Understanding the Impact of Network Dynamics on Mobile Video User Engagement

M. Zubair Shafiq[†], Jeffrey Erman[‡], Lusheng Ji[‡], Alex X. Liu[†], Jeffrey Pang[‡], Jia Wang[‡]
[†]Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA
[‡]AT&T Labs – Research, Bedminster, NJ, USA
{shafiqmu,alexliu}@cse.msu.edu, {erman,lji,jeffpang,jiawang}@research.att.com

ABSTRACT

Mobile network operators have a significant interest in the performance of streaming video on their networks because network dynamics directly influence the Quality of Experience (QoE). However, unlike video service providers, network operators are not privy to the client- or server-side logs typically used to measure key video performance metrics, such as user engagement. To address this limitation, this paper presents the first large-scale study characterizing the impact of cellular network performance on mobile video user engagement from the perspective of a network operator. Our study on a month-long anonymized data set from a major cellular network makes two main contributions. First, we quantify the effect that 31 different network factors have on user behavior in mobile video. Our results provide network operators direct guidance on how to improve user engagement — for example, improving mean signal-to-interference ratio by 1 dB reduces the likelihood of video abandonment by 2%. Second, we model the complex relationships between these factors and video abandonment, enabling operators to monitor mobile video user engagement in real-time. Our model can predict whether a user completely downloads a video with more than 87% accuracy by observing only the initial 10 seconds of video streaming sessions. Moreover, our model achieves significantly better accuracy than prior models that require client- or server-side logs, yet we only use standard radio network statistics and/or TCP/IP headers available to network operators.

Categories and Subject Descriptors

C.2.3 [**Computer System Organization**]: computer communication networks—*network operations*; C.4 [**Performance of Systems**]: measurement techniques, performance attributes

Keywords

Cellular Network; Performance; Quality of Experience (QoE); Video Streaming

SIGMETRICS'14, June 16–20, 2014, Austin, Texas, USA.

http://dx.doi.org/10.1145/2591971.2591975.

1. INTRODUCTION

Online video services such as YouTube, Netflix, and Hulu are very popular on mobile networks. It has been estimated that video currently makes up more than half of all mobile data traffic and will grow by a factor of 16 by 2017 [5]. Therefore, it is crucial for mobile network operators to monitor the user experience, or Quality of Experience (QoE), of video streaming and understand how network characteristics and performance influence it.

Unfortunately, prior approaches for monitoring and understanding the user experience of video streaming are insufficient for mobile network operators. Recent seminal work [7–9, 15] investigated how video streaming quality influences important user engagement metrics, such as video abandonment rate. However, these studies rely on client-side instrumentation to measure video quality metrics such as buffering, startup delay, and bitrate. This instrumentation is not available to network operators, so the ability to measure user engagement using only network-side measurements is crucial from their perspective. Other work used network traffic analysis to study video streaming volume and abandonment rates in wired [12, 14] and wireless networks [10]. However, these techniques use deep-packet-inspection to extract information beyond TCP/IP headers, which requires significant computational resources to employ at the scale of network carriers and can pose privacy problems in practice. Moreover, these studies did not provide insight into how network characteristics and performance influence abandonment rates.

To redress these limitations, this paper presents the first large-scale study to characterize video streaming performance in cellular networks and its impact on user engagement. Our study is based on month-long anonymized data sets collected from the core network and radio access network of a tier-1 cellular network in the United States. We analyze 27 terabytes of video streaming traffic from nearly half a million users in this data set. Our analysis makes two main contributions.

First, to the best of our knowledge, our analysis is the first to quantify the impact that network characteristics have on mobile video user engagement in the wild. We quantify the effect that 31 different cellular network factors have on video abandonment rate and video skip (*e.g.*, fast forward) rate. In particular, we quantify user engagement by labeling video streaming sessions in our data set as `completed`/`abandoned` and `skipped`/`non-skipped`, and then evaluate the extent to which core network and radio network factors correlate with abandonment rate and skip rate. These factors include

TCP flow throughput, flow duration, handover rate, signal strength, and the physical location's land cover type. Our results provide network operators insights and direct guidance on how to improve user engagement. For example, improving mean signal-to-interference ratio by 1 dB reduces the likelihood of video abandonment by 2%. Moreover, reducing the load in a cell sector by 10 active users reduces the likelihood of video abandonment by 7%. Through these insights, network operators can identify and prioritize network factors that have the most impact on user engagement.

Second, we are the first to show how a network operator can monitor mobile video user engagement using only standard radio network statistics and/or TCP/IP flow records, a necessity for continuous monitoring at scale and for mitigating privacy concerns. Moreover, we show that our approach can predict video abandonment very early in a video session, which can help future networks decide which users to optimize performance for (*e.g.*, using LTE self-organizing networks [1]). Specifically, we model the complex relationships between network factors and video abandonment. We find that the C4.5/M5P algorithm with bootstrap aggregation can build decision/regression tree models that accurately predict video abandonment. Our results show that it can predict whether a video streaming session is `abandoned` or `skipped` with more than 87% accuracy by observing only the initial 10 seconds. Our model achieves significantly better accuracy than prior models that require video service provider logs [7, 8], while only using standard radio network statistics and/or TCP/IP headers readily available to network operators.

Paper Organization: The rest of this paper is organized as follows. In Section 2, we present a brief background and details of the data collection process. Section 3 presents the characterization of video mobile video streaming performance and its impact on user engagement. We develop a machine learning model for user engagement and present the results in Section 4. Section 5 reviews related work and the paper is concluded in Section 6 with an outlook to our future work.

2. DATA

To study mobile video streaming performance, we collected anonymized flow-level logs from a tier-1 cellular network in the United States. Next, we first provide a brief background of video streaming in cellular networks, description of our data collection methodology, and some high-level statistics of the collected data set.

2.1 Cellular Network Background

A typical UMTS cellular network, shown in Figure 1, can be visualized as consisting of two major components: Radio Access Network (RAN) and Core Network (CN). RAN consists of NodeBs and Radio Network Controllers (RNCs). Each NodeB has multiple antennas, where each antenna corresponds to a different cell sector. A user via user equipment (UE) connects to an *active set* of one or more cell sectors in the RAN. The UE periodically selects a primary or serving cell among the active set based on their signal strength information. From the active set, only the primary cell actually transmits downlink data to the UE. The traffic generated by a UE is sent to the corresponding NodeB by cell sectors. Each RNC controls and exchanges traffic with multiple NodeBs, each of which serves many users in its coverage area. RNCs manage control signaling such as Radio Access Bearer (RAB) assignments, transmission scheduling, and handovers. Each UE negotiates allocation of radio resources with the RAN based on a wide range of factors, such as available radio resources and signal strength [6].

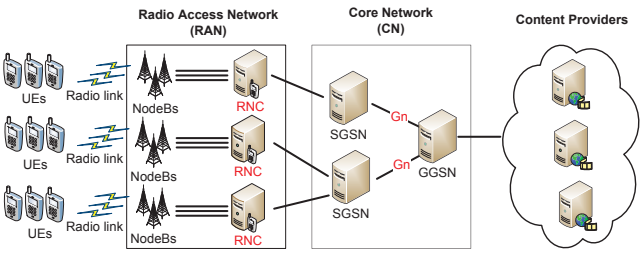


Figure 1: Cellular network architecture

CN consists of Serving GPRS Support Nodes (SGSNs) facing the user and Gateway GPRS Support Nodes (GGSNs) facing the Internet and other external networks. RNCs send traffic from NodeBs to SGSNs, which then send it to GGSNs. GGSNs eventually send traffic to external networks, such as the Internet. For data connections, the IP layer of a UE is peered with the IP layer of GGSNs in the form of tunnels known as Packet Data Protocol (PDP) contexts. These tunnels, implemented as GPRS Tunneling Protocol (GTP) tunnels, carry IP packets between the UEs and their peering GGSNs. From the perspective of an external network such as the Internet, a GGSN connecting CN to the Internet appears just like an IP router and the UEs that connect through the GGSN appear as IP hosts behind the router.

2.2 Data Collection and Pre-processing

For our study, we simultaneously collected two anonymized data sets from the RAN and CN of a tier-1 cellular network in the United States. Our data collection covers a major metropolitan area in the Western United States over the duration of one month in 2012. The RAN data set is collected at the RNCs and contains event-driven signaling information such as current active set, RAB state, handovers, bitrate, signal strength, and RRC requests from users and corresponding responses from the network. The CN data set is collected from the Gn interfaces between SGSNs and GGSNs, and contains flow-level information of video streaming traffic such as server IP and port, client IP and port, flow duration, TCP flags, anonymized user identifier (IMSI), and anonymized device identifier (IMEI). These fields require only TCP/IP or GTP level information, which is efficiently collected [22].

In order to determine the ground-truth of video abandonment, we also collected the following HTTP-level information: URL, host, user agent, content type, content length, and byte-range request from clients and response from servers. Large scale monitoring tools often do not collect HTTP information because it requires processing hundreds of bytes of text beyond the 40-byte TCP/IP header. Thus, it is important that day-to-day monitoring does not require its collection at scale. All device and user identifiers (*e.g.*, IMSI, IMEI) in our data sets are anonymized to protect privacy without affecting the usefulness of our analysis.

The data sets do not permit the reversal of the anonymization or re-identification of users.

To minimize the confounding factors that different content providers (live vs. video-on-demand), connectivity (cellular vs. cable), and device type (mobile vs. desktop) could have on our network-centric analysis, we chose to focus on the most popular video service provider in our cellular network data set. This provider (anonymized for business confidentiality) serves user generated content on demand, and according to a recent study [10], it serves over 37% of all video objects. This provider streams videos using progressive download with byte-range requests, which is the most common protocol currently in use. We believe the conclusions we draw in this paper apply to 9 of the 14 most popular mobile video content providers as they use the same protocol [10]. Previous work found the top providers that use this protocol behave similarly in wired networks [19].

Since our collected data contains traffic records for all types of content, we first need to separate video streaming traffic from the rest. Towards this end, we use the HTTP host and content-type headers to separate the video streaming traffic from other TCP/IP traffic. We can also separate video traffic based only on the server IP and port, since all video streaming traffic is served by a known block of CDN cache servers.

A video is progressively downloaded in one or multiple HTTP byte-range requests, which represent different portions of the video [10]. Figure 2 illustrates a video streaming session that involves multiple HTTP byte-range server response flows. The x-axis represents time, which starts with the first HTTP byte-range server response flow. The y-axis represents byte-range of the video file with maximum value same as the video file size, which is highlighted by the horizontal broken line. Consequently, each gray rectangle represents a distinct HTTP byte-range server response flow. Note that flows may have different byte-range lengths and durations, they may be overlapping, and there may be time gaps between consecutive flows.

For the purpose of our analysis, we group HTTP flows into video sessions based on a unique ID that is the same in the URLs of each video session. In practice, we found that we can group flows into sessions without any HTTP information. In particular, by looking for a change in the server IP to determine when a new video session for a user starts, we can detect session starts correctly with 98% accuracy. This is because videos are served from a CDN and different videos are likely served from different cache servers. Even if all videos were served from a single server, we found that we can still detect session starts with 97% accuracy using a simple decision tree classifier trained on the size of and inter-arrival time gap between HTTP flows. (We omit details due to space constraints.) Thus, we conclude that TCP/IP information would be sufficient to detect and group HTTP flows into video sessions.

2.3 Video Traffic Statistics

Next we present some details of our collected data set such as aggregate statistics, container types, encoding bitrates, and video player types. Overall, our data set consists of more than 27 terabytes worth of video streaming traffic, from more than 37 million flows, from almost half a million users over the course of one month.

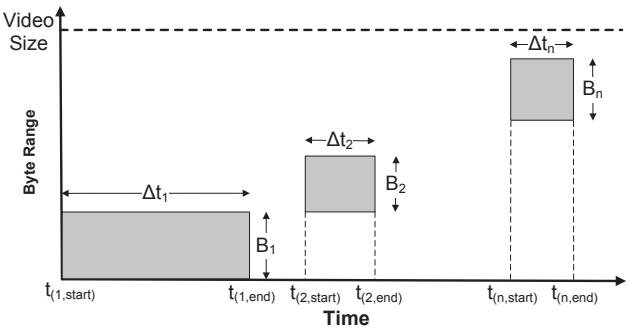


Figure 2: Illustration of a video streaming session. Gray rectangles represent distinct flows in a session.

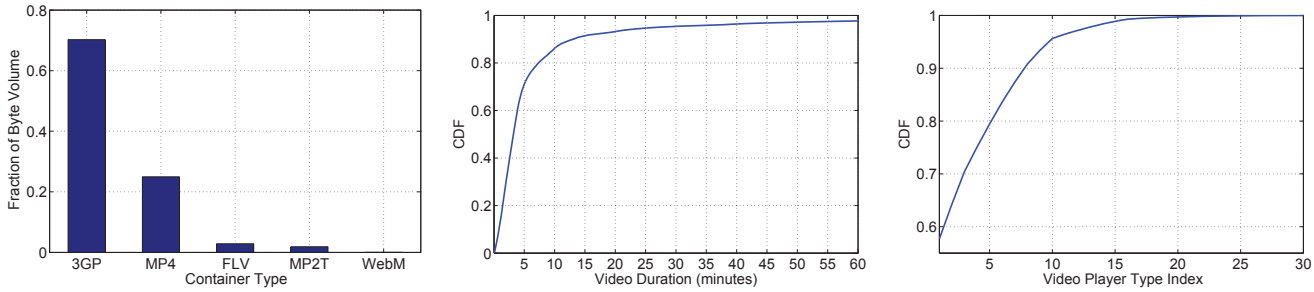
Our data set mostly contains standard definition video streaming streaming traffic. Figure 3(a) shows the distribution of video streaming traffic with respect to container types. The most common container types [3] are: (1) 3GP (3GPP file format), (2) MPEG-4, (3) FLV (Flash), (4) WebM, and (5) MP2T (MPEG transport stream). We observe that a vast majority, almost 70%, of video streaming traffic uses the 3GPP container type – followed by MPEG-4 and Flash container types as distant 2nd and 3rd most popular, respectively. Only a small fraction, less than 2%, of the video streaming traffic belongs to containers types used for live content. We exclude these from our analysis since our focus is on video-on-demand streaming. Further analysis of video encoding bitrate showed that a majority of video streaming traffic belongs to lower bitrates, which correspond to 144/240p video resolution. 240p is the most commonly used video resolution. Only a small fraction of video streaming traffic belongs to higher video resolutions. For example, less than 5% video streaming traffic belongs to high definition (HD) 720p content.

Short duration videos account for most of the streaming traffic in our data set. In Figure 3(b), we plot the cumulative distribution function (CDF) of video duration. We observe that more than 70% videos are less than 5 minutes long and only 10% videos are longer than 15 minutes. This type of skewed distribution is expected for content providers that serve user generated content [10].

Users employ only a few distinct video players to play video content in our data set. We plot the CDF of users across video player types (reverse-sorted with respect to fraction of users) in Figure 3(c). We identify video player types using the available user agent information [23, 24], which enables us to differentiate among video players on different hardware models, operating system versions, and web browsers. Our data set contains several dozen distinct video player types whose distribution is highly skewed, *i.e.*, a small fraction of video player types account for most users in our data set. Specifically, top-5 player types account for approximately 80% users in our data set and they represent both iOS- and Android-based devices.

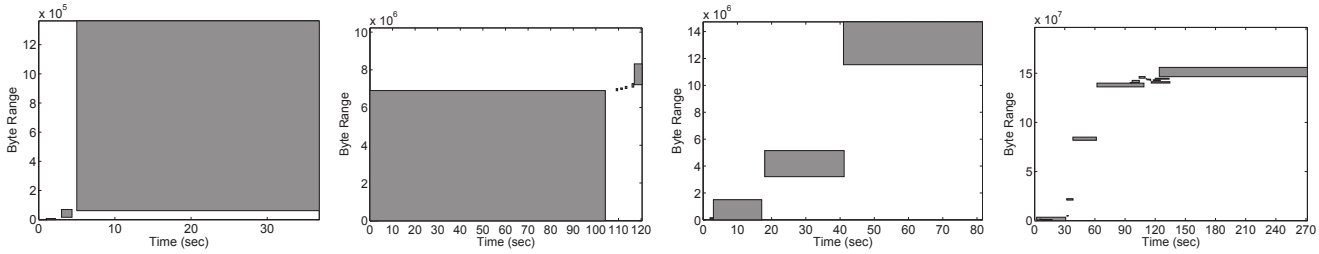
2.4 Quantifying User Engagement

As a first step towards analyzing user engagement, we discuss two ways to quantify it: discrete and continuous.



(a) Container type (b) Video duration (c) Video player type

Figure 3: Distributions of container type, video duration, and video player type



(a) `completed, non-skipped` (b) `abandoned, non-skipped` (c) `completed, skipped` (d) `abandoned, skipped`

Figure 4: Examples of video streaming session classes. Y-axis limits are set to the video sizes.

Discrete quantification of user engagement. For discrete quantification, we first use a nominal variable that represents the following classes: `completed` and `abandoned`. The `completed` class represents video streaming sessions in which the download process reaches the end-point. The `abandoned` class represents video streaming sessions in which the download process is abandoned before reaching the end-point. In our data set, 21.2% sessions belong to the `completed` class and 78.8% sessions belong to the `abandoned` class. Since users tend to skip videos when streaming gets stuck, we also use a nominal variable that represents the following classes: `skipped` and `non-skipped`. The `skipped` class represents video streaming sessions in which the download process includes at least one seek-forward between the start-point and the last byte downloaded. The `non-skipped` class represents video streaming sessions in which the download process does not include seek-forward between the start-point and the last byte downloaded. In our data set, 33.9% sessions belong to the `skipped` class and 66.1% sessions belong to the `non-skipped` class. Combining the aforementioned user engagement classification schemes, we can define the following four non-overlapping classes: (1) `completed, non-skipped`, (2) `abandoned, non-skipped`, (3) `completed, skipped`, and (4) `abandoned, skipped`. In our data set, 17.6% sessions belong to the `completed, non-skipped` class, 48.5% sessions belong to the `abandoned, non-skipped` class, 3.6% sessions belong to the `completed, skipped` class, and 30.3% sessions belong to the `abandoned, skipped` class. Figure 4 illustrates examples of video streaming sessions for the four user engagement classes. As mentioned earlier and observable in Figure 4, sessions generally consist of more than one flow. On average, a video streaming session in our data set consists of 11 flows, where earlier flows tend to be larger than the following flows. This trend is because video players tend to aggressively download larger chunks to fill up the available buffer during the initial buffering phase of a video streaming session [19]. The download rate in this initial phase is limited by the end-to-end available bandwidth. Afterwards in the steady state phase, the remaining video is generally downloaded in multiple smaller flows. The download rate in this phase depends on the video encoding rate and the playback progress.

Continuous quantification of user engagement. For continuous quantification, we use a continuous variable ($\in [0,1]$) representing the fraction of video download completion. Figure 5 shows the CDF of video download completion. Comparing videos of different durations, as expected we observe that shorter videos achieve higher download completion than longer videos [16]. For aggregate distribution, almost 15% of video streaming sessions are abandoned with less than 5% download completion. However, after the 5% completion mark, the distribution is fairly uniform until the 80% completion mark. The initial modality in the distribution indicates abandonment that is likely either because users tend to sample videos [7] or due to longer join times [9]. The later modality in the distribution (excluding the 100% completion mark) indicates abandonment that is likely either because users lose interest in the content (*e.g.*, due to video closing credits) or because shorter videos achieve higher download completion due to aggressive initial buffering.

We note that our definitions of user engagement detect abandonment and skips only during the download phase of a video. We cannot detect a video abandonment or skip if these events occur after a video has downloaded completely (*e.g.*, due to lack of user interest). However, network operators are typically not interested in those events because they are unlikely to be influenced by network factors.

Table 1: CN (top) and RAN (bottom) features. i denotes the flow index of a session with N flows.

Feature	Description
Flow volume	(B_i) The number of bytes transferred during the i^{th} flow. (Summary stats)
Flow duration	(t_i) The duration (in seconds) from the SYN packet to the last packet in the i^{th} flow. (Summary stats)
Flow TCP throughput	(T_i) The ratio of flow volume to flow duration in the i^{th} flow, in KB/s. (Summary stats)
Flow inter-arrival time	(I_i) Time (in seconds) between the end of the i^{th} flow and the start of the $i+1^{\text{th}}$ flow. (Summary stats)
Flow flags	FIN_i and RST_i respectively denote the number of packets with TCP-Finish (no more data from sender indicating completion) and TCP-Reset (reset the connection indicating some unexpected error) flags set in the i^{th} flow. Based on the direction of packet transfer, we distinguish between client-to-server (c→s) and server-to-client (s→c) flags. (Summary stats)
Largest flow volume	(B_j) The largest flow volume among all flow volumes, where j denotes the index of this flow.
Largest flow duration	(t_j) The duration of the j^{th} flow.
Largest flow TCP throughput	(T_j) The throughput of the j^{th} flow.
Largest flow flags	FIN_j and RST_j respectively denote the number of packets with TCP-Finish and TCP-Reset flags set in the j^{th} flow. We distinguish between c→s and s→c flags.
Number of flows	(N) The total number of flows in a session.
Session volume	(**B**) The sum of all flow volumes in a session.
Session duration	(**t**) The sum of all flow durations in a session. $\mathbf{t} = \sum_{i=1}^{N} t_i$.
Session TCP throughput	(**T**) The average throughput of a session. $\mathbf{T} = \sum_{i=1}^{N} B_i \ / \ \sum_{i=1}^{N} t_i$.
Session inter-arrival time	(**I**) The sum of all flow inter-arrival times (in seconds) in a session
Session flags	(**FIN** and **RST**) respectively denote the number of packets with TCP-Finish and TCP-Reset flags set in a session. We distinguish between c→s and s→c flags.
# soft handovers	(H_S) This handover occurs when a cell is added or removed from the active set [25]. (Session- and cell-level)
# inter-frequency handovers	(H_{IF}) This type of handover occurs when a UE switches to cell sector of the same or different NodeB with different operating frequency [25]. (Session- and cell-level)
# IRAT handovers	(H_{RAT}) This type of handover occurs when a UE switches between different radio access technologies (*e.g.*, UMTS and GPRS) [25]. (Session- and cell-level)
# RRC failure events	A RRC failure event is logged when a request by a user to allocate more radio resources is denied by the respective RNC due to network overload or other issues [21]. (Session- and cell-level)
# admission control failures	These events occur when a user cannot finish the admission control procedure often due to lack of available capacity. (Session- and cell-level)
Received signal code power	RSCP is the RF energy of the downlink signal obtained after the correlation and descrambling process [25]. It is usually measured in dBm. (Summary stats)
Signal energy to interference	This ratio (E_c/I_o) denotes the ratio of the received energy to the interference level of the downlink common pilot channel [25]. It is usually measured in dB. (Summary stats)
Received signal strength	RSSI takes into account both RSCP and E_c/I_o [25]. It is usually measured in dBm. It is defined as: RSSI = RSCP - E_c/I_o. (Summary stats)
Size of active set	(S_{AS}) The number of unique cell sectors in the active set. (Summary stats)
Radio access bearer state	Our measurement apparatus distinguishes among 84 different RAB states. RAB state encodes encodes information about RRC state (*e.g.*, FACH shared channel vs. DCH dedicated channel), RAB type (*e.g.*, interactive vs. streaming), and maximum data rate. Since a session may have multiple RAB states over time, we use the most common state for session-level and top-3 most common states for cell-level features.
Uplink RLC throughput	(T_U) The uplink data rate for UE in the DCH state (in kbps). (Session- and cell-level summary stats)
Downlink RLC throughput	(T_D) The downlink data rate for UE in the DCH state (in kbps). (Session- and cell-level summary stats)
# Users in DCH state	(U_{DCH}) Users served by the representative cell over a window of 1 hour.
Frequency	The operating frequency of the representative cell.
Landcover	A nominal variable that defines the geographical terrain of a cell. 2006 National Land Cover Database contains the 16-class geographical terrain categorization of the United States at a spatial resolution of 30 meters [2, 13]. The categories include developed-open space, developed-high intensity, perennial ice/snow, deciduous forest, open water, *etc.* We extract the top-3 most common landcover categories in terms of spatial area within 1 km of the representative cell.
Elevation	Elevation of a cell is extracted from the National Elevation Dataset (NED) at a spatial resolution of 30 meters [4]. We use average elevation of the representative cell as a feature.

3. ANALYSIS OF NETWORK FACTORS

Our main goal is to understand the influence of network factors on user engagement. Towards this end, this section presents an in-depth analysis of the relationships between network factors and video abandonment.

We first itemize a wide range of factors that can potentially impact or be influenced by mobile video user engagement. We compile a comprehensive list of features from the information available in both CN and RAN data sets. It is noteworthy that while features extracted from the RAN data set are only applicable for cellular networks, features extracted from the CN data set are applicable for other kinds of wired and wireless networks as well. Table 1 summarizes the features.

Core Network (CN) features. For each video streaming session, we can extract CN features for individual flows and the whole session (labeled as *Flow* and *Session* features in

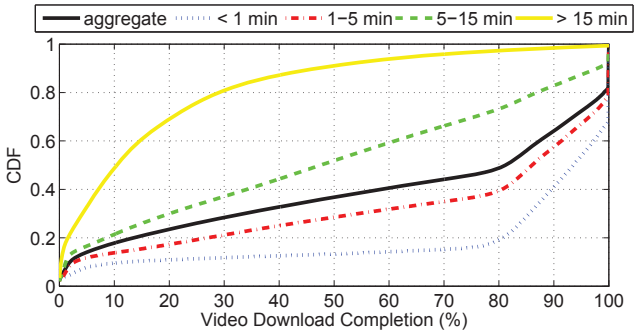


Figure 5: Distribution of video download completion

the top half of Table 1, respectively). Since sessions may have different number of flows, we compute the following statistical measures to summarize the flow-level features for whole sessions: mean, standard deviation, minimum, maximum, 25^{th} percentile, median (50^{th} percentile), and 75^{th} percentile. Hence each flow-level feature listed in Table 1 (labeled with "Summary stats") represents 7 summary values. We also extract these features for the largest flow (in terms of byte volume) of a video streaming session, as a single flow typically dominates each video session.

Radio Access Network (RAN) features. For each video streaming session, we also extract RAN features for the user and the cell sectors that service the user during the session. The RAN features are described in the bottom half of Table 1. For session-level features, the RAN data set records belonging to a user can be identified using the combination of IMSI and session start and end timestamp information. For cell-level features, however, the selection criterion of the representative cell for a session is not obvious because the active set and the primary cell may change between the start and end of a session. Towards this end, we select the most common cell sector (in terms of number of records) to be the representative cell for a session. For each session, cell-level features are computed for all users served by the representative cell in the time window at the session start. Features in Table 1 labeled with "Session- and cell-level" indicate features that we compute both a session-level value and cell-level value, as defined above. For example, for # soft handovers, we compute one feature as the number of soft handovers for the user during the session, and another as the number of soft handovers for all users in the representative cell of that session. For features that can hold multiple values during a session (*e.g.*, RSSI), we compute the same 7 summary statistic values listed above for flow features. These features are labeled with "Summary stats" in Table 1.

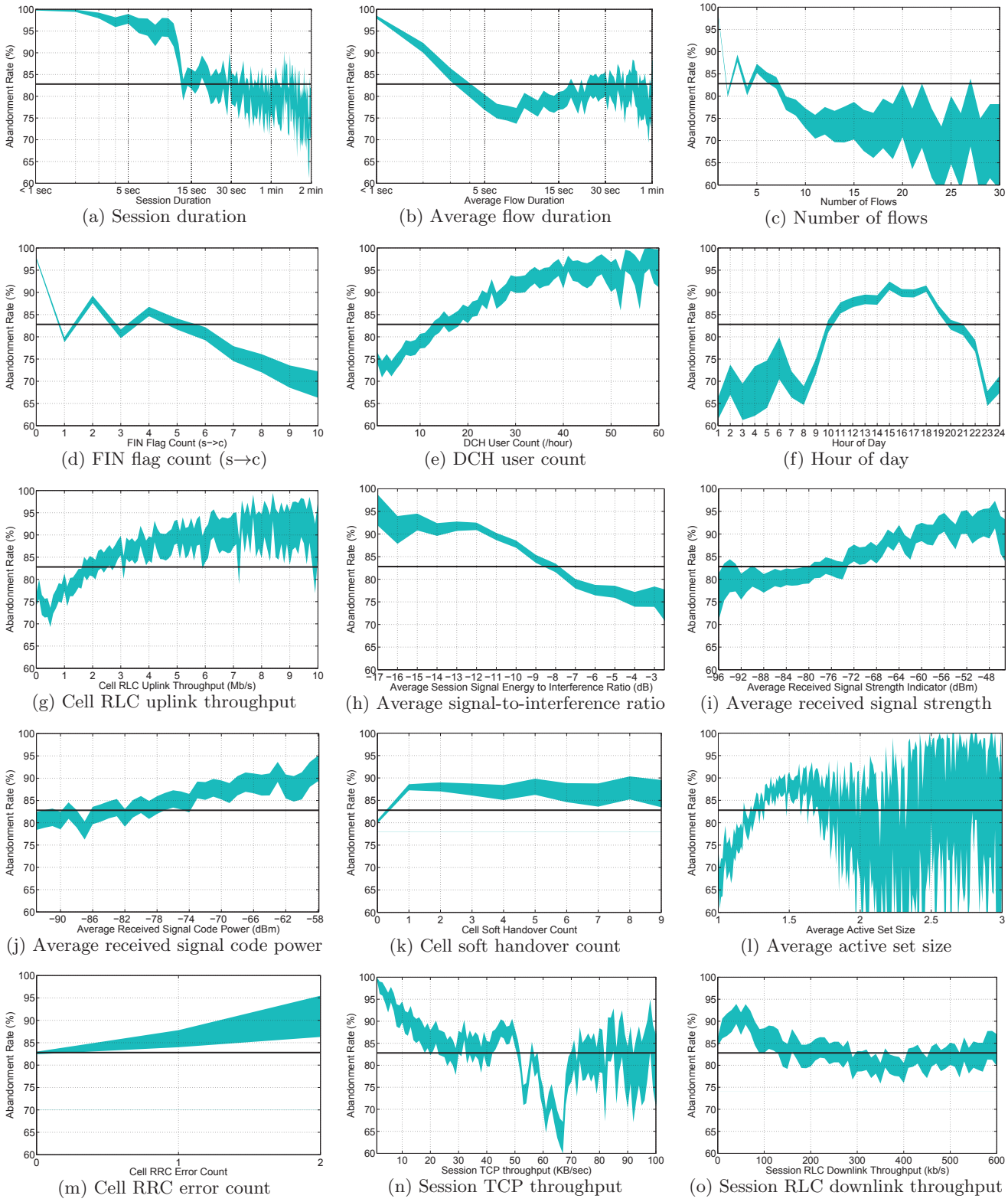
To better understand the relationship between features and user engagement, we plot the abandonment rate distributions of prominent features in Figure 6. The abandonment rate is defined as the fraction of sessions in the data set that are abandoned. The shaded areas represent the 95% confidence interval [26]. The horizontal line in each plot denotes the average abandonment rate. Figure 6 suggests the following implications:

Abandoned sessions are shorter. Although this result is expected, we find that each measure of session length provides unique information. Figure 6(a) shows sessions shorter than 15 seconds are significantly more likely to be abandoned. The sharp inflection point may be due to automated failure of sessions that do not complete the initial buffering phase. Similarly, Figure 6(b) shows a sharp drop in abandonment rate for sessions with average flow duration longer than 1-3 seconds. Figures 6(c) and 6(d), both measures of flow count, show that sessions with more flows are less likely to be abandoned. Thus, each of these features provides information useful for detecting abandonment.

Network load increases the abandonment rate. Despite the low bitrate of video streams relative to the capacity of a cell (∼500 kbps vs. 3-14 Mbps), we find there is a nearly linear relationship between various measures of RAN load and abandonment rate. For example, Figure 6(e) shows that the abandonment rate goes up by roughly 7% for each 10 active users in a sector, even though these resources are scheduled in a proportional fair manner every 2 ms [6]. This load relationship can also be seen in Figure 6(f), which shows that abandonment rate is highest during the peak load hours of the day and much lower during the off-peak hours. This effect can be explained by Figure 6(g), which shows that the abandonment rate begins to grow when aggregate cell uplink throughput is just 50 kbps, significantly less than the cell capacity. This is likely because even small amounts of uplink traffic can cause interference, and Figure 6(h) shows that abandonment rate decreases by 2% for each dB increase in the signal-to-interference ratio (E_c/I_o). Furthermore, Figures 6(i) and 6(j) show that the abandonment rate increases as RSSI and RSCP increase, contrary to the general belief that higher received power means a better user experience. These E_c/I_o, RSSI, and RSCP results strongly suggest that users with higher received power also experience more interference. Hence, user engagement in our data set is more limited by interference rather than poor coverage. In summary, these results suggest that measures a cellular operator takes to reduce per-sector load and interference will improve user engagement in a roughly linear manner.

Handovers increase the abandonment rate. Another important question for operators is whether cell handovers disrupt the user experience. Our results suggest that all handover types are correlated with a decrease in user engagement. Figure 6(k) shows that cells with soft handovers, which are "make-before-break" handovers, have significantly higher abandonment rates. This result is supported by Figure 6(l), which shows increase abandonment rates for non-integral mean active set values (i.e., sessions that incurred active set additions or deletions during soft handovers). These effects may be partially due to the RRC signalling required for handover. Figure 6(m) shows that when RRC signalling errors occur, abandonment rate increases as well.

Higher throughput does not always mean lower abandonment. Although measured throughput is often used as a proxy for network quality (*e.g.*, [14]), our results suggest higher average throughput does not always indicate lower abandonment. Figures 6(n) and 6(o) show that abandonment rate decreases as average TCP and RLC throughput increases up to a point. However, the abandonment rate is lowest at TCP throughput equal to the steady state streaming rate, and it grows for higher throughput values. This



(a) Session duration (b) Average flow duration (c) Number of flows

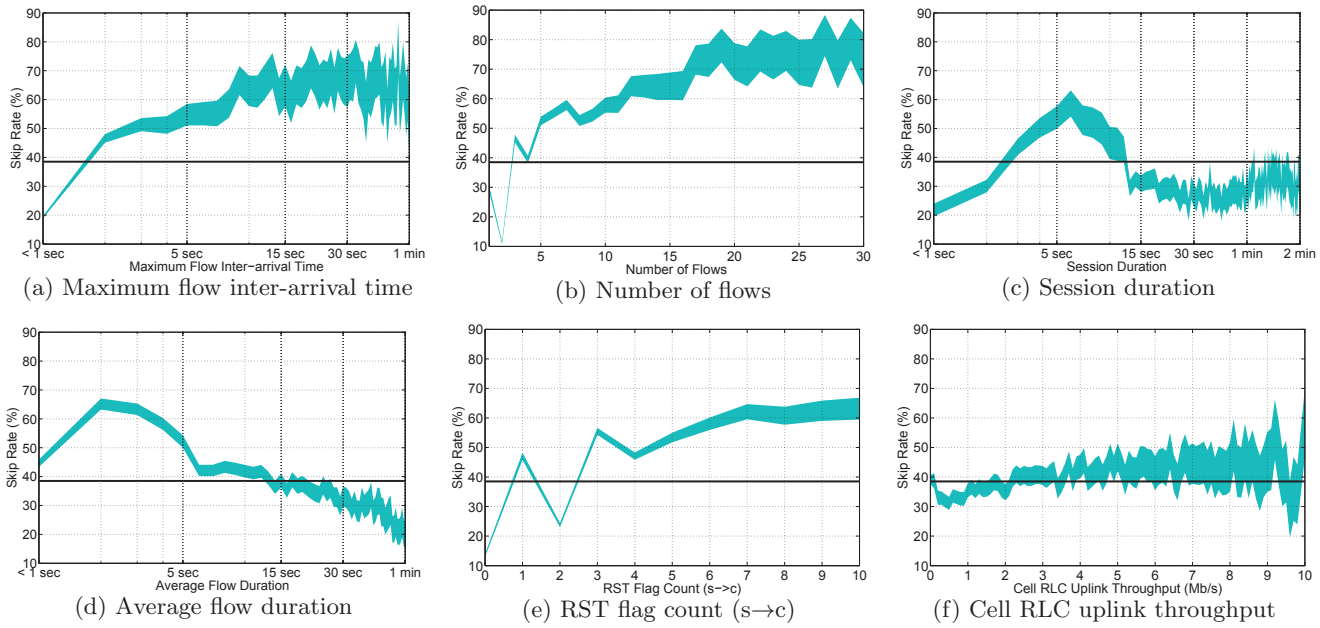
(d) FIN flag count (s→c) (e) DCH user count (f) Hour of day

(g) Cell RLC uplink throughput (h) Average signal-to-interference ratio (i) Average received signal strength

(j) Average received signal code power (k) Cell soft handover count (l) Average active set size

(m) Cell RRC error count (n) Session TCP throughput (o) Session RLC downlink throughput

Figure 6: Abandonment rate distributions. Shaded areas represent the 95% confidence interval.



(a) Maximum flow inter-arrival time (b) Number of flows (c) Session duration

(d) Average flow duration (e) RST flag count (s→c) (f) Cell RLC uplink throughput

Figure 7: Skip rate distributions. Shaded areas represent the 95% confidence interval.

pattern is because early abandonment, while the video is still in the non-rate-limited buffering phase, actually results in higher average throughput than watching a video through the rate-limited steady state phase.

In Figure 7, we plot the skip rate distributions of prominent features. The skip rate is defined as the fraction of sessions in the data set that are skipped. Due to space constraints, we only plot the skip rate curves for features that have different trends than the respective abandonment rate curves. The shaded areas represent the 95% confidence interval [26]. The horizontal line in these plots denotes the average skip rate.

We note that skip rate has a direct relationship with maximum flow inter-arrival time (Figure 7(a)) and number of flows (Figure 7(b)). This is likely because skips result in more flows and larger gaps between them. Skip rate peaks at session and flow durations of just a few seconds (Figures 7(c) and 7(d)), suggesting that users chose to skip early in a session, either due to network issues or lack of interest. Figure 7(e) shows larger RST flag count correlated with higher skip rate likely because skips cause connection resets. These contrasting patterns imply that it is more challenging to measure both skips and abandonment than a single engagement metric.

4. MODELING USER ENGAGEMENT

In this section, we develop models to accurately predict user engagement using only standard radio network statistics and/or TCP/IP header information.

4.1 Background and Problem Statement

Network operators would like to predict user engagement for three main applications. First, directly estimating these metrics from network traffic requires cost-prohibitive collection of sensitive data (requiring deep-packet-inspection) beyond TCP/IP headers. Thus, cost and privacy concerns would be alleviated with a model that accurately predicts these engagement metrics using only standard radio network statistics and/or TCP/IP header information that is already collected. A simple and efficient model would be able to monitor video engagement metrics over an entire network in real-time to facilitate trending and alarming applications. Second, self-organizing networks [1] (SON) enable mobile networks to adapt resource allocation dynamically. Thus, the ability to accurately predict video abandonment early in a video session can help guide SONs to provide more resources to the most vulnerable sessions. Third, an interpretable model that relates network factors to user engagement metrics can help network operators in prioritizing infrastructure upgrades by identifying the combination of factors that need to be adjusted for improving engagement.

Our goal is to jointly use the available features to accurately model both nominal and continuous measures of user engagement (defined in Section 2). Moreover, we want our models to make the prediction decisions as early as possible in a video session. Therefore, we define the modeling problem as follows: *given the feature set computed over the initial τ seconds ($\tau \leq \mathbf{t}$) of a video session, predict the user engagement metric.*

4.2 Proposed Approach

As we observed in Section 3, many network features are not independent of each other and the relationships among them can be non-linear. Therefore, modeling user engagement given all available features is a non-trivial prediction task. Furthermore, our modeling approach should answer pertinent questions such as: Which features are more useful for prediction? How many features do we need to reap a substantial accuracy gain?

To address these challenges, we use a machine learning approach for modeling the complex relationships between network features and user engagement metrics. The choice of learning algorithm is crucial to successfully modeling feature interdependence and non-linearity. After some pilot experiments, we found that decision tree algorithms with bootstrap aggregation (or bagging) [27] work well for both nominal (classification) and continuous (regression) user en-

gagement metrics. Other commonly used Bayes and linear regression algorithms were outperformed by the decision tree algorithms in our pilot experiments. Decision trees do not require feature independence assumption and can handle non-linearities by employing multiple splits/breaks for each feature. Furthermore, decision tree models comprise of simple if-then-else branches, which can process data efficiently. For our experiments, we used C4.5 decision tree algorithm [17] and M5P regression tree algorithm [18].

4.3 Experimental Setup

We evaluate the effectiveness of classification models in terms of the following standard Receiver Operating Characteristic (ROC) metrics [11]: (1) True Positives (TP), (2) True Negatives (TN), (3) False Positives (FP), and (4) False Negatives (FN). We summarize the classification results in terms of the following ROC metrics: True positive rate $= \frac{|\text{TP}|}{|\text{TP}|+|\text{FN}|}$, False positive rate $= \frac{|\text{FP}|}{|\text{FP}|+|\text{TN}|}$, and Accuracy $= \frac{|\text{TP}|+|\text{TN}|}{|\text{TP}|+|\text{TN}|+|\text{FP}|+|\text{FN}|}$. We also plot the standard ROC threshold curves in our evaluation. An ideal ROC threshold curve approaches the top-left corner corresponding to 100% true positive rate and 0% false alarm rate. The Area Under Curve (AUC $\in [0, 1]$) metric summarizes the classification effectiveness of an ROC threshold curve, where the AUC values approaching 1 indicate better accuracy. Besides, we evaluate the effectiveness of regression models in terms of the standard root-mean-square error (RMSE $\in [0, 1]$) metric.

To avoid class imbalance and over-fitting during model training, we use k-fold cross-validation with class resampling [27]. In our pilot experiments, different values of k yielded very similar results. All experimental results reported in this paper are presented for $k = 10$. Furthermore, we evaluate the feature sets on varying initial time window sizes: $\tau = \mathbf{t}$ (*i.e.*, use all available data), $\tau \leq 60$ seconds, and $\tau \leq 10$ seconds. We expect the classification accuracy to degrade for smaller initial time windows.

We separately evaluate the core network feature set (abbreviated as **CN**), the radio network feature set (abbreviated as **RAN**), and the combined feature set (abbreviated as **All**). The radio network and core network features are separately grouped because they require different types of instrumentation. Recall from Section 2, measuring radio network features requires instrumentation at RNCs and measuring core network features requires instrumentation at Gn interfaces in a cellular network.

4.4 Evaluation

Our experimental results demonstrate that the proposed machine learning model can predict both video abandonment and skips with high accuracy using only the initial 10 seconds of a session, while meeting the constraints of network operators. We find that although some features are more useful than the rest for prediction, using all available features results in significant accuracy gain as compared to using only a few top features. Moreover, our decision and regression tree models are interpretable; they inform us about the relative usefulness of features by ordering them at different tree levels and we can understand the specific set of network conditions that impact user engagement by following each path in the tree. Many of these conditions can be influenced by a network operator, and thus provide guidance on how to improve user engagement in different situations.

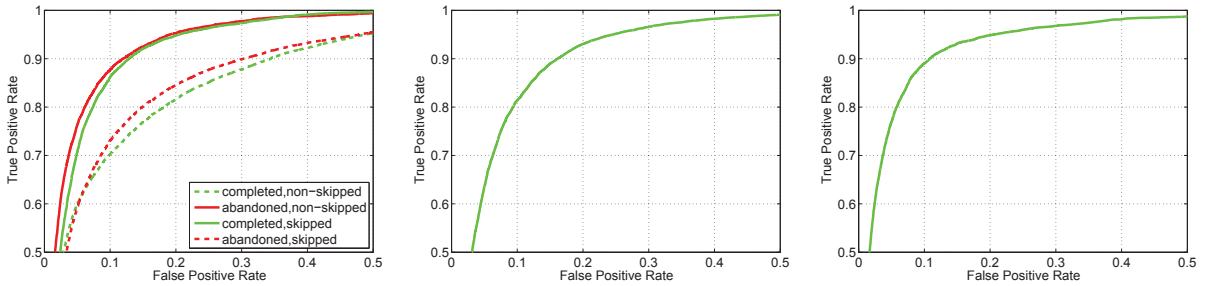
Table 2: Accuracy of 4-way classification

Feature Set	`completed non-ski.` (%)	`abandoned non-ski.` (%)	`completed skipped` (%)	`abandoned skipped` (%)	Avg. (%)
$\tau = \mathbf{t}$					
CN	72.0	78.4	76.2	73.4	75.0
RAN	64.1	53.7	73.2	55.7	61.7
All	73.1	77.8	77.4	74.4	**75.7**
$\tau \leq 60$ seconds					
CN	69.5	62.7	63.8	64.6	65.2
RAN	62.6	47.8	58.5	57.0	56.5
All	70.4	63.7	65.7	65.4	**66.3**
$\tau \leq 10$ seconds					
CN	69.5	59.6	63.3	65.3	64.4
RAN	60.5	46.6	59.0	57.4	55.9
All	69.6	60.7	64.9	65.5	**65.2**

Next we present detailed results for both classification and regression. For classification, we build decision tree models to predict both individual classes and their combinations. For individual classes, we train the decision tree algorithm for 4-way classification. By combining classes, we change the granularity at which the model predicts user engagement. We use the following two class pairs: `completed` vs. `abandoned` and `completed, non-skipped` vs. `rest`. For combined classes, we train the decision tree algorithm for 2-way classification. Naturally, we expect better accuracy for 2-way classification than 4-way classification because the model is trained at a coarser granularity.

For 4-way classification, we observe that the core network feature set outperforms the radio network feature set. Combining the core and radio network feature sets improves the average accuracy. In Table 2, we observe the best average accuracy of 75.7% for the combined feature set at $\tau = \mathbf{t}$ seconds. In practice, improvement in accuracy means that fewer sessions need to be measured before the network operator can be confident that a real change in user engagement has occurred and an alarm can be raised. For a cell sector serving only a handful of users simultaneously, this can mean a significant reduction in time to detection of issues since video sessions may not be frequent. For the combined feature set, ROC threshold curves are plotted in Figure 8(a). The ordering of ROC curves conforms with the class-wise accuracy results in Table 2. The best operating accuracy of 77.8% is observed for `abandoned, non-skipped` class, which corresponds to 95.5% AUC. As expected, in Table 2 we observe that the average accuracy degrades for smaller values of τ. We observe the best average accuracy of 65.2% for the combined feature set at $\tau \leq 10$ seconds, representing more than 10% accuracy reduction as compared to $\tau = \mathbf{t}$ seconds.

Since operators may only be interested in predicting video abandonment, it is also important to build models to accurately predict `completed` vs. `abandoned` and `completed, non-skipped` vs. `rest` class pairs (instead of all four classes). These class pairs compare the scenarios when users either abandon or skip the video streaming session. Table 3 presents the classification results for these two class pairs. As expected, we observe significant improvement in accuracy for both class pairs as compared to 4-way classification due to reduced number of classes. Moreover, we observe that the average accuracy suffers only minor degradation (less than 5%) as τ is reduced. For `completed` vs. `abandoned` class pair, we observe the best average accuracy of



(a) 4-way (b) `completed` vs. `abandoned` (c) `completed, non-skip.` vs. `rest`

Figure 8: ROC threshold plots for various class pairs

Table 3: Accuracy of `completed` vs. `abandoned` and `completed, non-skipped` vs. `rest` classification

Feature Set	comp. (%)	abandoned (%)	Avg. (%)	completed, non-ski. (%)	rest (%)	Avg. (%)
$\tau = \mathbf{t}$						
CN	80.5	85.9	83.2	77.2	92.3	**88.5**
RAN	73.9	77.9	75.9	71.9	88.8	84.5
All	80.5	86.5	**83.5**	76.9	92.4	**88.5**
$\tau \leq 60$ seconds						
CN	79.5	82.1	**80.8**	78.0	91.5	**88.1**
RAN	74.1	78.4	76.3	72.7	88.4	84.4
All	78.8	82.6	80.7	77.6	91.4	88.0
$\tau \leq 10$ seconds						
CN	79.6	80.7	**80.1**	77.1	90.7	**87.3**
RAN	74.2	78.9	76.5	73.2	89.2	85.1
All	77.8	82.1	79.9	76.6	90.7	87.2

83.5% for the combined feature set at $\tau = \mathbf{t}$ seconds. For the combined feature set, the ROC threshold curve is plotted in Figure 8(b), which corresponds to 93.4% AUC. For `completed, non-skipped` vs. `rest` class pair, we observe the best average accuracy of 88.5% for the combined feature set at $\tau = \mathbf{t}$ seconds. For the combined feature set, the ROC threshold curve is plotted in Figure 8(c), which corresponds to 95.1% AUC.

For regression, we build regression tree models to predict video download completion. Overall, we observe similar patterns across feature sets and varying initial window sizes for regression results as observed for classification results earlier. Table 4 presents the results of M5P regression tree algorithm and a simple linear regression algorithm. We note that M5P regression tree algorithm consistently outperforms the simple linear regression algorithm, indicating that M5P can successfully capture the non-linear dependencies between features and video download completion that are not modeled by the simple linear regression algorithm. RMSE is lower for larger τ values, and **All** feature set has the lowest RMSE as compared to individual **CN** and **RAN** feature sets. We observe the best RMSE of 0.14 for $\tau = \mathbf{t}$ and **All** feature set.

4.5 Discussion

How many features to use? Our evaluation highlighted that using all features together results in better classification/regression accuracy than using their subsets. To systematically analyze the utility of adding features to the classification/regression model, we plot accuracy versus feature set size for `completed` vs. `abandoned` classification in Figure 9. Towards this end, we iteratively rank the feature set using the following greedy approach: for k^{th} iteration, we evaluate the accuracy gain of the model by separately adding candidate features and selecting the $k+1^{th}$ feature which provides the best accuracy. The top features are related to session size and TCP throughput which we believe are correlated with user engagement, as sufficient throughput is required for video streaming and abandonment results in low volume sessions. The plot shows that a few top features provide most of the accuracy gain. However, the gains in accuracy we achieve from including 5% to 100% of features are not diminishing. Thus, it makes sense to use all available features because the computational overheads of feature extraction and testing for additional features is low (in order of milliseconds).

Actionable insights. The decision/regression tree models also provide actionable insights. The pruned versions of the

Table 4: Root-mean-square error of regression

Feature Set	Linear Regression	M5P Regression Tree
$\tau = \mathbf{t}$		
CN	0.25	0.15
RAN	0.30	0.27
All	0.23	**0.14**
$\tau \leq 60$ seconds		
CN	0.27	0.18
RAN	0.36	0.34
All	0.24	**0.17**
$\tau \leq 10$ seconds		
CN	0.29	0.22
RAN	0.37	0.34
All	0.28	**0.21**

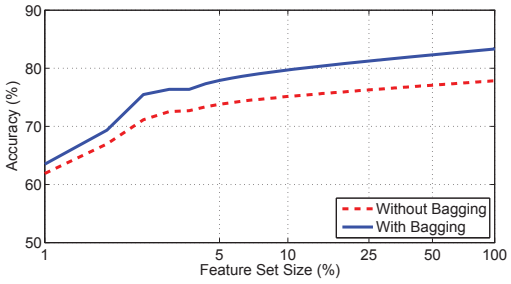
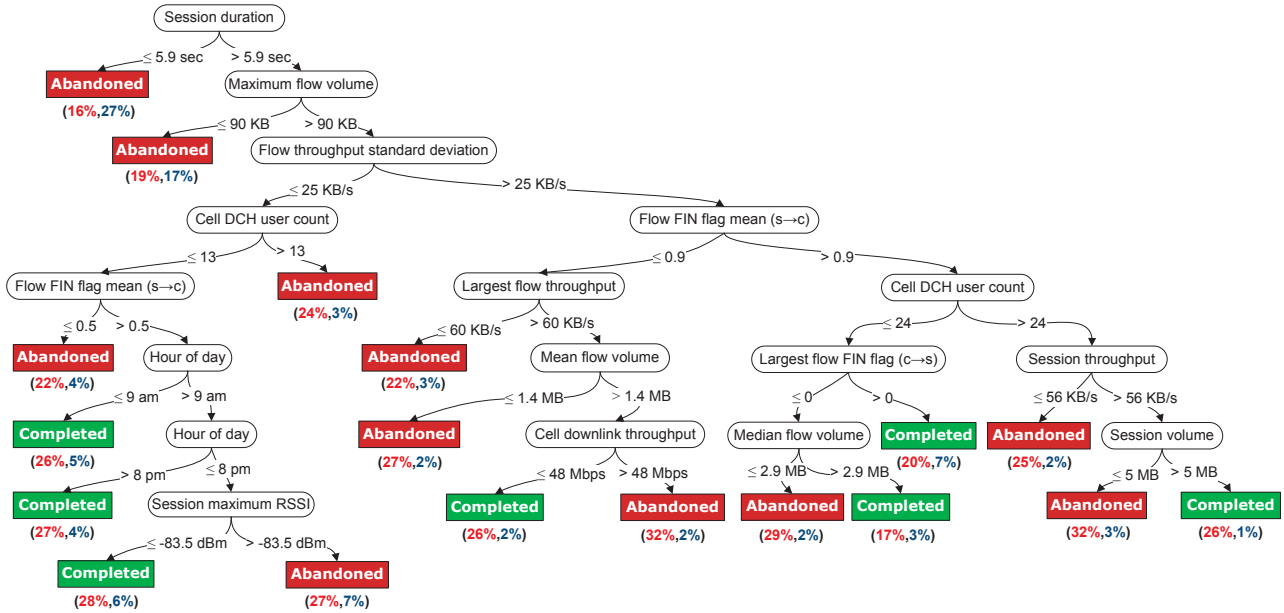
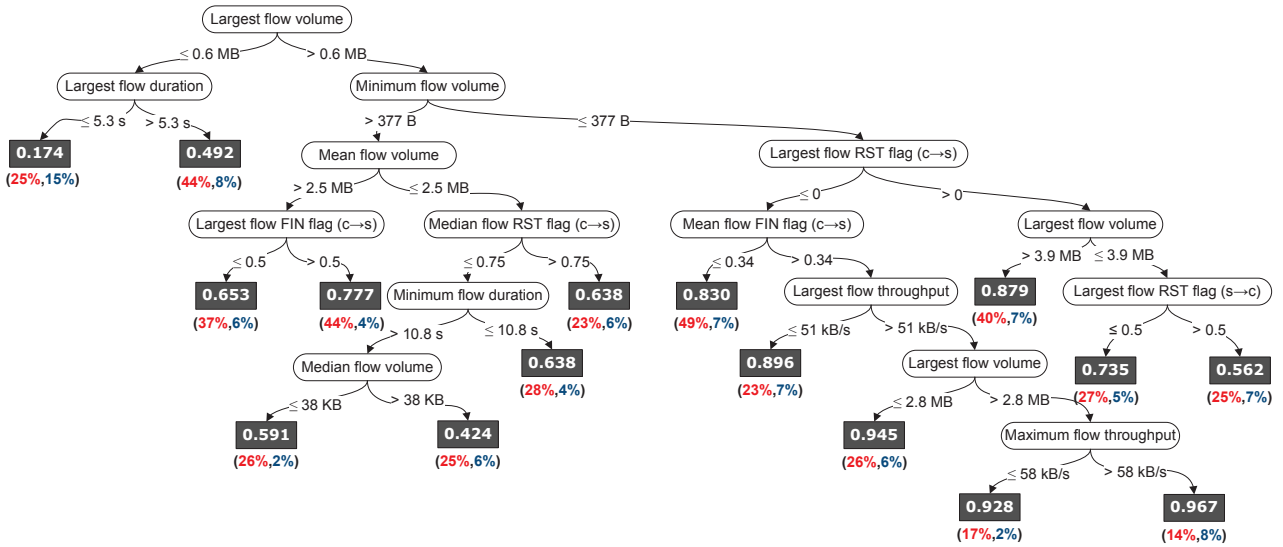


Figure 9: Accuracy vs. feature set size for `completed` vs. `abandoned` classification



(a) C4.5 decision tree model for `completed` vs. `abandoned` class pair



(b) M5P regression tree model for video download completion

Figure 10: Pruned decision and regression tree models for $\tau \leq 10$ seconds using All feature set. The tuples below leaf nodes represent (error%,population size%).

decision tree model for `completed` vs. `abandoned` class pair and the regression tree model for video download completion are plotted in Figures 10(a) and (b), respectively. The tuples below the rectangular leaf nodes represent their error and population size. Due to space constraints, we only plot the tree models for $\tau \leq 10$ seconds which are useful for network operators to predict user engagement by observing only the initial 10 seconds of video streaming sessions.

From the model, network operators can identify network factors that may have the most impact on user engagement and make decisions to prioritize certain infrastructure upgrades. The features at the higher levels of a tree tend to have more distinguishing power and account for more population size than lower level features. The root node is session duration for Figure 10(a) and largest flow volume in Figure 10(b). However, it is noteworthy that the ordering of network factors in Figure 10 does not strictly determine their importance. First, trees shown in Figure 10 represent one of many candidate tree models generated during bagging – other candidate trees have different feature ordering. Second, these features are not independent – session duration and maximum flow volume (top two features in Figure 10(a)) jointly account for session throughput and largest flow throughput to some extent.

The paths from the root node to leaves represent the equivalent rule sets, which inform network operators of the

interdependence among multiple features. For the regression tree in Figure 10(b), if largest flow volume is ≤ 0.6 MB and largest flow duration is ≤ 5.3 seconds then video download completion prediction is 17.4%. In contrast, if largest flow volume is ≤ 0.6 MB and largest flow duration is > 5.3 seconds then video download completion prediction is increased to 49.2%. Moreover, for the decision tree in Figure 10(a), if session duration > 5.9 seconds and maximum flow volume is ≤ 90 KB after the first 10 seconds then our model predicts that the video session will be abandoned with 19% error probability.

A network operator can influence many network factors to improve user engagement. The feature splits in Figure 10 provide network operators actionable insights for this purpose. For example, the decision tree predicts a session to be abandoned if cell DCH user count is larger than a threshold under certain conditions. Most cellular network users are covered by multiple cell sectors, and handover algorithms use signal quality and sector load to determine which sector each user should receive data from. The thresholds used for handover are typically fixed at a single global value. However, the feature splits in Figure 10(a) suggest that the network operator can tolerate a higher cell sector load threshold for sessions with higher throughput variance than sessions with lower throughput variance (24 vs. 13 users occupying DCH channels).

Limitations. Below, we discuss the limitations of our analysis and results. First, our results are based on traces from a single video service provider that uses progressive download with byte-range requests. Therefore, our findings may not be representative of video service providers that use other streaming methods. Second, our user engagement model cannot differentiate between video abandonment due to network-related issues and due to lack of user interest. Distinguishing between these two cases requires either client- or server-side information, which is not available to network operators.

5. RELATED WORK

Prior related studies can be categorized based on whether they use network-side or user-side instrumentation.

5.1 Network-side Instrumentation

Our study builds upon previous work by Gill *et al.* [14], Finamore *et al.* [12], and Erman *et al.* [10]. Each of these studies characterized the abandonment rate of video streaming sessions by collecting passive network traces at a campus edge network, 5 different wired edge locations, and a cellular network, respectively. To estimate video quality, these studies use deep-packet-inspection techniques (*e.g.*, [20]) to understand the video provider protocol. Our finding that 77% of video sessions are not completely streamed is closest to Finamore's result (80%), whereas Gill and Erman found lower abandonment rates (50% and 60%, respectively). All these results indicate that abandonment rates are high. Based on the ratio of download rate to encoding bitrate of video, Gill *et al.* concluded that approximately 20% of the video streaming sessions were interrupted due to poor performance. However, we find that average throughput is not always a good indicator of abandonment rate.

Our work makes two significant contributions on top of these studies. First, in order to measure abandonment, previous studies relied on deep-packet-inspection to extract information beyond TCP/IP headers, which requires prohibitive computational resources to employ at the scale of network carriers and can pose privacy problems in practice. Our work demonstrates that we can accurately measure abandonment without such information. Second, these studies did not provide insight into how network characteristics and performance impact abandonment rates. Our study is the first to examine the relationship between mobile network factors and user engagement and the first to provide guidance on how operators can reduce video abandonment.

5.2 Client-side Instrumentation

In [19], Rao *et al.* conducted an active measurement study of video streaming traffic from YouTube and Netflix. They proposed models to express various properties of completed and interrupted video streaming traffic as a function of the video parameters. However, the authors did not study user engagement because this work is based on active measurement data.

Dobrian *et al.* conducted a large scale, passive, user-side study to understand the impact of video streaming quality on user engagement [9]. They used video player instrumentation to quantify video streaming quality metrics such as join time, buffering ratio, average bitrate, rendering quality, and rate of buffering. Their analysis showed that buffering ratio has the largest impact on user engagement for non-live content and average bitrate significantly impacts user engagement for live content. Krishnan and Sitaraman also conducted a large scale, passive, user-side study to understand the impact of video streaming quality on user engagement [15]. They quantified the impact of video streaming quality metrics on user engagement using quasi-experimental designs. In [7, 8], Balachandran *et al.* developed a QoE model using various user-side video quality metrics. Specifically, they developed a decision tree based machine learning model to predict the extent of the video watched by users. For the two-class problem (completed vs. interrupted/abandoned), their trained model achieved up to 70% accuracy which progressively decreases as the number of classes is increased.

While these studies analyzed user engagement using data collected via video player instrumentation, our work focuses on characterizing and modeling user engagement using network-side measurements. Modeling user engagement using network-side data is particularly important for network operators because they do not have access to video player instrumentation data. Interestingly, our model based on network-side data can predict whether a user completely downloads a video with more than 87% accuracy, which is significantly better than the client-side model developed by Balachandran *et al.* Furthermore, since our model does not require client-side instrumentation, it can be used by any network operator, not just the video content provider.

6. CONCLUSIONS

This paper represents the first characterization of mobile video streaming performance and models its impact on user engagement from the perspective of network operators. We observed that many network features exhibit strong correlation with abandonment rate and skip rate. Our proposed model achieved more than 87% accuracy by observing only the initial 10 seconds of video streaming sessions. Overall, we conclude that the proposed model based on standard radio network statistics and/or TCP/IP header information can

be successfully used by network operators to predict video abandonment. Our model is useful for network operators to continuously monitor at scale to proactively mitigate the factors that can adversely impact user engagement.

In future, we plan to evaluate our proposed model on other types of networks (*e.g.*, DSL, WiFi). Recall from our evaluations that the core network feature set alone can provide most of the accuracy gain. Towards this end, we can reuse the core network feature set, which is not specific to cellular network infrastructure.

Acknowledgements

We thank our shepherd, Ranveer Chandra, and the anonymous reviewers for their useful feedback on this paper.

7. REFERENCES

[1] 3GPP Self-Organizing Networks. http://www.3gpp.org/SON.

[2] Multi-Resolution Land Characterization (MRLC) consortium, National Land Cover Database 2006 (NLCD 2006). `http://www.mrlc.gov/nlcd2006.php`.

[3] RFC 2046, Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types. `http://tools.ietf.org/html/rfc2046`.

[4] U.S. Geological Survey, National Elevation Dataset. `http://ned.usgs.gov/`.

[5] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2012–2017. Technical report, Cisco, 2013.

[6] V. Aggarwal, R. Jana, K. Ramakrishnan, J. Pang, and N. Shankaranarayanan. Characterizing fairness for 3G wireless networks. In *IEEE LANMAN*, 2011.

[7] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang. A quest for an internet video quality-of-experience metric. In *11th ACM Workshop on Hot Topics in Networks (HotNets-IX)*, 2012.

[8] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang. Developing a predictive model of quality of experience for internet video. In *ACM SIGCOMM*, 2013.

[9] F. Dobrian, V. Sekar, A. Awan, I. Stoica, D. Joseph, A. Ganjam, J. Zhan, and H. Zhang. Understanding the impact of video quality on user engagement. In *ACM SIGCOMM*, 2011.

[10] J. Erman, A. Gerber, K. Ramakrishnan, S. Sen, and O. Spatscheck. Over the top video: The gorilla in cellular networks. In *ACM IMC*, 2011.

[11] T. Fawcett. ROC Graphs: Notes and Practical Considerations for Researchers. Technical report, HP Laboratories, 2004.

[12] A. Finamore, M. Mellia, M. Munafo, R. Torres, and S. R. Rao. YouTube Everywhere: Impact of Device and Infrastructure Synergies on User Experience. In *ACM IMC*, 2011.

[13] J. Fry, G. Xian, S. Jin, J. Dewitz, C. Homer, L. Yang, C. Barnes, N. Herold, and J. Wickham. Completion of the 2006 National Land Cover Database for the Conterminous United States. *Photogrammetric Engineering & Remote Sensing (PE&RS)*, 77(9):858–864, 2011.

[14] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. YouTube Traffic Characterization: A View From the Edge. In *ACM IMC*, 2007.

[15] S. S. Krishnan and R. K. Sitaraman. Video stream quality impacts viewer behavior: Inferring causality using quasi-experimental designs. In *ACM IMC*, 2012.

[16] Z. Li, J. Lin, M.-I. Akodjenou, G. Xie, M. A. Kaafar, Y. Jin, and G. Peng. Watching videos from everywhere: a study of the PPTV mobile VoD system. In *ACM IMC*, 2012.

[17] J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann, 1993.

[18] R. J. Quinlan. Learning with continuous classes. In *5th Australian Joint Conference on Artificial Intelligence*, 1992.

[19] A. Rao, Y. sup Lim, C. Barakat, A. Legout, D. Towsley, and W. Dabbous. Network characteristics of video streaming traffic. In *ACM CoNEXT*, 2011.

[20] R. Schatz, T. Hobfeld, and P. Casas. Passive YouTube QoE Monitoring for ISPs. In *Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS)*, 2012.

[21] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, S. Venkataraman, and J. Wang. A first look at cellular network performance during crowded events. In *ACM SIGMETRICS*, 2013.

[22] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang. Characterizing geospatial dynamics of application usage in a 3G cellular data network. In *IEEE INFOCOM*, 2012.

[23] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang. A first look at cellular machine-to-machine traffic - large scale measurement and characterization. In *ACM SIGMETRICS/Performance*, 2012.

[24] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang. Characterizing and modeling Internet traffic dynamics of cellular devices. In *ACM SIGMETRICS*, 2011.

[25] L. Song and J. Shen, editors. *Evolved Cellular Network Planning and Optimization for UMTS and LTE*. CRC Press, 2010.

[26] E. B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

[27] I. H. Witten, E. Frank, and M. A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2011.