Youtube Revisited:  
On the Importance of Correct Measurement Methodology

Ossi Karkulahti, Jussi Kangasharju  
University of Helsinki
Introduction

• Measuring large systems is challenging
  • Full system analysis is expensive -> sampling
• The way sampling is conducted affects the results
  • Ideally a random and representative sample
  • Technological limitation may skew the sampling process
  • Biased sample may yield incorrect conclusions
  • Could also affect any derivative work
• We will show the effects of three different sampling methods on YouTube
Motivation

• Previously YouTube video metadata has been collected by:
  • selecting videos belonging to certain categories
  • crawling related videos
  • using most recent videos
• We argue that all these methods lead to a biased sample
• The result are not representative in all aspects
• Other work base their assumptions on these results
Our Contributions

- We have collected three datasets with three methods
- We compare the methods for collecting YouTube video metadata
- We demonstrate the differences in various metrics between the different datasets
Data Collection

• We have collected video metadata by three different methods, all are using YouTube API
  1. Most recent videos (MR)
  2. Related videos (BFS)
  3. Random string (RS)
• Fourth method is to use videos from a certain category, which is obviously biased
  • 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. IMC, 2007.
1. Most Recent Videos (MR)

- Collect periodically metadata of the most recent videos
  - Included information: video ID, view count, length, category, publish date etc.
- Obviously limited to new videos
- Previously used e.g. by
  - X. Cheng, J. Liu, and C. Dale. Understanding the characteristics of internet short video sharing: A youtube-based measurement study. Multimedia, IEEE Transactions on, 2013.
  - G. Szabo and B. A. Huberman. Predicting the popularity of online content. Communications of the ACM, 2010.
2. Related Videos (BFS)

- Select a video ID and then ask its related videos and then the related videos for all those videos and so on
- We limited related videos to 50 per one video
- In theory, one seed yields to \( \sim 125,000 \) videos \((50 \times 50 \times 50)\)
- \( N \) unique videos is lower, the related videos overlap
- Can be seen as similar to breadth-first search (BFS)
- Fast, most of the time one query returns metadata of tens of videos
- X. Cheng, J. Liu, and C. Dale. Understanding the characteristics of internet short video sharing: A youtube-based measurement study. Multimedia, IEEE Transactions on, 2013.
3. Random Strings (RS)

- Zhou et al. have used similar method to estimate YouTube’s size (“Counting YouTube Videos via Random Prefix Sampling”, IMC 2011)
- Generate a random character string and ask the API to return videos which IDs include the string
  - ‘a-Z’, ‘0-9’, ‘-’, ‘_’, four-letter strings work the best
  - On average a random string matched to 6.9 video IDs
  - For an unknown reason IDs include ‘-’
3. Random Strings (RS)

A random string w57j would match and return metadata for the following videos:

- W57J-21gSSo
- XcY-W57J-Uo
- w57j-VVNAg0
- W57J-msuors
## Datasets

| Dataset 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           | ~ 5 million |
| BFS          | Related videos       | Early 2014           | ~ 5 million |
Results

- Popularity
- Views
- Age
- Categories
- Length
RS and BFS: Very different view count distributions

- BFS has two-part distribution, with a quick-dropping tail
- RS follows more closely Zipf, with a truncated tail
- BFS data seems to over-estimate view counts
- RS: Top 10 -> 5% of all views, top 1000 -> 43%, top 10,000 -> 74%
Popularity after 30 days

- MR and BFS seem to over-estimate video popularity
- However, MR-09 resembles RS
Views

- The 5th percentile of BFS is higher than the median of RS and MR
- BFS view counts are at least one order of magnitude higher than the RS ones
Views

- The median, 5th and 95th percentiles for BFS and RS over eight years
Age Distribution

- BFS has less videos newer than two years, but a lot of very recent videos
- The drop in RS is an artifact of the method
- RS: 29% of videos are newer than a year, majority is newer than two years
Categories (share of videos)

- Most videos of:
  - RS: People & Blogs (Default category for an upload)
  - BFS: Music
Categories (share of views)

- Distribution of number of views is more similar
- Music videos get most views
Video Length

- RS and MR: Most common length is 60 s or less
- BFS: Most common 3-5 min, music videos?
- All: Videos of 3-5 mins length get most views
# Summary of the Methods

| Method | Description |
|--------|-------------|
| BFS    | - Tends to overestimate some of the metrics  
|        | - Fast, up to 100 per query  
|        | - Mostly popular music videos? |
| MR     | - Over-estimates the views  
|        | - Slow  
|        | - Limited to new videos  
|        | - Mostly news clips? |
| RS     | - Most ‘reliable’  
|        | - Not that fast, ~7 per query |
|        | - The mysterious ‘-’ curiosity |
Conclusion 1/2

• We have used YouTube as an example, using three data collection methods
• The datasets differ in many key metrics that have used in past research (MR, BFS)
• RS not previously used in this manner
• Differences between RS and the others raise questions about the general applicability of the previous results
• We believe the RS produces a representative sample
Conclusion 2/2

• As BFS dataset demonstrates even large datasets are not immune to bias introduced by the method
• Data collection method can have a significant impact on the results
• Whatever is the selected sampling method, be aware of it’s properties and weaknesses
• Be careful when adopting results from earlier work
• Time to accept more reappraisal work?