A YouTube Dataset with User-level Usage Data: Baseline Characteristics and Key Insights

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

***Abstract*—YouTube is the most popular video sharing platform with more than 2 billion active users and 1 billion hours of video content watched daily. The dominance of YouTube has had a big impact on the performance of Internet protocols, algorithms, and systems. Understanding the interaction of users with YouTube is thus of much interest to the research community. In this context, we collect YouTube watch history data from 243 users spanning a 1.5 year period. The dataset comprises of a total of 1.8 million videos. We use the dataset to analyze and present key insights about user-level usage behavior. We also show that our analysis can be used by researchers to tackle a myriad of problems in the general domains of networking and communication. We present baseline characteristics and also substantiated directions to solve a few representative problems related to local caching techniques, prefetching strategies, the performance of YouTube's recommendation engine, the variability of user's video preferences and application specific load provisioning.**

I. Introduction

YouTube is the world's largest video sharing site with more than 2 billion active users [1]. YouTube videos reportedly account for 38% of a mobile user's cellular data usage [2]. This represents the largest share of the cellular bandwidth usage among all applications on the mobile device. Given the prominence of YouTube in terms of share of wireless resources consumed, it is of much interest to understand the characteristics of YouTube usage that could be of use to researchers. However, beyond the macro-level statistics that YouTube publishes [1], there has been very little work done toward collecting any non-trivial data performing any meaningful analysis on such usage.

This forms the context for this work. We use Amazon's Mechanical Turk (mTurk) platform to collect a dataset for YouTube usage from 243 users. The dataset comprises of $1,826,075$ videos spanning a 1.5 year period of watch history. The videos, as an aggregate, represent approximately 65TB of videos watches over a 1.5 year period. We believe that the dataset will be of significant use for researchers working in a wide range of problems in the general area of Internet protocols, algorithms, and systems. We perform a baseline analysis of the dataset to identify some interesting standalone nuggets of information such as the average number of videos watched by a user per day, how long a typical video lasts, the typical number of categories videos are watched from by a user, the average number of playlists created by a user, and the typical number of channels a user subscribes to.

While we believe that the real value of the dataset lies in other researchers using it for their respective problems, the core contribution of this work includes considering a few representative problems in the domains of networking and communications, and analyzing the dataset to answer key questions pertaining to those problems. Note that the goal of this work is not to solve the problems, but instead provide substantiated directions for solutions *based on insights from the dataset*. Specifically, we consider the following sets of questions:

1) How often do users watch the same video again? If they do see certain videos again, how far apart are the redundant views? Are there any patterns in which videos are likely to be watched again?
2) How much of a user's watch behavior can be predicted? How much of a user's past watch behavior has to be considered to maximize the predictive accuracy while considering the associated costs?
3) How much of a user's watch behavior is influenced by recommendations? Are there certain categories of videos for which the users are more likely to be influenced by recommendations? Are there other attributes of a video (e.g. length, number of likes, etc.) that also influence recommended watch outcomes?
4) How static are a user's video preferences over time? Do they remain static over 1.5 years, or do they change drastically?
5) What are the typical data consumption patterns for YouTube usage for a user? Does this change based on time of day or day of week? How consistent or bursty is the usage?

For each of these sets of questions, we delve into the collected dataset, extract insights, and provide a summary analysis. The rest of the paper is organized as follows: In Section II, we discuss the significance of YouTube as the application of interest, the data collection methodology, and present a baseline analysis. In Section III, we present the analysis pertaining to the aforementioned sets of questions. In Section IV, we discuss some related works, and in Section V we present conclusions of the paper.

II. YouTube and Data Collection

A. YouTube and its significance

YouTube is a content community that was founded in 2005 which allows users to post, view, comment and share videos on the site. It is the most visited website in the world, with

978-1-7281-5089-5/20/$31.00 ©2020 IEEE

just over 2 billion monthly visitors and more than 300 hours of content uploaded every minute [3]. It consumes nearly 12% of global network traffic share (following Netflix and HTTP media streaming), and benefits from being the most commonly embedded video on other sites, including Facebook [4]. YouTube content currently dominates mobile data traffic, and is reported to account for 38% of all mobile traffic [2]. Furthermore, YouTube's data traffic usage is the highest among all other mobile apps. As reported by Cisco, the average mobile YouTube data traffic consumed per smartphone per month is 2.3 GB, and the average usage for PC/tablets is 3.3 GB per month [5].

B. Dataset Collection

To collect the dataset, we rely on mTurk to gather anonymized watch-history from the users [6]. The mTurk platform allows a task to be posted for a fee, which in turn can be completed by users known as mTurkers. Previous studies have shown that mTurk samples can be accurate when studying technology use in the broader population [7]. The task we posted required mTurkers to navigate to Google's *Takeout* page and download their YouTube related data. The mTurker would then extract the archive file and select the files related to their watch-history, playlists and subscriptions data; these files were then anonymously uploaded via a dropbox link (we were advised by the IRB that IRB approval was not required as no private or personally identifiable information was collected).

The archived file that was uploaded contained the following files: *watch-history.html*, a JSON file for each playlist created by the user, and *subscriptions.json*. The *watch-history.html* file contains a list of all video titles, where the title of the video is a hyperlink to the video URL, viewed by the mTurker, and the associated time it was viewed. The JSON file for each user-created playlist contains a list of the video IDs for all videos added to that playlist. Similarly, the *subscriptions.json* file contains a list of all channels the user is subscribed to.

C. Baseline Characteristics

A high level overview of the statistics of the per-user watch-history data is presented in Table I. In the collected dataset, there are $1,826,075$ videos watched by 243 users. Each video is categorized by the uploader according to 18 predefined categories and added to a particular channel; users can subscribe to the channel (known as subscriptions) and add the video to user-created playlists. The videos watched per user per day (videos/day), the number of categories the user has watched videos from (categories), the number of playlists the user has created (playlists), and the number of channels the user has subscribed to (subscriptions), is shown in the table.

The total number of *unique* videos watched by the users is $1,172,111$ videos. Using YouTube's data API, we obtained meta-data associated with each video in our dataset regarding its video duration, the number of views, the number of likes, the number of dislikes and the number of comments each video has at the time of data collection. Table II summarizes the metrics associated with the videos watched by the users.

TABLE I: User statistics

Attribute	Mean	Std Dev.	Min	Max
Videos/day	15.01	6.24	0	48
Categories	4.2	0.7	3	13
Playlists	1.4	5.8	0	24
Subscriptions	10.9	12.8	0	57

TABLE II: Videos statistics

Attribute	Mean	Std Dev.	Min	Max
Duration (min.)	13.2	30.1	0.02	820
Views ($\times 10^6$)	3.2	26.9	3	560
Likes ($\times 10^3$)	20.6	124.3	0	30,079
Dislikes ($\times 10^3$)	1.4	16.9	0	9,518
Comments ($\times 10^3$)	1.9	13.2	0	52,639

III. Analysis and Key Insights

In the following subsections, we attempt to answer 5 questions to gain insights regarding our users' YouTube watching behavior. Based on the insights, we also present implications and the feasibility of the applications and development of associated technologies. We present results on a per-user basis and also study the effect of various attributes. In particular, we are concerned with the following attributes, namely: 1) video related attributes- the category that a video belongs to, the duration of the video, the number of views the video has; and 2) user related attributes- the hour of day that the video was watched by a particular user, and the day of week that the video was watched by the user. For each of the aforementioned video related attributes, the percentage of videos, from our videos dataset, belonging to each category, video duration window and view count range is computed and is shown as the *overall distribution* in relevant results that follow. Similarly, for user related attributes, we show the overall distribution of videos watched in each hour of the day and day of the week, across all users.

A. How often do users watch the same video again?

Local caching attempts to speed the access to data by storing data that has recently been accessed by the client. Caching plays a vital role for web traffic and can effectively decrease network traffic volume, lessen server workload, and reduce the latency perceived by end users [8]. A fundamental prerequisite for successful caching is the presence of redundancy in a user's behaviour i.e. do users watch the same video again? We seek to capture this redundancy by analyzing how often, as well as when and what types of videos, are watched again by a particular user. We also present the feasibility of local caching based on our findings.

Methodology and Metrics: To explore this aspect, for each user in the dataset, we compute the percentage of videos from their watch-history that are watched more than once by the user. Furthermore, we see whether a video that was watched again belonged to a channel that the user subscribed to or appeared in any of their playlists. We also compute the time difference between subsequent watches. It is also beneficial to understand the characteristics of the videos that are watched again; to this end, for each category, duration window, and views range, we calculate the percentage of videos that are watched again for that parameter value. It is important to

mention that for our analysis, a video is considered to be watched again only if the video content is retrieved from YouTube servers and not stored on their device.

Analysis and Discussion: Fig. 1 shows the percentage of videos that are watched again by each user, arranged in ascending order. The average percentage of videos that are watched again is 10.9%, and ranges from 0.8% to 33.7%, with a standard deviation of 6.4% and median of 9.2%. With respect to their subscriptions and playlists, we found that 8.4% of user's repeated views are from channels that the user has subscribed to, and 4.3% are from their playlists. We also computed the average time difference between each repeated watch of a video on a per user basis. We found that the average difference between such watches across all users, is approximately 2.8 months, and ranges from 2.7 days to approximately a year, with a standard deviation of 2 months and a median of 2.3 months.

To understand how the video related attributes impact the repeatability of video watches, we further look into this redundancy expressed as a function of video category, duration window and views count. In Fig. 2, we see that videos belonging to the "movies" category are the most likely to be watched again (with nearly 10% of all "movies" being watched again), as compared to other categories; we should bear in mind that the fraction of videos in our dataset belonging to the movies category is less than 1%. Following the "movies" category, the videos that are categorized as "shows", "comedy", and "music" are more likely to be watched again. We also analyze whether the video duration affects if a video is watched again or not; this is shown in Fig. 3. We see that video duration does not have a large impact on the repeatability, however, the repeated video watches percentage generally increases as the duration increases. Fig. 4 shows the percentage of repeated watches of videos with various view counts. We see that in general, videos that have a higher view count are more likely to be watched again.

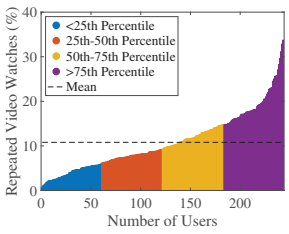

Fig. 1: Repeated video watches per user

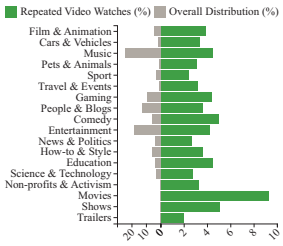

Fig. 2: Repeated watches per category

Key Insights: With the use of local caching, even a minute reduction of YouTube traffic volume can lead to savings of tens of millions of dollars for carriers which operate under severe resource constraints [9]. There are also several benefits from the user's perspective; two notably being an improved quality of experience, and reduction in costs associated with network data transfers. Typically, YouTube content is not locally cached beyond caching only video chunks as stipulated by YouTube's

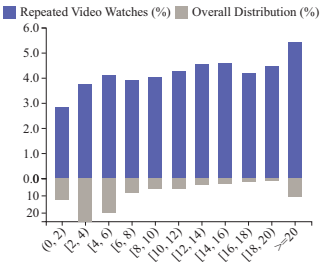

Fig. 3: Repeated watches for varying duration

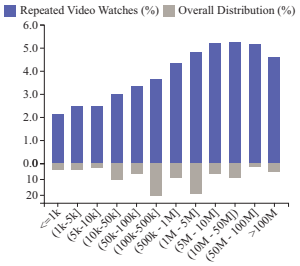

Fig. 4: Repeated watches for changing view counts

employment of the MPEG-DASH protocol [10]. From the results presented in this section, we see that there is a scope for local caching of video content with nearly 11% of videos watched again by the user; algorithms to determine what videos and for how long they should be kept in the cache, can be developed.

> *With approximately 11% of videos being watched again by a user, local caching can yield an acceptable hit rate; however, the cached content will need to be stored for 2.8 months on average.*

B. Can a user's YouTube watch behaviour be predicted?

Being able to predict what a user will watch in the future is particularly useful for prefetching strategies. Prefetching content has extensively been used to reduce user-perceived latency when loading web pages across the internet [11], [12]. These strategies anticipate the content a user is likely to consume, download the content ahead of time, and make the content available at the time of consumption. To explore the feasibility of prefetching, we consider how a user's YouTube watch behaviour is influenced by videos they have seen in the past. Specifically, we see whether videos that are related to videos that has been seen by a user in the past, is consumed by the user in the future.

Methodology and Metrics: YouTube algorithmically determines videos that are related to one another using the video's meta-data, and also by employing collaborative filtering methods. We use YouTube API's *relatedToVideoId* endpoint to retrieve a list of videos which is related to a particular video. For a particular user, we fetch 50 related videos of every video that has been watched by the user, and then see if any of the related videos were watched later; we term this set as the "related set". We perform this analysis for all the users in our collected dataset for their entire watch-history, and present the per-user results, as well as the results pertaining to several video related attributes.

Analysis and Discussion: Fig. 5 shows the percentage of videos that are found in the related set of videos they have seen in the past. We find that the average percentage is 59.1%, and ranges from 36.5% to 91.6%, with a standard deviation of 16.2% and median of 58.6%. In addition, 9.6% of these videos are from channels the user has subscribed to, while 2.9% appears in their playlists. In addition, we also study the time

difference between when a video was consumed, and when a video related to it, was watched in the future. The average time difference between such watches is 25.9 days, and ranges from 6.4 days to 45.7 days, with a standard deviation of 7.3 days and median of 25.9 days.

The video related attributes we investigate are the category, duration and view count. In Fig. 6, we compute the percentage of videos watched in each category that was in the related set of a video that a user watched in the past. We observe that related videos belonging to the "entertainment" category are more likely to be watched as compared to any other category. In Fig. 7, we similarly perform the analysis for videos of varying duration; here we see that related videos that are between 4-6 minutes long are most likely to be watched; the least likely are videos from 0 to 2 minutes. Fig. 8 shows that, in general, related videos with a higher view count are watched more.

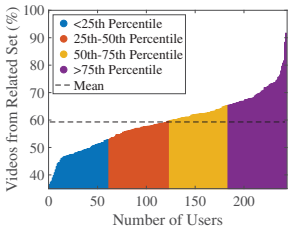

Fig. 5: % Videos from related set per user

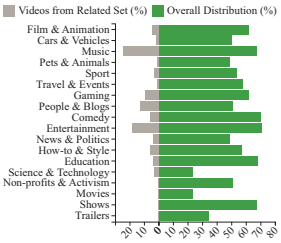

Fig. 6: Related video watches per category

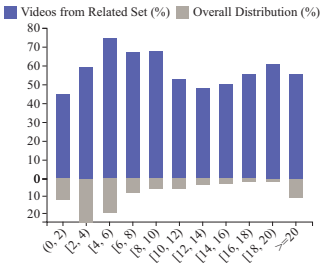

Fig. 7: Related watches for varying duration

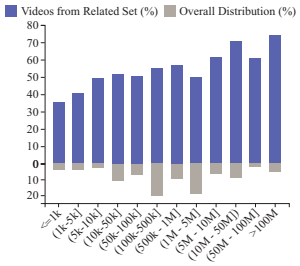

Fig. 8: Related watches for changing view counts

Key Insights: The motivation for prefetching videos stems from one of two reasons: 1) to reduce network usage during peak times, and 2) to enable high video viewing QoE by prefetching content to avoid unstable network connections. The results presented in this section show that YouTube content is indeed predictable, across categories and especially for more popular videos. Hence, there is a potential for developing successful prefetching systems which can be used to fetch content during low-cost periods (such as over WiFi or off-peak periods).

With 59% of videos watched by a user being present in the related set of videos that the user has previously watched, YouTube watch behaviour is predictable and can be used in the development of effective prefetching systems.

C. Do users consume videos suggested through YouTube's recommendation engine?

YouTube's recommendation engine (RE) uses sophisticated algorithms to understand user preferences and suggest videos that the user is likely to watch. The understanding of how video views are driven through the RE is beneficial for not only the research community (in serving as a case study of how video content is discovered), but also advertisers and content providers [13]. Using our collected dataset, we attempt to independently quantify the effectiveness of YouTube's RE. Furthermore, we provide insights about the types of videos that are better referrers.

Methodology and Metrics: The recommended videos are based on the user's past watch-history and videos identified as "related videos" (as discussed in the previous section) through collaborative filtering and other association algorithms. Specifically, YouTube's RE consists of two neural networks: the candidate generation network and the ranking network [14]. For each video watched by a user, there are recommended videos shown alongside; we term this as the "recommended set". Due to the RE being dependent on the user's live actions and the prioritization of fresh content, there is no simple approach to obtain the recommended set for a user.

In order to approximate the RE's behavior for a user, we create a test YouTube account (account that had no prior watch-history) and programmatically re-played the user's watch-history for 1 year (for the full video length). Even though the recommended videos will not be an exact match of the videos shown to the user at their time of viewing, we see that across 10 randomly selected users, a relatively large fraction (67%) of their future video watches have previously appeared as a recommendation. We also compute the percentage of videos watched from their related video sets, and find that on average, 63% of their future video views appear in the related video sets; this is only 4% lower than their recommended video sets. Due to the complexity and computational inefficiency associated with emulating the RE, we use the related videos as a close proxy for the recommendation videos set.

We obtain the RE effectiveness by computing the percentage of videos which appeared in the related videos set of the video previously watched; this provides an indication of whether the user clicked on a video that appeared in the recommendation list of a video they were currently watching. We also study how the recommendation system performs across video categories, duration, number of views and also, the hour of day the video was consumed.

Analysis and Discussion: The RE effectiveness per user is shown in Fig. 9; the average effectiveness among all users is 21.4%, and ranges from 3.2% to 47.7%, with a standard deviation of 7.5% and median of 21.2%. 8.1% of the recommended videos were watched from the user's subscribed channels, and 2.2% from their user-created playlist. Fig. 10 shows the RE effectiveness for different video categories; we see that the RE for the "shows" category is found to be the most effective where the recommended videos from just over

35% of videos watched from this category, is watched by users. Fig. 11 shows the effectiveness as a function of view count; in general, recommended content from more popular referrer videos are likely to be be seen. Fig. 12 shows how the time of day a particular video was watched affects if a recommended video was watched; we find that the recommendation system tends to be more effective from 5am to 2pm. This corresponds to a decrease in the overall distribution of video traffic. Furthermore, we compute the effectiveness for different video duration, and find that this has no significant impact on the performance of the RE.

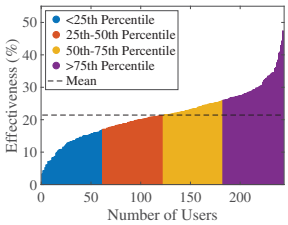

Fig. 9: RE effectiveness per user

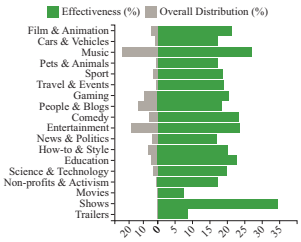

Fig. 10: RE effectiveness per category

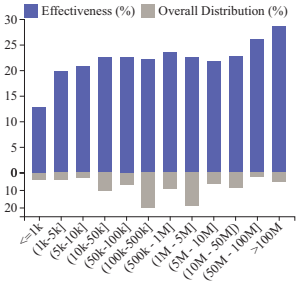

Fig. 11: RE effectiveness for changing view counts

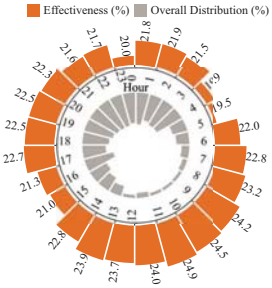

Fig. 12: RE effectiveness per hour of day

Key Insights: The RE plays a vital role in attracting and retaining users, and also increasing video popularity. Advertisers and content providers will be able to plan their strategies to increase visibility and predict their effectiveness by understanding how and when video's recommended content is consumed. We find that we can anticipate that a user will watch a recommended video over one fifth of the time, even though it is not watched within the same watch session. We also find that there are certain videos that serve as better referrers than others (e.g. more popular videos in "shows" category), and that the RE is more effective during off-peak periods.

For a video currently being watched by a user, approximately 21.4% of videos watched next, appear as a recommendation. Furthermore, the more popular a referrer video is, the more likely a recommended video will be consumed thereafter.

D. Do user's YouTube video preferences change over time?

The immense prevalence and widespread consumption of YouTube has influenced advertisers to design their strategies incorporating this platform. Advertising revenue on YouTube is estimated to be up to $4.5 billion [15]. User preferences and how this evolves would thus be of interest to advertisers for targeting and personalizing adverts. To gauge how dynamic user's preferences are, we explore how video duration, channel and category preferences change with time.

Methodology and Metrics: We study how the duration of a video influences a user's preferences and whether this changes depending on when they watch the videos. Furthermore, we analyze the user's category and channel preferences, and how this changes over time. The preference strength is proportional to the volume of video content consumed i.e. the more video content that is consumed from a particular channel or category, the more preferred that channel or category is.

Analysis and Discussion: Fig. 13 shows the average duration of videos watched across all users; the average per video duration across all users is 12.8 minutes, and ranges from 9.9 minutes to 13.8 minutes, with a standard deviation of 0.7 minutes and median of 13 minutes. Performing this analysis on a per month basis, we see that the average duration differs by only 0.8 minutes from month-to-month; this is equivalent to a 6.3% change across 1.5 years. Fig. 14 and Fig. 15 shows how the average duration of videos watched by users differ for the time of the day they are watching it, and the day of week the video is watched. We find that during off-peak periods, the average video length is approximately 2.5 minutes longer per video than during peak periods. Over weekends (Friday to Sunday), the average video duration is only slightly higher than during the rest of the week.

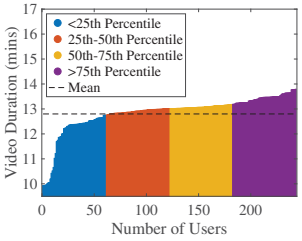

Fig. 13: Video duration preference per user

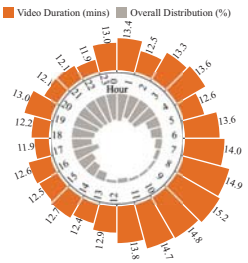

Fig. 14: Video duration preference per hour

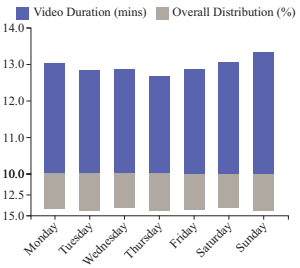

Fig. 15: Video duration preference per day

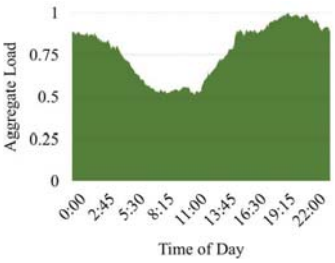

Fig. 16: Aggregate load across all users

We also evaluate the dynamic nature of user's preferences by studying how their channel and category preferences

change over time. We find that users are actually fairly static with 95% of all their watched videos, belonging to their 3 most preferred categories. Similarly, we find that 38% of all video watches are from user's 10 most preferred channels, while 63% are from their 30 most preferred channels. When we perform this analysis on a month-to-month basis for each user, and compute the percentage change of videos watched from their 3 most preferred categories and 30 most preferred channels of the previous month, we find that their consumed video content changes by 4.6% and 32.4% per month for category and channel preferences, respectively.

Key Insights: Learning about user preferences makes it possible to model user information needs and adapt services to meet these needs. Our results suggest that users tend to watch videos between 12 to 13 minutes of length. We also see that user preferences related to the types of videos they watch (characterized by their category and channels) does not vary significantly across time, and so there is potential for time-invariant personalized advertising.

User preferences in terms of the video duration, their 3 most preferred categories, and 30 most preferred channels change by 6.3%, 4.6% and 32.4%, respectively over their 1.5 years of watch history.

E. What are the typical data consumption patterns for YouTube usage for a user?

Internet access provisioning or network load provisioning is the process of preparing and equipping a network to allow it to handle the anticipated load and provide new services to its users. Predicting the peak workload of an Internet application and capacity provisioning based on these estimates is notoriously difficult [16]. This is because typically, the peaks of individual users are uncorrelated, and so, the network peak load grows much more slowly than the sum of the peak loads of the individual users. To investigate the how user peak load affects overall traffic, we provide results to show the distribution of YouTube traffic across time and how bursty the usage is.

Methodology and Metrics: To understand how YouTube specific network load is distributed through the day, for each user in our dataset, for each minute of day a video was seen, we check to see whether a video is being watched during that minute (here we assume that the video was watched in its entirety unless the start time of the next video watched by the user is before the current video has finished playing). We also calculate the length and gap between watch sessions for each user; we deem a watch session as the period of time that videos are watched within 5 minutes of each other. In addition, we show how the hour of day, and the day of week that videos are watched, affect the watch session length.

Analysis and Discussion: Fig. 16 shows the normalized aggregate load across all users for each minute of the day. Here we see that from approximately 5am and 12pm, the load drops significantly. During the rest of the day, the load is nearly twice as much. With regard to the watch session length, Fig. 17 shows this on a per user basis. The mean is 26.9 minutes, and ranges from 5.6 minutes to 106.7 minutes, with a standard deviation of 14.4 minutes and a median of 24.8 minutes. In addition, the average time difference between such watch sessions for each user is shown in Fig. 18. The mean is 61.3 hours, and ranges from 1.1 hour to 592.8 hours with a standard deviation of 114.9 hours and a median of 20.6 hours.

Fig. 19 shows how the watch session length varies for hour of the day, and Fig. 20 shows how it varies for day of the week. We see that there is an increase in the watch session length during off-peak periods (5am to 12pm), and also a slight increase on Fridays and Saturdays.

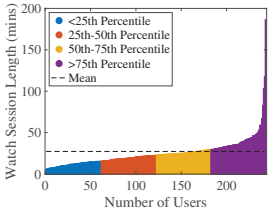

Fig. 17: Watch session length per user

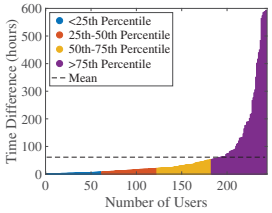

Fig. 18: Watch session difference per user

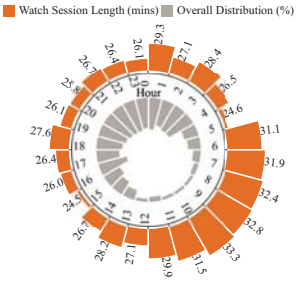

Fig. 19: Watch session length per hour

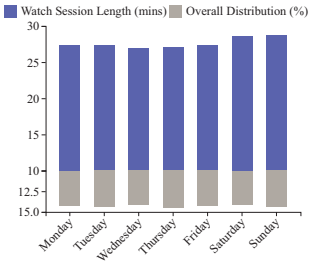

Fig. 20: Watch session length per day

Key Insights: With YouTube having a severe influence on network traffic, it is beneficial to understand the distribution of YouTube traffic and users' access patterns. From our analysis, we see that user's access patterns are similar in that their aggregate usage results in a clear distinction between off-peak and peak periods (shown in Fig. 16). We find that YouTube specific network traffic almost doubles during peak hours, while the average watch session length increases by 17% during off-peak hours. Hence, we see that the time of day plays an important role in the burstiness and overall traffic load. Network service providers would thus need to take this into account when developing their provisioning strategies.

YouTube traffic is nearly 2x as much during peak periods as compared to off-peak periods, and the average watch session length increases by 17% during off-peak hours.

IV. Related Works

Since it's inception in 2005, several studies have been performed aimed at understanding and characterizing YouTube traffic. One of the earlier works investigating the platform

was performed by Cheng *et al.* [17] by crawling YouTube's site and obtaining video meta-data. The authors found that YouTube streaming videos have noticeably different statistics to traditional streaming videos ranging from length, caching strategies to their access pattern and active life space. Similar findings were presented by authors in [18]-[20]. These studies were also alike in their approach of collecting data, by either scraping data from the network edge, or by crawling YouTube's site for publicly available content. The focus of studies from the perspective of users has been limited. Halvey *et. al* [21] examined user's social behaviour with YouTube by analyzing their publicly available online interactions such as commenting and sharing videos. 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 YouTube, and what the implications are for the research community. This allows us to effectively capture a user's viewing pattern and behavior, rather than primarily through their online interactions.

V. Conclusion

In this paper, we collected and analyzed a real-life dataset of YouTube watch history from 243 users comprised of over 1.8 million videos spanning over a 1.5 year period. Using this data, we provided a number of insights and associated implications by answering 5 questions regarding a user's interaction with YouTube: i) How often do users watch the same video again? ii) Is a user's watch behaviour predictable? iii) What role does YouTube's recommendation engine play in influencing users? iv) How dynamic are user's video preferences? and v) What are user's typical YouTube data consumption patterns? These questions pertain to certain representative problems and our associated analysis provided key insights related to those problems. Furthermore, the results and analysis provided attempt to serve as a basis for tackling several problems in the general area of Internet protocols, algorithms and systems.

Acknowledgment

This work was supported in part by the Wayne J. Holman Endowed Chair and the National Science Foundation under grant CNS-1813242.

References

[1] (2019) Youtube for press. [Online]. Available: https://www.youtube.com/about/press/

[2] (2019) 2019 mobile internet phenomena. [Online]. Available: https://www.sandvine.com/2019-mobile-internet-phenomena-report

[3] (2019) The latest youtube stats on audience demographics: Who's tuning in. [Online]. Available: https://www.thinkwithgoogle.com/data-collections/youtube-viewer-behavior-online-video-audience/

[4] (2016) 2016 global internet phenomena. [Online]. Available: https://www.sandvine.com/hubfs/downloads/archive/2016-global-internet-phenomena-report-latin-america-and-north-america.pdf

[5] (2016) Cisco visual networking index: Global mobile data traffic forecast update, 2017-2022. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.pdf

[6] (2019) Amazon mechanical turk. [Online]. Available: https://www.mturk.com

[7] F. R. Bentley, N. Daskalova, and B. White, "Comparing the reliability of amazon mechanical turk and survey monkey to traditional market research surveys," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '17. New York, NY, USA: ACM, 2017, pp. 1092–1099. [Online]. Available: http://doi.acm.org/10.1145/3027063.3053335

[8] J. Erman, A. Gerber, M. T. Hajiaghayi, D. Pei, and O. Spatscheck, "Network-aware forward caching," in *Proceedings of the 18th International Conference on World Wide Web*, ser. WWW '09. New York, NY, USA: ACM, 2009, pp. 291–300. [Online]. Available: http://doi.acm.org/10.1145/1526709.1526749

[9] F. Qian, K. S. Quah, J. Huang, J. Erman, A. Gerber, Z. Mao, S. Sen, and O. Spatscheck, "Web caching on smartphones: Ideal vs. reality," in *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '12. New York, NY, USA: ACM, 2012, pp. 127–140. [Online]. Available: http://doi.acm.org/10.1145/2307636.2307649

[10] I. Sodagar, "The mpeg-dash standard for multimedia streaming over the internet," *IEEE MultiMedia*, vol. 18, no. 4, pp. 62–67, April 2011.

[11] V. N. Padmanabhan and J. C. Mogul, "Using predictive prefetching to improve world wide web latency," *SIGCOMM Comput. Commun. Rev.*, vol. 26, no. 3, pp. 22–36, Jul. 1996. [Online]. Available: http://doi.acm.org/10.1145/235160.235164

[12] C.-Y. Chang and M.-S. Chen, "A new cache replacement algorithm for the integration of web caching and prefectching," in *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, ser. CIKM '02. New York, NY, USA: ACM, 2002, pp. 632–634. [Online]. Available: http://doi.acm.org/10.1145/584792.584903

[13] R. Zhou, S. Khemmarat, and L. Gao, "The impact of youtube recommendation system on video views," in *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, ser. IMC '10. New York, NY, USA: ACM, 2010, pp. 404–410. [Online]. Available: http://doi.acm.org/10.1145/1879141.1879193

[14] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for youtube recommendations," in *Proceedings of the 10th ACM Conference on Recommender Systems*, New York, NY, USA, 2016.

[15] (2019) Nielsen social report 2019. [Online]. Available: https://www.nielsen.com/wp-content/uploads/sites/3/2019/04/The-Social-Media-Report-2012.pdf

[16] B. Urgaonkar, P. Shenoy, A. Chandra, and P. Goyal, "Dynamic provisioning of multi-tier internet applications," in *Second International Conference on Autonomic Computing (ICAC'05)*, June 2005, pp. 217–228.

[17] X. Cheng, J. Liu, and C. Dale, "Understanding the characteristics of internet short video sharing: A youtube-based measurement study," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1184–1194, Aug 2013.

[18] 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

[19] 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

[20] V. K. Adhikari, S. Jain, Y. Chen, and Z. Zhang, "Vivisecting youtube: An active measurement study," in *2012 Proceedings IEEE INFOCOM*, March 2012, pp. 2521–2525.

[21] M. J. Halvey and M. T. Keane, "Exploring social dynamics in online media sharing," in *Proceedings of the 16th International Conference on World Wide Web*, ser. WWW '07. New York, NY, USA: ACM, 2007, pp. 1273–1274. [Online]. Available: http://doi.acm.org/10.1145/1242572.1242804