Increasing the Coverage of Vantage Points in Distributed Active Network Measurements by Crowdsourcing

Valentin Burger, Matthias Hirth, Christian Schwartz, Tobias Hoßfeld,
Phuoc Tran-Gia

University of Würzburg, Germany
Chair of Communication Networks

Abstract. Internet video constitutes more than half of all consumer traffic. Most of the video traffic is delivered by content delivery networks (CDNs). The huge amount of traffic from video CDNs poses problems to access providers. To understand and monitor the impact of video traffic on access networks and the topology of CDNs, distributed active measurements are needed. Recently used measurement platforms are mainly hosted in National Research and Education Networks (NRENs). However, the view of these platforms on the CDN is very limited, since the coverage of NRENs is low in developing countries. Furthermore, campus networks do not reflect the characteristics of end user access networks. We propose to use crowdsourcing to increase the coverage of vantage points in distributed active network measurements. In this study, we compare measurements of a global CDN conducted in PlanetLab with measurements assigned to workers of a crowdsourcing platform. Thus, the coverage of vantage points and the sampled part of the global video CDN are analyzed. Our results show that the capability of PlanetLab to measure global CDNs is rather low, since the vast majority of requests is directed to the US. By using a crowdsourcing platform we obtain a diverse set of vantage points that reveals more than twice as many autonomous systems deploying video servers.

1 Introduction

Internet video constitutes more than half of all consumer Internet traffic globally, and its percentage will further increase [4]. Most of the video traffic is delivered by content delivery networks (CDNs). Today the world's largest video CDN is YouTube. Since Google took over YouTube in 2006 the infrastructure of the video delivery platform has grown to be a global content delivery network. The global expansion of the CDN was also necessary to cope with growing demand of user demands and the high expectations on the video playback. Therefore, content delivery networks try to bring content geographically close to users. However, the traffic from content delivery networks is highly asymmetric and produces a large amount of costly inter-domain traffic [11]. Especially Internet Service Providers (ISPs) providing access to many end users have problems to deal with the huge amount of traffic originating from YouTube. Furthermore, the Google CDN is constantly growing and changing, which makes it difficult for access providers to adapt their infrastructure accordingly.

To understand and monitor the impact of YouTube traffic on ISPs and the topology of CDNs appropriate measurements are aquired. Due to YouTube's load-balancing and caching mechanisms the YouTube video server selection is highly dependent on the location of the measurement points. Hence, we need a globally distributed measurement platform to perform active measurements to uncover the location of YouTube servers. Recent work [1, 2] has performed such measurements in PlanetLab [13], a global test bed that provides measurement nodes at universities and research institutes. The problem is that probes disseminated from PlanetLab nodes origin solely from National Research and Education Networks (NRENs). This may not reflect the perspective of access ISPs which have a different connection to the YouTube CDN with different peering or transit agreements.

To achieve a better view on the YouTube CDN from the perspective of end users in access networks we use a commercial crowdsourcing platform to recruit regular Internet users as measurement probes. Thus, we increase the coverage of vantage points for the distributed measurement of the YouTube CDN. To evaluate the impact of the measurement platform and the coverage of their vantage point, we perform the same measurements using PlanetLab nodes and crowdsourcing users and compare the obtained results.

Our measurements show that distributed measurements in PlanetLab are not capable to capture a globally distributed network, since the PlanetLab nodes are located in NRENs where the view on the Internet is limited. We demonstrate that recruiting users via crowdsourcing platforms as measurement probes can offer a complementary view on the Internet, since they provide access to real end users devices located out side of these dedicated research networks. This complementary view can help to gain a better understanding of the characteristics of Video CDNs. Concepts like ALTO or economic traffic management (ETM) [7] need a global view of the CDN structure to optimize traffic beyond the borders of ISPs. Finally, models for simulation and performance evaluation of mechanisms incorporating CDNs need to apply the characteristics identified by crowd sourced network measurements.

This paper is structured as follows: Section 2 explains the basic structure and functionality of the YouTube CDN, give as short overview of the concept of crowdsourcing, and the reviews work related. The measurements conducted in the PlanetLab and via crowdsourcing are described in Section 3. In Section 4 we details on the measurement results and their importance for the design of distributed network measurements. We conclude this work in Section 5.

2 Background and Related Work

In this section, we briefly describe the structure of the YouTube video CDN and give a short introduction in the principles of crowdsourcing. Further, we summarize related work in the field of distributed active measurements of CDNs as well as work related to crowdsourcing aided network measurements.

2.1 Evolution and Structure of Content Delivery Networks

Since the launch of the YouTube service content delivery has drastically changed. The number of users watching videos on demand has massively increased and the bandwidth to access videos is much higher. Furthermore, the increased bandwidth enables web services to be interactive by using dynamic server- or client-side scripts. The appearance of dynamic services and the increasing quality of multimedia content raised user expectations and the demand on the servers. To bring content in high quality to end-users with low latency and to deal with increasing demand, content providers have to replicate and distribute the content to get it close to end-users. Thus, content delivery networks such as the Google CDN evolved.

The global expansion of the CDNs also changes the structure of the Internet. Google has set up a global backbone which interconnects Google's data centers to important edge points of presence. Since these points of presence are distributed across the globe, Google can offer direct peering links to access networks with many end users. Such, access network providers save transit costs, while Google is able to offer services with low latency. To bring content even closer to users ISPs can deploy Google servers inside their own network to serve popular content, including YouTube videos [5].

To select the closest server for a content request and to implement load balancing CDNs use the Domain Name System (DNS). Typically a user watches a YouTube video by visiting a YouTube video URL with a web browser. The browser then contacts the local DNS server to resolve the hostname. Thereafter, the HTTP request is directed to a front end web server that returns an HTML page including URLs for default and fallback video servers. These URLs are again resolved by DNS servers to physical video servers, which stream the content. The last DNS resolution can happen repeatedly until a server with enough capacity is found to serve the request. Thus, load balancing between the servers is achieved [1].

2.2 Crowdsourcing

Crowdsourcing is an emerging service in the Internet that enables outsourcing jobs to a large, anonymous crowd of users [16]. So called *Crowdsourcing platforms* acts as mediator between the users submitting the tasks, the *employers*, and the users willing to complete these tasks, the *workers*. All interactions between workers and employers are usually managed through these platforms and no direct communication exits, resulting in a very loose worker-employer relationship. The complexity of Crowdsourcing tasks varies between simple transcriptions of single words [17] and even research and development tasks [10]. Usually, the task description are much more fine granular than in comparable forms in traditional work organization [8]. This small task granularity hold in particular for *micro-tasks*, which can be completed within a few seconds to a few minutes. These tasks are usually highly repetitive, e.g., adding textual descriptions to pictures, and are grouped in larger units, so called *campaigns*.

2.3 Related Work

There already exist a number of publications which study the structure of the YouTube CDN and its selection of video servers. A distributed active measurement platform is necessary for these evaluation, because the CDN mechanisms consider the client locations, both geographical as well as in terms of the connected access network. In [15] two university campus networks and three ISP networks were used to investigate the YouTube CDN from vantage points in three different countries. The results show that locality in terms of latency is not the only factor for video server selection.

While the view of five different ISPs on a global CDN is still narrow, the authors of [2] used PlanetLab to investigate the YouTube server selection strategies and load-balancing. They find that YouTube massively deploys caches in many different locations worldwide, placing them at the edge of the Google autonomous system or even at ISP networks. The work is enhanced in [1], where they uncover a detailed architecture of the YouTube CDN, showing a 3-tier physical video server hierarchy. Furthermore, they identify a layered logical structure in the video server namespace, allowing YouTube to leverage the existing DNS system and the HTTP protocol.

However, to assess the expansion of the whole YouTube CDN and its cache locations in access networks, the PlanetLab platform, which is located solely in NRENs, is not suitable, since it does not reflect the perspective of end users in ISP access networks. Therefore, a different distributed measurement platform is used in [14] which runs on end user equipment and thus implies a higher diversity of nodes and reflects the perspective of end user in access networks. However, the number of nodes that was available for the measurement is too small to obtain a global coverage of vantage points

To achieve both, the view of access networks and a high global coverage with a large number of measurement points, the participation of a large number of end users in the measurement is necessary. Bischof et al. [3] implemented an approach to gather data form peer-to-peer networks to globally characterize the service quality of ISPs using volunteers.

In contrast to this we propose using a commercial crowdsourcing platform to recruit users running a specially designed measurement software and therewith act as measurement probes. In comparison to other approaches using volunteers, this approach offers better scalability and controllability, because the number and origin of the participants can be adjusted using the recruiting mechanism of the crowdsourcing platform. This is confirmed by Table 1 which compares

Table 1. Quantitative Comparison: Crowdsourcing / Social Network Study.

	Crowdsourcing (C)	**Social network (S)**
Implementation time	about 2 weeks; test implemented via dynamic web pages, application monitoring	same as for (C)
Time for acquiring people	5 minutes	2 hours, as users (groups) were asked individually
Campaign submission cost	16 Euro	0 Euro
Subjects reward	0.15 Euro	0 Euro
Number of test conditions	3	3
Advertised people	100	350
Campaign completion time	31 hours	26 days; strongly depends on advertised user groups however
Participating users	100	95
Reliable users (very strict filtering of users)	30	58
Number of different countries of subjects	30	3; strongly depends on users groups however

a crowdsourcing study with a social network study quantitatively. The crowdsourcing study is described in [9]. The study is designed to assess the subjective QoE for multimedia applications, like video streaming. The same study was conducted additionally in a social network environment for recruiting test users. Table 1 shows that acquiring people in crowdsourcing platforms takes very short time compared to asking volunteers in a social network, which allows adding participants easily. Furthermore, the completion time of the campaign of 31 hours is much shorter compared to the 26 days for the social network campaign. Finally, in the crowdsourcing campaign workers can be selected according to their country, which allows distributing the campaign on many different countries. In the social network the coverage of countries depends on the network of user groups, which spread the campaign. Hence, it is easy to control the number and origin of subjects participating in a crowdsourcing campaign and the completion time is considerably fast, which makes the campaign scalable and controllable. The price you pay is the reward for the workers that summed up to a total of 16 Euro for that campaign.

To the best of our knowledge this is the first work which uses crowdsourcing for a distributed active measurement platform.

3 Measurement Description

To assess the capability of crowdsourcing for distributed active measurements we conduct measurements with both PlanetLab and the commercial Crowdsourcing platform Microworkers [18]. We measure the global expansion of the YouTube CDN by resolving physical server IP-addresses for clients in different locations.

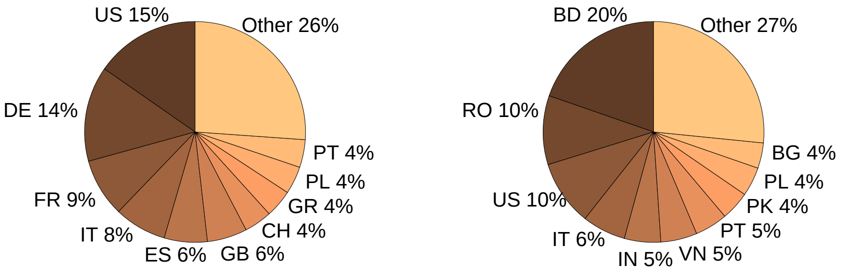
3.1 Description of the PlanetLab Measurement

PlanetLab is a publicly available test bed, which currently consists of 1173 nodes at 561 sites. The sites are usually located at universities or research institutes. Hence, they are connected to the Internet via NRENs. To conduct a measurement in PlanetLab a slice has to be set up which consists of a set of virtual machines running on different nodes in the PlanetLab test bed. Researchers can then access these slides to install measurement scripts. In our case the measurement script implemented in Java extracted the server hostnames of the page of three predetermined YouTube videos and resolved the IP addresses of the physical video servers. The IP addresses of the PlanetLab clients and the resolved IP addresses of the physical video servers were stored in a database. To be able to investigate locality in the YouTube CDN, the geo-location of servers and clients is necessary. For that purpose the IP addresses were mapped to geographic coordinates with MaxMinds GeoIP database [12]. The measurement was conducted on 220 randomly chosen PlanetLab nodes in March 2012.

3.2 Description of the Crowdsourcing Measurement

To measure the topology of the YouTube CDN from an end users point of view who is connected by an ISP network we used the crowdsourcing platform Microworker [18]. The workers were asked to access a web page with an embedded Java application, which automatically conducts client side measurements. These include, among others, the extraction of the default and fallback server URLs from three predetermined YouTube video pages. The extracted URLs were resolved to the physical IP address of the video servers locally on the clients. The IP addresses of video servers and of the workers client were sent to a server which collected all measurements and stored them in a database.

In a first measurement run, in December 2011, 60 different users of Microworkers participated in the measurements. Previous evaluation have shown, that the majority of the platform users is located in Asia [6], and accordingly most of the participants of there first campaign were from Bangladesh. In order to obtain wide measurement coverage the number of Asian workers participating in a second measurement campaign, conducted in March 2012, was restricted. In total, 247 workers from 32 different countries, finished the measurements successfully identifying 1592 unique physical YouTube server IP addresses.



(a) PlanetLab

(b) Crowdsourcing

Fig. 1. Distribution of measurement points on countries in a) PlanetLab and b) Crowdsourcing platform.

4 Results

In this section we show the results of the distributed measurement of the global CDN. The obtained results show the distribution of clients and servers over different countries. Furthermore, the mapping on autonomous systems gives insights to the coverage of the Internet.

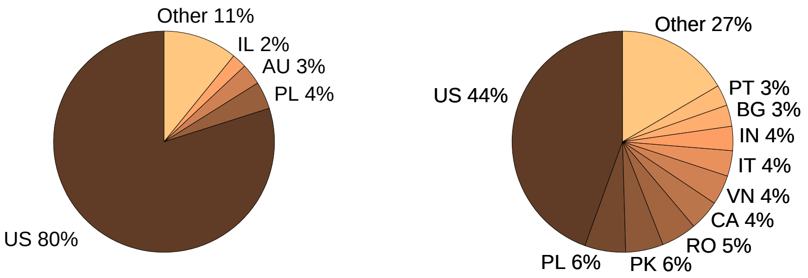
4.1 Distribution of Vantage Points on Countries

To investigate the coverage of measurement points we study the distribution of the PlanetLab nodes and Crowdsourcing workers. Figure 1(a) shows the distribution of PlanetLab nodes on countries over the world. The pie chart is denoted with the country codes and the percentage of PlanetLab nodes in the respective country. Most of the 220 clients are located in the US with 15% of all clients. However, more than 50% of the clients are located in West-Europe. Only few clients are located in different parts of the world. The tailored distribution towards Western countries is caused by the fact, that the majority of the PlanetLab nodes are located in the US or in western Europe.

Figure 1(b) shows the geo-location of workers on the crowdsourcing platform. In contrast to PlanetLab, most of the 247 measurement points are located in Asia-Pacific and East-Europe. The majority of the participating workers 20% are from Bangladesh followed by Romania and the US with 10%. This bias is caused by the overall worker distribution on the platform [6]. However, this can be influences to a certain extend by limiting the access to the tasks to certain geographical regions.

4.2 Distribution of Identified YouTube Servers on Countries

To investigate the expansion of the YouTube CDN we study the distribution of YouTube servers over the world. Figure 2(a) shows the location of the servers



(a) PlanetLab

(b) Crowdsourcing

Fig. 2. Distribution of physical YouTube servers on countries accessed from a) PlanetLab nodes and b) workers of a crowdsourcing platform.

identified by the PlanetLab nodes. The requests are mainly directed to servers in the US. Only 20% of the requests were directed to servers not located in the US.

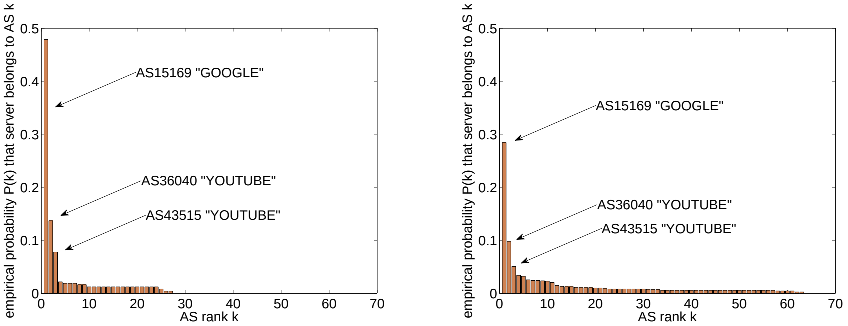
The servers identified by the crowdsourcing measurement are shown in Figure 2(b). The amount of requests being directed to servers located in the US is still high. 44% of clients were directed to the US. However, in this case the amount of requests resolved to servers outside the US is higher. In contrast to the PlanetLab measurement many requests are served locally in the countries of clients. Furthermore, the decrease of 80% to 44% of request being directed to the US shows a huge difference.

Hence, network probes being overrepresented in the US and Europe leads to a limited view of the content delivery network and the Internet. This shows the impact of different locations of measurement points on the view of the CDN. It also demands a careful choice of vantage points for a proper design of experiments in distributed network measurements. Although both sets of measurement points are globally distributed the fraction of the CDN which is discovered by the probes has very different characteristics.

The amount of servers which is located in the US almost doubles for the PlanetLab measurement. While 44% of the requests are resolved to US servers in the Crowdsourcing measurement, nearly all requests of PlanetLab nodes are served by YouTube servers located in the US. Although less than 15% of clients are in US, requests are frequently directed to servers in the US. That means that there is still potential to further distribute the content in the CDN.

4.3 Coverage of Autonomous Systems with YouTube Servers

To identify the distribution of clients on ISPs and to investigate the expansion of CDNs on autonomous systems we map the measurement points to the corresponding autonomous systems.



(a) PlanetLab

(b) Crowdsourcing

Fig. 3. Distribution of YouTube servers on autonomous systems from a) PlanetLab and b) Crowdsourcing perspective.

Figure 3(a) shows the autonomous systems of YouTube servers accessed by PlanetLab nodes. The autonomous systems were ranked by the number of YouTube servers located in the AS. The empirical probability $P(k)$ that a server belongs to AS with rank k is depicted against the AS rank. The number of autonomous systems hosting YouTube servers that are accessed by PlanetLab nodes is limited to less than 30. The top three ranked ASes are AS15169, AS36040 and AS43515. AS15169 is the Google autonomous system which includes the Google backbone. The Google backbone is a global network that reaches to worldwide points of presence to offer peering agreements at peering points. AS36040 is the YouTube network connecting the main datacenter in Mountain-View which is also managed by Google. AS43515 belongs to the YouTube site in Europe which is administrated in Ireland. Hence, two thirds of the servers are located in an autonomous systems which is managed by Google. Only few requests are served from datacenters not being located in a Google AS. The reason that request from PlanetLab are most frequently served by ASes owned by Google might be a good interconnection of the NRENs to the Google ASes.

Figure 3(b) depicts the autonomous systems where requests to YouTube videos from the crowdsourcing workers were directed. The empirical probability that a server belongs to an AS has been plotted dependent on the AS rank. The YouTube servers identified by the crowdsourcing probes are located in more than 60 autonomous systems. Hence, the YouTube CDN is expanded on a higher range of ASes from the crowdsourcing perspective compared to PlanetLab. Again the three autonomous systems serving most requests are the ASes managed by Google, respectively YouTube. But the total number of requests served by a Google managed AS is only 41%. Hence, in contrary to the PlanetLab measurement, requests are served most frequently from ASes not owned by Google. Here, caches at local ISPs managed by YouTube could be used to bring the content close to users without providing own infrastructure. This would also explain the large number of identified ASes providing a YouTube server. The results show that the PlanetLab platform is not capable to measure the structure of a global CDN, since large parts of the CDN are not accessed by clients in NRENs.

5 Conclusion

In this study we proposed the usage of crowdsourcing platforms for distributed network measurements to increase the coverage of vantage points. We evaluated the capability to discover global networks by comparing the coverage of video server detected using a crowdsourcing platform as opposed to using the PlanetLab platform. To this end, we used exemplary measurements of the global video CDN YouTube, conducted in both the PlanetLab platform as well as the crowdsourcing platform Microworkers.

Our results show that the vantage points of the concurring measurement platforms have very different characteristics. In the PlanetLab measurement the country with most measurement points is the US, while more than 50% of measurement points are located in West-Europe. In contrary most measurement points are located in Asia-Pacific and East-Europe in the crowdsourcing measurement. Further we could show that the distribution of vantage points has high impact on the capability of measuring a global content distribution network. The capability of PlanetLab to measure a global CDNs is rather low, since 80% of requests are directed to the United States.

Finally, our results confirm that the coverage of vantage points is increased by crowdsourcing. Using the crowdsourcing platform we obtain a diverse set of vantage points that reveals more than twice as many autonomous systems deploying video servers than the widely used PlanetLab platform. Part of future work is to determine if the coverage of vantage points can be even further increased by targeting workers from specific locations to get representative measurement points for all parts of the world.

Acknowledgement

The authors would like to thank Philipp Amrehn and Christian Moldovan for conducting the measurements and implementing the evaluation scripts.

This work was partly funded by Deutsche Forschungsgemeinschaft (DFG) under grants HO 4770/1-1 and TR257/31-1 and in the framework of the EU ICT Project SmartenIT (FP7-2012-ICT-317846). The authors alone are responsible for the content.

References

1. Adhikari, V., Jain, S., Chen, Y., Zhang, Z.: Vivisecting YouTube: An Active Measurement Study. In: Proceedings IEEE INFOCOM (2012)
2. Adhikari, V., Jain, S., Zhang, Z.: Where Do You "Tube"? Uncovering YouTube Server Selection Strategy. In: IEEE ICCCN (2011)

3. Bischof, Z.S., Otto, J.S., Sánchez, M.A., Rula, J.P., Choffnes, D.R., Bustamante, F.E.: Crowdsourcing isp characterization to the network edge. In: Proceedings of the first ACM SIGCOMM workshop on Measurements up the stack (2011)
4. Cisco: Forecast and Methodology, 2012–2017. Cisco Visual Networking Index (2013)
5. Google: Peering & Content Delivery. https://peering.google.com/
6. Hirth, M., Hoßfeld, T., Tran-Gia, P.: Anatomy of a Crowdsourcing Platform - Using the Example of Microworkers.com. In: Workshop on Future Internet and Next Generation Networks (FINGNet). Seoul, Korea (2011)
7. Hoßfeld, T., Hausheer, D., Hecht, F., Lehrieder, F., Oechsner, S., Papafili, I., Racz, P., Soursos, S., Staehle, D., Stamoulis, G.D., Tran-Gia, P., Stiller, B.: An Economic Traffic Management Approach to Enable the TripleWin for Users, ISPs, and Overlay Providers. IOS Press Books Online, Towards the Future Internet - A European Research Perspective (2009)
8. Hoßfeld, T., Hirth, M., Tran-Gia, P.: Modeling of Crowdsourcing Platforms and Granularity of Work Organization in Future Internet. In: In Proceedings of the International Teletraffic Congress (ITC) (2011)
9. Hoßfeld, T., Schatz, R., Biersack, E., Plissonneau, L.: Internet Video Delivery in YouTube: From Traffic Measurements to Quality of Experience. In: Ernst Biersack, Christian Callegari, M.M. (ed.) Data Traffic Monitoring and Analysis: From measurement, classification and anomaly detection to Quality of experience. Springers Computer Communications and Networks series, Volume 7754 (2013)
10. InnoCentive, Inc: Innocentive. http://www.innocentive.com/
11. Labovitz, C., Iekel-Johnson, S., McPherson, D., Oberheide, J., Jahanian, F.: Internet Inter-Domain Traffic. In: ACM SIGCOMM Computer Communication Review (2010)
12. MaxMind: GeoLite Databases. http://dev.maxmind.com/geoip/geolite/
13. PlanetLab: An open platform for developing, deploying, and accessing planetary-scale services. http://www.planet-lab.org/
14. Rafetseder, A., Metzger, F., Stezenbach, D., Tutschku, K.: Exploring youtube's content distribution network through distributed application-layer measurements: a first view. In: Proceedings of the 2011 International Workshop on Modeling, Analysis, and Control of Complex Networks (2011)
15. Torres, R., Finamore, A., Kim, J.R., Mellia, M., Munafo, M.M., Rao, S.: Dissecting Video Server Selection Strategies in the YouTube Cdn. In: 31st International Conference on Distributed Computing Systems (ICDCS) (2011)
16. Tran-Gia, P., Hoßfeld, T., Hartmann, M., Hirth, M.: Crowdsourcing and its Impact on Future Internet Usage. it - Information Technology 55 (2013)
17. Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., Blum, M.: recaptcha: Human-based character recognition via web security measures. Science (5895) (2008)
18. Weblabcenter, Inc.: Microworkers. http://microworkers.com/