform a measurement study to understand Netflix architecture. These studies, however, are agnostic to users’ individual behavior, and provide a macro-view of Netflix traffic. Huang *et al.* [16] performed an analysis of streaming user behavior on a large-scale VoD platform in China. The authors studied user’s request patterns and their viewing interests for TV shows watched. In contrast, we not only study user’s preferences for certain TV series, we also study how the user interacts with specific series at an episodic level. Our work fundamentally differs with previous works in that we are able to present an in-depth, long-term study of how user’s interact with Netflix and derive key insights.

VI. CONCLUSION

In this paper, we collected and analyzed a real-world Netflix dataset which consisted of 1-year viewing activity from 1060 users amounting to over 1.7 million watched episodes and movies. Equipped with this data, we derived key insights pertaining to the user’s watch patterns, watch-session length, user preferences, predictability and series continuation tendencies. We also implemented and evaluated prediction models that is used to predict if a user will continue watching a series or not. We found that we were able to achieve an overall accuracy of 68% with the Random Forest classifier. Given the dominance of Netflix on Internet traffic, the results and analysis serves

to contribute to not only the research community, but to network operators, content providers, marketing agencies, content creators and users themselves.

REFERENCES

- [1] (2017) Cisco vni complete forecast highlights. [Online]. Available: https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecast-highlights/pdf/Global_2022_Forecast_Highlights.pdf
- [2] (2020) The global internet phenomena report covid-19 spotlight. [Online]. Available: <https://www.sandvine.com/phenomena>
- [3] (2020) User behavior analytics. [Online]. Available: <https://research.netflix.com/research-area/analytics>
- [4] (2019) Amazon mechanical turk. [Online]. Available: <https://www.mturk.com>
- [5] F. R. Bentley, N. Daskalova, and B. White, “Comparing the reliability of amazon mechanical turk and survey monkey to traditional market research surveys,” in *CHI 2017*, ser. CHI EA ’17. New York, NY, USA: ACM, 2017, pp. 1092–1099.
- [6] (2019) Tmdb api. [Online]. Available: <https://developers.themoviedb.org/3/getting-started/introduction/>
- [7] (2020) Netflix help center. [Online]. Available: <https://help.netflix.com/en/node/87>
- [8] K.-I. Goh and A.-L. Barabási, “Burstiness and memory in complex systems,” *EPL (Europhysics Letters)*, vol. 81, no. 4, p. 48002, jan 2008.
- [9] G. Bonaccorso, *Machine Learning Algorithms: A Reference Guide to Popular Algorithms for Data Science and Machine Learning*. Packt Publishing, 2017.
- [10] (2019) 2019 global internet phenomena. [Online]. Available: <https://www.sandvine.com/press-releases/sandvine-releases-2019-global-internet-phenomena-report>
- [11] V. K. Adhikari, S. Jain, Y. Chen, and Z. Zhang, “Vivisectioning youtube: An active measurement study,” in *2012 Proceedings IEEE INFOCOM*, March 2012, pp. 2521–2525.
- [12] M. Zink, K. Suh, Y. Gu, and J. Kurose, “Characteristics of youtube network traffic at a campus network – measurements, models, and implications,” *Computer Networks*, vol. 53, no. 4, pp. 501 – 514, 2009, content Distribution Infrastructures for Community Networks. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1389128608003423>
- [13] S. Lall, M. Agarwal, and R. Sivakumar, “A youtube dataset with user-level usage data: Baseline characteristics and key insights,” in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, 2020, pp. 1–7.
- [14] A. Rao, A. Legout, Y.-s. Lim, D. Towsley, C. Barakat, and W. Dabbous, “Network characteristics of video streaming traffic,” in *Proceedings of the Seventh Conference on Emerging Networking EXperiments and Technologies*, ser. CoNEXT ’11. New York, NY, USA: ACM, 2011, pp. 25:1–25:12. [Online]. Available: <http://doi.acm.org/10.1145/2079296.2079321>
- [15] V. K. Adhikari, Yang Guo, Fang Hao, M. Varvello, V. Hilt, M. Steiner, and Z. Zhang, “Unreeling netflix: Understanding and improving multi-cdn movie delivery,” in *2012 Proceedings IEEE INFOCOM*, 2012, pp. 1620–1628.
- [16] L. Huang, B. Ding, Y. Xu, and Y. Zhou, “Analysis of user behavior in a large-scale vod system,” in *Proceedings of the 27th Workshop on Network and Operating Systems Support for Digital Audio and Video*, ser. NOSSDAV’17. New York, NY, USA: Association for Computing Machinery, 2017, p. 49–54. [Online]. Available: <https://doi.org/10.1145/3083165.3083171>