The Pareto principle on YouTube

M. Bent
Faculty of Electrical Engineering, Mathematics and Computer Science
University of Twente, Enschede, The Netherlands
m.bent@student.utwente.nl

ABSTRACT
The Pareto principle (also known as the 80/20 rule) states that about 80% of the consequences are created by 20% of the causes. In this paper we describe the distribution of several aspects of YouTube, a website used for online video sharing, to see if the Pareto principle applies to these aspects. These aspects include the views and comments of videos, and the videos uploaded and comments posted by users. After obtaining data from more than 2,500 videos, 1,700 users and 900,000 comments, we conclude that 20% of the videos causes 89% of the traffic (i.e. views) and contain 91% of the comments. In both cases these values have a relatively close resemblance to the Pareto principle. We also conclude that 20% of the users is responsible for 42% of the comments and 45% of the uploaded videos, indicating there is no distribution similar to the Pareto principle in these cases. After comparing the results with previous research we conclude that the distribution of traffic and comments over videos has changed, while the age of YouTube has doubled. These result can benefit YouTube (and presumably similar websites), because research has shown that caching popular videos can reduce server traffic by 50% and server load by 75%.

Categories and Subject Descriptors
H.3.5 [Information Storage and Retrieval]: On-line Information Services – Web based services; I.4 [Social and Behavioral Sciences]: Sociology

General Terms
Human Factors, Measurement

Keywords
Pareto principle, 80/20 rule, YouTube

1. INTRODUCTION
YouTube is a website where people can watch and share online videos with other users of YouTube [14]. The website was founded in February 2005, and was officially launched in December 2005. In November 2006 YouTube was purchased by Google Inc. Besides watching videos on YouTube, other features include embedding videos in websites, creating RSS-feeds and watching videos on a mobile phone. Registered users can create their own homepage and upload videos, add friends and send messages to other registered users. But also post comments at videos, rate videos, create playlists and subscribe to channels and other users. YouTube also offers an API to interact with the website via applications or other websites.

The Pareto principle (also known as the 80/20 rule) states that about 80% of the consequences are created by 20% of the causes [8] [9]. This rule is named after Vilfredo Pareto, an Italian economist, sociologist and philosopher, by Joseph M. Muran, a management consultant. Pareto observed that 80% of the land was owned by 20% of the population. The Pareto principle is often used in management and economics, but also in computer science and human activity. The Pareto principle is a simplified version of the Pareto distribution, a mathematical power law probability distribution. The numbers 20 and 80 are not mathematically fixed, but are used as a rule of thumb application. If results are around these numbers, it gives a good indication of the distribution of these results.

In this paper we will research the distribution between several aspects of YouTube, and see if the Pareto principle applies to these aspects. These aspects include the number of views of videos and the number of comments posted at a video, and also the number of comments a user has posted and the videos a user has uploaded. We limit our research only to these features, because these are key features of the site and because of time constrains. We will also compare our results to results from previous research by Cheng, Dale and Liu [3], to see if the distribution has changed over time. To find an answer to the problem stated above, the research question is:

"Does the Pareto principle apply to aspects of YouTube?"

We will answer this question, by answering the following sub questions:

1) What relevant research is available?
2) How is the traffic (i.e. number of views) distributed over the videos?
3) How are the comments distributed over the videos?
4) How are the uploaded videos distributed over the users?
5) How are the comments distributed over the users?
6) How has the distribution of the traffic and the comments over the videos changed over time?

To answer the first sub question we will obtain relevant literature using databases like Scopus, Web of Knowledge and Scholar. We then discuss several of these papers. To answer the next four sub questions we will first obtain data from YouTube using the application programming interface (API). This data is then stored in a database so it can be analyzed, using SQL-queries. To visualize the results, the different distributions will be plotted in graphs. To answer the last sub
question we will plot our results in a graph, compare them with the results from Cheng et al. and discuss the similarities and differences.

This research is interesting because it has the potential to decrease the internet traffic. In a recent forecast of internet traffic for 2013, Cisco Systems Inc. (Designer of electronics, networking and communications technology and services) expects 90% of the traffic will be online video, reaching over 18 Exabyte’s per month [4]. YouTube is number four of the most visited websites worldwide, and the most popular video website [1]. With the recent support for high definition video resolutions (like 1080p) the traffic per video more than doubles [13]. This suggests a large part of the video traffic will be caused by YouTube. If popular videos are cached, the server and internet traffic can be reduced [2]. Also, changes over time in distributions can be predicted, based on the differences and similarities between our results and the results from Cheng et al.

The rest of the paper is organized as follows. In section 2 related work is discussed. In section 3 we describe how we will collect and analyze the data. After analyzing the data the results will be described in Section 4, with sub sections for sub questions two to five. In section 5 we will answer sub question six by comparing our results with the results from Cheng et al. and in section 6 we conclude, including suggestions for future work.

2. RELATED WORK
Since the launch of YouTube back in 2005, the website has been subject to lots of different kinds of research. Not only in the area of computer science, but also in social and cultural studies. In this section some previous research related to YouTube is discussed. The papers discussed are obtained using online databases like Scopus, Web of Knowledge and Scholar. By searching for keywords like YouTube and Pareto, links to relevant papers are found. By following citations in the papers, or by finding literature where the papers themselves are cited, even more literature is found.

In 2007, Gill, Arlitt, Li and Mahanti presented a traffic characterization study of YouTube, by observing 25 million transactions between users on a network and YouTube, and by monitoring the popular videos [6]. They concluded there are many similarities to traditional web and media streaming workloads. The most significant difference was a sustainable growth in the publishing of content, because everyone is allowed to publish content. They looked for evidence of the presence of the Pareto principle in the video workload which was observed in web and media server workload studies, but did not find it. They explain this by pointing at the diversity and the amount of videos. Also, in the video requests they did not find evidence of the Pareto principle. The top 20% of videos accounts for 52.4% of the video requests, not 80%.

In 2008, Cheng, Dale and Liu have collected a large amount of data over a 3-month period [3]. After a detailed investigation they demonstrated that, while sharing certain similar features with traditional video repositories, YouTube exhibits many unique characteristics, especially in length distribution, access pattern and growth trend. They also investigated the social network among videos, and found characteristics of small-world linking, and large clustering. In their research, they discovered that the active life span of a video follows a Pareto distribution, implying most videos are only popular during a short span of time. Statistics of videos were also researched. These statistics include size, bit rate, category, ratings, views and comments. Cheng et al. plotted the number of views and the number of comments versus the rank of the video. To see if the distributions of the views and comments over the videos have changed over time, we will compare our results with theirs in section 5.

Cha, Kwak, Rodriguez, Ahn and Moon have analyzed popularity of videos on YouTube [2]. They studied the lifecycle of videos, statistical properties of request and their relation with video age, aliasing (same video several times) and illegal content. In their research they found evidence of the Pareto principle between the number of views and the least r-th popular videos, the 10% most popular videos account for nearly 80% of the views. They also explored the benefits of alternate distribution schemes, namely, caching and peer-to-peer. In a hybrid finite cache, a cache which contains the long-term popular videos and the daily most popular videos, only 17% of the requests cannot be handled by the cache, the other 83% can.

Crane observed, by studying the spreading of YouTube videos over time, that about 90% of the videos behave like a Poisson process or do not experience much activity [5]. The other 10% can be divided into three subgroups; junk, quality and viral videos. Thus the kind of spreading of videos over time behaves according to the Pareto principle.

In the above discussed research the researchers found several aspects of YouTube which behave like the Pareto principle, and how a server can benefit from this, using caches. While the research focuses primarily on advanced aspects of YouTube, our research will focus primarily on basic features, like viewing and uploading videos and posting comments. This does not mean the basic features are not discussed in the related work. To the extend these features are already discussed, we will comment on the differences and similarities between our and their results.

3. MEASUREMENT
To answer the research questions, data is collected and analyzed. This section describes what data is collected, how the data is collected, and how the data will be analyzed.

3.1 Measurement setup
A YouTube video has a lot of attributes which can be collected. We only need the attributes which are relevant to the research questions. These include for a video:

- The number of views
- The number of comments
- The uploader
- The comments

And for the comments:

- The users who posted them

There are two methods to collect the data that we need. We can manually retrieve the data by visiting the website, or we can use a tool to automatically collect the data. Because the first method is very time-consuming, and because there are tools to ‘crawl’ a website and collect the data, we will use the second method. We will use TubeKit, a toolkit to create an YouTube crawler [11]. TubeKit can create a crawler which can collect up to 16 attributes from a video, including the attributes
we need [10] [12]. We did not find other tools or scripts which have a non-commercial license and have the required functionality.

TubeKit works as follows. In the setup the relevant attributes which need to be collected are selected. Then, it creates a database where the data will be stored and it creates a PHP-script which can be used to collect the data. Next we need to provide the crawler with search terms which will be used to find videos.

After this setup we can run the PHP-script. The script creates an RSS-feed for a search term and uses the YouTube API to collect data from the videos in the RSS-feed. The data from a video is stored in an XML document, which is obtained using the API. The RSS-feed can contain up to 1,000 videos, ordered by relevance. For every video, the script stores the data which need to be collected in the database.

Because YouTube has well over 100 million videos [3], collecting data from all the videos is not possible. Therefore we will use a representative sample of the videos. We cannot collect data from a random sample of videos, because TubeKit uses search terms to find videos. We use five different search terms. These search terms are:

- Obama
- Google
- Windows
- Accident
- Country

These terms are randomly chosen. Because some of the terms are nouns, and others can be considered brands, the results could be influenced. For example, the term ‘Google’ may have a large number of videos uploaded by Google Inc. Therefore, we will also discuss the differences between (types of) terms in section 4.

The number of videos each search term has does not matter, because the used sample size (explained next) is large enough for every amount of videos a search term can have.

For every search term, we will collect data of about 500 videos, because some preliminary tests have shown that the number of videos in the RSS-feed for the different search terms is between 500 and 600. We use equation 1 [7] to determine if the sample size is statistically correct for a total population of one million videos and with a 95% confidence interval and a sample error of 5%. In equation 1, S is the sample size, $X^2$ is the confidence level value obtained from the Chi-Square table (3.84 for 95% confidence), n is the population size, p is the population proportion (default .50) and d is the confidence interval (.05 for 5%).

$$S = \frac{X^2 \cdot n \cdot p \cdot (1 - p)}{d^2 \cdot (n - 1) + X^2 \cdot p \cdot (1 - p)}$$

Equation 1: Sample size for a known population.

Equation 1 gives us a minimum sample size of 384 videos. For smaller populations, the sample size also becomes smaller. It can be proven that for larger populations the minimum sample size always is 384. Because we have a smaller population and because we collect more samples than needed, the sample size is certainly correct. The crawler created with TubeKit only collects data from videos. To collect data from comments TubeKit provides another script. This script collects and stores for a video the comments and for every comment the name of the user who posted the comment. It uses the videos from the first script. Per video up to 1,000 comments can be collected, due to the limitations of the RSS-feed.

3.2 Analysis approach

We will analyze the obtained data to find answers to the research questions. For every sub question we will create a graph in which the cumulative distribution is plotted. The horizontal axis shows the percentage of videos or the percentage of users, while the vertical axis shows the percentage of the number of views, comments or videos. On the horizontal axis, we will use a domain of 0% to 100% with intervals of 5%. On the vertical axis we will calculate the corresponding value in a range from 0% to 100%. We will also plot the Pareto principle together with the results to see if the results are distributed the same way as the Pareto principle.

3.2.1 Sub question 2

Because the data is stored in a database, we use SQL-queries to calculate the corresponding value. For sub question 2 we use the following SQL query to calculate the results:

```sql
SELECT * FROM pr_once p
WHERE p.query_id = X
ORDER BY p.view_count DESC LIMIT Y;
```

This query selects for every user the number of videos, with the most views first. Then it selects the first Y elements. Y is calculated by multiplying the total number of views with the percentage. Next a view is created from the result:

```sql
SELECT sum(i.view_count) FROM view_i i;
```

This is done 21 times (0; 0.05; 0.10; ...; 1). For all these views, we calculate the number of views for Y. Finally, we will change this value into a percentage by normalizing it.

3.2.2 Sub question 3

For sub question 3 we use the following query:

```sql
SELECT * FROM pr_once p
WHERE p.query_id = X
ORDER BY p.comment_count DESC LIMIT Y;
```

This query works the same as the query of sub question 2, but instead of the number of views, it uses the number of comments. Subsequently, we use the same method as in sub question 2.

3.2.3 Sub question 4

For sub question 4 we use the following query:

```sql
SELECT count(p.video_id) AS nr, p.username
FROM pr_once p WHERE p.query_id = X
GROUP BY p.username
ORDER BY nr DESC LIMIT Y;
```

This query selects for every user the number of videos he has uploaded, the user with the most uploads first. Subsequently, we use the same method as in sub question 2.

3.2.4 Sub question 5

For sub question 5 we use the following query:
Table 1: Statistics of the collected data for the different search terms, and the total amount. Obtained November 4th 2009.

<table>
<thead>
<tr>
<th></th>
<th>Obama</th>
<th>Google</th>
<th>Windows</th>
<th>Accident</th>
<th>Country</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of videos</td>
<td>513</td>
<td>538</td>
<td>546</td>
<td>503</td>
<td>538</td>
<td>2.638</td>
</tr>
<tr>
<td>Total no. of views</td>
<td>285.389.224</td>
<td>179.461.711</td>
<td>782.454.412</td>
<td>478.328.343</td>
<td>413.512.350</td>
<td>2.139.146.040</td>
</tr>
<tr>
<td>No. of uploaders</td>
<td>231</td>
<td>277</td>
<td>477</td>
<td>431</td>
<td>401</td>
<td>1.817</td>
</tr>
<tr>
<td>Total no. of comments</td>
<td>1.129.422</td>
<td>300.072</td>
<td>1.786.840</td>
<td>676.261</td>
<td>766.155</td>
<td>4.658.750</td>
</tr>
<tr>
<td>No. of commenters</td>
<td>102.962</td>
<td>67.481</td>
<td>179.354</td>
<td>130.4</td>
<td>156.837</td>
<td>637.083</td>
</tr>
</tbody>
</table>

SELECT COUNT(p.comment) AS nr, p.author
FROM pr_comments_X p GROUP BY p.author
ORDER BY nr DESC LIMIT Y;

This query works the same as the query of question 4. Also, the same method is used to calculate the percentages.

3.2.5 Sub question 6
To see if the distributions have changed over time, we will compare our results with the results from Cheng et al. Because Cheng et al. have plotted their results in a graph with a log-log scale, we will also plot our results in this way, so we can easily compare them. The horizontal axis contains the rank of the video, ordered by number of views / comments. The vertical axis contains the number of views / comments.

4. RESULTS
In this section the results obtained by the method described in section 3 are discussed. Table 1 contains some statistics of the obtained data which is used to calculate the results, including the number of videos, views, comments, uploaders and commenters.

In every sub section the results for one of the sub questions will be discussed using the graph of the distribution. Also the 80/20 Pareto principle function obtained from [8] is plotted in the graphs, so the distributions can be compared. There are no error bars in the graphs, because they greatly reduce the visibility, but the errors are at most 5%.

4.1 Characteristics of views
The distribution of the number of views over the videos is plotted in Figure 1. For all five search terms the distribution is plotted, together with a plot of the Pareto principle.

![Figure 1: The number of views distributed over the videos.](image)

The plots for all the search terms have the same form. The plots also have the same form as the plot of the Pareto principle, but approach the 100% asymptote earlier. The mean values of the search terms suggests 10% of the videos generate 77% of the views and 20% of the videos generate 89% of the views. This is higher than the Pareto principle states (20% causes 80%), but this does not change the fact that a few percentage of videos generate a large percentage of views.

4.2 Characteristics of comments at videos
In Figure 2, the distributions of the comments over the videos are plotted.

![Figure 2: The number of comments distributed over the videos.](image)

The distribution is almost identical to the distribution of the views, except the plots increase slightly faster. Thus, the plots have the same form as the plot of the Pareto principle. Again, the plots reach the 100% asymptote earlier than the plot of the Pareto principle. The mean values suggests 10% of the videos contain 81% of the comments, and 20% of the videos contain 91% of the comments.

4.3 Characteristics of video-uploads by users
The number of uploads distributed over the uploaders is plotted in Figure 3, together with the plot of the Pareto principle.

![Figure 3: The number of uploads distributed over the users.](image)
It is immediately obvious the plots do not have the same form as the plot of the Pareto principle. While the Pareto principle is an exponential function, the plots of the search terms are almost linear. Linearity indicates that users have uploaded roughly the same amount of videos (in our measurements most users have uploaded 1 or 2 videos). There is a difference between several search terms in the increase in the first 10%, caused by the type of term. Terms which can be considered brands have a greater increase than terms based on nouns. The top videos obtained by search terms Accident and Country (nouns) are uploaded mostly by home-users, while the top videos from Obama and Google (brands) are uploaded by professional companies (Associated Press, the White House and Google). The term Windows which is also a brand (and a noun), does not have the same characteristics because there are no professional companies (e.g. Microsoft) who upload videos for this term. Even so, the distribution is not similar to the Pareto principle. The mean values suggest 10% of the users have uploaded 36% of the videos, and 20% of the users have uploaded 45% of the videos. 80% of the uploads is caused by 71% of the users (and not 20% as the Pareto principle states).

4.4 Characteristics of comments posted by users

Figure 4 shows the distribution of the number of comments over the users.

![Figure 4: The number of comments distributed over the users.](image)

Similar to the distribution of the uploads, the distribution of the comments over the users does not have a Pareto-like distribution. After the initial increase of comments for the first 20% of users, the plots of the search terms become linear. There is no notable difference between the search terms based on their type. The mean values of the search terms suggests 10% of the users have posted 30% of the comments, and 20% of the users have posted 42% of the comments. 80% of the comments are posted by 73% of the users (versus 20% stated by the Pareto principle).

5. CHANGES OVER TIME

As mentioned in section 2, Cheng et al. have researched statistics of YouTube, including the distribution of views over videos and comments over videos. The data Cheng et al. use in their plot is collected on April 16th 2007. This is slightly more than two years after YouTube was founded. Our data is collected on November 4th 2009, about four and a half years after YouTube was founded. Roughly said, YouTube’s age has doubled. In this section we will compare their results with our results and comment on the differences and similarities.

5.1 Changes in the distribution of views

Figure 5 shows the number of views as a function of the rank of the video, as obtained by Cheng et al. The figure also contains plots of Weibull, Gamma and Zipf distributions, but we will ignore them. Our data is plotted in a similar way in Figure 7.

The curves have the same basic form. They start with a linear decrease and have a tail which decreases tremendously. There are also notable differences. The maximum number of views the most watched video has, increased from about 5 million to almost 100 million. Also the tail starts earlier. The tail in Figure 5 starts in the middle of the y-axis (10³). In Figure 7, the tail starts at 3/4 of the y-axis (10⁶). The results indicate that a subset of videos has become more popular. One explanation is that there are a lot more users who watch videos, and the lifetime of a video may be larger (two vs. four and a half year). The larger tail indicates that there are less unpopular videos. One explanation for this is that because of the increase of users, also unpopular videos get more views. The larger tail could also be due to the absence of very unpopular videos in the used RSS-feed.

If the observed changes are applied to a plot of the Pareto principle, this would lead to a slower increase in the first part of the plot, and a later approach of the 100% asymptote. If we apply this to the results in Figure 1 in section 4.1, the plots of the results will approach the plot of the Pareto principle.

Cha et al. [2] found 10% of the videos account for nearly 80% of the views. This is consistent with our findings, where 10% of the videos generate 77% of the views.

5.2 Changes in the distribution of comments

Like the distribution of the views over the videos, the distribution of the comments over the videos has also previously been researched by Cheng et al. Figure 6 shows their findings, together with a plot of the ratings (which we will ignore). Our results are plotted in Figure 7.

The plots in Figure 6 and 7 have roughly the same basic form. They start with a linear decrease and finish with a tail which decreases tremendously. The maximum number of comments posted at a video has increased from 4,000 to 400,000 comments. An explanation for this is that the number of users who post comments has increased, and the videos may have a greater lifetime. The tail in Figure 7 is larger than the tail in Figure 6, meaning there are fewer videos with few comments. This could be due to the increase of users, because more users increase the chance of a user posting a comment at a video without comments. Again, the absence of very unpopular videos in the RSS-feed could also influence the tail of the plot.

These finding will have the same effect on a plot of the Pareto principle as the findings in the previous section. The initial increase slows down, and the asymptote is reached later. If applied to the results in Figure 2 in section 4.2, this would bring the plots of the results closer to the plot of the Pareto principle.

The plot of the distribution of views over videos and the plot of the distribution of the comments over videos (Figure 7) also have the same form. This is consistent with the results obtained in section 4.1 and 4.2, because in those sections the plots of the distributions (Figure 1 and Figure 2) also have the same form.
6. CONCLUSIONS

In this paper we have analyzed the distribution of several aspects of the video sharing website YouTube. After collecting data from videos, users and comments for five different search terms, we have observed several aspects of YouTube behaving like the Pareto principle, while other aspects behave in a different way, mostly linear. For the different search terms of an aspect, the distributions are roughly equal. Only in the distribution of the video-uploads over the users is a clear difference between types of terms visible. Brand-based search terms have a higher percentage of videos uploaded by the top 10% uploaders than noun-based terms, caused by professional uploaders.

The traffic of videos (i.e. number of views) is distributed over the analyzed videos closely resembling the Pareto principle. As shown in Figure 1, the plot of the distribution is similar to the plot of the Pareto principle. The number of views is distributed this way, because only a few popular videos cause almost all of the views.

The number of comments posted by a video is also distributed over the analyzed videos like the Pareto principle. This is shown in Figure 2. Like the number of views, almost all of the comments are created for a few very popular videos. These popular videos are not the same videos as the videos creating the most views, because 5% of the 20% most popular videos have disabled the ability to post comments.

The number of videos uploaded is not distributed over the users like the Pareto principle (Figure 3). Only a small percentage of the users upload a lot of videos, but most of the users upload approximately the same number of videos, resulting in a linear plot.

The distribution of the comments over the users who posted them is also not distributed like the Pareto principle. This is shown in Figure 4. A small percentage of the users post a lot of comments, but most of the users post a few comments, resulting in a linear plot.

The distribution of the traffic (i.e. number of views) over the videos and the number of comments over the videos changes over time. In comparison with data obtained two and a half years ago (half the age of YouTube), analysis shows popular videos have more views and more comments. Also there are less very unpopular videos.

6.1 Future work

Future work could focus on how the results we obtained can be used to improve aspects of YouTube. We already pointed out that caching has potential to improve server traffic and load, but there are probably more solutions. Also, the benefits for end-users can be researched. The distributions on other community based websites like flickr or Facebook could be researched as well. We showed distributions change over time. Future work could also focus on methods to predict changes in distributions.

7. ACKNOWLEDGMENTS

We would like to thank the members of the Bachelor Referaat track “Broadband for all” for their comments and reviews that helped to improve this paper. And we thank Aiko Pras for guiding us through the progress of writing a paper.

8. REFERENCES


Proceedings of the 7th ACM SIGCOMM conference on Internet measurement (San Diego, California, USA, October 24 - 26, 2007), p. 1-14.


