Dataset on Usage of a Live & VoD P2P IPTV Service

Yehia Elkhatib, Mu Mu, Nicholas Race
School of Computing and Communications, Lancaster University, LA1 4WA, United Kingdom
Email: {i.lastname}@lancaster.ac.uk

***Abstract*—This paper presents a dataset of user statistics collected from a P2P multimedia service infrastructure that delivers both live and on-demand content in high quality to users via different platforms: PC/Mac, and set top boxes. The dataset covers a period of seven months starting from October 2011, exposing a total of over 94k system statistic reports from thousands of user devices at a fine granularity. Such rich data source is made available to fellow researchers to aid in developing better understanding of video delivery mechanisms, user behaviour, and programme popularity evolution.**

I. Introduction

The increasing amount of multimedia content available online, the evolution of digital entertainment devices, and the ever growing popularity of social media has led to explosive consumption of audio-visual content in our daily lives. Video accounted for 66% of all consumer traffic in 2013, and is predicted to be 79% by 2018 [1]. By August 2013, 36% of UK households accessed television (TV) content over the internet at least once every week [2]. This continuous move towards internet TV has challenged the value chain of traditional linear broadcasting by allowing virtually any IP-based network, wired and wireless, to deliver audio-visual content.

Simultaneously, user expectations for quality have dramatically increased leaving any content of less than standard definition not acceptable [4]. It is hence becoming increasingly difficult to ignore the challenges of distributing multimedia content over best-effort packet-based IP networks. This has motivated significant work on designing better multimedia delivery systems. Of all content delivery mechanisms, peer-to-peer (P2P) IPTV has become an ideal candidate for energy efficient and low-cost delivery for commercial and user-generated multimedia content.

We introduce the Lancaster Living Lab [8], [10], a long-running P2P IPTV service infrastructure developed and maintained in Lancaster University. The Living Lab supports high quality live and on-demand content distribution and a converged service platform which covers consumer, personal and mobile devices. From this infrastructure, we collected a vast amount of fine-grain statistics reported by user devices. We present a large subset of this database, covering the period of October 2011 – April 2012 with a total of over 94k entries.

The remainder of this paper is organised as follows. §II introduces the system from which the dataset emanates. §III presents the dataset characteristics and collection methodology. §IV presents some examples of statistical analysis results produced from the data. Finally, §V concludes.

II. Lancaster Living Lab

The Lancaster Living Lab serves as a small IPTV service provider, ensuring the end-to-end delivery of high quality audio-visual services to university staff and students for the purpose of research and real-life evaluation of state-of-the-art technologies. This encompasses the development and operation of technical services, the provision of user devices, and the subsequent measurement and evaluation procedures. Content is made available under the Educational Recording Agency (ERA) licensing scheme. In this section, we describe the Living Lab service chain starting with content admission through to consumption by the users. The architecture of the Living Lab is summarised in Figure 1.

A. Content Ingestion

At the beginning of the IPTV service chain is the process of content ingestion. A headend node receives and de-multiplexes live terrestrial and satellite signals from a total of 82 TV and radio channels, including UK and European channels as well as Lancaster University student services. In addition to audio-visual content, Event Information Table (EIT) data is received and used for the production of Electronic Programme Guides (EPGs) which contain basic programme information (title, description and duration). This is used for enhancing programme listing and discovery.

B. Content Processing

Beyond content ingestion, the service infrastructure predominantly operates as a private cloud. Overall, there are 25 independently operating virtual machines (VMs) supporting key Living Lab services like video transcoding, live streaming, VoD streaming, and statistics service (capturing, parsing and maintaining millions of statistics reports from all clients).

Two groups of VMs prepare the content for live and on-demand viewing. The Live Service injects source video streams into the P2P platform, producing the P2P version of the stream along with the associated torrent file. Each Live Service instance functions as the primary seed for content, and announces its torrents to a local tracker, the address of which is embedded in the produced torrent file.

The VoD Service records programmes from 12 TV channels according to the EPG feed and stores each recorded TV programme as a standalone file onto a storage area network (SAN) back-end. As part of the P2P distribution, a torrent file for each programme is automatically created and announced to the

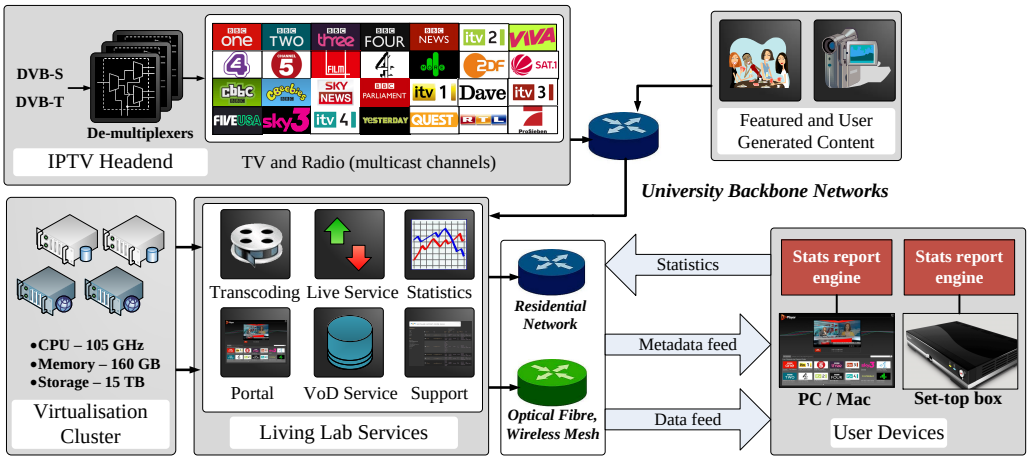


Fig. 1: Service infrastructure

tracker. These processes enable fully automated programme recording, ensuring the availability of content through the P2P VoD platform minutes after it has finished airing live.

The VoD Service maintains a repository of 8000 content items, which corresponds to around 14 days of on-demand content across all 12 VoD channels. The repository is managed following a FIFO strategy.

C. Delivery & User Base

At the other end of the chain, users receive the Living Lab services via the NextShare platform [3]. The key NextShare component is NextShare^{Core} which encapsulates the inner workings and underlying protocols of the content delivery platform. NextShare features are exposed through a common API which abstracts away the underlying signalling, distribution and networking logic. Using this API, a number of derivative end user applications have been made available. These are depicted in Figure 2 and described as follows.

- NextShare^{PC} - A multi-platform component which provides web browser integration with a light-weight plugin to interact with the NextShare^{Core} API.
- NextShare^{TV} - An integrated consumer electronics set-top box (STB) that is based upon the STB7200 system-on-chip technology from ST Microelectronics. Beside playback and video processing functions, a number of social networking features have also been integrated.
- NextShare^{Mobile} - A smartphone app that gives users rich control of STB features (i.e. discover content, manage playback session, etc.).

The distribution of Living Lab IPTV services covers a diverse spectrum of access networks, end devices and content consumers within and beyond the Lancaster University campus. The campus users include university students (13,000) and staff members (3,000). Content services on campus are reachable via Ethernet (100 Mbit/s) or 802.11g WiFi (54 Mbit/s) on both NextShare^{TV} and NextShare^{PC} platforms. Furthermore, the Living Lab IPTV services are made available to rural communities in Wray, a small village in north Lancashire, via a second generation wireless mesh network [7].

III. Dataset Characteristics

This section describes the dataset we present in this paper, which is publicly available at http://www.comp.lancs.ac.uk/~elkhatib/p2p14/ in CSV and SQL formats.

A. Data Collection

To provide functional and performance tests, a full chain of IPTV statistics services is constructed and maintained. Statistics report engines are developed and integrated in both the NextShare^{TV} and NextShare^{PC} platforms. The report engine listens to all system events, generates predefined status reports and submits them periodically (every 15 minutes) and triggered by certain events (e.g., start of media playback). Status reports include system status (e.g., IP address, CPU load), user activities (e.g., request for content, viewing time), video statistics (e.g., playback starts and playback stops), and granular P2P statistics (e.g., connected peers and the download and upload rates). A central load balanced frond-end web service receives statistical reports from all user clients which are subsequently parsed at the statistics processing server and maintained in dedicated statistics databases, holds over 15 million records of service information dating back to 2010.

To capture user activities, three time-coded events of media playback; *media_play_request*, *media_play_started* and *media_play_stop* are reported by all end devices to the statistics service. The *media play request* event records the timestamp and media information of user's request. This is usually triggered by a click on the icon of a live channel or VoD item. The *media play started* event is defined as the time that the first video frame is rendered for display. When playback is stopped, either explicitly by the stop button or implicitly by the *media play request* event for another content, the *media play stop* event is registered. An example of the raw report for a *media play request* event is shown below. In this particular case, the

(a) NextShare^{PC} (b) NextShare^{TV} (c) NextShare^{Mobile}

Fig. 2: The different NextShare derivative application platforms

user requested a VoD item (with the on-demand content ID 229535) distributed by our P2P service.

```
<event>
   <attribute>media_play_request</attribute>
   <timestamp>1332202564</timestamp>
   <status>torrent</status>
   <infohash>v91ZvG/BxGaHqydhJ/YgHDWWczw=</infohash>
   <value>
      http://w.x.y.z/vod/tstream/229535.mpegts.tstream
   </value>
</event>
```

The time tracked *media play request*, *media play started* and *media play stop* reports enable a number of comprehensive analysis such as the video loading time (the time difference between *media play request* and *media play started*), viewing time (the time difference between *media play started* and *media play stop*), and user behaviour and programme popularity (combining playback events with the programme title).

B. Data Format

In order to facilitate user behaviour and preference analysis in relation to TV viewing experience, all irrelevant records are filtered out of the raw statistics and user playback sessions are linked to the corresponding programme information (such as the original broadcast time and programme title). Due to user agreements and data protection measures, users' MAC addresses and IP addresses are anonymised using an MD5 hash function. Because of similar reasons, the data submitted includes user requests made between October 2011 and April 2012. This covers the time span of two consecutive University terms separated by a Christmas break. This feature also enables the study of user activity shifts around holiday periods.

The fields are introduced by name, format and description:

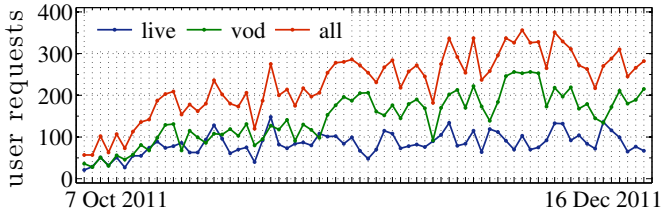
1) *uid* [integer] - unique stats ID.
2) *main_uid* [integer] - External key; could be used to link the aggregated statistics to the original records. We keep this key for future reference if members of the P2P community ask about specifics beyond this submission.
3) *starttime* [datetime] - The time that a requested piece of content (either as live or on-demand) starts playing.
4) *stoptime* [datetime] - The time that requested content stops playing, either due to the user clicking the stop button or requesting different content (which terminates the current playback).
5) *programstartpoint* [characters] - The time at which content is scheduled to start, gathered from the original DVB EPG feed.
6) *programstarttime* [datetime] - Normalised *programstartpoint* field to facilitate time-based analysis.
7) *programstoptime* [datetime] - The normalised date and time that the content is scheduled to finish in the EPG.
8) *mediaduration* [characters] - The duration of the programme as given by the DVB EPG in the format of 'PTaHbMcS' where a, b, and c represent the number of hours, minutes and seconds respectively.
9) *mac* [32 digit hex] - MD5 hash of user device MAC address. Originally used to discern PC from STB clients.
10) *ip* [32 digit hex] - MD5 hash of user device IP address.
11) *type* [integer] - A digit denoting the nature of user requests: '0' = live streaming live via P2P overlay; '1' = streaming on-demand programmes; '2' = live streaming via IP multicasting; and '3' = other testing data.
12) *channel* [characters] - The name of TV channel (in lower case) where the programme was initially broadcasted on. Examples are 'bbcone', 'itv1', and 'channel5'.
13) *channelid* [integer] - An unique ID assigned for each TV channel. For instance, '120' identifies 'bbcone'.
14) *program* [character] - The name of the programme (in lower case), as originally reported in the DVB-T EPG.
15) *mediatype* [integer] - A digit that signifies either a TV (field value: 0) or radio (field value: 1) programme.
16) *recordingid* [character] - An unique recording ID assigned to every VoD item when it is made available. Not valid for user requests of live content.

IV. Service Analysis

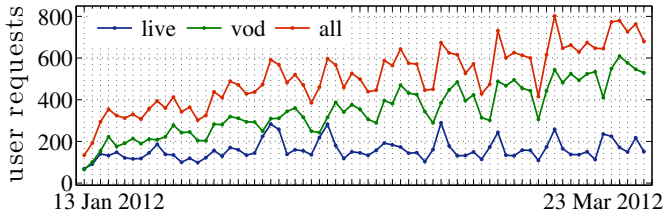
Service measurement and data mining are essential to provide valuable insights for improving service quality and user engagement. Using the specifically designed statistics service, service aspects such as user behaviour, program popularity and social information are captured in high fidelity which enables detailed service analysis. We present some examples here to highlight a few insights that can be drawn from the presented dataset. Please note that the submitted dataset is a subset of the entire Living Lab database, which is used for some of the examples given below.

Figure 3 shows the daily playback requests on live, on-demand and overall content of the first (Oct-Dec 2011) and second (Jan-Mar 2012) university terms. We notice VoD becoming increasingly popular while the requests for live remaining fairly steady. By December, more than one out of

every two viewing requests in the entire service were attributed to the on-demand service. By March 2012, the number of VoD requests is more than double that of live programmes.



(a) First university term



(b) Second university term

Fig. 3: Live and VoD user requests

Figures 4 and 5 offer snapshots of channel and programme popularity over the NextShare^{PC} platform for live and VoD, respectively. For live, we observe that BBC One, BBC Three, BBC Two, BBC News and BBC Radio 1 are the most popular channels in descending order. VoD popularity is of a contrasting nature: E4, a channel of little live viewing popularity, is the most popular channel, followed by BBC Three, Channel 4, BBC One and ITV1. In both cases, "Family Guy" is the most watched programme. Using the submitted dataset, the popularity distribution can also be studied on hourly, daily or monthly basis, giving more insights into the seasonality of user preferences and how it is influenced by system and external factors (e.g. [5]). Such information could be leveraged to improve P2P swarm robustness, smart caching, programme schedules, personalised marketing, and recommendation.

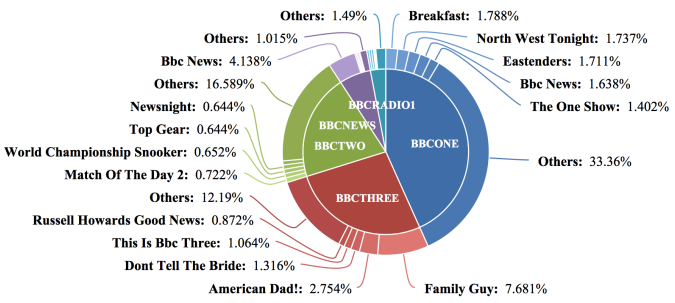


Fig. 4: Live programme popularity on NextShare^{PC}

We also studied the characteristics of on-demand requests, grouped by item age in hours since it is made available in the VoD service. From Figure 6a, we observe huge quantities of requests within the first few hours of an item's lifetime, followed by a drop to its first trough at around the 9th hour. The figure then exhibits a number of major peaks which

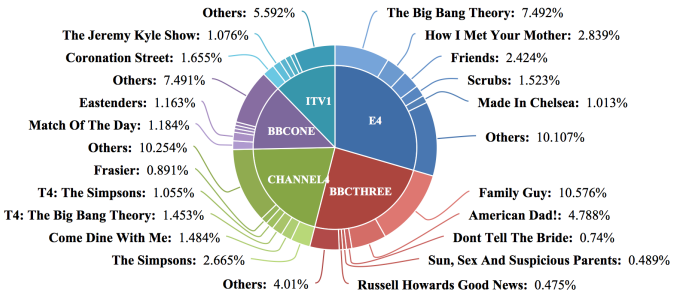
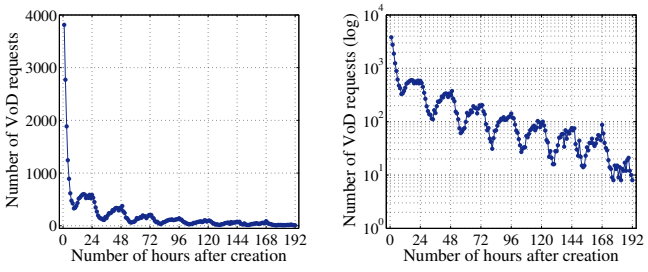


Fig. 5: VoD programme popularity on NextShare^{PC}

appear to be equally spaced while the global descending trend remains. The shape of the major peaks slightly varies and there also seems to be small peaks. These observations are more obvious in Figure 6b where VoD requests are plotted in logarithm scale. There is also a periodic component residing in the temporal distribution as major peaks appear to be separated by around 24 hours, suggesting a correlation between user behaviour and the nature of on-demand programmes. Overall, Figure 6 proffers that the popularity of VoD items attenuates as content ages, with a fairly regular daily pattern.

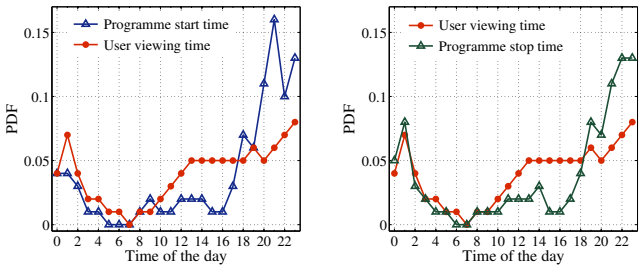


(a) (b)

Fig. 6: VoD requests on hour after creation

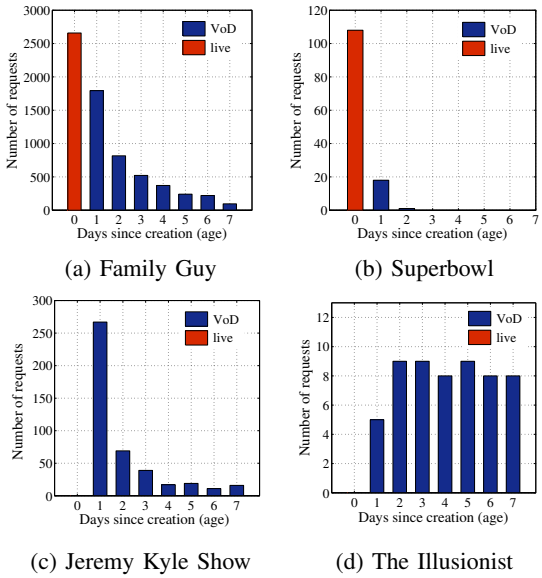
Diving deeper into this temporal distribution reveals how students, who make up the majority of our user base, are very likely to watch TV programmes at specific times in the day. Certain TV programmes are broadcasted every day with a fixed schedule. When these programmes end, the corresponding VoD items are created immediately which inherit the same pattern in TV schedule. Using the dataset, we extract and compare the user viewing time of VoD items and the original start / stop time of the requested programmes as defined in the live programme schedule (Figure 7). The stop time of a live programme is approximately the time the VoD version is available in the on-demand service. Figure 7 visualises user activities in VoD services, highlighting the time spans of popular programmes. The viewing time of our student users in the Living Lab spans over 12 hours from noon to midnight with two peaks at 23:00 and 01:00. Compared with the relatively flat viewing time distribution, the requested programmes exhibit a highly skewed distribution in terms of both programme start and stop times. This suggests that a group of VoD items (or programmes) which are broadcasted live in the late evening

are always popular as VoD item regardless of the user viewing time. We conjecture that this phenomenon is explained by the example of a student who watches, say, "The Big Bang Theory" in the evening and requests more episodes of the same programme in late afternoon on the following day. The combination of steady daily viewing pattern and the highly screwed popular programme distribution is believed to be the source of the diurnal pattern observed in VoD popularity.



(a) (b)

Fig. 7: Comparing user viewing with programme times



(a) Family Guy (b) Superbowl

(c) Jeremy Kyle Show (d) The Illusionist

Fig. 8: Temporal popularity of different programmes

Analysing individual programmes offers further insight. The charts in Figure 8 present a clear contrast in the popularity trends of different programmes when viewed live and on-demand on days subsequent to publishing. The figures reflect our previous observations that "Family Guy" (comedy series) is particularly popular as both live and VoD content: a total of 1608 requests are received within a day of the episodes being made available as on-demand content, and gradually dropping from the second day onwards. In contrast, "American Football: Super Bowl" (live sports) attracted many viewers when first broadcasted but very few and rapidly diminishing VoD requests thereafter. Further example include "The Jeremy Kyle Show" (daytime talk show) and "The Illusionist" (film) are two of the many programmes that are hardly ever watched live but become relatively popular as VoD content. The charts also indicate varying distributions of VoD popularity: the "The Jeremy Kyle Show" shows a more skewed distribution compared with that of "Family Guy". Moreover, the temporal distribution of the "The Illusionist" is distinctively different from others genres like comedy and sports. In fact this is a characteristic shared between films and documentaries which are less influenced by the age of the content as other genres.

V. Conclusions

This paper presented a dataset from the Lancaster Living Lab, an infrastructure serving both live and on-demand high quality content to users via consumer set-top devices, personal computers and mobile devices. The data collected offers a rich look into how users interact with such a converged service, a few examples of which have been presented to highlight the value of this contribution. This dataset could be mined for different research insights in a myriad of research fields. For instance, the dataset provides fine grained user activity records identifying playback length and absolute playback start and stop times. Such information is a valuable asset in understanding media request patterns on different platforms, user viewing behaviour in relation to time of day and season [5], the effect on quality of experience (QoE) [9] and resource waste [6], and feasibility of alternative delivery mechanisms [11], to name but a few examples.

Acknowledgment

This work was supported by the European Commission FP7 grant agreements 216217 (P2P-Next), 318343 (STEER) and 603662 (FI-Content2).

References

[1] Cisco visual networking index: Forecast and methodology, 2013-2018. Technical report, Cisco, 2013.

[2] The communications market report. Technical report, Ofcom, 2013.

[3] N. Capovilla, M. Eberhard, S. Mignanti, R. Petrocco, and J. Vehkapera. An architecture for distributing scalable content over peer-to-peer networks. In *Conf. on Advances in Multimedia (MMEDIA)*, June 2010.

[4] F. Dobrian, A. Awan, D. Joseph, A. Ganjam, J. Zhan, V. Sekar, I. Stoica, and H. Zhang. Understanding the Impact of Video Quality on User Engagement. *ACM SIGCOMM CCR*, 41(4):362–373, 2011.

[5] Y. Elkhatib, R. Killick, M. Mu, and N. Race. Just browsing? Understanding user journeys in online TV. In *ACM Multimedia*, Nov. 2014.

[6] K.-W. Hwang, V. Gopalakrishnan, R. Jana, S. Lee, V. Misra, and K. K. Ramakrishnan. Abandonment and its impact on P2P VoD streaming. In *IEEE P2P*, 2013.

[7] J. Ishmael, S. Bury, D. Pezaros, and N. Race. Deploying rural community wireless mesh networks. *IEEE Internet Computing*, 12(4):22–29, 2008.

[8] M. Mu, J. Ishmael, W. Knowles, M. Rouncefield, N. Race, M. Stuart, and G. Wright. P2P-Based IPTV Services: Design, Deployment, and QoE Measurement. *IEEE Trans. on Multimedia*, 14(6):1515–1527, 2012.

[9] M. Mu, J. Ishmael, K. Mitchell, N. Race, and A. Mauthe. Multimodal QoE evaluation in P2P-based IPTV systems. In *ACM Multimedia*, pages 1321–1324, 2011.

[10] M. Mu, W. Knowles, and N. Race. Understanding Your Needs: An Adaptive VoD System. In *IEEE International Symposium on Multimedia (ISM)*, pages 255–260, 2012.

[11] G. Tyson, S. Kaune, S. Miles, Y. Elkhatib, A. Mauthe, and A. Taweel. A trace-driven analysis of caching in content-centric networks. In *IEEE ICCCN*, 2012.