Geographical Characterization of YouTube: a Latin American View

Fernando Duarte, Fabricio Benevenuto, Virgilio Almeida, Jussara Almeida
Computer Science Department
Federal University of Minas Gerais
Brazil
{fernando, fabricio, virgilio, jussara}@dcc.ufmg.br

Abstract

Online social media applications have exploded in popularity in the Web. In most of these applications, users interact with other users and create content that becomes available on the Web, such as textual information, photos, and videos. YouTube is the largest online social video sharing service, that generates a huge amount of Web traffic. This paper presents a geographical characterization of YouTube usage. It analyzes video and user characteristics for different geographical regions, concentrating mainly on Latin America. We develop efficient crawlers for collecting data about videos and users. Because the number of videos and users is very large, we sample from the object spaces, sampling over 2 million videos and 5 million users. Based on the collected data, we show that there exists relationships between geography and the social network features available in YouTube. We present evidences that indicate that geography creates a locality space in YouTube, which could be used to explore infrastructure improvements, such as caching, content distribution networks and broadband pricing mechanisms.

1 Introduction

The use of the Internet as a channel for the delivery of multimedia content is gaining widespread popularity. Particularly, online social video sharing services are becoming very popular, allowing users to generate and distribute their own videos to large audiences.

In this paper we present a geographical characterization for YouTube [2], a popular online social video sharing service which generates high-volumes of Internet traffic [1]. Our goal is to characterize the influence of geographical localization on traffic and social relationship among users. We are particularly interested in highlighting the differences between Latin American and non-Latin American activity. Such a characterization is of interest for two reasons. The first is a technical reason, stemming from the necessity to understand geographical factors that affect the large amount of traffic generated by video services. The second is sociological, relating to geographical localization and cultural differences which influence the behavior of users interacting in a social video sharing community.

Our approach to collect data from YouTube consists of developing two crawlers for sampling data about popular videos and users. Because the number of videos and users is very large, we sample over 2 million videos and 5 million users. We are not aware of any other study that focus on geographical issues of a large online social video sharing service, such as YouTube.

Our results highlight differences in user behavior, depending on their geographical region. For instance, we show the usage of social network features, such as placing comments or video response to a video or choosing favorite videos are strongly influenced by the geographical origin of the users. Interestingly, we find that more than 90% of the comments made to videos uploaded by Latin American users were posted by Latin American users. We also compare the behavior of users from different countries of Latin America.

The rest of the paper is organized as follows. Section 2 describes YouTube and how we crawled and sampled it. Section 3 presents our geographical characterization of YouTube and discusses the main findings. Section 4 discusses related work. Section 5 offers conclusions and directions for future work.

2 YouTube and Sampling Mechanism

Streaming about 300 million videos a day (as of May 2007), YouTube is perhaps the largest and the most popular online social video sharing service today, generating high-volumes of Internet traffic [1]. The wide variety of YouTube content includes movies, documentaries, political campaigning, TV clips and music videos, as well as amateur content such as videoblogging and short original videos.

YouTube, as well as typical online social media networks such as MySpace and Yahoo Videos, exhibits the following key characteristics: (i) users can contribute with an unlimited number of videos, which the contributor typically annotates with a title, description, and tags. (ii) users evaluate content by rating and text commenting. Moreover, YouTube provides a video response feature, allowing users to respond with a video to other videos, creating asynchronous multimedia dialogs within the YouTube site; (iii) users maintain lists of friends, favorite videos and can subscribe to other users to receive updates when new content is posted. A user is anyone who has created an account with YouTube and visitors without an account can only watch videos on the Web site.

2.1 Crawling YouTube

To analyze the geographical characteristics of YouTube users, ideally we would like to have at our disposal data for each existing YouTube video and user. Without having direct access to these proprietary data, we can instead attempt to crawl the YouTube site to obtain it. However, YouTube stores a huge amount of videos and thousands of new videos are daily uploaded at YouTube, making it difficult to crawl the entire site with constrained resources. In this context, we decide to sample YouTube and obtain data for a subset of videos and users. Our crawl and sample strategy consists of collecting information of popular videos and analyzing the user interactions around these videos.

Each YouTube video page provides a number of different mechanisms to discover other YouTube videos. These mechanisms include related videos (which are videos identified by a YouTube algorithm), the favorite videos of the contributor of the current video, videos with the same tags as the current video, and so on. We chose to use related videos, beginning the crawling with the most all-time watched video, provided by YouTube. The related videos are influenced by number of views, and consequently, this crawling strategy is biased towards popular videos, as desired. We employed Snowball sampling, which has been shown to exhibit a number of desirable properties for sampling social networks [14]. Snowball sampling is a breadth-first scheme. As applied here, after sampling the root video at tier 1, we sampled each of the 20 most related views at tier 2; we then sampled each of the 20 most related videos of each of the tier-2 videos, and so on.

For each video that was crawled, we collected the following information: video id, owner id, title, category, description, tags, upload time, video duration, list of top 20 related videos, number of ratings, average rating, number of views, number of users who set the video as favorite, number of comments received and number of video responses received. We also collected the author of the comments and responses of each video.

Our crawler consists of seven Linux boxes at the Federal University of Minas Gerais in Brazil. We implement a parallel crawler framework for crawling YouTube, which is similar to the structure for crawling online social networks presented by [10]. Our distributed crawler has (i) a master node which maintains a centralized list of videos to be visited; and (ii) slave nodes, which obtain video identifiers from the master, crawl YouTube to collect details about each video, and return to the master the top 20 related videos for each crawled video. The master coordinates the operation of all the slaves to prevent redundant crawling. We crawled the YouTube site to obtain information about over 2 million videos, exhausting 6 tiers in 11 days (from 04/03/2007 to 04/14/2007).

To test whether our sampling scheme is covering the majority of the popular videos, as desired, we checked to see if the 100 most all-time popular videos (as provided by the YouTube site) were included in our sample. We found that 96 of 100 of these videos were part of the sample, thereby confirming that our scheme provides good coverage of the more popular videos.

We also built a second crawler for the purpose of collecting information about the YouTube users which we found in our first crawler. In other words, we collect each user who uploaded at least one video, one comment, or one video response. The crawler collected information about over 5.9 million users. The information obtained about each crawled user includes user id, first name, last name, age, number of videos uploaded, number of videos watched, gender, country, number of friends, and number of favorite videos.

3 Geographical Characterization

In order to provide a geographical characterization of the YouTube main entities (i.e., users and videos), we first look at the basic statistics of the data collected with our two crawls. In Table 1 we group the statistics into geographical regions: i) United States (USA), that solely is responsible for 28% and 38% of the videos and users collected in our data; ii) Latin America (LA); iii) the rest of the world (Other), composed of the different countries found in our data. An extra column in Table 1 shows the percentage of users who have not provided country information. These users, corresponding to 13% of the users collected, were excluded from the analysis presented in this paper.

Our crawlers collected data about 2.12 million videos and 5.92 million users. This sample of popular videos were streamed more than 17 billion times, received almost

Characteristic	Geographical region				
	United States (%)	Latin America (%)	Others (%)	Empty (%)	Total
Number of videos visited	28.5	5.9	34.2	31.4	2,126,584
Number of videos with comments	27.5	5.2	32.8	34.5	1,489,806
Number of videos with responses	25.9	3.2	23.0	47.9	50,354
Number of views	24.8	5.7	30.8	38.7	17,924,461,783
Number of comments	35.0	4.6	35.3	25.1	29,062,323
Number of responses	27.0	3.0	23.5	46.5	98,949
Number of contributors of videos	31.5	6.9	39.0	22.6	917,810
Number of authors of comments	36.3	7.0	43.7	13.0	4,433,617
Number of authors of responses	29.8	3.4	27.7	39.1	49,523
Number of users	38.2	6.8	41.9	13.1	5,915,630
Number of uploaded videos of users	26.3	7.2	35.9	30.6	16,798,997
Number of watched videos of users	38.2	5.5	34.5	21.8	3,427,930,407
Number of friends of users	35.1	2.7	24.2	38.0	10,813,159
Number of favorites of users	37.2	6.1	32.8	23.9	92,617,813

Table 1. Statistics of the Video and Users Collected from YouTube

30 million comments and almost 100 thousand responses. The users we collected have uploaded more than 16 million videos and watched more than 3 billion videos.

As we can see from the collected videos, there are many more views than comments, and more comments than responses. Moreover, there are more videos with comments than videos with responses. A possible explanation is that a video response tends to require more effort from the user to be produced than a simple text comment. As an example, only 0.4% from the Latin American users have posted video responses, whereas 77% of Latin American users has posted at least one comment. Moreover, video response is a new feature of YouTube, realized in May 2006.

From the total number of users collected about 7% are Latin Americans, responsible for 7% of the total of uploaded videos and 6% of the number of watched videos. We now discuss the data in more detail, focusing on characteristics of Latin American users.

3.1 Latin American Users

YouTube represents one of the most popular sites among Latin American users. For example, statistics [1] show that YouTube is the 6^{th} most popular Web site in Argentina, Brazil, and Paraguay, the 5^{th} in Mexico, Chile and Peru, and the 4^{th} in Ecuador and Venezuela. Table 2 shows statistics of the data collected distributed over the Latin American countries[1] which appear in our data. The numbers shown in parenthesis represent the average measurement per active user. As we can notice from the table, Brazil, Mexico, Argentina and Chile have the largest number of YouTube users. In Latin America, users from Brazil, Mexico and Argentina have contributed with the largest fraction of videos to YouTube. In terms of uploads per user, Peru leads the rank, with 8 uploads/user. In terms of traffic, measured by the number of watched videos, Brazil, Mexico, Argentina and Chile generate the largest portion of YouTube traffic from Latin America. Note that the number of watched videos reported include both complete and incomplete views, that occur when users stopped viewing after a few seconds or more. We do not have data about incomplete views. However, we conjecture that due to the low effort and cost of viewing a video and the rich interconnections between videos on the site, there is a large amount of fairly random surfing and exploration where visitors and users check out various videos.

From Table 2, we observe that Latin American users have an average of 22 favorite videos. Users from Peru, Puerto Rico and French Guiana have selected more than 27 favorite videos and users from Bolivia and Paraguay less than 19. Based on the collected data, it seems that LA users do not make heavy use of the available features of social networking. The average number of friends of active users is much lower than the number observed in other social online communities. Latin American users have an average of two friends. In [4], the authors report that Orkut users have an average of 30 friends and MySpace users have an average of 137 friends. The observations above suggested that Latin American users are not exploiting all the social features available at YouTube. Most part of the users have uploaded few videos, do not have a large number of friends, do not have a sizable list of favorite videos and send few responses and comments. We guess that most users interacts with their friends in other online social networks, such as Orkut and MySpace and use YouTube only for watching videos.

[1] A list of countries in Latin America and the Caribbean can be found at http://lanic.utexas.edu/subject/countries

		Total number (average number of the active users)					
Member	**# Users**	**uploaded videos**	**watched videos**	**favorite videos**	**friends**	**responses**	**comments**
Brazil	132710	479458 (6.7)	56305678 (429.6)	1885369 (23.3)	100716(2.6)	1215 (2.1)	315312 (3.3)
Mexico	72506	228655 (7.2)	40947662 (570.0)	1132289 (25.7)	52475 (2.6)	486 (1.6)	272651 (4.7)
Argentina	38300	146748 (7.8)	15776606 (417.0)	440420 (22.3)	21213 (2.7)	305 (1.7)	129471 (4.1)
Virgin Islands	37458	42667 (4.8)	21468684 (582.6)	533976 (26.1)	40670 (2.7)	198 (1.7)	163009 (6.0)
Chile	32736	95772 (6.1)	14979434 (461.3)	401696 (22.8)	10980 (2.2)	158 (1.5)	108053 (3.8)
Peru	22115	70471 (8.0)	9258126 (425.5)	340979 (27.6)	15491 (2.7)	128 (1.7)	73781 (4.2)
Venezuela	17074	45721 (6.8)	8411106 (497.6)	247447 (25.4)	10546 (2.5)	110 (1.3)	69819 (4.9)
Colombia	9401	20739 (6.5)	3162933 (343.7)	124411 (23.7)	6004 (2.4)	54 (1.4)	32153 (4.2)
Puerto Rico	6951	14523 (6.3)	3470636 (508.5)	122137 (27.3)	9207 (3.2)	73 (1.9)	30703 (5.9)
Bolivia	5599	7311 (4.8)	2756508 (502.6)	47286 (18.3)	3285 (2.4)	14 (1.2)	22453 (5.0)
Costa Rica	4327	7834 (5.4)	1855650 (434.9)	60087 (22.6)	3411 (2.5)	23 (1.3)	16419 (4.8)
Dom. Republic	3768	10005 (6.5)	1960553 (526.8)	48618 (22.6)	3162 (2.5)	19 (1.1)	14184 (5.0)
Ecuador	3101	7465 (6.7)	954928 (315.8)	33152 (19.3)	1847 (2.3)	20 (1.8)	10198 (4.1)
Guatemala	2842	7461 (6.4)	1112305 (398.8)	39526 (22.7)	2343 (2.4)	14 (1.3)	9096 (4.2)
El Salvador	2806	6394 (5.8)	1562366 (562.2)	44113 (25.1)	2532 (2.7)	30 (2.1)	10542 (4.8)
Uruguay	2392	6324 (6.6)	914943 (388.7)	23301 (19.8)	1663 (2.9)	11 (1.2)	9146 (4.7)
Aruba	2287	3243 (5.1)	1281484 (569.0)	25037 (21.7)	2103 (3.2)	34 (2.4)	10932 (5.9)
Panama	1915	4308 (6.4)	976219 (515.4)	28219 (25.6)	1570 (2.6)	16 (2.0)	8037 (5.3)
Honduras	1312	2937 (6.0)	477751 (370.6)	15690 (19.7)	920 (2.0)	10 (1.4)	4439 (4.4)
Cuba	1051	2081 (6.2)	430644 (418.1)	11706 (22.1)	603 (2.5)	21 (2.6)	5705 (6.5)
Nicaragua	581	1324 (6.3)	216220 (382.7)	6991 (20.8)	433 (2.2)	6 (3.0)	3076 (6.7)
Haiti	507	781 (6.3)	248261 (508.7)	5766 (22.4)	395 (2.4)	6 (1.5)	1881 (5.1)
Paraguay	435	1161 (7.4)	145394 (338.9)	4243 (17.8)	702 (6.5)	3 (1.0)	1281 (3.8)
Guadeloupe	263	727 (6.1)	116512 (448.1)	3237 (23.1)	201 (2.4)	0 (0.0)	822 (4.4)
Martinique	242	630 (6.9)	105165 (440.0)	3145 (23.8)	203 (3.1)	2 (1.0)	708 (3.9)
French Guiana	239	523 (6.5)	111890 (472.1)	3473 (29.4)	136 (2.6)	0 (0.0)	828 (4.4)

Table 2. YouTube Statistics of Some Latin American Countries

3.2 Video Popularity

Figure 1 shows the popularity of videos in terms of number of views, number of comments, and number of responses received by videos of contributors from United States, Latin America, and other regions. The curves show the number of views, comments, and responses from the most popular video to the less popular video.

It has been shown that accesses to files on web sites exhibit a significant skew, as a function of the popularity of some files. The popularity of Web objects has been widely modeled by Zipf's Law [5, 7, 8, 12]. Zipf's law states that frequency of occurrence of some event (P) as a function of the rank (r), is a power-law represented by the following relation: $P \sim r^{-\alpha}$, where the exponent α is close to 1. Thus, a natural question with respect to videos in YouTube is whether they exhibit a similar popularity profile. We note in Figure 1 that the curve for the number of views does not descend linearly as would be expected from a typical power law distribution. In fact, our crawl collected only a subset of the entire set of YouTube videos, and our subset is biased towards popular videos, which are more likely to be reached than less popular ones. We also note that the curves of number of comments and video responses do not suffer from this effect. Moreover, we can observe the body of the curve of popularity of videos in terms of number of views follows a straight line. The same skewed popularity profile holds if one considers the number of comments and video responses as the measure of video popularity.

Figure 1 shows that access is highly concentrated on popular videos. Among Latin American videos, we found that 10% of the top popular videos concentrate 76% of the views. It suggests that storing top popular videos could be a good strategy for caching and produce high hit rates.

Looking at the number of videos from Latin American contributors, we can see that they are less popular in terms of views, comments and responses compared to videos from users from United States and other countries. We also note that there are about 30 videos uploaded by Latin American users that have received more than 1,000 comments each, and few responses were sent to Latin American videos.

3.3 Duration of Videos and Responses

We also analyze the differences on video duration due to geographical localization of the video owner or contributor. Figure 2 shows the cumulative distribution of duration of the videos (top) and video responses (bottom) of con-

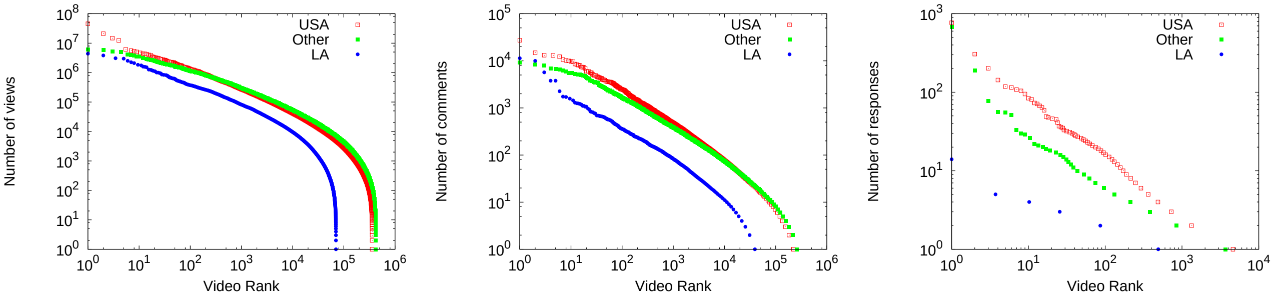

Figure 1. Popularity of videos uploaded by users from United States, Latin America, and other regions in terms of views (left), comments (middle) and responses (right)

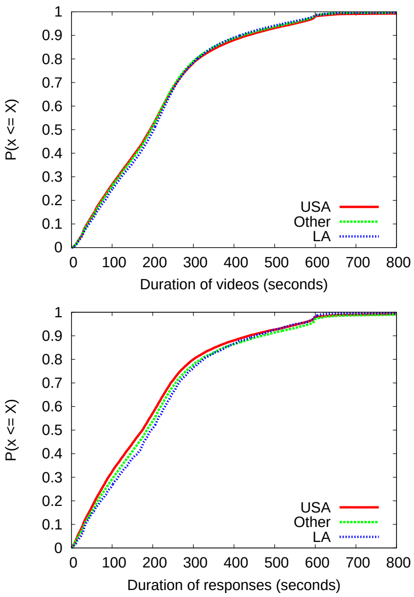

Figure 2. Distribution of the duration of videos (top) and video responses (bottom)

tributors from United States, Latin American, and the other regions. About 80% of the videos and video responses from the three geographical localizations are smaller than five minutes (YouTube limits in ten minutes the maximum duration of a video upload for a common user). We do not observe strong differences in the duration of videos from different countries. Observing the duration of video responses, we can note that there is a small difference between videos of contributors from Latin America, United States, and other regions. Moreover, comparing the duration of videos and video responses, we found that responses are slightly smaller than videos.

3.4 Use of Social Network Features

Using the collected data, we evaluate how YouTube users make use of the social network features available in the site. To characterize the user social profile of the YouTube population, Figure 3 shows the level of social interactivity in terms of number of comments, responses, number of videos selected as favorites, number of friends, level of contribution in terms of number of videos uploaded, and level of interest in terms of number of videos watched. The curves show the various user attributes versus user rank, where the i^{th} ranked user is the one issuing the i^{th}-most participation on YouTube. The graphics in Figure 3 underscore a power law distribution for number of comments, responses, friends, and uploaded videos. However, we note that the curves for the number of favorite and watched videos do not descend linearly as would be expected in a power law distribution. In fact, our crawl collected users based on information of the videos collected. Most part of the users collected come from comments collected with our video crawler. We conjecture that the users that post comments tend to be users with high level of social activity.

In common, these plots show that Latin American users interact less at YouTube than the other users. Besides less interactive, there are Latin Americans who uploaded more than 1,400 videos, sent more than 1,200 comments and almost 100 responses. Moreover, there are Latin American users who have more than 2,400 friends and more than 3,000 videos as favorite.

3.5 Textual and Video-based Interactions on YouTube

The different levels of user involvement shown in Figure 3 raises interesting questions about the influence of geographical localization on the interactions of users. We next study the textual (comments) and video-based (video responses) interactions among users from different geographical localizations.

We quantify the influence of geographical localization on the interactions between contributors of videos and their visitors in two ways: (1) Given that users commented/responded videos from USA, LA and other regions, we analyze the probability of these users being from LA,

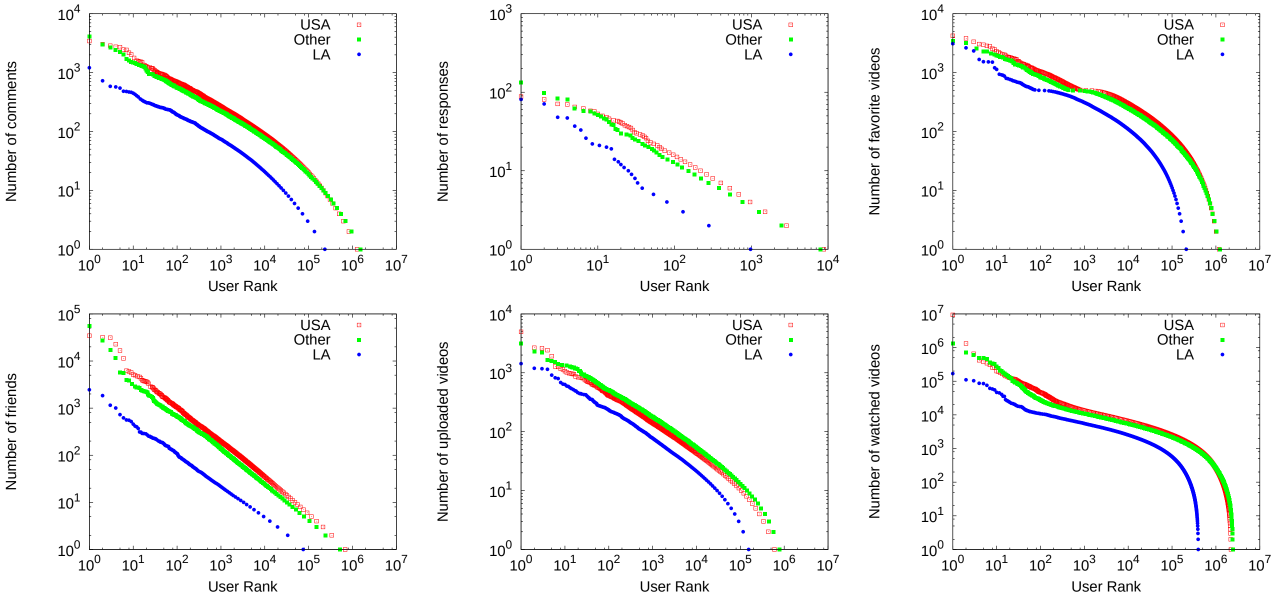

Figure 3. Use of Social Network Features: comments, responses, favorite videos, number of friends, uploaded videos and watched videos

video id	% of users of each region		
	LA	**USA**	**Other**
1	80	10	10
2	75	5	20
3	90	6	4
...	...	...	...

Table 3. Example showing where the users that has commented a video are from

USA and other regions. (2) Given that videos were commented/responded by users from USA, LA and other regions, we analyze the probability of these videos being from LA, USA and other regions. To do it, we calculate, for each video, the percentage of users from United States, Latin America, and other regions that has commented or responded it. In order to illustrate how we provide these analysis, Table 3 shows an example of how users who has commented videos are distributed among LA, USA and other regions. In order to analyze the participation of each region, we study the distribution of the values of the first, second and third column. In our results we ignore comments and responses sent by the contributor of the video.

We first analyze the textual interactions between users. Figure 4 (top) shows the cumulative distribution of the percentage of users from different geographical localizations that comment on videos uploaded by contributors from Latin America, United States, and other regions. We provide several view points by considering a range from 0 to 100% of users. For example, we can see that the probability of Latin American videos have more than 60% of the users who posted comments from Latin America is 0.32 whereas the probability for users from United States is 0.08. Moreover, the probability of USA videos have at least one comment posted by users from USA is 0.8 whereas the probability for LA users is 0.2. As we can see, the probability of a Latin American video be commented by Latin Americans is considerably higher than to receive a comment from United States or other regions. Moreover, videos uploaded by contributors from United States and other regions are also more commented by users from the same geographical localization of the owner of the video.

Analyzing from another point of view, Figure 4 (bottom) shows the cumulative distribution of the percentage of videos from different geographical localizations that has been commented by users from Latin America, United States, and other regions. Clearly, the probability of Latin American users comment a video uploaded by other Latin Americans is higher than a Latin American comment a video upload by USA or other regions.

Figure 5 shows the video-based interactions between users from Latin America, United States, and other regions. The cumulative distribution of the percentage of users from different geographical localizations that responded videos uploaded by contributors from Latin America, United States, and other regions is shown in Figure 5 (top). As in textual interaction, we note that videos receive

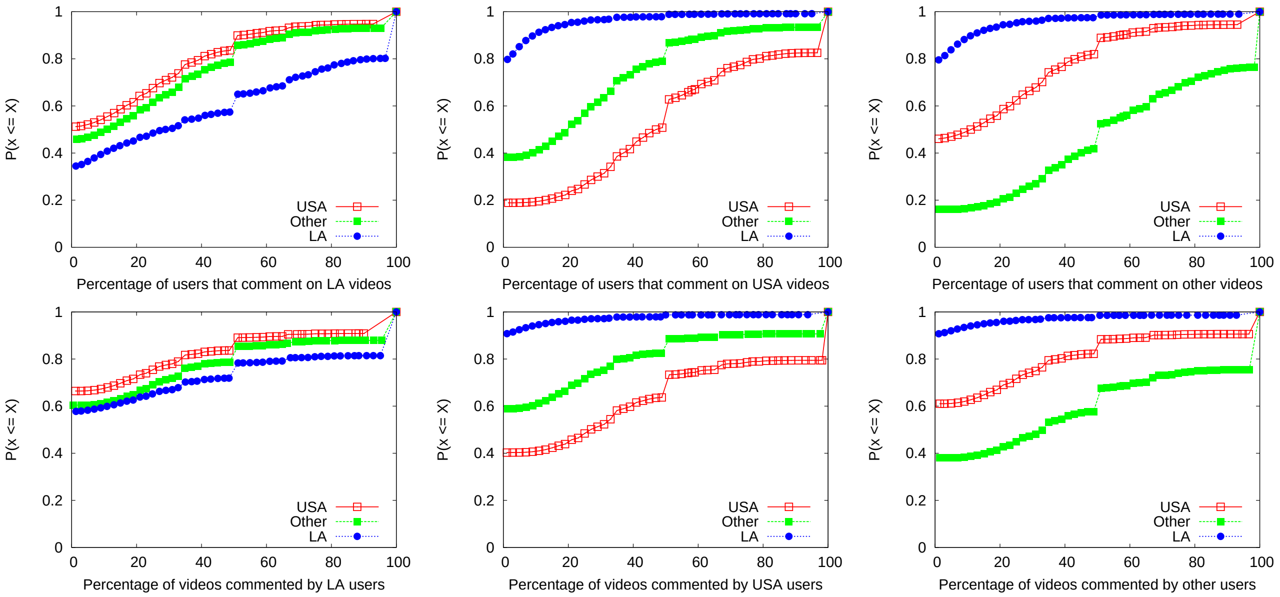

Figure 4. Textual interactions: percentage of users that comment on Latin American/USA/Other users (top) and percentage of videos commented by Latin American/USA/Other users (bottom)

more video-based messages from users of the same geographical localization of the owner than from users from different localizations. Figure 5 (bottom) shows the cumulative distribution of the percentage of videos from different geographical localizations that are responded by users from Latin America, United States, and other regions. We now note the probability of Latin American users respond videos of users from different localizations is approximately equal. This effect might be due to the small number of responses sent by Latin Americans.

Interestingly, we observe in Figures 4 and 5 that Latin American users communicate more with videos of different regions than users from United States and other regions. Almost none of the users from United States and other regions post comments or responses to videos of Latin American users.

Finally, Figures 4 and 5 show that textual and video-based communication suffer a strong geographical influence. Then we conjecture that, if comments and responses traffic are strongly affected by geographical influence (e.g. the language), the number of views of the videos and other user interactions should also be influenced by geographical location of users. More importantly, it shows that Internet traffic is influenced by geographical factors, suggesting the benefit of potential use of content distribution networks for online video sharing services.

4 Related Work

Workload characterization is fundamental to the understanding and engineering of Internet systems. Many studies focused on the characterization of Web workloads [5, 7, 8, 12]. Some of the important findings of these studies include establishing Zipf-like popularity of Web objects and the temporal and spatial reference locality in request stream. We found similar profiles for video and user popularity; however, we showed that the stream-based nature of interactions between users and objects in online video sharing service is fundamentally different than that observed in traditional Web content, based on text and image.

There has been a number of studies about stored and live media streaming. Acharya *et al.* [3] characterized user access to video objects on the Web and found that half of the requests were for a partial access of the object, indicating early stoppage of transfers by users. Costa *et al.* [11] analyzed workloads from two media servers. They found that client session arrival process follows a Poisson distribution, the time between interactive requests follows a Pareto distribution, and the popularity of the considered media objects can be modeled by the concatenation of two Zipf-like distributions. Live streaming media workload was initially characterized by Veloso *et al.* [17] and Sripanidkulchai *et al.* [16]. The former study characterized a live streaming media workload in three increasingly granular levels: client, sessions and transfers. They show that access to live objects is object driven and different from access to stored objects

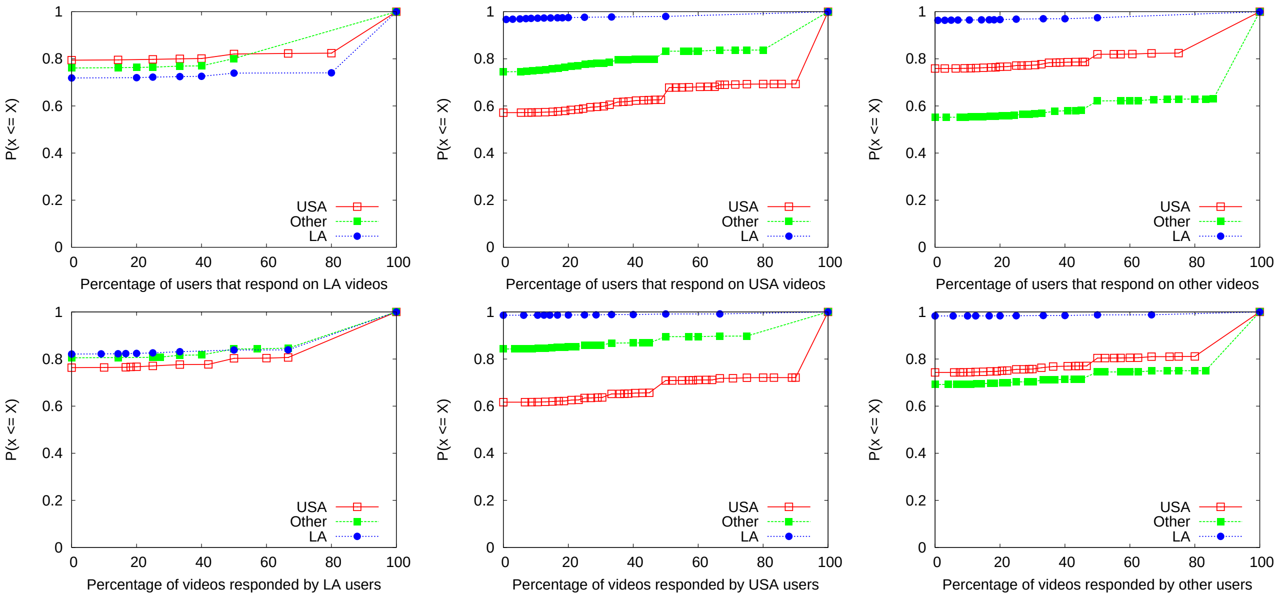

Figure 5. Video-based interactions: percentage of users that respond on Latin American/USA/Other users (top) and percentage of videos responded by Latin American/USA/Other users (bottom)

which is user driven. The latter study characterized popularity, arrival process, session duration, and transport protocol in a live streaming workload from a large content delivery network.

More recently, Li *et al.* [15] characterized streaming audio and video stored on Web pages from diverse geographic localizations. They found the distribution of the durations of streaming audio and video clips are long-tailed and that more than half of the streaming media clips encountered are video, encoded primarily for broadband connections and at resolutions considerably smaller than the resolutions of typical monitors. In 2006, Yu *et al.* [18] presented a measurement study of a large video-on-demand system deployed by China Telecom. Their study focused on user behavior, content access patterns, and their implications on the design of multimedia streaming systems.

Based on the analysis of different proxy server logs, Almeida *et al.* [6] shown evidences of the influence of regional, cultural and social issues on the performance of a caching proxy server. Moreover, many information resources on the Web are relevant primarily to limited geographical communities. In [9], Buyukkokten *et al.* exploited the geographical location information of Web sites so that search engines could rank resources in a geographically sensitive fashion, in addition to using more traditional information-retrieval strategies.

We are not aware of any other study that has considered the geographical characterization of users of a large online social video sharing service such as YouTube. This year, Halvey and Keane [13] presented a preliminary study of a much smaller set of 57 thousand users crawled from YouTube site. They showed that many users do not form social networks in the online community and a very small number do not appear to contribute to the wider community. Our work is the first that studied traffic and social interactions due to geographical localization of users and, particularly, the first focused on Latin American users of YouTube.

5 Conclusions and Future Work

In this paper we have presented what we believe to be the first geographical characterization of YouTube. Our characterization has highlighted a number of interesting differences between YouTube users from Latin America and other countries. Our main findings are summarized as follows.

- There are Latin American users that have contributed a considerable number of videos to YouTube community and are actively using all YouTube features; however, a great number of users has uploaded few videos, does not have a large number of friends, does not have a sizable list of favorite videos and sends few responses and comments.

- Videos uploaded by Latin American users present different characteristics than videos uploaded by users from other countries, being less visualized and discussed through comments or even video responses.

- Latin American users interact more with videos of different regions than other users. Almost none of the users from United States and other regions send comments or responses to videos of Latin American users.

- We conjecture the YouTube behavior of Latin American users may be constrained by the existing broadband infrastructure in Latin America. For example, the small number of uploaded videos could be limited by the asymmetric capacity of the broadband networks, that have greater download capacity compared to their upload capacity.

- Textual (through comments) and video-based (through responses) interactions on YouTube present strong influence of geographical localization. We conjecture that views of videos and other user interactions are also influenced by geographical localization. These conclusions suggest that caching and content distribution networks (CDNs) should be used to improve the performance and scalability of online social video sharing services, and also reduce Internet traffic.

As future work we plan to collect more data that to allow us to carry out a more complete geographical characterization of YouTube, analyzing the impact of language on traffic and user behavior. Another direction is to understand the various characteristics of social networks that emerge from the interactions between users and videos in YouTube across different regions of the world.

Acknowledgments

This research was sponsored by UOL (www.uol.com.br), through its UOL Bolsa Pesquisa program, process number 20060520221328a.

References

[1] The Alexa Web Site. `http://www.alexa.com`.

[2] The YouTube Web Site. `http://www.youtube.com`.

[3] C. Acharya, B. Smith, and P. Parnes. Characterizing User Access to Videos on the World Wide Web. In *Proc. Multimedia Conferencing and Networking*, 2000.

[4] Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of Topological Characteristics of Huge Online Social Networking Services. In *Proc. WWW Conference*, May 2007.

[5] V. Almeida, A. Bestavros, M. Crovella, and A. Oliveira. Characterizing Reference Locality in the WWW. In *Proc. Conf. Parallel and Distributed Information Systems*, 1996.

[6] V. Almeida, M. Cesario, R. Fonseca, W. Meira, and C. Murta. The influence of Geographical and Cultural Issues on the Cache Proxy Server Workload. In *Proc. Intl. World Wide Web Conference*, 1998.

[7] M. Arlitt and C. Williamson. Web Server Workload Characteristics: The Search for Invariants. *IEEE/ACM Trans. on Networking*, 5(5), 1997.

[8] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker. Web Caching and Zipf-like Distributions: Evidence and Implications. In *Proc. INFOCOM*, 1999.

[9] O. Buyukkokten, J. Cho, H. Garcia-Molina, L. Gravano, and N. Shivakumar. Exploiting Geographical Location Information of Web Pages. In *Proc. Workshop on Web Databases*, 1999.

[10] D. Chau, S. Pandit, S. Wang, , and C. Faloutsos. Parallel Crawling for Online Social Networks. In *Proc. WWW Conference*, 2007.

[11] C. Costa, I. Cunha, A. Borges, C. Ramos, M. Rocha, J. Almeida, and B. Ribeiro-Neto. Analyzing Client Interactivity in Streaming Media. In *Proc. Intl. World Wide Web Conference*, 2004.

[12] F. Duarte, B. Mattos, A. Bestavros, V. Almeida, and J. Almeida. Traffic Characteristics and Communication Patterns in Blogosphere. In *Proc. Intl. Conf. on Weblogs and Social Media*, 2007.

[13] M. Halvey and M. Keane. Exploring Social Dynamics in Online Media Sharing. In *Proc. WWW Conference*, 2007.

[14] S. Lee, P. Kim, and H. Jeong. Statistical Properties of Sampled Networks. *Phys. Rev. E*, 73, 2006.

[15] M. Li, M. Claypool, R. Kinicki, and J. Nichols. Characteristics of streaming media stored on the web. *ACM Trans. on Internet Technology*, 5(5), 2005.

[16] K. Sripanidkulchai, B. Maggs, and H. Zhang. An Analysis of Live Streaming Workloads on the Internet. In *Proc. Internet Measurement Conference*, 2004.

[17] E. Velloso, V. Almeida, W. Meira, A. Bestavros, and S. Ji. A Hierarchical Characterization of a Live Streaming Media Workload. *IEEE/ACM Trans. on Networking*, 14(1), 2006.

[18] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng. Understanding User Behavior in Large-Scale Video-on-demand Systems. In *Proc. Eurosys Conference*, 2006.