HIGH SCHOOL EDITION
Journal of Student Research
Volume 10 Issue 3 (2021)

Investigation of Benford's Law with YouTube Social Media Statistics

Chi-Wei Chen[1], Shao-Yu Yu[1] and Hsin-Ye Chen[#]

[1]Taipei Fuhsing Private School, Taipei, Taiwan, R.O.C.
[#]Advisor

ABSTRACT

In this study, we used social media data to investigate Benford's Law. In our experimental analysis, we used three control variables: Total Subscriptions, Total Views, and Video Uploads of YouTube channels to verify if the data is artificial and whether or not it fits Benford's Law. We noticed how Total Subscriptions does not fit Benford's Law for the top 5000 most-subscribed channels, and Total Views doesn't fit for the top 5000 most-viewed channels. The reasons that cause this difference are further investigated in this paper. We also proposed a mathematical model to verify if other datasets fit Benford's Law. After curve fitting the experimental data, results revealed that closer a and b values in our mathematical model indicate that a dataset fits Benford's Law.

Introduction

The advent of modern technology and the Internet of Things has provided the public with more convenient methods for obtaining data and knowledge, thus increasing the importance of one's ability to discern true data from false data and rumors. During the past few decades, experts in the fields of Accounting and Statistics have utilized Benford's Law[1] to assess the reliability of data. Benford's Law is the result of physicist Frank Benford's observations across diverse sets of real life data in 1938, including the surface area of 335 rivers, the size of 3259 populations in the US, and the molecular mass of 1800 molecules, all of which show results that conform with the first-digit frequencies as described by Benford's Law: in a randomly generated and evenly distributed dataset, the frequencies of data having 1, 2, and 3 as their leading digit are 30.1%, 17.6%, and 12.5%, respectively, with the frequencies of all following numbers decreasing[2][3]. Today, Benford's Law is used to test the reliability of Presidential election voting counts, financial statements, and many other types of data[2][4]. Being an easily implemented test for evaluating data credibility, Benford's Law has been the topic of many research papers. However, the limits to the applicability of Benford's Law are much less than specific, resulting in many misinterpretations of data and misconceptions [4][5]. The consensus is that Benford's Law is applicable only to "randomly and uniformly distributed data". According to this rule of thumb, it makes sense for cheque numbers and ID numbers to not conform with Benford's Law, but this definition is often blurred when applied to other datasets, causing randomly generated data to not follow Benford's Law while others do. In addition, many researchers have proven Benford's Law under the assumption that the growth rate of data is proportionate to its current value, but not all datasets that follow Benford's Law have this property. In this research, we further discuss the applicability of Benford's Law to data from social media platforms and the mechanics behind Benford's Law.

Review of Literature

We wish to derive a mathematical model of Benford's Law through mathematical analysis. The following is the theoretical proof of Benford's Law:

First, we hypothesize that there is a set of data, whose rate of growth is proportional to its value, which can be expressed with the following equation:

$$\frac{\Delta N}{N \times \Delta t} = C$$

We can analyze this equation to reach two conclusions:
The value of this data will grow exponentially, as presented by the following equation:

$$N = N_0 \times e^{ct}$$

The time required for this data value to grow from N1 to N2 can be estimated by the following equation:

$$t = c \times log \frac{N_2}{N_1}$$

This means that the time required for this data to grow from having 1 as its first-digit to having 2 as its first-digit is:

$$t_1 = c \times log \frac{2}{1}$$

According to the same equation, we can estimate the time needed for the first-digit to change from 2 to 3:

$$t_2 = c \times log \frac{3}{2}$$

Following this pattern, if the value starts with n as its first-digit, the time needed for its first-digit to grow from n to n+1 is:

$$t_n = c \times log \frac{(n+1)}{n}$$

The time needed for this data to grow from 1 to 10 is:

$$t = t_1 + t_2 + t_3 + \cdots + t_9 = c \times log10 = c$$

From the information above, we can deduce that the probability of a value's first-digit being 1 while it's growing from a single-digit number to a double-digit number or more can be calculated from the following equation:

$$P_1 = \frac{t_1}{t} = log2$$

The probability of the first digit being 2 will follow this equation:

$$P_2 = \frac{t_2}{t_1} = log\frac{3}{2}$$

The probability of each number appearing at the first-digit is:

$$P_1 = log2 P_2 = log\frac{3}{2}$$
$$P_3 = log\frac{4}{3}$$
$$\vdots$$
$$P_9 = log\frac{10}{9}$$

The above is the theoretical proof of Benford's Law.

From the proof above, we can know that the values of P_1,P_2,P_3,P_4,P_5,P_6,P_7,P_8, and P_9 are 0.301, 0.176, 0.125, 0.097, 0.079,0.067, 0.058, 0.051, and 0.046 and create **Figure 1**.

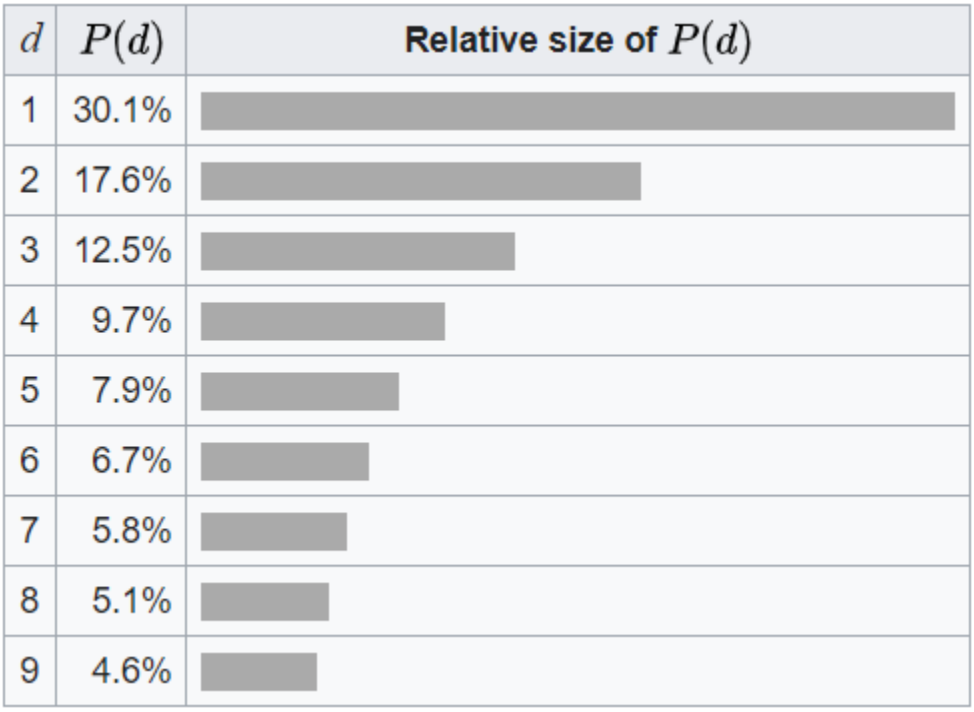


Figure 1. The first digit distribution of Benford's Law

Methods

Collect data of social media platforms

In order to prove Benford's Law, we used YouTube statistics from the analytics website SocialBlade (https://socialblade.com) as an example. The three variables taken into consideration are Total Video Views, Video Uploads, and Total Subscriptions of top YouTube channels. With these three parameters as variables, we would discuss the applicability of Benford's Law by cross comparing the first-digit distribution of datasets with different parameters. First, we used Total Subscriptions as the control variable and found the top 5000 channels with the highest subscription numbers and their respective Video Uploads, Total Video Views, and Total Subscriptions, graphing the dataset's first digit distribution. Next, we gathered and graphed data with Video Uploads and Total Subscriptions as independent variables. Because of the online website's large dataset, we collected data with a Python Crawler program and recorded the data on Microsoft Excel sheets. The Crawler code is shown in **Figure 2**.

```
from selenium import webdriver
from selenium.webdriver.chrome.options import Options
from bs4 import BeautifulSoup, Tag, NavigableString
import time
import os
import csv
import time

options= Options()
options.add_argument('--headless')
options.add_argument('--disable-gpu')
driver=webdriver.Chrome(os.getcwd()+"/chromedriver",chrome_options=options)
driver.get('https://socialblade.com/youtube/top/5000/mostviewed')
#https://socialblade.com/youtube/top/5000/mostsubscribed

sourceCode=BeautifulSoup(driver.page_source,"html.parser")
youtubers=[]
youtubers.extend(sourceCode.find_all('div',{'style':'width: 860px; background: #fafafa; padding: 10px 20px; color:#444;
                                        font-size: 10pt; border-bottom: 1px solid #eee; line-height: 40px;'}))
youtubers.extend(sourceCode.find_all('div',{'style':'width: 860px; background: #f8f8f8;; padding: 10px 20px; color:#444;
                                        font-size: 10pt; border-bottom: 1px solid #eee; line-height: 40px;'}))
youtubers.extend(sourceCode.find_all('div',{'style':'width: 860px; background: #fafafa; padding: 0px 20px; color:#444;
                                        font-size: 10pt; border-bottom: 1px solid #eee; line-height: 30px;'}))
youtubers.extend(sourceCode.find_all('div',{'style':'width: 860px; background: #f8f8f8;; padding: 0px 20px; color:#444;
                                        font-size: 10pt; border-bottom: 1px solid #eee; line-height: 30px;'}))

count=0
print("Start")
with open('views5000_May_raw.csv', 'w', newline='',encoding="utf-8") as file:
    writer = csv.writer(file)
    writer.writerow(["Rank","Name","Uploads","Subscribers","Views"])
    for youtuber in youtubers:
        time.sleep(0.1)
        try:
            name=youtuber.find_all("a")[0].text
        except IndexError:
            name="N/A"
        try:
            upload=youtuber.find_all("span",{"style":"color:#555;"})[0].text.strip()
        except IndexError:
            upload="N/A"
        try:
            subscribe=youtuber.find_all("div",{"style":"float: left; width: 150px;"})[0].text.strip()
        except IndexError:
            subscribe="N/A"
        try:
            view=youtuber.find_all("span",{"style":"color:#555;"})[1].text.strip()
        except IndexError:
            view="N/A"

        data=[str(count+1),name,upload,subscribe,view]
        count+=1
        writer.writerow(data)

print(count)
```

Figure 2. Python Crawler Code

Mathematically fitting social media data

In our research, we would like to develop a mathematical model to describe the trend of datasets that fit Benford's Law. We have chosen to fit the first-digit distribution data with $y = ae^{-bx}$ based on the hypothesis, as mentioned before, that the growth speed of a dataset's values is proportional to its current values.

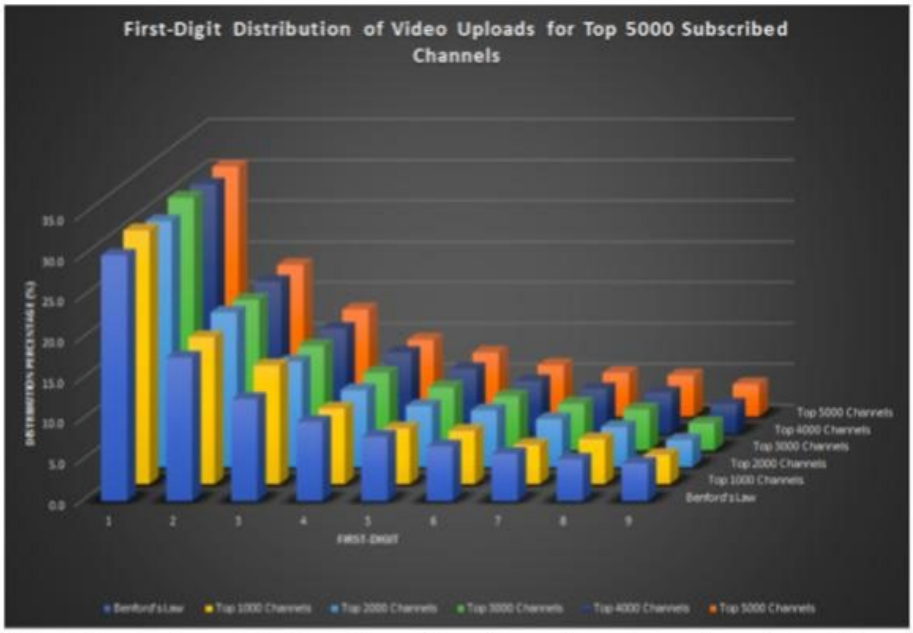
By graphing the first-digit distribution of different groups and exponentially fitting them, we can find the a and b values of the groups and compare them to Benford's Law. We hypothesize that groups with a and b values closer to calculated a and b values for Benford's Law will have a higher chance of fitting Benford's Law. After graphing the results and mathematically fitting them, we can analyze the errors of first-digit distribution data and a and b values in order to evaluate whether or not a dataset follows Benford's Law, as a way of verifying whether or not a dataset is artificial.

Results

With Total Subscriptions as the control variable

Firstly, we record the respective Video Uploads of the top 5000 Total Subscriptions channels and calculate the first-digit distribution of these 5000 channels' Video Uploads. **Figure 3a** shows the first-digit distribution of Video Uploads for the top 5000 most-subscribed channels, showing distribution data for the top 1000 channels, top 2000 channels, top 3000 channels, top 4000 channels, top 5000 channels, and expected values according to Benford's Law. **Figure 3b** shows the raw data of distribution percentages, which shows that the first-distribution of the dataset has a decreasing trend that closely follows the expected values as calculated by Benford's Law.

(3a)



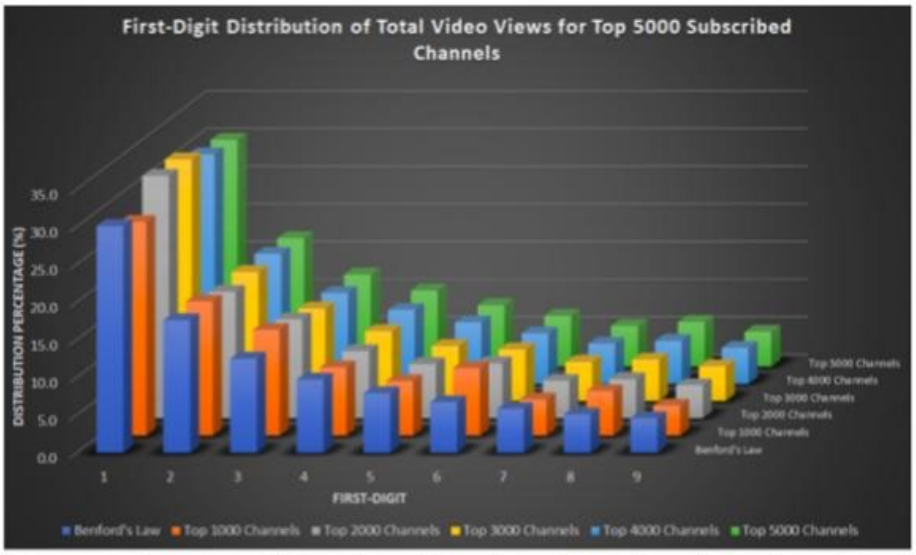
(3b)

First-Digit Distribution (%)	1	2	3	4	5	6	7	8	9
Benford's Law Distribution	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6
Top 1000 Distribution	31.1	18.0	14.5	9.2	6.8	6.5	4.8	5.5	3.6
Top 2000 Distribution	30.2	19.0	12.9	9.4	7.5	6.9	5.9	4.9	3.3
Top 3000 Distribution	30.9	18.4	12.8	9.5	7.7	6.6	5.7	5.0	3.3
Top 4000 Distribution	30.5	18.6	12.8	9.9	7.9	6.4	5.4	5.0	3.5
Top 5000 (All data)	30.6	18.6	13.1	9.4	7.8	6.3	5.3	5.0	3.9

Figure 3 (a) First-digit distribution of video uploads compared to Benford's Law distribution for top 1000, 2000, 3000, 4000, and 5000 subscribed channels (upper) and **(b)** the experimental result corresponding to upper graph (lower).

Secondly, we record the respective Total Video Views of the top 5000 Total Subscriptions channels and calculate the first-digit distribution of these 5000 channels' Total Video Views. **Figure 4a** shows the first-digit distribution of Total Video Views for the top 5000 most-subscribed channels, showing distribution data for the top 1000 channels, top 2000 channels, top 3000 channels, top 4000 channels, top 5000 channels, and expected values according to Benford's Law. **Figure 4b** shows the raw data of distribution percentages, which shows that the first-distribution of the dataset has a decreasing trend that closely follows the expected values as calculated by Benford's Law.

(4a)



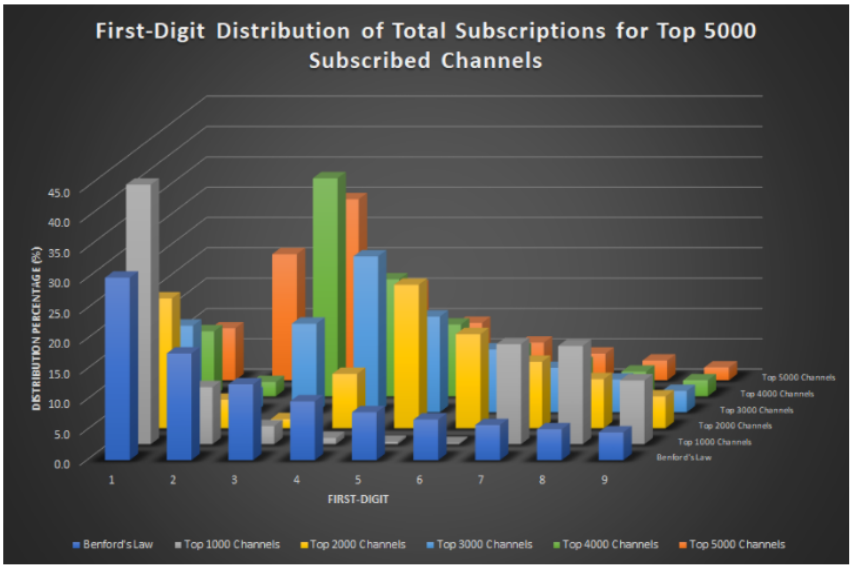
(4b)

First-Digit Distribution (%)	1	2	3	4	5	6	7	8	9
Benford's Law Distribution	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6
Top 1000 Distribution	28.4	17.8	14.0	9.0	7.2	8.9	4.8	5.9	4.0
Top 2000 Distribution	32.2	16.8	13.1	8.9	7.2	7.3	5.0	5.2	4.4
Top 3000 Distribution	32.1	17.1	12.3	9.2	7.3	6.8	5.2	5.5	4.6
Top 4000 Distribution	30.5	17.2	12.0	9.8	8.1	6.7	5.3	5.7	4.7
Top 5000 (All data)	30.1	17.0	12.2	10.1	8.1	6.7	5.4	5.9	4.6

Figure 4 (a) First-digit distribution of total video views compared to Benford's Law distribution for top 1000, 2000, 3000, 4000, and 5000 subscribed channels (upper) and **(b)** the experimental result corresponding to upper graph (lower).

Thirdly, we record the respective Total Subscriptions of the top 5000 Total Subscriptions channels and calculate the first-digit distribution of these 5000 channels' Total Subscriptions. **Figure 5a** shows the first-digit distribution of Total Subscriptions for the top 5000 most-subscribed channels, showing distribution data for the top 1000 channels, top 2000 channels, top 3000 channels, top 4000 channels, top 5000 channels, and expected values according to Benford's Law. **Figure 5b** shows the raw data of distribution percentages, which shows that the first-distribution of the dataset does not follow Benford's Law. For example, the highest distribution percentage of the top 1000 channels group did occur at 1, but it didn't fully follow the decreasing trend of Benford's Law. The distribution percentage decreased from first-digit 1 to first-digit 6, but the distribution percentage increased back to 16.5% at first-digit 7. The other groups had similar deviations from Benford's Law, sometimes with the highest distribution percentage not occurring at first-digit 1 or with both increasing and decreasing trends while the first-digit number increased.

(5a)



(5b)

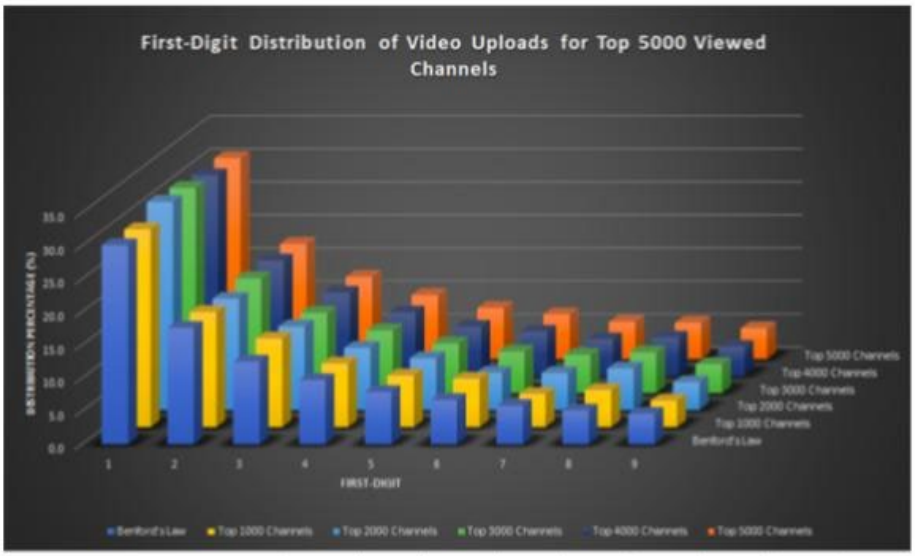
First-Digit Distribution (%)	1	2	3	4	5	6	7	8	9
Benford's Law Distribution	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6
Top 1000 Distribution	42.8	9.4	3.0	1.0	0.4	0.2	16.5	16.2	10.5
Top 2000 Distribution	21.4	4.7	1.5	9.0	23.7	15.5	11.0	8.1	5.3
Top 3000 Distribution	14.3	3.1	14.6	25.7	15.8	10.3	7.3	5.4	3.5
Top 4000 Distribution	10.7	2.4	36.0	19.3	11.8	7.8	5.5	4.1	2.6
Top 5000 (All data)	8.6	20.8	29.8	15.4	9.5	6.2	4.4	3.2	2.1

Figure 5 (a) First-digit distribution of total subscriptions compared to Benford's Law distribution for top 1000, 2000, 3000, 4000, and 5000 subscribed channels (upper) and **(b)** the experimental result corresponding to upper graph (lower).

With Total Views as the control variable

Firstly, we record the respective Video Uploads of the top 5000 Total Views channels and calculate the first-digit distribution of these 5000 channels' Video Uploads. **Figure 6a** shows the first-digit distribution of Video Uploads for the top 5000 most-viewed channels, showing distribution data for the top 1000 channels, top 2000 channels, top 3000 channels, top 4000 channels, top 5000 channels, and expected values according to Benford's Law. **Figure 6b** shows the raw data of distribution percentages, which shows that the first-distribution of the dataset has a decreasing trend that closely follows the expected values as calculated by Benford's Law.

(6a)



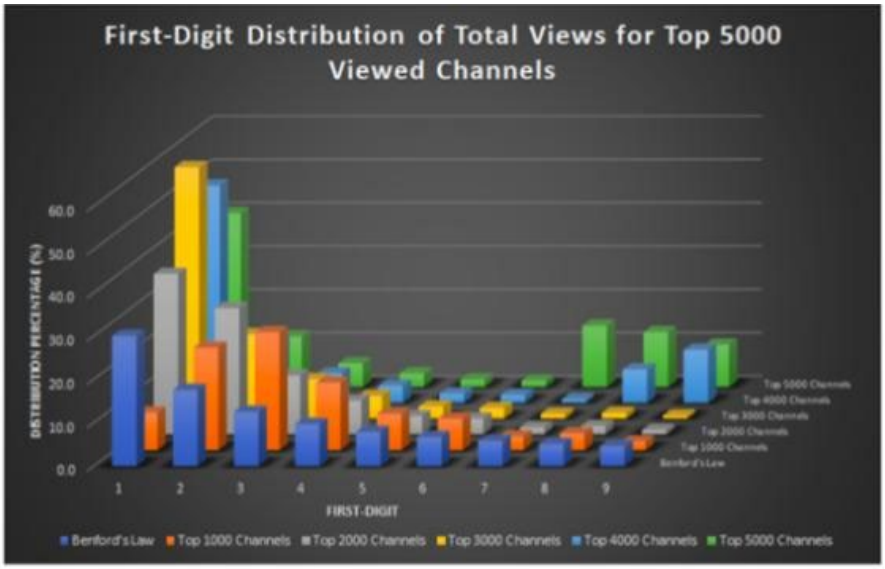
(6b)

First-Digit Distribution (%)	1	2	3	4	5	6	7	8	9
Benford's Law Distribution	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6
Top 1000 Distribution	29.9	17.4	13.4	9.6	7.8	7.3	5.0	5.6	4.0
Top 2000 Distribution	31.5	16.9	12.7	9.4	7.9	5.7	5.6	6.3	4.2
Top 3000 Distribution	31.0	17.3	12.1	9.5	7.5	6.3	5.8	6.2	4.4
Top 4000 Distribution	30.4	17.5	12.6	9.6	7.5	6.7	5.6	5.8	4.4
Top 5000 (All data)	30.4	17.4	12.4	9.6	7.7	6.7	5.6	5.5	4.6

Figure 6 (a) First-digit distribution of video uploads compared to Benford's Law distribution for top 1000, 2000, 3000, 4000, and 5000 subscribed channels (upper) and **(b)** the experimental result corresponding to upper graph (lower).

Secondly, we record the respective Total Views of the top 5000 Total Views channels and calculate the first-digit distribution of these 5000 channels' Total Views. **Figure 7a** shows the first-digit distribution of Total Video Views for the top 5000 most-viewed channels, showing distribution data for the top 1000 channels, top 2000 channels, top 3000 channels, top 4000 channels, top 5000 channels, and expected values according to Benford's Law. **Figure 7b** shows the raw data of distribution percentages, which shows that the first-distribution of the dataset does not follow Benford's Law. For example, the highest distribution percentage of the top 1000 channels group occurred at first-digit 3 instead of 1, and it had an increasing trend from 1 to 3 and a decreasing trend from 3 to 9. The other groups had similar deviations from Benford's Law, sometimes with the highest distribution percentage not occurring at first-digit 1 or with both increasing and decreasing trends while the first-digit number increased.

(7a)



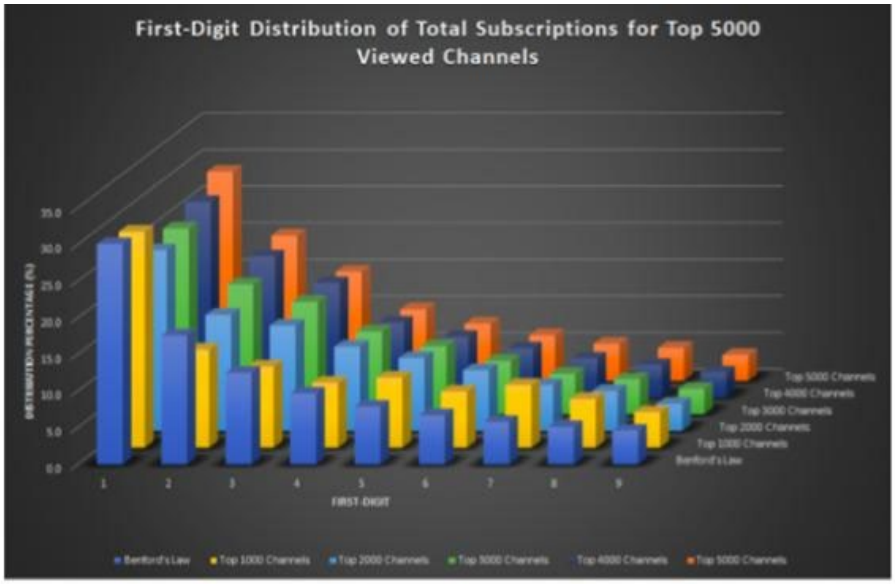
(7b)

First-Digit Distribution (%)	1	2	3	4	5	6	7	8	9
Benford's Law Distribution	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6
Top 1000 Distribution	29.5	13.4	11.1	8.9	9.6	7.6	8.6	6.6	4.8
Top 2000 Distribution	24.7	15.8	14.4	11.5	9.9	8.3	6.3	5.4	3.7
Top 3000 Distribution	25.4	17.7	15.3	11.3	9.3	7.3	5.5	4.8	3.5
Top 4000 Distribution	26.7	19.3	15.6	10.2	8.3	6.7	5.3	4.4	3.4
Top 5000 (All data)	40.1	11.7	5.4	3.1	1.7	1.5	14.1	12.6	9.7

Figure 7 (a) First-digit distribution of total views compared to Benford's Law distribution for top 1000, 2000, 3000, 4000, and 5000 viewed channels (upper) and **(b)** the experimental result corresponding to upper graph (lower)

Thirdly, we record the respective Total Subscriptions of the top 5000 Total Views channels and calculate the first-digit distribution of these 5000 channels' Total Subscriptions. **Figure 8a** shows the first-digit distribution of Total Subscriptions for the top 5000 most-viewed channels, showing distribution data for the top 1000 channels, top 2000 channels, top 3000 channels, top 4000 channels, top 5000 channels, and expected values according to Benford's Law. **Figure 8b** shows the raw data of distribution percentages, which shows that the first-distribution of the dataset has a decreasing trend that closely follows the expected values as calculated by Benford's Law.

(8a)



(8b)

First-Digit Distribution (%)	1	2	3	4	5	6	7	8	9
Benford's Law Distribution	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6
Top 1000 Distribution	29.5	13.4	11.1	8.9	9.6	7.6	8.6	6.6	4.8
Top 2000 Distribution	24.7	15.8	14.4	11.5	9.9	8.3	6.3	5.4	3.7
Top 3000 Distribution	25.4	17.7	15.3	11.3	9.3	7.3	5.5	4.8	3.5
Top 4000 Distribution	26.7	19.3	15.6	10.2	8.3	6.7	5.3	4.4	3.4
Top 5000 (All data)	28.5	19.8	14.9	9.7	7.8	6.2	5.0	4.5	3.5

Figure 8 (a) First-digit distribution of total subscriptions compared to Benford's Law distribution for top 1000, 2000, 3000, 4000, and 5000 viewed channels (upper) and (b) the experimental result corresponding to upper graph (lower)

With Video Uploads as the control variable

Because SocialBlade doesn't provide rankings for the top 5000 channels with the highest Video Uploads, we are unable to conduct analysis with Video Uploads as the control variable. Nevertheless, we have observed the pattern that parameter A will not conform with Benford's Law while parameter B and C will, where A is used as the groups' control variable. Therefore, we can hypothesize that Total Subscriptions and Total Views will conform with Benford's Law while Video Uploads will not conform with Benford's Law, where Video Uploads is the control variable.

Discussion

Error of First-Digit Distribution in Comparison with Expected values

Most-Subscribed Channels

channels had a minimum of 0.189% error and a maximum of 28.381% error, with the average error being 6.237%. The error of the first-digit distribution of the top 3000 most-subscribed YouTube channels had a minimum of 1.270% error and a maximum of 28.099% error, with the average error being 5.260%. The error of the first-digit distribution of the top 4000 most-subscribed YouTube channels had a minimum of 0.261% error and a maximum of 22.891% error, with the average error being 5.362%. The error of the first-digit distribution of the top 5000 most-subscribed YouTube channels had a minimum of 1.268% error and a maximum of 14.559% error, with the average error being 5.226%. These results are also shown in **Figure 9**.

Most-Subscribed Channels	Minimum Error (%)	Maximum Error (%)	Average Error (%)
Top 1000	2.203	22.676	10.122
Top 2000	0.189	28.381	6.237
Top 3000	1.270	28.099	5.260
Top 4000	0.261	22.891	5.362
Top 5000	**1.268**	**14.559**	**5.226**

Figure 9. Error of first-digit distributions of Video Uploads for top 5000 most-subscribed channels

Next, we evaluated the error of observed Total Views first-digit distribution for the top 1000 most-subscribed YouTube channels. From the digits 1 to 9, the error of each digit distribution was between 1.049% and 33.479%, with the average error being 12.779%. The error of the first-digit distribution of the top 2000 most-subscribed YouTube channels had a minimum of 1.368% error and a maximum of 14.361% error, with the average error being 7.051%. The error of the first-digit distribution of the top 3000 most-subscribed YouTube channels had a minimum of 0.219% error and a maximum of 10.971% error, with the average error being 4.869%. The error of the first-digit distribution of the top 4000 most-subscribed YouTube channels had a minimum of 0.548% error and a maximum of 11.278% error, with the average error being 3.811%. The error of the first-digit distribution of the top 5000 most-subscribed YouTube channels had a minimum of 0.049% error and a maximum of 15.597% error, with the average error being 3.991%. These results are also shown in **Figure 10**.

Most-Subscribed Channels	Minimum Error (%)	Maximum Error (%)	Average Error (%)
Top 1000	1.049	33.479	12.779
Top 2000	1.368	14.361	7.051
Top 3000	0.219	10.971	4.869
Top 4000	0.548	11.278	3.811
Top 5000	**0.049**	**15.597**	**3.991**

Figure 10. Error of first-digit distributions of Total Views for top 5000 most-subscribed channels

Thirdly, we evaluated the error of observed Total Subscriptions first-digit distribution for the top 1000 most-subscribed YouTube channels. From the digits 1 to 9, the error of each digit distribution was between 42.193% and

217.167%, with the average error being 108.535%. The error of the first-digit distribution of the top 2000 most-subscribed YouTube channels had a minimum of 7.732% error and a maximum of 199.367% error, with the average error being 76.710%. The error of the first-digit distribution of the top 3000 most-subscribed YouTube channels had a minimum of 5.882% error and a maximum of 164.948% error, with the average error being 58.446%. The error of the first-digit distribution of the top 4000 most-subscribed YouTube channels had a minimum of 5.603% error and a maximum of 187.600% error, with the average error being 5.362%. The error of the first-digit distribution of the top 5000 most-subscribed YouTube channels had a minimum of 7.463% error and a maximum of 138.720% error, with the average error being 47.771%. These results are also shown in **Figure 11**.

Most-Subscribed Channels	Minimum Error (%)	Maximum Error (%)	Average Error (%)
Top 1000	42.193	217.167	108.535
Top 2000	7.732	199.367	76.710
Top 3000	5.882	164.948	58.446
Top 4000	5.603	187.600	63.544
Top 5000	**7.463**	**138.720**	**47.771**

Figure 11. Error of first-digit distributions of Total Subscriptions for top 5000 most-subscribed channels

Most-Viewed Channels

First, we evaluated the error of observed Video Uploads first-digit distribution for the top 1000 most-viewed YouTube channels. From the digits 1 to 9, the error of each digit distribution was between 0.664% and 13.793%, with the average error being 6.321%. The error of the first-digit distribution of the top 2000 most-viewed YouTube channels had a minimum of 0.633% error and a maximum of 23.529% error, with the average error being 7.423%. The error of the first-digit distribution of the top 3000 most-viewed YouTube channels had a minimum of 0.575% error and a maximum of 20.915% error, with the average error being 5.326%. The error of the first-digit distribution of the top 4000 most-viewed YouTube channels had a minimum of 0.001% error and a maximum of 14.216% error, with the average error being 3.530%. The error of the first-digit distribution of the top 5000 most-viewed YouTube channels had a minimum of 0.597% error and a maximum of 7.451% error, with the average error being 2.004%. These results are also shown in **Figure 12**.

Most-Subscribed Channels	Minimum Error (%)	Maximum Error (%)	Average Error (%)
Top 1000	0.664	13.793	6.321
Top 2000	0.633	23.529	7.243
Top 3000	0.575	20.915	5.326
Top 4000	0.001	14.216	3.530
Top 5000	**0.597**	**7.451**	**2.004**

Figure 12. Error of first-digit distributions of Video Uploads for top 5000 most-viewed channels

Next, we evaluated the error of observed Total Views first-digit distribution for the top 1000 most-viewed YouTube channels. From the digits 1 to 9, the error of each digit distribution was between 6.329% and 117.600%, with the average error being 47.149%. The error of the first-digit distribution of the top 2000 most-viewed YouTube channels had a minimum of 8.800% error and a maximum of 78.261% error, with the average error being 46.878%. The error of the first-digit distribution of the top 3000 most-viewed YouTube channels had a minimum of 10.417% error and a maximum of 92.802% error, with the average error being 60.771%. The error of the first-digit distribution of the top 4000 most-viewed YouTube channels had a minimum of 17.188% error and a maximum of 164.674% error, with the average error being 70.749%. The error of the first-digit distribution of the top 5000 most-viewed YouTube channels had a minimum of 33.289% error and a maximum of 147.843% error, with the average error being 83.519%. These results are also shown in **Figure 13**.

Most-Subscribed Channels	Minimum Error (%)	Maximum Error (%)	Average Error (%)
Top 1000	6.329	117.600	47.149
Top 2000	8.800	78.261	46.878
Top 3000	10.417	92.802	60.771
Top 4000	17.188	164.674	70.749
Top 5000	**33.289**	**147.843**	**83.519**

Figure 13. Error of first-digit distributions of Total Views for top 5000 most-viewed channels

Thirdly, we evaluated the error of observed Total Subscriptions first-digit distribution for the top 1000 most-viewed YouTube channels. From the digits 1 to 9, the error of each digit distribution was between 2.070% and 47.436%, with the average error being 18.217%. The error of the first-digit distribution of the top 2000 most-viewed YouTube channels had a minimum of 6.844% error and a maximum of 25.816% error, with the average error being 16.178%. The error of the first-digit distribution of the top 3000 most-viewed YouTube channels had a minimum of 0.739% error and a maximum of 24.811% error, with the average error being 13.012%. The error of the first-digit distribution of the top 4000 most-viewed YouTube channels had a minimum of 0.047% error and a maximum of 23.815% error, with the average error being 11.474%. The error of the first-digit distribution of the top 5000 most-viewed YouTube channels had a minimum of 0.497% error and a maximum of 23.815% error, with the average error being 10.604%. These results are also shown in **Figure 14**.

Most-Subscribed Channels	Minimum Error (%)	Maximum Error (%)	Average Error (%)
Top 1000	2.070	47.436	18.217
Top 2000	6.844	25.816	16.178
Top 3000	0.739	24.811	13.012
Top 4000	0.047	25.566	11.474
Top 5000	**0.497**	**23.815**	**10.604**

Figure 14. Error of first-digit distributions of Total Subscriptions for top 5000 most-viewed channels

Analysis

From the error analysis above, it can be seen that the error of first-digit distributions for Total Subscriptions when using Total Subscriptions as the control variable is much higher than the error of first-digit distributions for Total Views and Video Uploads, which leads us to believe that Total Subscriptions, when using the data from the top 5000 most-subscribed channels, can be seen as "artificially" generated and therefore doesn't fit Benford's Law, which would make sense because we are artificially selecting the top 5000 most-subscribed channels. However, the other two variables (the ones that aren't used as the control variable) have first-digit distributions that fit Benford's Law even though the two other variables are also extracted from the top 5000 artificially selected channels. This phenomenon is also seen for the analysis of the top 5000 most-viewed channels.

This phenomenon was further investigated through the use of smaller groups of channels within the top 5000 channels, including the top 1000, 2000, 3000, and 4000. From these tests, we noticed how there was a general trend of average errors decreasing as the number of channels increased for groups that fit Benford's Law, but groups that analyzed the same variable as the control variable still had larger errors while the groups that analyzed a variable different from the control variable generally had smaller errors. This leads us to believe that the two other variables (Total Views and Total Variables in the case of the top 5000 most-subscribed channels and Total Uploads and Total Subscriptions in the case of the top 5000 most-viewed channels) are not influenced by how the channels are selected, and are therefore not correlated with the control variable, allowing them to be "randomly" generated, which fits the requirements of Benford's Law.

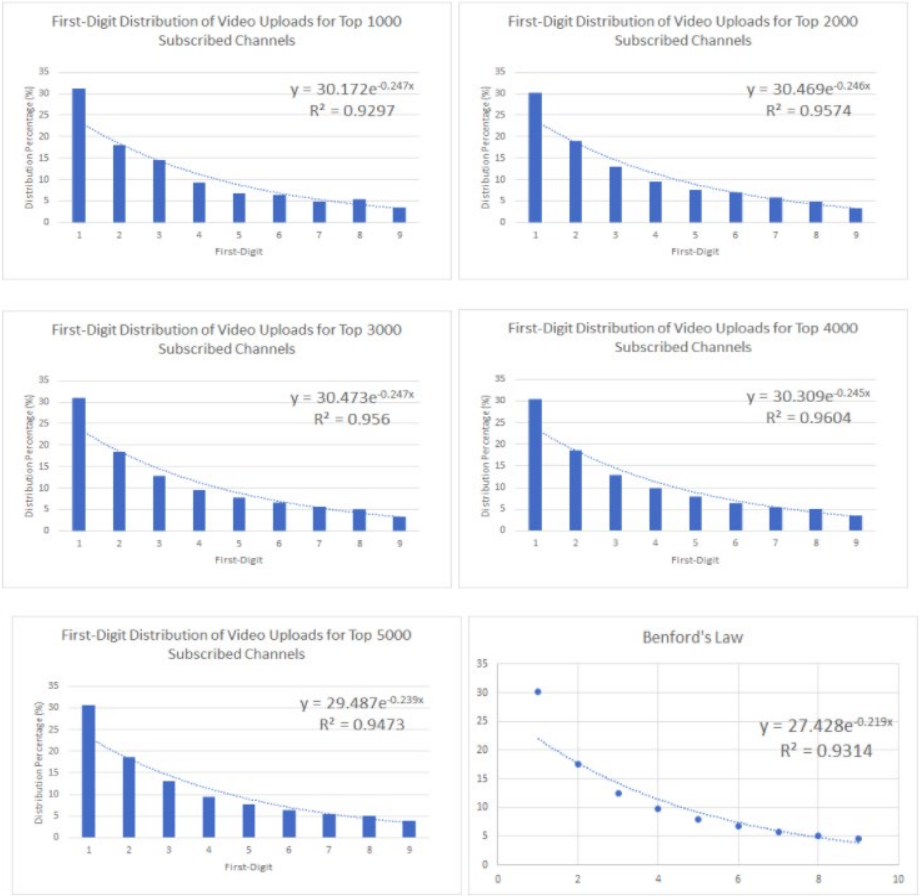


Figure 15. First-digit distribution of top 5000 video uploads (ranked) curve fitted with exponential models. Five graphs above are respectively the result of top 1000, 2000, 3000, 4000, 5000 subscribed channels and Benford's Law.

Verifying Benford's Law with exponential model

Curve fitting with exponential model

Figures 15, 16, and 17 show the results of curve fitting the data, and it can be seen that Total Subscriptions of the top-5000 most subscribed channels had many different curving results that highly deviated from the expected results of Benford's Law while Video Uploads and Total Views generally have a and b values within a close range of the expected values.

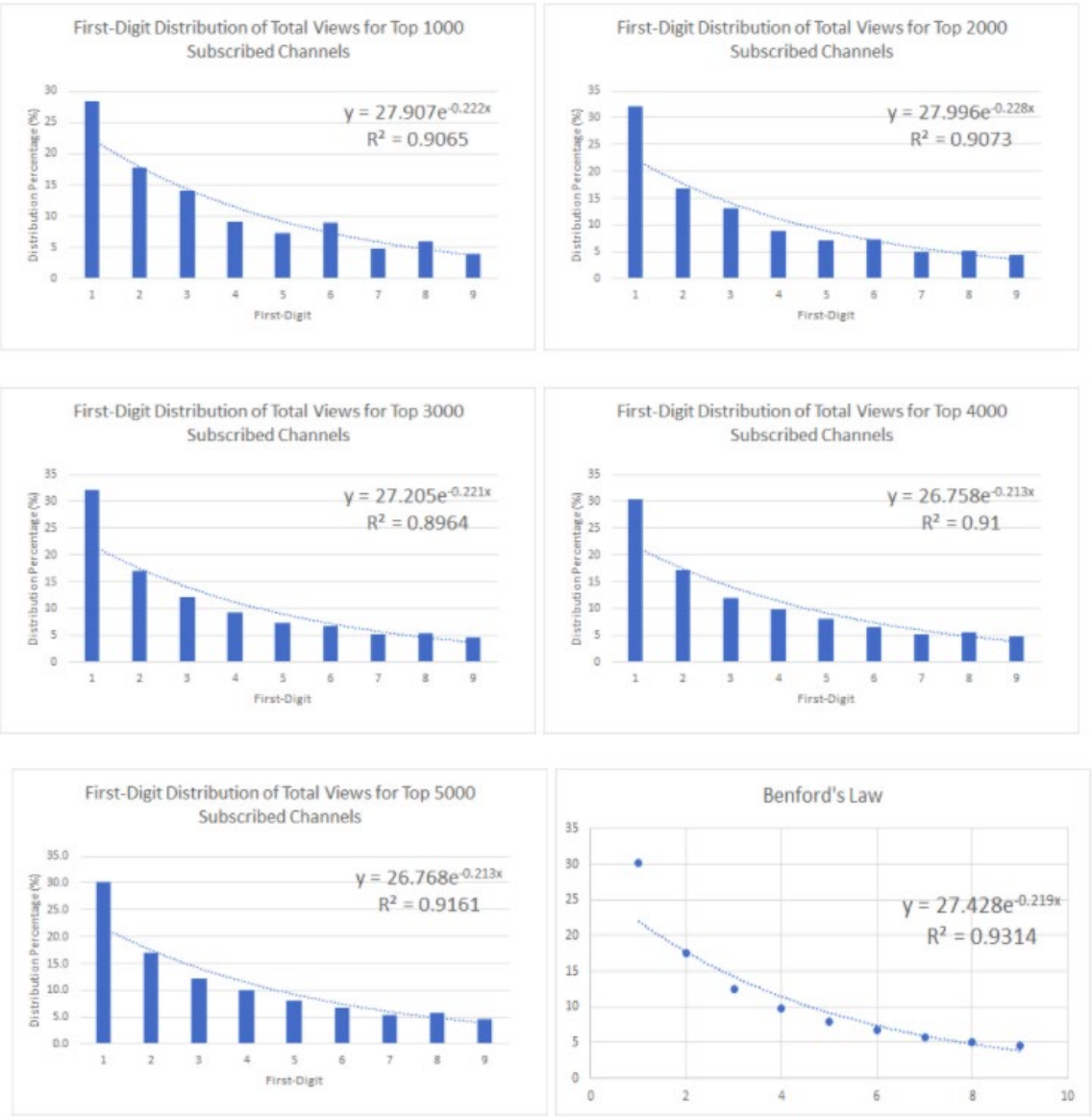


Figure 16. First-digit distribution of top 5000 total video views (ranked) curve fitted with exponential models. Five graphs above are respectively the result of top 1000, 2000, 3000, 4000, 5000 subscribed channels and Benford's Law.

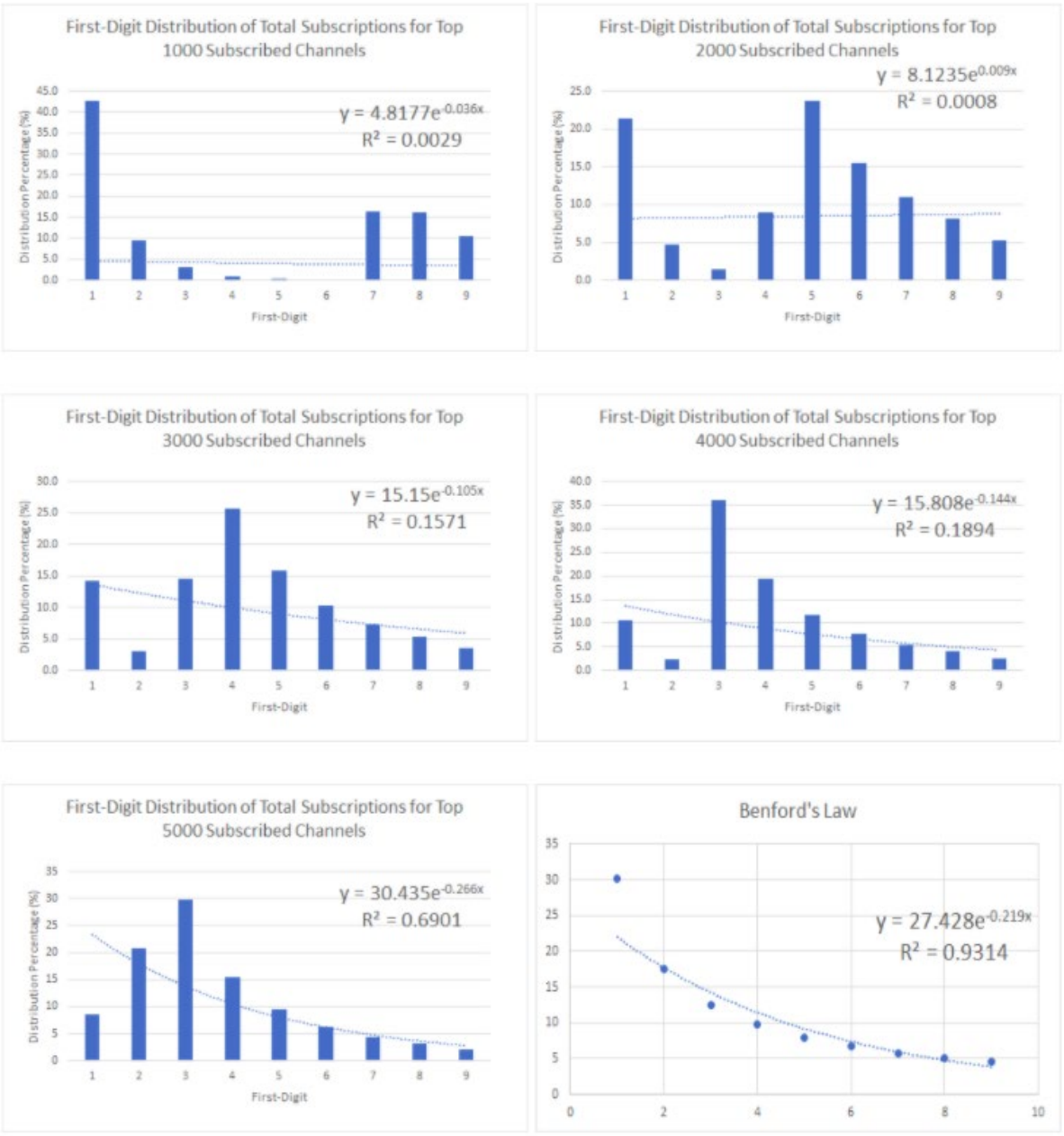


Figure 17. First-digit distribution of top 5000 subscribed channels (ranked) curve fitted with exponential models. Five graphs above are respectively the result of top 1000, 2000, 3000, 4000, 5000 subscribed channels and Benford's Law.

Figures 18, 19, and 20 show the results of curve fitting the data for the top 5000 most-viewed channels. It can be seen the Total Views has distributions that generally don't fit the expected curve of Benford's Law, while Total Subscriptions and Video Uploads slightly deviated from the expected curve but still distribute patterns fitting Benford's Law.

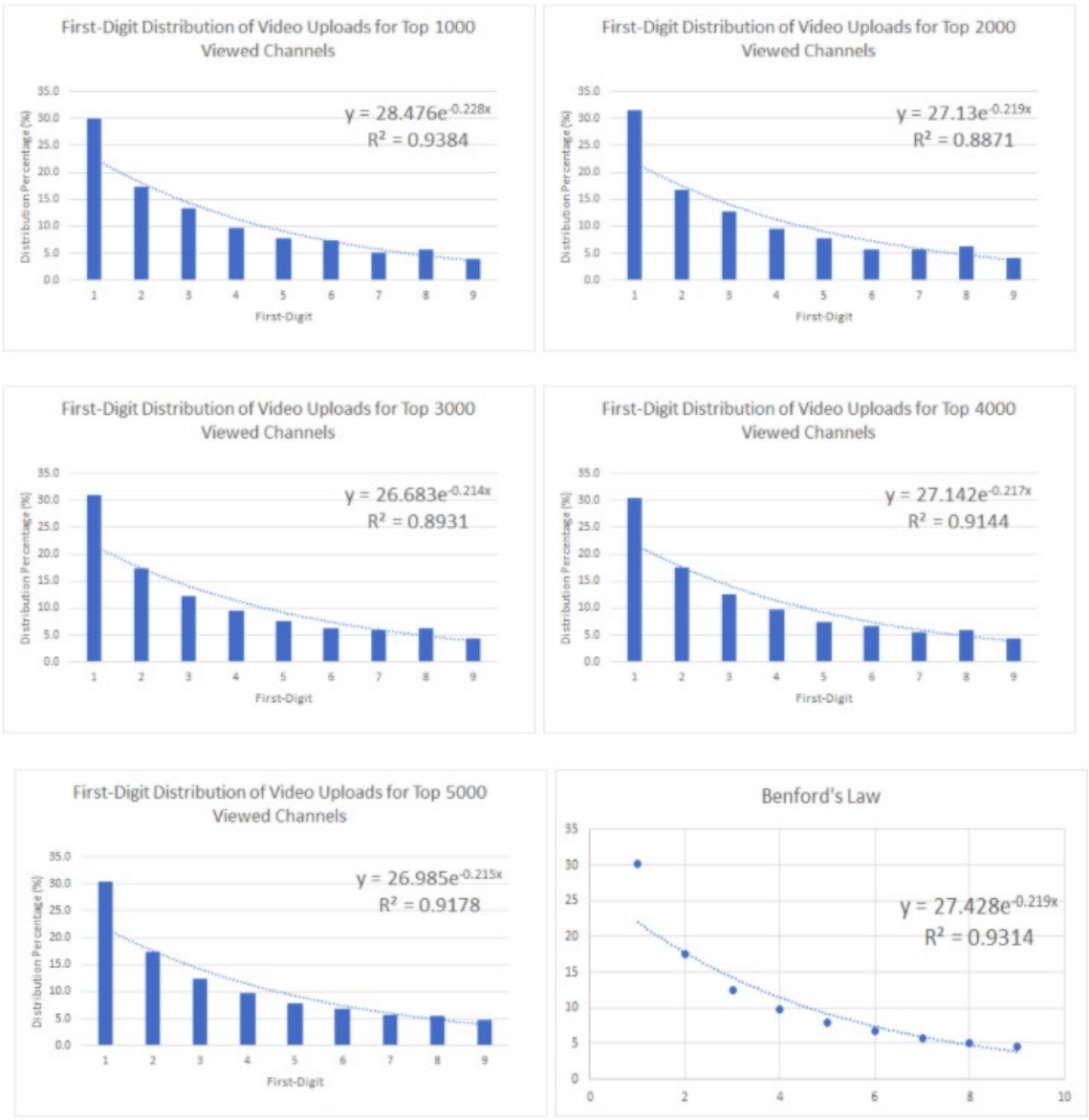


Figure 18. First-digit distribution of top 5000 video uploads (ranked) curve fitted with exponential models. Five graphs above are respectively the result of top 1000, 2000, 3000, 4000, 5000 total video views and Benford's Law.

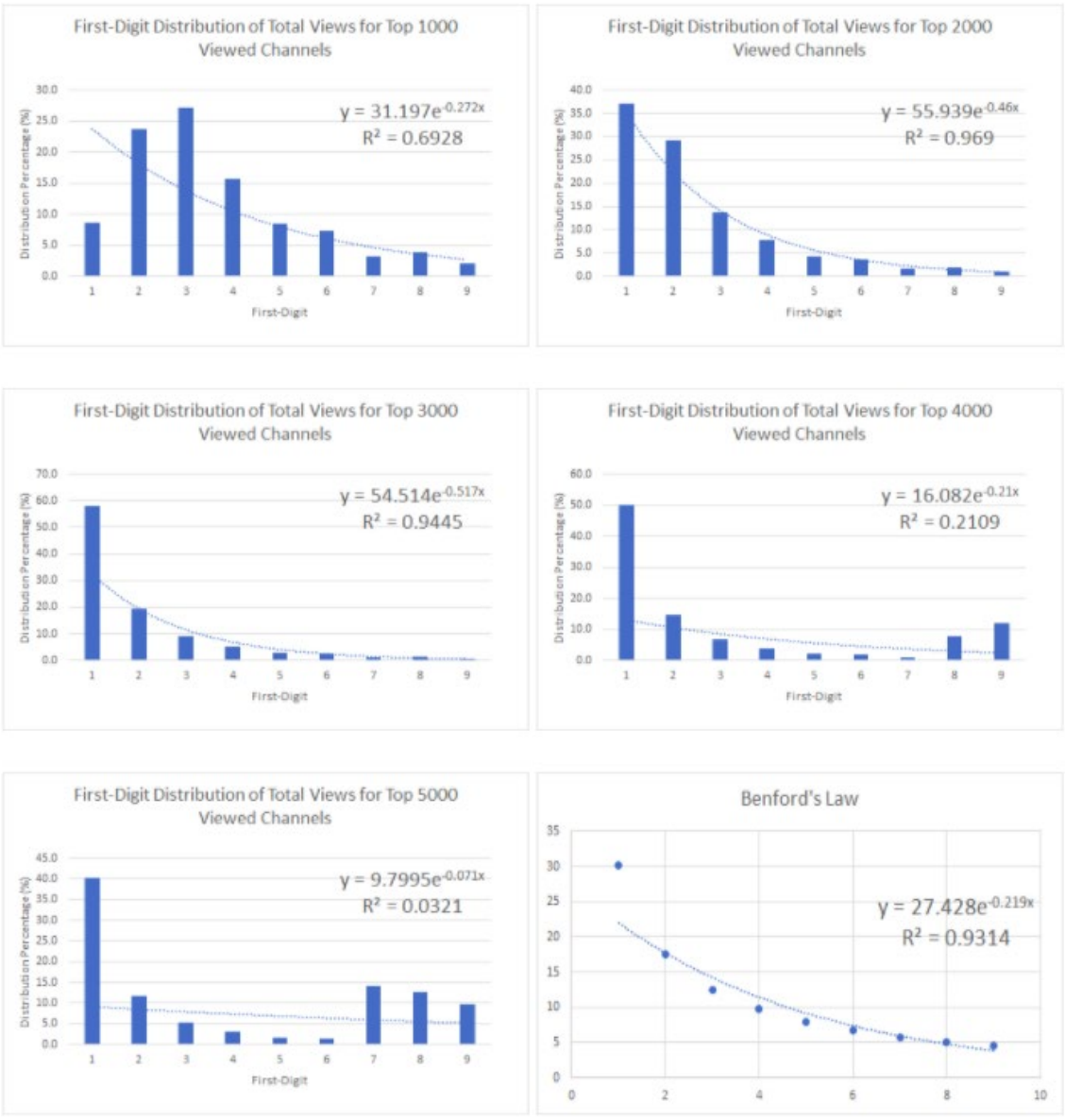


Figure 19. First-digit distribution of top 5000 total video views (ranked) curve fitted with exponential models. Five graphs above are respectively the result of top 1000, 2000, 3000, 4000, 5000 total video views and Benford's Law.

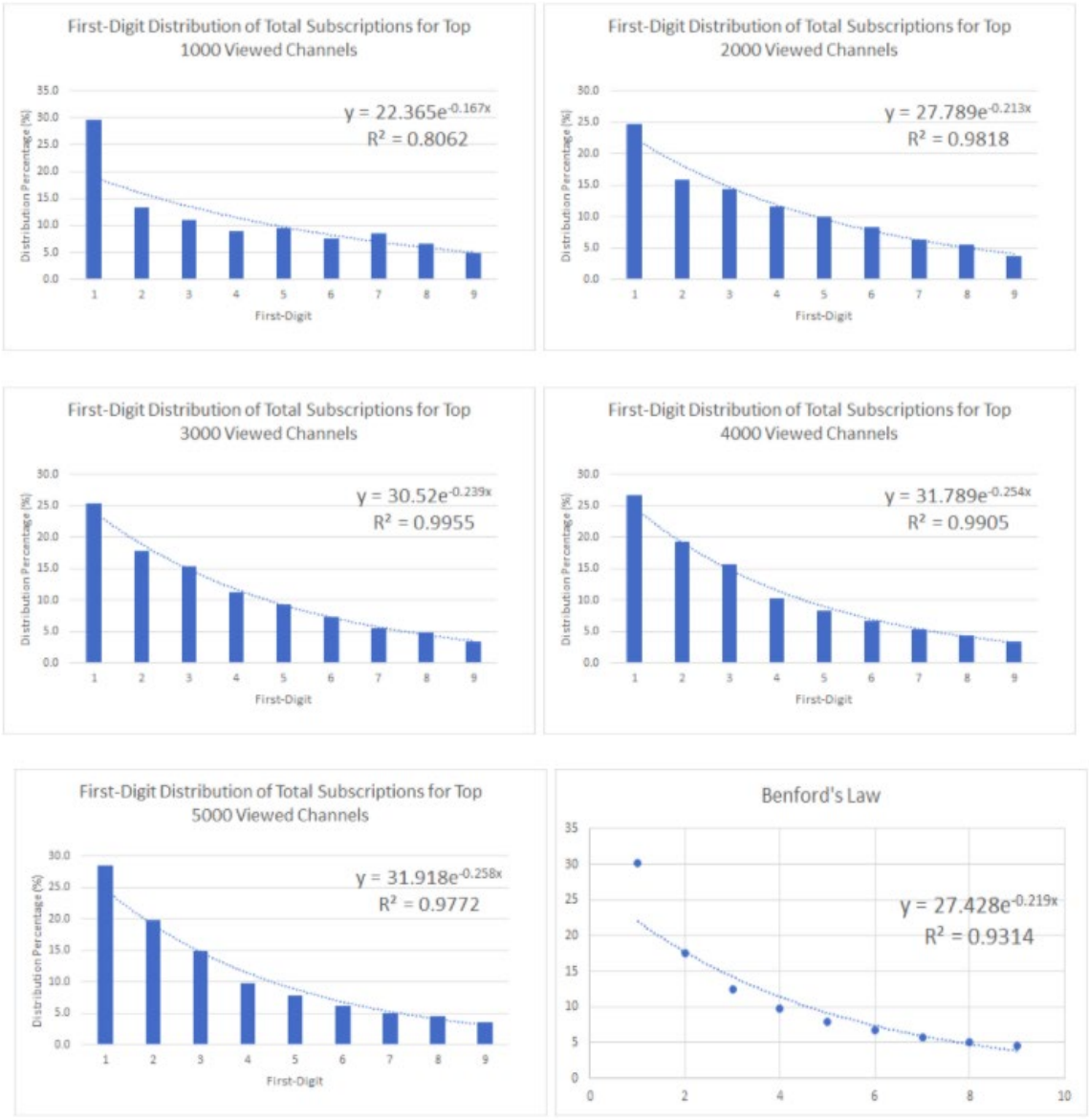


Figure 20. First-digit distribution of top 5000 subscribed channels (ranked) curve fitted with exponential models. Five graphs above are respectively the result of top 1000, 2000, 3000, 4000, 5000 total video views and Benford's Law.

Error of a and b values in Comparison with Expected values

From **Figures 21** and **22**, it can be seen that errors of a and b values resulting from curve fitting are smaller for Total Views and slightly larger for Video Uploads but all less than 20% percent, while the errors of a and b values for Total Subscriptions as shown in **Figure 23** easily exceeded 20%. This corroborates our analysis of first-digit distributions above that groups analyzing the same variables as their control variables won't fit Benford's Law.

Most-Subscribed Channels	Error of a value (%)	Error of b value (%)
Top 1000	10.004	12.785
Top 2000	11.087	12.329
Top 3000	11.102	12.785
Top 4000	10.504	11.872
Top 5000	**7.507**	**9.132**

Figure 21. Error of a and b values of groups taking Video Uploads for top 5000 most-subscribed channels in comparison with expected a and b for Benford's Law

Most-Subscribed Channels	Error of a value (%)	Error of b value (%)
Top 1000	1.746	1.370
Top 2000	2.071	4.110
Top 3000	0.813	0.913
Top 4000	2.443	2.740
Top 5000	**2.406**	**2.740**

Figure 22. Error of a and b values of groups taking Total Views for top 5000 most-subscribed channels in comparison with expected a and b for Benford's Law

Most-Subscribed Channels	Error of a value (%)	Error of b value (%)
Top 1000	82.434	83.562
Top 2000	70.381	95.890
Top 3000	44.764	52.055
Top 4000	42.365	34.247
Top 5000	**10.963**	**21.461**

Figure 23. Error of a and b values of groups taking Total Subscriptions for top 5000 most-subscribed channels in comparison with expected a and b for Benford's Law

From **Figures 24** and **26**, it can be seen that errors of a and b values resulting from curve fitting are smaller for Video Uploads and slightly larger for Total Subscriptions but are mostly smaller than 20%%. This corroborates with our analysis above. For **Figure 25**, the Total Views for the top 5000 most-viewed channels, the errors exceeded 20%% easily, with some reaching above 100%, showing that they don't fit Benford's Law.

Most-Subscribed Channels	Error of a value (%)	Error of b value (%)
Top 1000	3.821	4.110
Top 2000	1.086	0.000
Top 3000	2.716	2.283
Top 4000	1.043	0.913
Top 5000	**1.615**	**1.826**

Figure 24. Error of a and b values of groups taking Video Uploads for top 5000 most-viewed channels in comparison with expected a and b for Benford's Law

Most-Subscribed Channels	Error of a value (%)	Error of b value (%)
Top 1000	13.741	24.201
Top 2000	103.949	110.046
Top 3000	98.753	136.073
Top 4000	41.366	4.110
Top 5000	**64.274**	**67.580**

Figure 25. Error of a and b values of groups taking Total Views for top 5000 most-viewed channels in comparison with expected a and b for Benford's Law

Most-Subscribed Channels	Error of a value (%)	Error of b value (%)
Top 1000	18.459	23.744
Top 2000	1.316	2.740
Top 3000	11.273	9.132
Top 4000	15.900	15.982
Top 5000	**16.370**	**17.808**

Figure 26. Error of a and b values of groups taking Total Subscriptions for top 5000 most-viewed channels in comparison with expected a and b for Benford's Law

Analysis

The experimental results show that fitting first-digit distributions of a dataset with an exponential model can be used to evaluate how closely a dataset fits Benford's Law. Also, for most of the datasets that follow Benford's Law, as shown in Figures 21, 22, 24, and 26, the lowest error of a and b values typically occurs at groups with higher amounts of channels (4000 or 5000). This most likely indicates that datasets with larger quantities of data are more likely to follow the expected values of Benford's Law more closely. From Figure 23 and Figure 25, it can also be seen that even for datasets that don't fit Benford's Law according to its first-digit distributions, modeling these distributions with an exponential model will allow us to observe whether it has a similar trend to Benford's Law (a decreasing trend of frequency as the digit increases from 1 to 9).

Conclusion and Implications

In this paper, we have successfully utilized social media data to investigate Benford's Law. Using YouTube channel data taken from SocialBlade, we analyzed three variables- Total Subscriptions, Total Views, and Video Uploads- for each channel to verify if YouTube data fits Benford's Law and whether it is artificial or not. When taking Total Subscription data for the top 5000 most-subscribed channels, the first-digit distribution of Total Subscriptions doesn't fit Benford's Law, but the other two variables, Video Uploads and Total Views obtained fit. The same happens when taking Total Views for the top 5000 most-viewed channels. Thus, we can hypothesize that when analyzing variable A's first digit distribution for the top channels ranked with variable A, the first-digit distribution of variable A will not fit Benford's Law, while variables B and C's first-digit distribution obtained from variable A will fit and therefore are not artificial. In order to prove this hypothesis, we changed our number of channels to the top 1000, 2000, 3000, 4000, in addition to 5000 and found out that groups with different numbers of channels produce the same results. Otherwise, we also utilize an exponential model $y = ae^{-bx}$ to mathematically fit all the data. Results show that the a value of Benford's law is 27.428 and the b value is 0.219. If the results of fitting first-digit distribution graphs produce a and b values that are closer to the expected a and b from Benford's Law, it is more likely that the data fits Benford's Law and isn't artificial. This method can be adopted to verify whether the first-digit distribution of data fits Benford's Law. In the future, we will use this proposed model to verify whether or not other datasets fit Benford's Law, and whether they are artificial or not.

Acknowledgments

We would like to thank our advisor Hsin-Ye (David) Chen for helping us with this project.

References

[1] Berger, Arno, and Theodore P. Hill. "A Basic Theory of Benford's Law." Probability Surveys, vol. 8, no. none, 2011, doi:10.1214/11-ps175.

[2] Kruger, Paul, and Sarma Yadavalli. "THE POWER OF ONE: BENFORD'S LAW." South African Journal of Industrial Engineering, vol. 28, no. 2, 2017, doi:10.7166/28-2-1753.

[3] Jamain, Adrien. "Benford's Law." Sept. 2001.

[4] Berger, Arno, and Theodore P. Hill. An Introduction to Benford's Law. Princeton U.P., 2015.

[5] Romero-Roch'in, V.. "A derivation of Benford's Law ... and a vindication of Newcomb." (2009).