Youtube Revisited: On the Importance of Correct Measurement Methodology

Ossi Karkulahti, Jussi Kangasharju

Youtube Revisited: On the Importance of Correct Measurement Methodology

Ossi Karkulahti (✉) and Jussi Kangasharju

University of Helsinki, Finland
Department of Computer Science
karkulah@cs.helsinki.fi, jakangas@cs.helsinki.fi

Abstract. Measurements of large systems typically rely on sampling to keep the measurement effort practical. For example, Youtube's video popularity has been measured by crawling either related videos or videos belonging to certain categories or by using a list of, e.g., the most recent videos as the data-source. In this paper we demonstrate that all these methods lead to a biased sample of data when compared to a random sample. We demonstrate the bias by comparing the differently sampled data sets in terms of different commonly used metrics, such as video popularity, age, length, or category. The results show that different sampling methods lead to significantly different values in the metrics, thus potentially leading to very different conclusions about the system under study. The goal of the paper is not to provide yet-another-set-of-numbers for YouTube; instead we seek to emphasize the importance of using correct measurement methodologies and understanding the inherent weaknesses of different methodologies.

1 Introduction

Measuring large systems or services is challenging and typically measurements are performed via sampling since analyzing the complete system is either prohibitively expensive or even impossible. Naturally, the way the sampling is performed has a strong effect on the measurement results and the conclusions that can be drawn from them. Ideally, the sampling should be done in a way as to produce a random, representative sample of the total system, but in many cases technological limitations on the sampling may skew the process away from getting a representative sample. Using such a biased sample may yield incorrect conclusions about the properties of the system and further affect any derivative work which uses those results as its basis.

In this paper we show the effects of three different sampling methods on YouTube. YouTube is the largest and most popular video service on the Internet and has been an active focus in research for many years. Previously, YouTube's video popularity has been measured, for example, by crawling related videos [2], selecting videos belonging to certain categories [1], or by using a list of, e.g., the

most recent videos [6] as the data-source. The problem with these methods is that, while the corresponding results of the measurements are valid as such, the methods lead to a biased sample, and thus, the results are not representative of YouTube in all respects. Since other works may base their assumptions on the measured values, it is important that they indeed do represent the whole service and not a subset of it.

To demonstrate our case, we have collected three datasets, two by using methods from earlier research, and one by using a method that is based on random video IDs that has previously been used to estimate the number of videos on YouTube. We will show that, even though all data is obtained from the same source, via the YouTube API, there are noticeable discrepancies in the video popularity and other metrics depending on the method used.

Our main goal is to highlight the importance of using proper sampling techniques and show how different sampling methods can lead to different conclusions. The main contributions of the paper are the following:

- We review prior YouTube measurements and data collection methodologies and show their differences.
- We compare three existing methods for collecting YouTube video metadata.
- We demonstrate the differences in various metrics between the different sampling methods.

We also argue that, while out of the scope of this paper, the value of the result and the implications drawn from results span multiple research areas such as storage, replication, bandwidth and even wider disciplines such as marketing, user experience and user behavior.

The rest of the paper is organized as follows. In Section 2 we discuss related work and review previous measurement methods that have been used on YouTube. Section 3 presents our data collection process. The results are presented in Section 4 where we compare several key metrics obtained by the different methods and demonstrate their differences. Finally, Section 5 concludes the paper.

2 Related Work

Cha et al. [1] analyzed the video popularity of YouTube in 2006-2007. Their dataset consists of video metadata formed by crawling the indexed pages and getting videos belonging to certain categories. They had 1.7 million videos from Entertainment category and another 250,000 from Science category. Their results showed that the video popularity ranking of both categories exhibited power-law behavior "across more than two orders of magnitude" with "truncated tails" but "the exact popularity distribution seems category-dependent." The authors called for further research on the subject. The traces collected by the study have been a source for [7].

Cheng et al. [2] also measured and examined, among other things, the popularity of YouTube videos. They collected metadata for three million videos in 2007 and for further five million in 2008, using bread-first search (BFS) starting

with initial video and asking its related videos and then their related videos until the fourth depth. Looking at video popularity they observed that: "though the plot has a long tail on the linear scale, it does not follow the well-known Zipf distribution." and found "that the Gamma and Weibull distributions both fit better than the Zipf, due to the heavy tail that they have".

Since the authors were concerned that the BFS method would be biased towards more popular videos, they formed another dataset by collecting metadata of videos from the recently added list for four weeks. Comparing the two datasets they concluded that also the videos from the recently added list exhibit popularity where: "There is a clear heavy tail" and "verifying that our BFS crawl does find non-popular videos just as well as it finds popular ones".

Szabo and Huberman took a slightly different approach and wanted to see whether it is possible to predict content popularity. In the case of YouTube they measured the popularity and view counts of new videos for 30 days [6]. Their data is from 2008 and consists of 7,146 videos selected daily from the recently added list. They chose the list over other alternatives in order to get "an unbiased sample". They concluded that the popularity of a YouTube video on the 30th day can be predicted with a 10 % relative error after 10 days.

In the research mentioned above, the data has been collected either by BFS crawling, or by selecting videos of a certain category or by picking most recent videos. We will show in the results section the problems that are associated with the methods and popularity distributions they produce.

Another method is used e.g. by Gill et al. [3] who analyzed the traffic between a university campus and Youtube servers. They concluded that "video references at our campus follow a Zipf-like distribution". They reasoned it to be partly because Youtube did not allow video downloading, meaning that a user had to issue another request to see the same video again. They also found out that on a longer time frame the most popular categories were Entertainment, Music, and Comedy. Zink [9] et al. also measured the Youtube viewing and traffic patterns on a campus level and studied the effects of proxy caches to reduce traffic.

On a more general level, the importance of a correct sampling method has been noted e.g. by Krishnamurthy et al. [4] who used three different data collection methods and analyzed their strengths and weaknesses in order to examine Twitter and improve the prior research, and by Stutzbach et al. [5] who introduced a technique for a more accurate and unbiased sampling for unstructured peer-to-peer networks.

3 Data Collection

We have collected data using three different approaches. In the first approach, we started by periodically asking a list of the 50 most recently published videos using the YouTube API version 2 and later version 3. The list included information of the videos such as ID, view count, and publish date. Having obtained the IDs of the videos, we later collected their view counts after 30 days. We had done similar surveys in 2009 and 2011 and we wanted to compare the results by doing

the same procedure again in late 2013 and early 2014. We refer to this method as MR (Most Recent). The inherent problems of the MR method are that it is a slow way of collecting data and that videos for which data is collected are limited to similar age. The method is similar to one used in [6] and [2].

However, as it is not known in which manner videos end up on the MR list and thus it is not possible to know whether they constitute a representative sample, we simultaneously started collecting data using a different method in order to the verify our results. In this approach, we generated random character strings and requested through the API a list of video IDs which include the string. Hence we call this method RS (Random Strings). In more detail, the method can be described as follows. We formed four characters long strings using random characters from 'a-Z', '0-9', '-', and '_'. As the YouTube video IDs are 11-character long strings generated with the same character set, we used the strings as keywords to request video IDs containing the random strings (4 characters were the shortest strings that returned matches consistently via the search). Resulting data also included video metadata such as duration, category, etc., and on average a random string yielded 6.9 video IDs. Besides randomness, the benefits of the method are that we were able to collect a very large number of video IDs with corresponding metadata and it provided a way to get a comprehensive sample of different-aged videos. Given that different strings might match to same ID, we further pruned out the duplicates.

Interestingly, for reasons unknown to us, with this method the YouTube API only returns video IDs that have at least one '-' in them, even though, in general, video IDs do not need to contain a '-'. The "-" was usually the fifth character of the ID. However, we argue that as the search strings are randomly generated (and the IDs are likely similarly generated, although this cannot be proven), statistically the sample obtained in this manner is equivalent to a random sample over all the videos; obviously this is a potential weakness of this method. Incidentally, Zhou et al. [8] provide a detailed description and discussion of the same method, with evidence to support that it indeed provides a random sample of the videos. However, their focus is on estimating the number of videos on YouTube and they do not investigate different metrics for the videos. They also mention a potential bias in other collection methodologies, such as BFS, but do not present any evidence of that. While we strongly conjecture that the RS method provides a random sample, for the purposes of this paper, i.e., to demonstrate the differences between different sampling methods, it is not strictly necessary for the method to actually produce a random sample. A further limitation of this method is that it will not return videos with 0 views or deleted videos.

Our third method to collect data was to randomly select a video ID and then ask for its related videos and after that the related videos for all those videos up until to the fourth level. We set a limit of 50 related videos per one video, so theoretically one seed video could return up to 125,000 videos (50x50x50). The actual number of unique videos is naturally lower, due to overlap in the related videos. This can been seen as similar to breadth-first search and we shall

refer to the method as BFS. As mentioned in Section 2 this method has been used earlier by [2]. This method is a fast way of obtain a large set of IDs, since the API allows getting the information of 50 videos with just one API request compared to the average of 6.9 obtained with the random strings. Because a video can be, and usually is, related to multiple videos, the method also needs pruning to remove duplicates.

Table 1: Description of datasets

Set name	Method	Time period	N
MR-09	Most recent videos	summer 2009	9,405
MR-11	Most recent videos	summer 2011	8,766
MR-14	Most recent videos	late 2013 - early 2014	10,000
RS	Random id	early 2014	5M
BFS	BFS related videos	early 2014	5M

Table 1 shows an overview of the different datasets that we collected using the three methods described above. All of the data we have collected will be made available. In the following, we refer to the different datasets by their names and in some cases combine all three MR datasets into a single set, called MR.

4 Results

As described in the previous section, we have three datasets collected using three different methods. Now we are going to show how the datasets differ according to different typical metrics that have been used in previous research on YouTube. We start with the video popularity ranking and then use number of views, age, length, and categories to further compare the datasets. Obviously, as the MR dataset is much smaller and the videos are by definition very recently uploaded (to the time when the dataset is collected), thus it does not allow one-to-one comparison with the other two methods in some metrics.

4.1 Popularity

Figure 1 plots the videos of RS and BFS datasets ranked based on the view count in log-log scale. Both datasets have 5 million videos. As can be obviously seen, there is a clear difference in the view count distributions provided by the two methods. The data collected using BFS method has a clear two-part distribution, with a quick-dropping tail. The RS data follows more closely a Zipf distribution, with a truncated tail. Across the board, the distribution of BFS data exhibits much higher popularity (higher view counts), being in parts four orders of magnitude higher (around the millionth most viewed video). Since RS represents a random sample, it can be argued that the BFS method provides videos which significantly over-estimate the actual view counts in YouTube. We

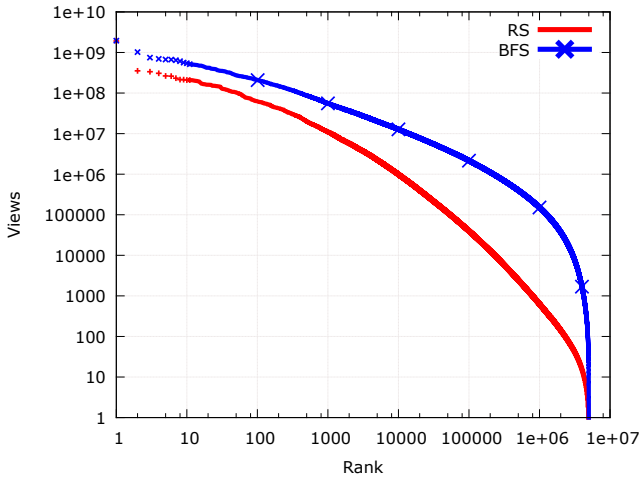


Fig. 1: Video popularity

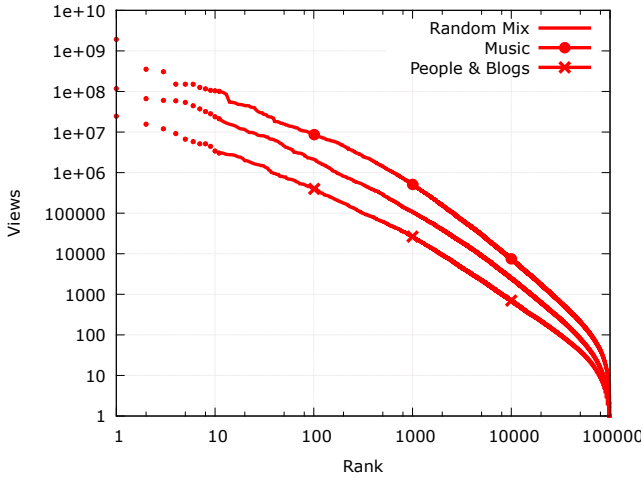


Fig. 2: Video popularity per category

suspect that when determining which videos to show as related videos, YouTube proposes videos that are more popular than average, and, thus, BFS datasets are prone to have inflated number of videos with high view counts.

A simple analysis reveals that the 10 most viewed videos in RS dataset account for 5 % of the total views, 100 most viewed for 17 %, 1000 for 43 %, and 10,000 (0.2 % of the total sample) for 74 %.

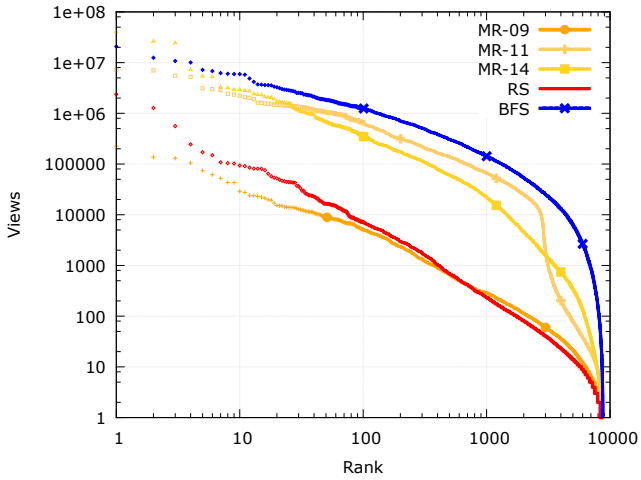


Fig. 3: 30-day view count ranking comparison

Popularity per Category Figure 2 plots the popularity distributions of different categories. We show view counts for categories Music and People & Blogs as well as the view counts for a random selection; all other categories fall somewhere between Music and People. The data is taken from RS dataset and the sample size is 100,000. While the shapes of the curves are qualitatively similar, the actual numerical values (between the categories shown here) can differ by an order of magnitude or more in terms of number of views. This illustrates that while a category-based video selection may yield qualitatively correct results, it cannot be relied to provide quantitatively correct results.

These results highlight the pitfalls in sampling method selection. Different methods may yield qualitatively, even quantitatively, similar results on some metrics, but fail on other metrics as we demonstrate below.

Popularity after 30 Days Figure 3 shows the view counts of videos 30 days after their uploading, on a log-log scale, i.e., the plot captures the popularity of one month old videos. We show all three MR datasets separately and the x-axis is limited to 8766 which is the size of the MR-11 dataset (the smallest dataset in our study) to make the curves comparable. As can be seen, the datasets have noticeably different popularity distributions. In general, both MR and BFS methods seem to overestimate the video popularity when compared to RS (Recall Figure 1 which shows the same result between BFS and RS across a larger dataset). Interestingly, the MR-09 shows a relatively straight line, close to that of RS, with a truncated tail, resembling the observations of Cha et al. [1], whereas the MR-11 would seem at least bipartite, pivoting around 12,000 views.

The view counts of MR-11, MR-14, and BFS are orders of magnitude higher than those of RS. We suspect that this is because either a) new videos on the

Table 2: View count statistics of the datasets

	N	Mean	Std. Dev	Median	Min	Max
RS	5M	16,260	1,115,835	81	1	1,920,284,708
BFS	5M	260,019	2,595,870	19,217	1	1,950,573,461
MR	21K	68,553	1,205,992	461	1	111,762,034

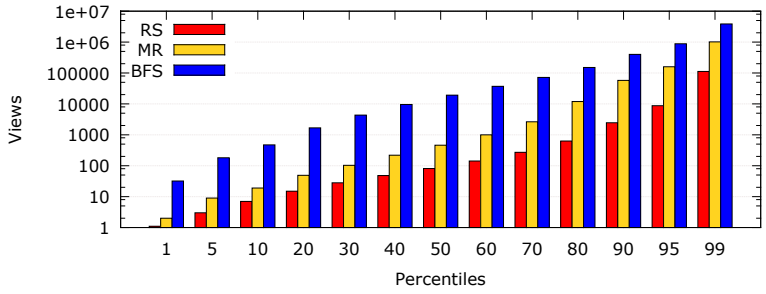


Fig. 4: View count percentiles

most recent list are such that are more likely attract more views or b) being on the list will make the videos gain more views. The same conjecture applies also more or less to the related videos.

4.2 Views

Table 2 list the view count statistics for the datasets. It should be noted that the numbers for the MR dataset are not directly comparable with the others, since the dataset includes mostly new videos and thus they have had a shorter time to accumulate views. As already stated, the BFS method favors more popular videos, which can be seen in the much higher mean and median values. In other words, in general, the videos of the BFS dataset are more viewed than those of RS. Figure 4 shows the different percentiles of the view counts. We can see that e.g. the 5th percentile of BFS is higher than the median of RS and across the board the BFS view counts are at least one order of magnitude higher than the RS ones. Figure 5 further illustrated this point by showing the median and the 5th and 95th percentiles of the RS and BFS datasets for eight years. For example, in the RS dataset the median value of 730-day-old videos is approximately 100 views. Looking at the percentiles we can see that there is overlap in the datasets, but the median of BFS is most of the time two orders of magnitude higher than the median of RS.

4.3 Age

Figure 6 illustrates the age distribution of the videos gathered by the RS and BFS methods. The MR data is left out as the age is already determined by the

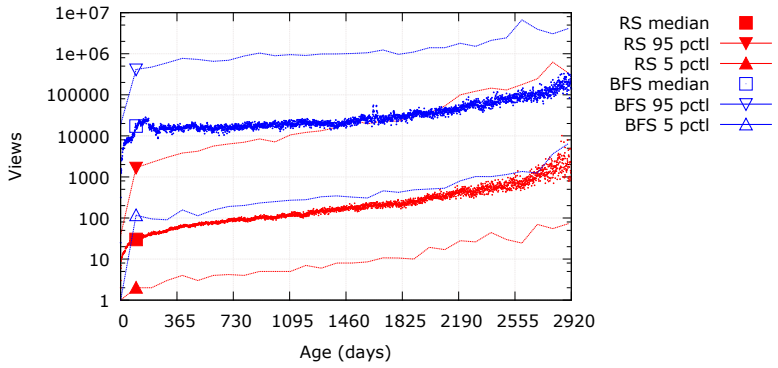


Fig. 5: Median and 5th and 95th percentiles of RS and BFS

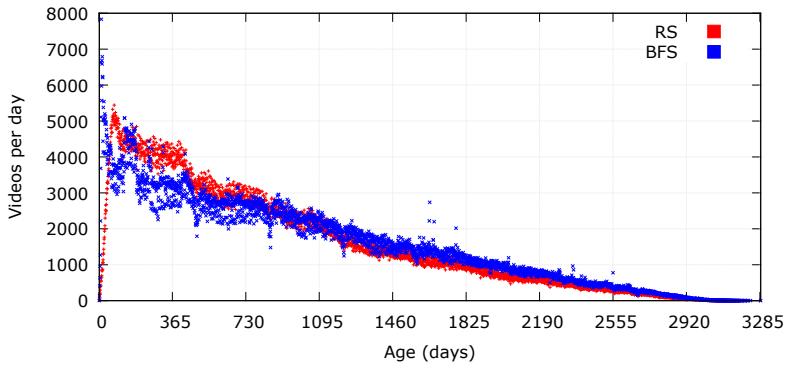
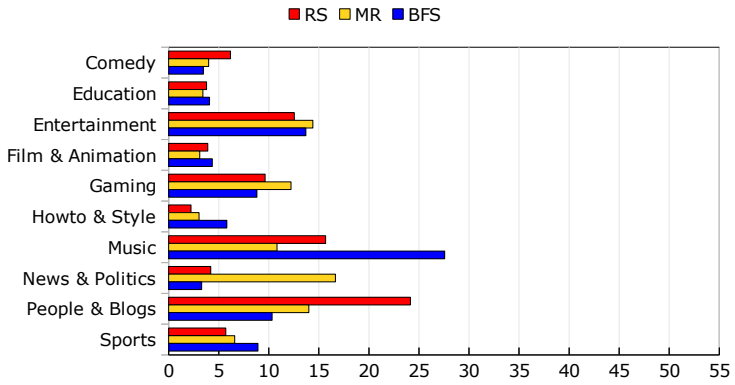


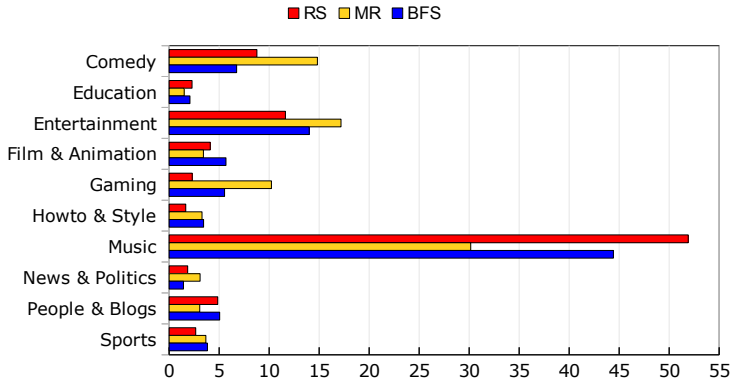
Fig. 6: Video age distribution

way the method works, limiting the data to new videos only. The plot is made by calculating the number of videos published on each day. The BFS set has less videos that are newer than three years, when compared to the RS dataset. However, for very recent videos, the BFS dataset shows a considerable increase, reaching up to more than three times the number of videos with similar age in the RS set. It therefore appears that the selection of related videos is biased towards recent videos and implies that the BFS dataset has a disproportionate number of recent videos, when compared to the RS set.

RS dataset shows a sharp decrease in the number of recent videos, but this is an artifact of the sampling method. This is because the method can only match existing videos and therefore videos that were uploaded after the data collection began have had a smaller probability of being selected, thus artificially reducing their number in the set. This effect can be eliminated simply by not counting the videos published during the data collection period.



(a) Percentage of videos



(b) Percentage of views

Fig. 7: Video categories

On a more general note, looking at the RS data, we can see that that number of videos has grown rapidly, (even exponentially in some points), and continues to do so. Videos that are less than six months old make up 14 % of all video, less than one year 29 % and less than two years 53 %. In other words, majority of the YouTube content is newer than two years and 80 % newer than four years. Hence, the rate at which videos are uploaded to YouTube is still increasing and majority of videos have been published in the past two years.

Table 3: Length statistics of the datasets

	N	Mean	Std. Dev	Median	Min	Max
RS	5M	296	614	157	1	131,516
BFS	5M	512	1,181	247	1	800,492
MR	21K	545	1,535	190	1	45,122

4.4 Categories

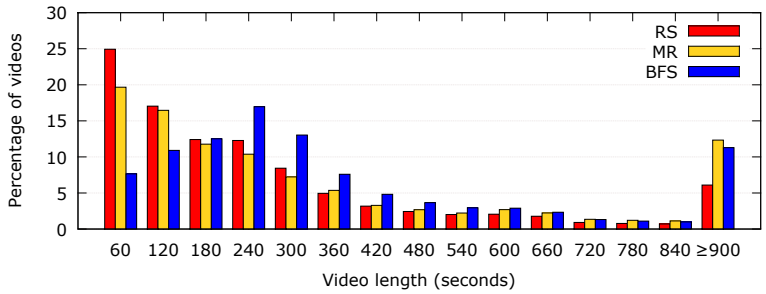
Figure 7a shows the fraction of videos in different categories in the different datasets. The bars for MR combine all the three MR datasets MR-09, -11, and -14. Interestingly, the category with most videos is different in each dataset and the differences are significant. RS has most videos from the People & Blogs category, MR's biggest category is News & Politics, and Music is the largest category for BFS. When uploading a video, YouTube requires that the user sets a category for the video. If user does not not explicitly define a category, YouTube sets the video's category to the category of the last video that the user uploaded. If no prior upload exists, YouTube sets the video's category to People & Blogs, which is a very likely explanation why the RS dataset has the most videos in the People & Blogs category. Likewise, since MR takes the videos from the (curated) most recent list, it is not surprising that topical events dominate the list. For BFS, the high number of music videos is also not surprising since suggesting another music video as a related video to another music video seems intuitive.

However, even though the number of videos in different categories is very different for the three datasets, Figure 7b shows that the distribution of number of views across categories in the three datasets is very similar. Music is the most watched category for all three datasets, followed by Entertainment and then Comedy. Again, this highlights that the results from different methods may end up looking similar on some metrics, but not on others.

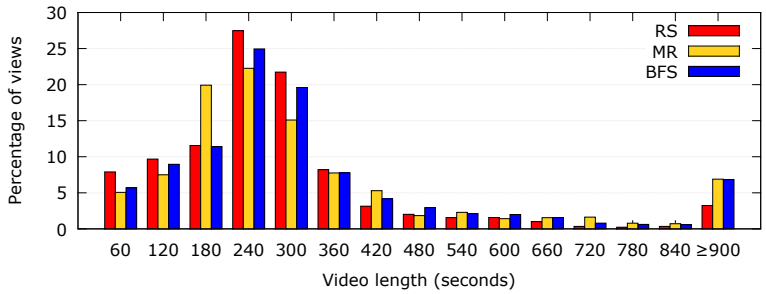
4.5 Length

YouTube used to cap the video duration to 10 minutes, but now the default limit has been extended to 15 minutes and a user can remove the limit completely by verifying the account. Table 3 shows the length statistics. The lengths are in seconds. We have checked that the maximum value for the BFS dataset is valid. The median video length is the highest for the videos of the BFS dataset, followed by MR and RS, whereas MR has the highest mean and standard deviation.

Figure 8a shows how the lengths of the videos in the different datasets vary; the videos have been rounded to the next minute for plotting. Both RS and MR show that the most common length of a YouTube video is 60 seconds or less and that majority of video are less than three minutes long. The BFS in turn indicates that most videos between three and five minutes. This can be considered further evidence that BFS promotes certain type of videos forming



(a) Percentage of videos



(b) Percentage of views

Fig. 8: Video length

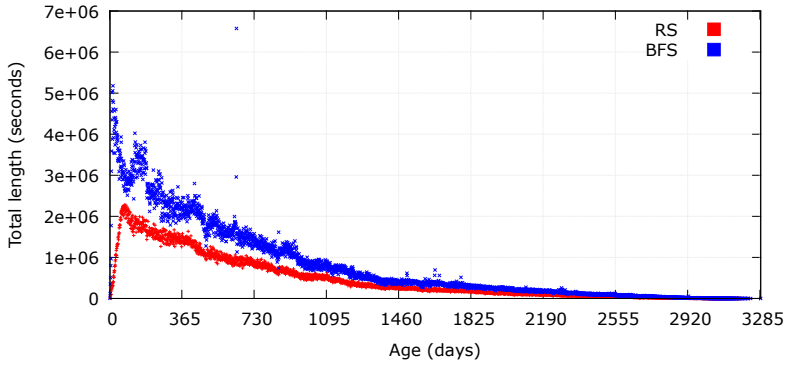


Fig. 9: Total video length per day

a biased sample; as we already saw that BFS contains more music videos which are typically three to five minutes long. Interestingly, MR and RS differ only in that MR has more videos over 15 minutes whereas RS has more videos of one minute or less.

However, Figure 8b shows videos between three and five minutes have the most views in all datasets. If this data were used to produce an estimate of how much traffic YouTube sees, all three datasets would yield similar values, with MR being likely slightly below the others as it contains proportionally more videos of around 3 minutes.

Figure 9 show total duration of videos uploaded per day as a function of the age of the videos. This could also be used to obtain a rough estimate of total storage requirements of YouTube service. Again, BFS over-estimates the video length. As the figure shows, the amount of data has risen almost exponentially for years. 40 % of the amount consists of less than one year old videos and 80 % of videos newer three years.

4.6 Summary of Results and Methods

When comparing the three methods among themselves, BFS tends to over-estimate most of the metrics we used and cannot therefore be considered a reliable method; however, it is the fastest of the three for collecting a large dataset. MR, on the other hand, is a very slow method, limited to new videos only, and it also tends towards over-estimation of the metrics. While we consider the RS method to be the most reliable, its weakness is that it is not very fast (recall that it returns on average 6.9 videos per query). Also, since all returned videos contain '-', there is potential for a bias in the returned videos, in case video IDs are not assigned randomly.

5 Conclusion

In this paper we have argued that data collection methodology can have a significant impact on what kinds of results can be obtained from measurements. We have used YouTube as an example and considered three different data collection methods, two from existing research and one adapted from previous work. By comparing the datasets obtained with the three different methods, we have shown that they differ, sometimes greatly, in many of the key metrics used in past research on YouTube. Even a large sample is not immune to the bias introduced by a particular measurement methodology, as the results of the BFS dataset demonstrate.

The random sampling method behind the RS dataset has not been used to measure different metrics on YouTube whereas MR and BFS have been used in previous research to characterize YouTube. Given the large difference between RS and the others on several key metrics, it is natural to raise questions about the general applicability of previously obtained results on YouTube done via MR or BFS methods. As we have shown in this paper, depending on the metric and the collection methodology, results may differ either qualitatively, quantitatively, or both, or they might not differ from the RS dataset. While we have strong reasons to believe that the RS method produces a representative sample of YouTube,

we cannot exclude a potential bias in its selection methodology; further research would be needed to ascertain that.

In essence, our results demonstrate that there is a need to understand the strengths and weaknesses of the different measurement methodologies in order to understand their impact on the measurement results. We believe that on the whole, a more critical approach to measurement methodologies is required in order to ensure that the measurements capture the essence of the measured system, to the extent that it is feasible.

References

1. M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 1–14. ACM, 2007.
2. X. Cheng, J. Liu, and C. Dale. Understanding the characteristics of internet short video sharing: A youtube-based measurement study. *Multimedia, IEEE Transactions on*, 15(5):1184–1194, Aug 2013.
3. P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Youtube traffic characterization: a view from the edge. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 15–28. ACM, 2007.
4. B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about twitter. In *Proceedings of the first workshop on Online social networks*, pages 19–24. ACM, 2008.
5. D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger. On unbiased sampling for unstructured peer-to-peer networks. *IEEE/ACM Transactions on Networking (TON)*, 17(2):377–390, 2009.
6. G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.
7. V. Valancius, N. Laoutaris, L. Massoulié, C. Diot, and P. Rodriguez. Greening the internet with nano data centers. In *Proceedings of the 5th international conference on Emerging networking experiments and technologies*, pages 37–48. ACM, 2009.
8. J. Zhou, Y. Li, V. K. Adhikari, and Z.-L. Zhang. Counting youtube videos via random prefix sampling. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 371–380. ACM, 2011.
9. M. Zink, K. Suh, Y. Gu, and J. Kurose. Characteristics of youtube network traffic at a campus network–measurements, models, and implications. *Computer Networks*, 53(4):501–514, 2009.