Large Scale Data Analytics of User Behavior for Improving Content Delivery

Amanpreet Kaur[1] Sukhpreet Kaur[2]
[1,2]Department of Computer Science and Engineering
[1,2]Shaheed Udham Singh College of Engineering & Technology, Tangori, Punjab, India

Abstract— The Internet is fast becoming the de facto content delivery network of the world, supplanting TV and physical media as the primary method of distributing larger files to ever-increasing numbers of users at the fastest possible speeds. Recent trends have, however, posed challenges to various players in the Internet content delivery ecosystem. These trends include exponentially increasing traffic volume, increasing user expectation for quality of content delivery, and the ubiquity and rise of mobile traffic. For example, exponentially increasing traffic—primarily caused by the popularity of Internet video—is stressing the existing Content Delivery Network (CDN) infrastructures. Similarly, content providers want to improve user experience to match the increasing user expectation in order to retain users and sustain their advertisement based and subscription-based revenue models. Finally, although mobile traffic is increasing, cellular networks are not as well designed as their wireline counterparts, causing poorer quality of experience for mobile users. These challenges are faced by content providers, CDNs and network operators everywhere and they seek to design and manage their networks better to improve content delivery and provide better quality of experience.

Keywords: Data Analytics, Machine Learning, User Behavior, User Experience, Content Delivery, Peer-To-Peer, Video Streaming, Web Browsing

I. INTRODUCTION

Internet today is largely a content driven network. Starting from simple data transfer between two computers directly connected by a wire, the complexity of content delivery over the Internet has come a long way to include several complex applications such as adaptive video streaming, peer-to-peer file sharing, massively multiplayer online gaming, cloud storage, and cloud-based computation. Over the years, there have been several innovations to support the growth of content delivery, both in protocols used for delivering content, as well as in the infrastructure to support and improve new content delivery applications. Today the main challenges faced by the content delivery ecosystem include:

A. Exponentially Increasing Traffic:

Traffic over the Internet has been exponentially increasing over the past few years with predictions that it will quadruple by 2016 [9]. In 2011, 51% of the traffic on the Internet consisted of video. Market predictions suggest that more than 90% of the traffic on the Internet will be video in 2015. This development is hugely due to the very low costs of obtaining video over the Internet. In fact, we have come to a point where Internet video would replace traditional TV viewership. However, recent studies have shown signs of the CDN infrastructure being stressed by the increasing traffic [95]. This is placing an onus on the CDNs to distribute content efficiently. Some of the techniques to augment the existing infrastructure to handle the increasing load, that have received significant industry attention, include hybrid P2P-CDN design which combines the P2P and client-server models for content delivery [63, 64], and federated CDN models [48, 105].

B. Increasing User Expectation:

Another side effect of the decreasing costs of obtaining content, specifically video, over the Internet includes the rise of several content providers competing for users' attention. With multiple competitors in the market, user expectation for the quality of the content has been steadily growing [3]. Content providers also want to maximize user engagement in order for better gains from their advertisement-based or subscription-based business models. This has led to improving user experience as one of the primary goals of content delivery today. This trend has spawned several third-party optimization and analytics services that operate for optimizing user experience given the limited resources using techniques such as cross-CDN optimization.

C. Rise of Mobile:

Another major development in the past few years in the content delivery scenario is the rise of mobile traffic. Today, mobile traffic constitutes 28% of the overall Internet traffic with one in four visits to websites arising from a smartphone. Mobile web usage is estimated to increase eleven-fold between 2013 and 2016 with around 50 billion connected mobile devices on the Internet by 2020 [9]. However, today, cellular networks are slower than wireline networks primarily because the mobile architecture was not designed for the web. Recent studies on top websites showed that loading them via wireline connectivity lead to an average of 2 seconds when compared to 9 seconds via mobile connectivity [3]. Increasing user expectation is posing challenges to the cellular network operators to configure their networks by adding the right infrastructure in order to provide wireline compatible user experience over cellular networks.

D. Complex Multiple Party Involvement:

The content delivery ecosystem today consists of several parties including content providers, CDNs, network operators, third party analytics and CDN optimization services etc. The increasing complexity and multiple party involvement makes it much harder to pinpoint problems in delivering content to users [85]. Harder still is to estimate whether a proposed feature or fix to a system or protocol will have a measurable impact on users or revenue. If a user complains about a video not loading on her smartphone, what can the content provider do? Perhaps the problem is the buffering time or the bitrate selected by the CDN or third-party optimization services. Or it could be caused by overload at the CDN server. It could otherwise be simply because the user's 2 device switched from a 4G to a 3G connection. Even if one or more of these issues are addressed, content providers, network operators, CDNs etc. traditionally have no

way to know if the problem is widespread or if the fix had the desired impact.

II. BACKGROUND AND SCOPE

A. Content Delivery Ecosystem:

We begin with a brief overview of the different players in the content delivery ecosystem in the Internet today. Each of these players have access to rich data that can be used towards improving content delivery.

1) Content providers encompass a wide variety of media and e-commerce players who provide content on the Internet primarily for revenue. These include news websites (e.g., CNN), social networking websites (e.g., Facebook, Yelp), and also video providers (e.g., HBO, ABC). Content providers want to maximize their revenues from subscription-based and advertisement-based business models while trying to minimize content distribution costs. To this end, content providers have business arrangements with CDNs (e.g., Akamai, Limelight) to distribute their content across different geographical locations. Similarly, more recently they also have contracts with third-party analytics services (e.g., Google Analytics, Ooyala [28]) and optimization services (e.g., Conviva [11]) to understand and improve user experience.
2) Content Distribution Networks consist of distributed system of servers allocated across different geographical regions for serving content to end users with high performance and availability. CDNs provide content providers a cost-effective mechanism to offload content from their infrastructure. Hence CDNs need to allocate their resources efficiently across user population. CDNs aim to design their delivery infrastructure to minimize their delivery costs while maximizing their performance. Towards this end there have been many studies and proposals on efficient design of the CDN infrastructure. Although CDNs primarily serve content using dedicated servers operated by them, more recently there have been proposals for other designs including hybrid models that make use of peer-to-peer mechanisms on user-owned devices and also federation across multiple CDNs. CDNs collect a large amount of logs daily on user behavior. Tailoring CDN design based on the user behavior to improve content delivery with minimal costs is an interesting problem faced by CDNs.
3) Internet Service Providers (ISPs) form the backbone of the Internet by delivering the content from the CDNs and content providers to the end users. Traffic on the Internet has been increasing exponentially over the years. In particular, with the advent of smartphones and new wireless technologies such as 3G and 4G, mobile traffic is on the rise. But, unlike the other players, ISPs do not have access to detailed client-side or server-side logs making it more challenging to extract user behavior information from network traces alone. However, extracting user behavior information from network traces can help ISPs can use this data towards improving content delivery by configuring their network better.
4) Cross-CDN optimization services help content providers work with multiple CDNs towards delivering content for better resilience. Recent studies have also argued that crossCDN optimization can lead to improved content delivery [95]. There are also commercial players in the market that offer cross-CDN optimization services for content providers, especially in Internet video providers. These services also have access to user behavior at a fine grained level at a large-scale and make real-time decisions on which CDNs to serve a content based on current network conditions. Such optimizations are towards improving user experience while mainting low content-delivery costs. Similarly, there are also several third party analytics services that collect user access logs information at a large-scale to translate them into insights towards improving revenue for the user at low content delivery costs.
5) Users are ultimately the source of all revenue and the sink for all content produced by this ecosystem. Users prefer services that give them a better cost-experience tradeoff and hence content providers need to deliver the best possible quality of experience to the users at the minimum cost. At the same time, we now have the ability to collect, store and analyze finegrained access patterns and behavior from the users. This information, even at an aggregate level can help the content delivery system to minimize costs by provisioning resources appropriately and also improve individual user's quality of experience by potentially even personalizing content delivery based on their preferences.

B. Big Data Analytics:

Big data analytics is now extensively used in fields of computer science such as recommendation systems, search and information retrieval, computer vision and image processing, and is making its foray into the real world in terms of business intelligence, healthcare and supply chain analysis. It is also used even within the domain of networks in areas such as network security. Several technology innovations in the past decade were essential in being able to analyze massive volumes of data. The MapReduce framework [22] is perhaps the innovation that heralded the area of big data analytics, and open-source versions of MapReduce such as Hadoop [18] and the distributed HDFS [19] filesystem allow researchers to rapidly gather insights from more data that can fit on any single machine. Hadoop, Hive [20] and recent advancements such as Spark [36] make short work of analyzing massive quantities of data. Keeping up with the infrastructure developments, there have been algorithms and libraries that are specifically suited to data mining and machine learning, ranging from the more traditional tools such as Weka [42] and Scikit-learn [33] to tools built for big data such as Graphlab [17]. In our work, we make heavy use of the Hadoop and HDFS infrastructure to process logs from various sources. Hive helped us easily interact with the data.

Fig. 1: Flow of information during content delivery from CDNs to ISPs to Users.

C. Thesis Scope:

There are several ways in which big data analytics could potentially help improve content delivery. In this dissertation the main focus is on how big data analytics on user behavior can be of use to improve content delivery over the Internet. Big data analytics can also be used to predict network throughput, to build better control plane and data plane mechanisms and hence improve content delivery.

III. THESIS STATEMENT AND APPROACH

Big data analytics tools can be used to improve content delivery not just by informing system design decisions, but also by building automatic models that can be directly used in decision processes. Based on this idea, this thesis argues the following:

It is possible for different players to use big data analytics on user behavior data for learning predictive models that can be used to improve content delivery even when the data collected does not explicitly contain user behavior information.

We performed large-scale studies conducted on data from three different perspectives:

A. Inform CDN Resource Management:

We first look at how we can use large scale data analytics on user behavior data for gaining insights that can be used to inform system design changes for handling the ever-increasing traffic at CDNs. With exponentially increasing traffic, CDNs wish to minimize bandwidth costs while improving performance in terms of availability. This implies that they must provision and select servers appropriately. Most commonly, CDNs serve content from dedicated servers. However, there have been recent proposals to improve this scheme using hybrid P2P CDNs and federated CDNs. We show that it possible to analyze data already collected from existing CDNs to evaluate the advantages of these new proposals. We also observe several interesting user behavior patterns that have important implications to these designs. Using aggregate client access logs, we analyze proposals to improve provisioning that have seen significant industry traction in the recent times. This study shows that even simple data analytics can be very useful for improving content delivery.

B. Develop Internet Video Quality of Experience (QoE) models:

Next we look at the problem of increasing user expectation for quality of content. Content providers are very keen on understanding and improving users' QoE since it directly affects their revenue. For instance, previous studies in Internet video viewing have shown that one percent increase in buffering can lower viewing time by as much as three minutes [72]. We use big data analytics, particularly machine learning algorithms to build predictive models that capture users QoE. Using simulation studies, we also show that using this model to pick the best bitrate and CDN server for a video session can lead to as much as 20% improvement in user engagement. This study shows that in addition to informing system design decisions, big data analytics can also be used to build predictive models that can potentially be directly used in systems.

C. Model Cellular Web Browsing Quality of Experience:

Although QoE for various content delivery applications over the Internet has been improving over the years, QoE in certain spaces are still not as good. Cellular networks do not provide good QoE even for simple applications such as web browsing. Further, unlike other players in the ecosystem cellular network operators do not have access to detailed server-side and client-side logs of user behavior. We show that machine learning algorithms can be used to extract user behavior information from network traces and build predictive models for web browsing QoE from this extracted user behavior information. We show that these models can be used by network operators to monitor and configure their networks for better QoE. This study shows that big data analytics can be useful even in scenarios where the player does not have explicit user behavior information.

IV. LARGE-SCALE DATA ANALYTICS FOR CDN RESOURCE MANAGEMENT

Traffic on the Internet has been steadily growing, and it is predicted to quadruple by 2018 [9]. Dealing with exponentially increasing traffic volumes is a significant challenge for different play- ers in the content delivery ecosystem. Video accounts for a large fraction of the traffic on the Internet today, and its share is growing over time. In 2011, around 51% of Internet traffic was video [9], and market predictions suggest that video will account for over 90% of the traffic on the Internet in 2015. There are already signs that the CDN infrastructure is being stressed [94, 95] by the increasing traffic and this has placed the onus on CDNs for managing their resources more efficiently.

Hybrid P2P-CDNs and telco-CDN federation are two CDN infrastructure augmentation strategies to alleviate the stress caused by increasing traffic. These two approaches have received significant industry attention recently.

1) Telco-CDN federation is based on the recent development amongst various CDNs oper- ated by telecommunication companies to federate by interconnecting their networks and compete directly with the traditional CDNs [8, 48, 70, 105]. This would enable users to reach CDN caches that are closer. Interconnecting resources across telco-CDNs would also ensure better availability and will benefit the participating ISPs in terms of provisioning costs [8].
2) A hybrid strategy of serving content from dedicated CDN servers using P2P technology (e.g., [63, 64]) has been around for a while in the research literature but has only recently seen traction in the industry [4, 118]. A hybrid P2P-CDN approach would provide the scalability advantage of P2P along with the reliability and manageability of CDNs.

Given that several industry efforts and working groups are underway for both these approaches [4, 48, 70, 105, 118], it is crucial to analyze the potential benefits that these CDN augmentation strategies can offer for Internet video workloads. Our main contribution in this chapter is in using large-scale data analytics for identifying video access

patterns that have significant implications to these two strategies and analyzing the potential benefits of these two strategies. To the best of our knowledge, there has not been any previous large-scale study on the benefits of federated telco-CDN infrastructures. While there is prior work on analyzing the benefits of P2P augmentation, these were done long before Internet video became main- stream [63, 64], and hence were ahead of their times. Moreoever, the significant improvement in big data analytics approaches and the ability to collect large amounts of data puts us in a better position to do this study today. We leverage on these to revisit the benefits of P2P augmentation on today's traffic and suggest new improvements.

Using a dataset of around 30 million VOD and live sessions collected over two months from viewers across the United States, we identify several video viewing patterns that have implications to these two designs including:

A. Regional Interest:

Typically, we observe significant population induced difference in load across different regions. But, for live events with regional biases like a local team playing a match, we observe significantly skewed access rates from regions that exhibit low load in the typical case.

B. Temporal Shift in Peak Load:

We observe strong diurnal effects in access patterns and also confirm temporal shifts between regions in the demand for VOD objects using cross-correlation analysis. The temporal shift in access pattern is caused by time zone differences. The video access load peaks at around 8pm local time for each region.

C. Evolution of Interest:

We observe that peak demand for VOD objects occur on the day of release and the decay in demand in the subsequent days can be modeled using an exponential decay process. Interestingly, overall user viewing patterns are very different across genres. For example, decay rates of news shows are much higher than TV series episodes. Also, TV series episodes have highly predictable and stable demand from week to week.

D. Synchronized Viewing Patterns:

While we expect synchronous viewing behavior for live video, we unexpectedly observe synchrony in the viewership of VOD objects. This is especially true for popular shows during the peak demand period.

E. Partial Interest in Content:

We reconfirm prior observations that users watch only part of the video during a session [50, 76]. For instance, in the case of VOD, a significant fraction of the viewers typically watch only the first 10 minutes of the video before quitting. We observe that around 4.5% of the users are "serial" early-quitters while 16.6% of the users consistently watch videos to completion.

We develop simple models to capture the deployment of federated telco-CDNs and analyze the potential benefit of federation to increase availability and reduce provisioning required to serve video workloads. We also revisit the potential benefits that P2P-assisted architectures pro- vide in the light of these video access patterns. Our key findings are:

1) Telco-CDN federation can reduce the provisioning cost by as much as 95%. VOD work- loads benefit from federation by offloading daily peak loads and live workloads benefit by offloading unexpected high traffic triggered by regional events.
2) Using P2P can lead up to 87% bandwidth savings for the CDNs during peak access hours. Employing a strategy to filter out users who quit early by serving them using P2P can alone lead to 30% bandwidth savings for VOD traffic and 60% savings for live traffic.

1) Dataset:

The data used for this analysis was collected by conviva.com in real time using a client-side instrumentation library in the video player that collects information pertaining to a session. This library gets loaded when the user watches video on conviva.com's affiliate content providers' websites. The library also listens to events from the player (e.g., seek, pause). The data is then aggregated and processed using Hadoop [18].

We focus on two of the most popular content providers (based in the US). These two providers appear consistently in the Top 500 sites in overall popularity ranking. Our analysis is based on data queried over two months—January 2012 and March 2012—and consists of over 30 million video viewing sessions during this period. We classify the video content into two categories:

VOD: The first provider serves VOD objects that are between 35 minutes and 60 minutes long. These comprise TV series episodes, news shows, and reality show episodes.

Live: The second provider serves sports events that are broadcast while the event is happening, and hence the viewing behavior is synchronized.

The VOD dataset consists of approximately 4 million users and 14 million viewing sessions and covers 1,000 video shows. The live dataset consists of around 4.5 million users and 16 million video viewing sessions covering around 10,000 different events. As in several prior studies on content popularity [51, 91], we also observe a heavy tailed Zipf distribution for overall popularity of objects for both VOD and live. While most objects have few accesses over the two months, some extremely popular objects had significant viewership. On average, users viewed 4 VOD objects and 2 live events during the course of a month, which amounts to 85 minutes of VOD objects and 65 minutes of live events per month. We also observed a few heavy hitters who watched upwards of 500 videos per month on these websites.

Session characteristics: In order to understand user behavior, we look at several characteristics of individual video sessions. Specifically, for each session we collected the following information:

ClientID: The first time a client watches a video on the player, a unique identifier is assigned to the player and stored in a Flash cookie to be used by subsequent views.
Geographical location: Country, state, and city of the user.

Provider: Information on the AS/ISP from which the request originated.

Session events: Start time and duration of the session along with details on other user interaction events like

pausing and stopping. Session Performance: Average bitrate, estimated bandwidth etc. during the playback.

Content: Information on the content being watched, in particular, the name of the video and the actual duration of the content.

2) Analyzing Telco-CDN federation:

The tremendous increase in video traffic on the Internet over the past few years has caused great challenges for ISPs. The increasing traffic has strained the ISP networks leading to higher costs and maintenance issues. However, this trend has not significantly contributed too much increase in revenue for ISPs since most of the video content is served by content providers using CDNs. As a result, several ISPs have started deploying proprietary CDNs inside their own network, providing services to content providers to serve content from caches closer to customers. This could result in increased revenue for the ISPs along with traffic reduction caused by content caching [8].

There has also been recent developments that point to interest among ISPs to deploy telco CDN federations by consolidating their CDN capacity and offering services to users in other ISPs [48, 70, 105]. By interconnecting telco-CDNs, consumers can reach CDN caches that are closer and are also ensured of better availability and service in case of local network conges- tion. Pooling resources across ISPs could potentially benefit the participating ISPs in terms of provisioning costs. It also enables ISPs to provide a global "virtual CDN" service to the content providers [8].

3) User Access Patterns:

We observed video access patterns for live and VOD content that have implications to telco- CDN federation. For instance, in our live dataset, we observed unexpected surges in demand for certain objects from regions which can potentially be served using spare capacity in servers in other regions if CDNs federate. Similarly, we observed strong temporal shifts in when specific regions hit peak load in the VOD dataset opening up new possibilities for handling peak loads using federation. We finally also present statistics on ISP coverage and their relative performance which also have important implications when ISPs decide to federate.

a) Regional Interests:

Typically, the number of accesses to a particular content from a geographical region is strongly correlated with the total population of the region. However, in our live dataset, we observed anomalies in the case of content with region-specific interest (e.g., when a local team is playing a game). Such unexpected surges in demands triggered by regional interests can potentially be served from servers in other regions if CDNs federate.

b) Implications:

The skeweness in access rates caused by regional interest is an important factor to be considered while provisioning the delivery infrastructure to handle unexpected high loads. Federation can potentially help offload such unexpected surges triggered by regional interests by using spare capacity in CDNs in other regions.

V. Developing a Predictive Model for Internet Video Quality-of-Experience

Video streaming forms the majority of traffic on the Internet today and its share is growing exponentially with time [9]. This growth has been driven by the confluence of low content delivery costs and the success of subscription-based and advertisement-based revenue models [10]. At the same time, users expectations for video quality are steadily rising [5]. Content providers want to maximize user engagement in order for better gains from their advertisement-based and subscription-based revenue models. Given this context, there is agreement among leading industry and academic initiatives that improving users' quality of experience (QoE) is crucial to sustain these revenue models [72, 87].

Despite this broad consensus, our understanding of Internet video QoE is limited. This may surprise some, especially since QoE has a very rich history in the multimedia community [23, 24, 40]. The reason is that Internet video introduces new effects with respect to both quality and experience. First, traditional quality indices (e.g., Peak Signal-to-Noise Ratio (PSNR) [29]) are now replaced by metrics that capture delivery-related effects such as rate of buffering, bitrate delivered, bitrate switching, and join time [5, 57, 72, 96, 113]. Second, traditional methods of quantifying experience through user opinion scores are replaced by new measurable engagement measures such as viewing time and number of visits that more directly impact content providers' business objectives [5, 113]. Meeting these requirements, however, is challenging because of three key factors:

A. Complex relationship between quality and engagement:

Prior measurement studies have shown complex and counter-intuitive effects in the relationship between quality met- rics and engagement. For instance, one might assume that increasing bitrate should in- crease engagement. However, the relationship between bitrate and engagement is strangely non-monotonic [72].

B. Dependencies between Quality Metrics:

The metrics have subtle interdependencies and have implicit tradeoffs. For example, bitrate switching can reduce buffering. Similarly, aggressively choosing a high bitrate can increase join time and also cause more buffering.

C. Confounding Factors:

There are several potential confounding factors that impact the relationship between quality and engagement: the nature of the content (e.g., live vs. Video on Demand (VOD), popularity), temporal effects (e.g., prime time vs. off-peak), and user- specific attributes (e.g., connectivity, device, user interest) [87].

1) Motivation and Challenges:

In this section, we provide a brief background of the problem space and highlight the key chal- lenges in developing a unified QoE model using data-driven techniques.

2) Problem scope:

Multiple measurement studies have shown that video quality impacts user engagement [72, 87]. Given that engagement directly affects advertisement- and subscription-based revenue streams, there is broad consensus across the different

players in the Internet video ecosystem (content providers, video player designers, third-party optimizers, CDNs) on the need to optimize video quality according to these metrics. In this study, we focus on the fraction of video that the user viewed before quitting as the measure of engagement and the following industry-standard quality metrics:

3) Average bitrate:

Video players typically switch between different bitrate streams during a single video session. Average bitrate, measured in kilobits per second, is the time average of the bitrates played during a session weighted by the time duration each bitrate was played.

4) Join time:

This represents the time it takes for the video to start playing after the user initiates a request to play the video and is measured in seconds.

5) Buffering ratio:

It is computed as the ratio of the time the video player spends buffering to the sum of the buffering and play time and is measured in percentage.

6) Rate of buffering:

It captures the frequency at which buffering events occur during the session and is computed as the ratio of the number of buffering events to the duration of the session.

VI. Conclusion

In this dissertation we showed that applying large scale data analytics is a step forward towards solving some of the main challenges faced by the various players in the content delivery. We showed that large scale data analytics and machine learning algorithms can be used as an effective tool to characterize user behavior in the wild to inform various content delivery system design decisions. This chapter concludes the dissertation with a summary of the approach and contributions followed by a discussion of the lessons learned and remaining open problems in this space.

References

[1] Tom Mitchell. Machine Learning. McGraw-Hill.
[2] Jeffrey C. Mogul. The Case for Persistent-Connection HTTP. In SIGCOMM, 1995.
[3] Ricky K. P. Mok, Edmond W. W. Chan, Xiapu Luo, and Rocky K. C. Chang. Inferring the QoE of HTTP Video Streaming from User-Viewing Activities. In SIGCOMM W-MUST, 2011.
[4] Tongqing Qiu, Zihui Ge, Seungjoon Lee, Jia Wang, Qi Zhao, and Jun Xu. Modeling channel popularity dynamics in a large IPTV system. In Proc. SIGMETRICS, 2009.
[5] R. Powell. The federated cdn cometh. May 2011. TelecomRamblings.com.
[6] S. Guha, S. Annapureddy, C. Gkantsidis, D. Gunawardena, and P. Rodriguez. Is HighQuality VoD Feasible using P2P Swarming? In Proc. WWW, 2007.
[7] Mehran Sahami, Susan Dumais, David Heckerman, and Eric Horvitz. A Bayesian Approach to Filtering Junk E-Mail. In AAAI, 1998.
[8] M. Zubair Shafiq, Lusheng Ji, Alex X. Liu, Jeffrey Pang, and Jia Wang. Characterizing Geospatial Dynamics of Application Usage in a 3G Cellular Data Network. In INFOCOM, 2012.
[9] F. Donelson Smith, Felix Hernndez Campos, Kevin Jeffay, and David Ott. What TCP/IP Protocol Headers Can Tell Us About the Web. In SIGMETRICS, 2001.
[10] Han Hee Song, Zihui Ge, Ajay Mahimkar, Jia Wang, Jennifer Yates, Yin Zhang, Andrea Basso, and Min Chen. Qscore Proactive Service Quality Assessment in a Large IPTV System. In Proc. IMC, 2011.
[11] Srikanth Sundaresan, Nick Feamster, Renata Teixeira, and Nazanin Magharei. Measuring and Mitigating Web Performance Bottlenecks in Broadband Access Networks. In IMC, 2013.
[12] Thomas Tullis and William Albert. Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2008. ISBN 0123735580, 9780123735584.
[13] Mark Watson. Http adaptive streaming in practice. In MMSys - Keynote, 2011.
[14] I H Witten and E Frank. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2000.
[15] C Wu, B Li, and S Zhao. Diagnosing Network-wide P2P Live Streaming Inefficiencies. In Proc. INFOCOM, 2009.
[16] Wei Xu, Ling Huang, Armando Fox, David Patterson, and Michael Jordan. Detecting large-scale system problems by mining console logs. In Proc. SOSP, 2009.
[17] H Yin et al. Inside the Bird's Nest: Measurements of Large-Scale Live VoD from the 2008 Olympics.
[18] Hao Yin, Xuening Liu, Tongyu Zhan, Vyas Sekar, Feng Qiu, Chuang Lin, Hui Zhang, and Bo Li. Design and Deployment of a Hybrid CDN-P2P System for Live Video Streaming: Experiences with LiveSky. In Proc. ACM Multimedia, 2008.
[19] H Yu, D Zheng B Y Zhao, and W Zheng. Understanding User Behavior in Large-Scale Video-on-Demand Systems. In Proc. Eurosys, 2006.
[20] Renjie Zhou, Samamon Khemmarat, and Lixin Gao. The impact of YouTube recommendation system on video views. In Proc. IMC, 2010.

