Welcome to

DS504/CS586: Big Data Analytics

Data acquisition and measurement

Prof. Yanhua Li

Time: 6:00pm –8:50pm THURSDAY
Location: AK 232
Fall 2016

Data acquisition and measurement

via Sampling and Estimation

measurement distortions

"World Map" in 1459

- proved incomplete (Columbus et al. 1492)
- wrong proportions (Africa & Asia)

The Fra Mauro world map (1459)

source: Wikipedia

outline

- Why sampling?

- Sampling methods

Motivation

- Measurement studies aid understanding existing systems and user behaviors.

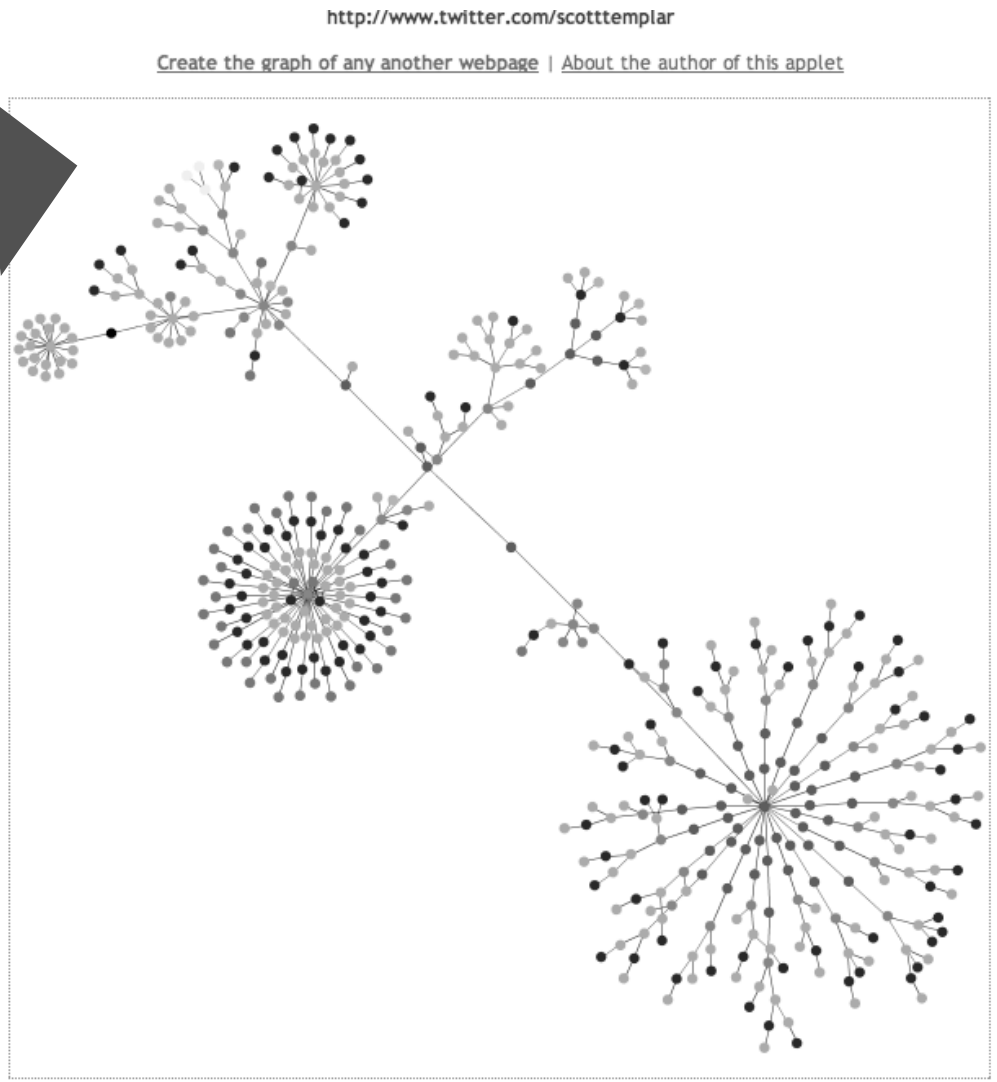
- Capturing an accurate global "snapshot" is often infeasible.

  - *How can we collect representative samples?*

Motivation

Sample data
to estimate the
statistics, i.e., size
degree distribution,
etc.

sample of social networks



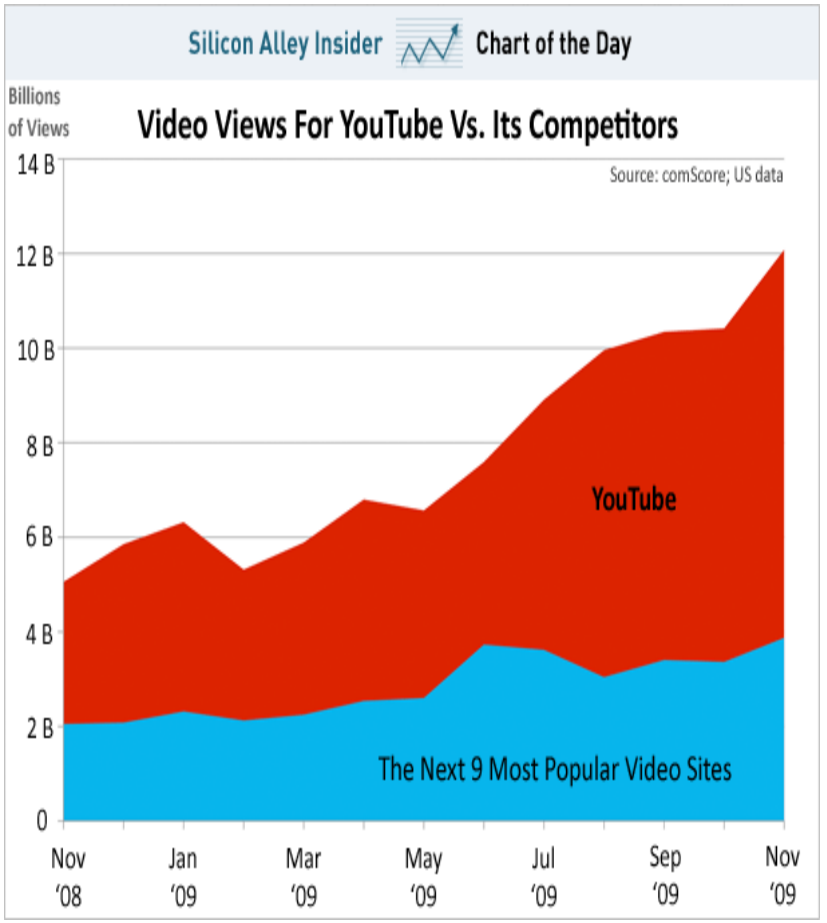
- Capturing an accurate global "snapshot" is often infeasible.
- *How can we collect representative samples?*

Counting YouTube Video via Random Prefix Sampling

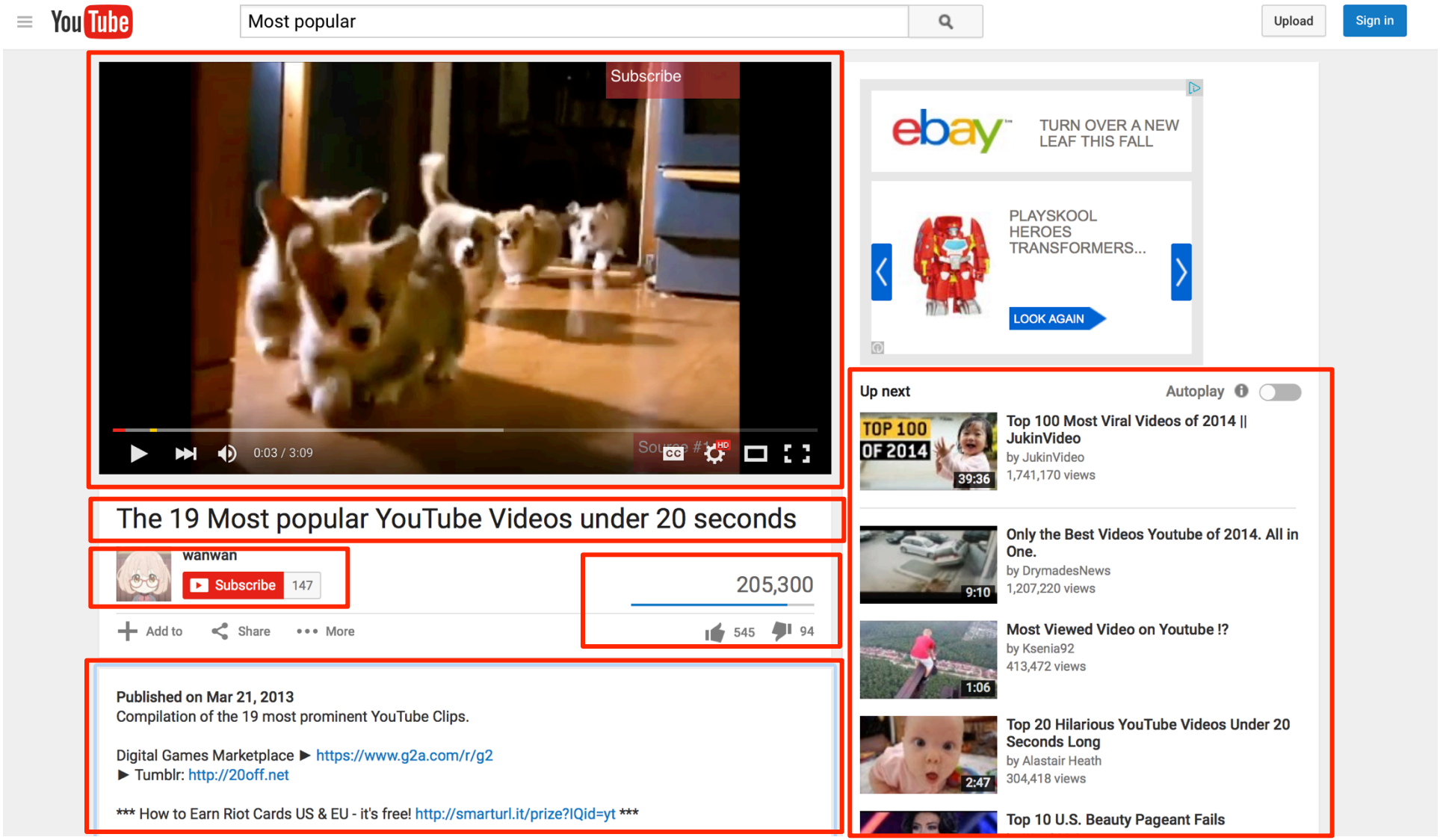
Why YouTube?

World's largest (mostly user-generated) global (excl. China) video delivery service

- More than 13 million hours of video were uploaded during 2010 and 35 hours of video are uploaded every minute.
- More videos are uploaded to YouTube in 60 days than the 3 major US networks created in 60 years
- 70% of YouTube traffic comes from outside the US
- YouTube reached over 700 billion playbacks in 2010
- YouTube mobile gets over 100 million views a day



YouTube Video



Comments from other YouTube users

Socio-technical Aspects of YouTube: Counting Videos & Views

Why Counting YouTube Videos and Views::

- YouTube traffic contributes to a significant portion of inter-domain network traffic
- Knowing the total number of videos and view counts per day can shed light on
  - the total amount of storage
  - as well as the system capacity needed to store and deliver YouTube videos

Challenges:

- These statistics are *not* made available publicly by YouTube
- Even for YouTube, it is costly to get an exact answer.

Challenges for Counting Videos & Views

- Video id space is extremely large, of the order $O(64^{11})$
  - brute-force survey of the entire YouTube video population will be too costly
  - direct application of (uniform) random sampling to the video id space will be ineffective
- Existing methods for *collecting* YouTube videos following the “related videos” links produce a biased sample

Contributions of the IMC 11 paper

- A theoretical model to derive an *unbiased* estimator for estimating the total number of YouTube videos
  - Bounds on variance and confidence interval
  - Cross-validation using two distinct collections of YouTube video id's
- Apply the random prefix sampling method to
  - Estimate the total number of videos and analyze its dynamics
  - Estimate the views counts and study its properties
  - Large bias introduced by traditional related videos based sampling

Sampling Techniques to Count Population

- **German Tank Problem**
- Panther tanks, 1943.
- World War II
- Estimate # German Tanks (*N*)
- the problem of estimating

the maximum of a discrete uniform distribution from Sampling without replacement

- *m* : the max series number
- *k* : total number of tanks observed
- Estimator: $\hat{N} = m(1 + k^{-1}) - 1$
- the sample maximum plus the average gap between observations in the sample.

Sampling Techniques to Count Population

- **Mark and recapture**
- a method commonly used in ecology to estimate an animal population's size *N*.

- Step 1: A portion of the population *K* is captured, marked, and released.

- Step 2: Later, another portion *n* is captured and the number of marked individuals within the sample is counted *k*.
- Estimation: $$\hat{N} = \frac{Kn}{k}$$

Sampling Techniques to Count Population

- **Mark and recapture**

- *N* = Number of animals in the population
- *K* = Number of animals marked on the first visit
- *n* = Number of animals captured on the second visit
- *k* = Number of recaptured animals that were marked

- Assumption: Each animal has an equal probability *p* being captured
- Thus, $p = \frac{k}{K} = \frac{n}{N}$

- The estimator is obtained, as $\hat{N} = \frac{Kn}{k}$.

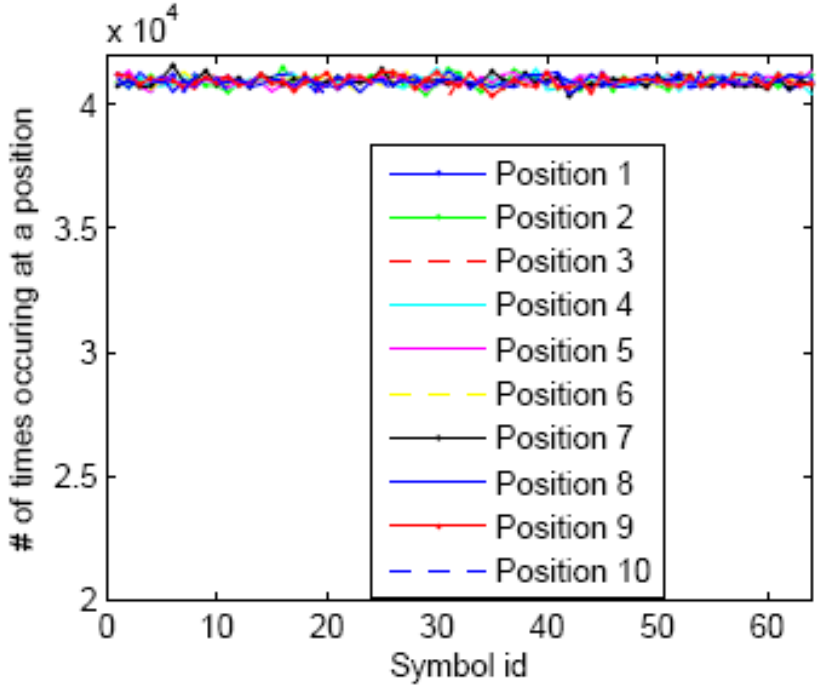
YouTube Video ID Space

Each YouTube id consists of 11 characters

The first 10 characters of a valid id contain any of the characters in S = {0–9,_,–,A–Z, a–z}

The last (11-th) character only comes from T = {0, 4, 8, A,E, I,M, Q,U, Y, c, g, k, o, s,w}

A YouTube video id is randomly generated from the id space $\mathcal{S}$



Prefix Search in YouTube

- Key unique property of YouTube search API we accidentally stumble on
  - When searching using a keyword string of the format ”watch?v=xy-...z”
  - YouTube returns a list of videos whose id’s begin with “xy-”, if they exist.
- The above property is well validated by three real datasets

Certain return limits apply, e.g., maximum # of videos returned.

can we use German Tank and Mark-recapture method to estimate the YouTube video population size, and why?

Random Prefix Sampling

- Let p_L denote the probability that a randomly generated id matches a given L-length prefix

$$p_L = 1/|S|^L = 1/64^L, \text{ if } L=1,\ldots,10$$

$$p_L = 1/(|S|^{10}|T|) = 1/(64^{10}*16), \text{ if } L=11$$

- Generate m prefixes of length L.
- Let X_i^L be the total number of videos with a prefix i of length L, and N the total number of videos

then, $X_i^L \sim \text{Binomial}(N, p_L)$;

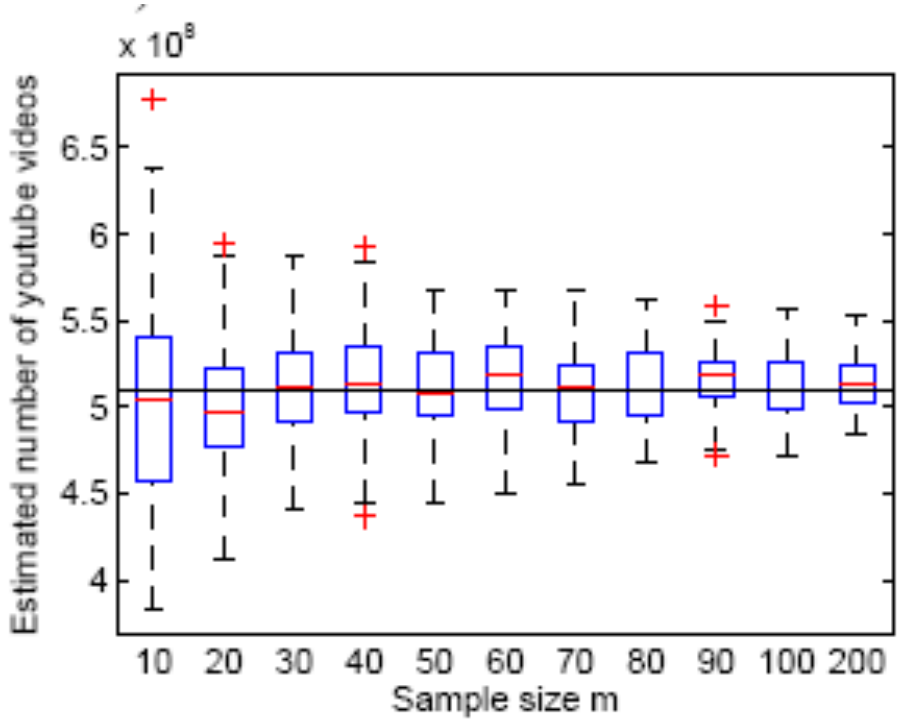
Unbiased Estimator for the Total Number of Videos

- Given *m* samples X_i^L by querying randomly generated prefixes of the same length in [1,11], we have the unbiased estimator of total number of videos

$$\hat{N} = \frac{1}{mp_L} \sum_{i=1}^{m} X_i^L$$

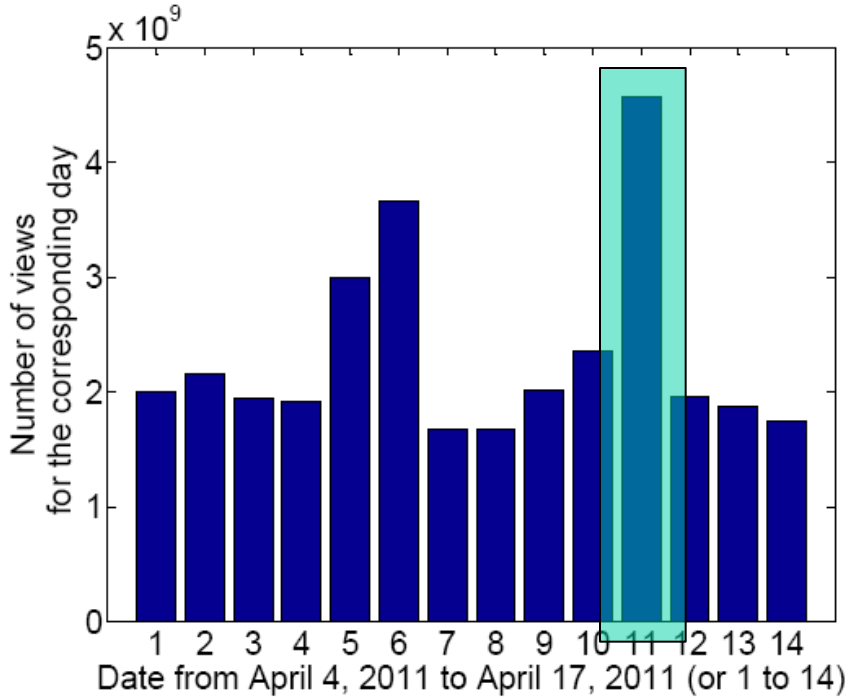
(See paper for the confidence interval and variance)

Estimated number of YouTube videos by 05/12/2011



- The estimated result becomes more stable with more samples
- Around half a billion videos by May 2011

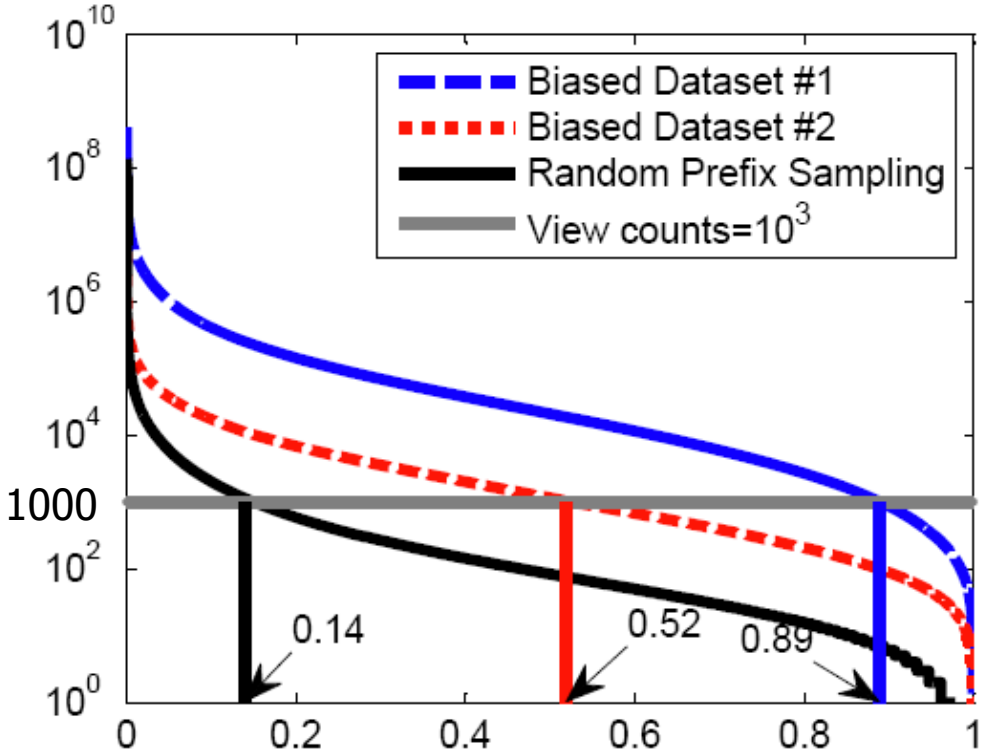
Number of Views for a two week period



On average it is 2.3 billion per day

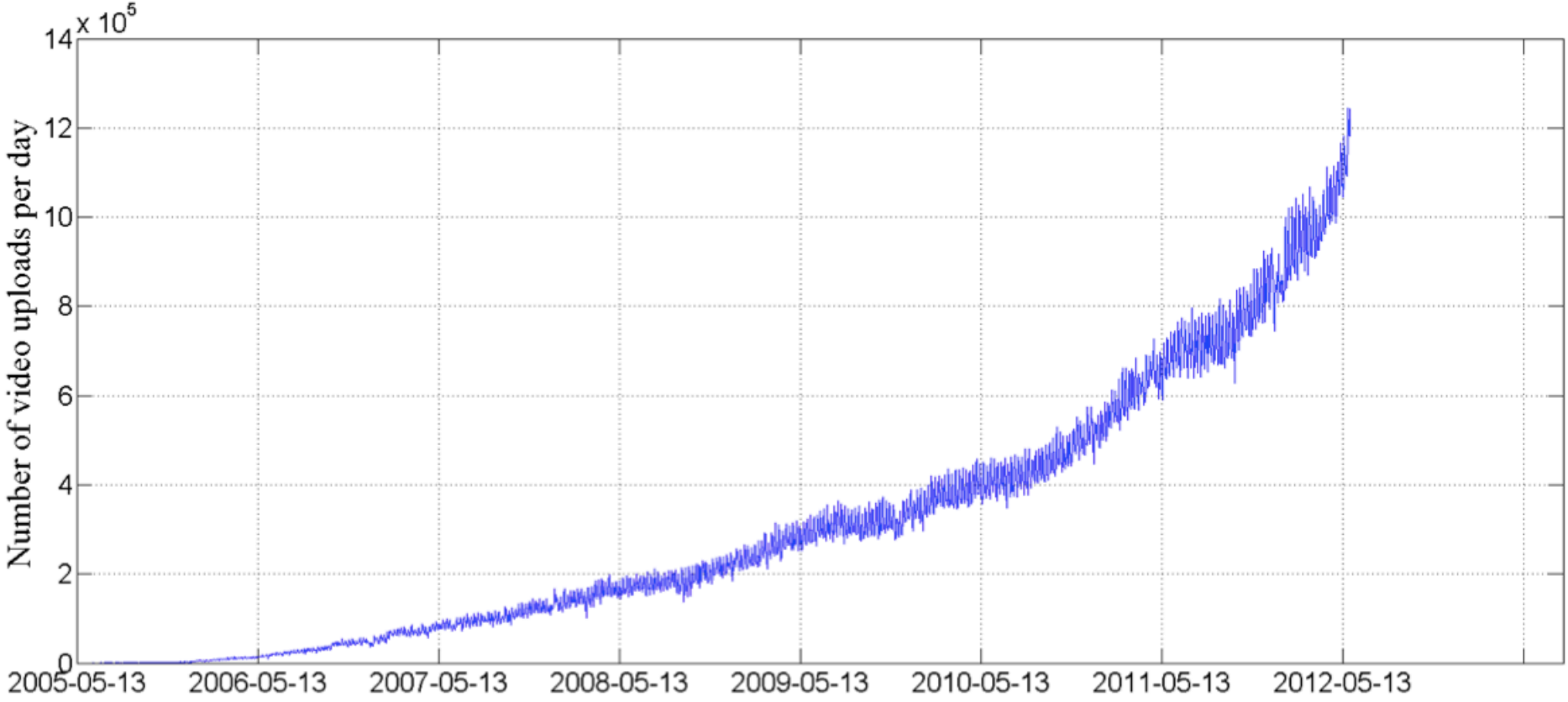
For some day it can be as large as over 4.6 billions or over twice of the average, e.g., April 11, 2011

Number of Views by different DataSets



- X-axis: proportion of videos in each dataset
- Y-axis: view counts
- DataSets based on related videos show high biases toward hot videos
- Datasets based on related videos ignore a large portion of videos with view counts less than 1000

Daily YouTube video uploads



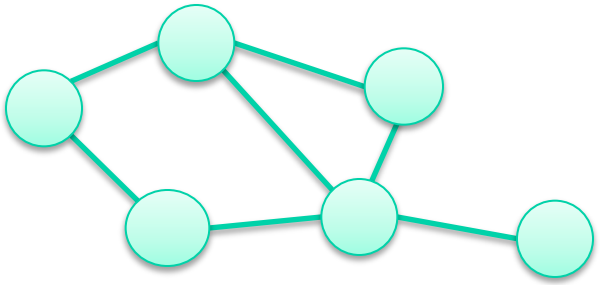
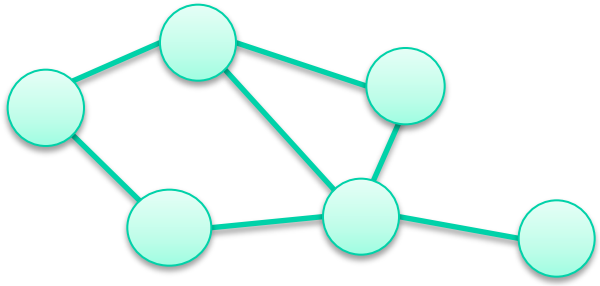
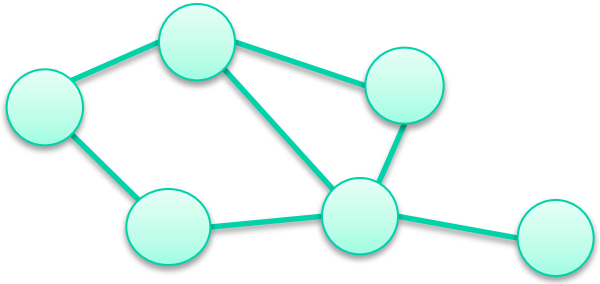
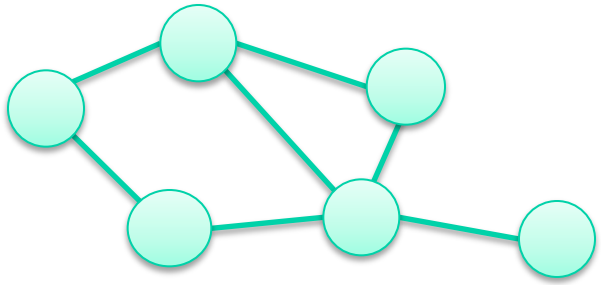
Slow in the first two years but increase more and more quickly in the following years;

Sampled Data

- Q00I-y9iePw|Tech|2008-08-19T02:52:52.000Z|23|blessingsolarenergy
- q00i--f2s4s|Entertainment|2008-10-12T18:29:22.000Z|602|corester69
- q00j-Zrs730|Music|2009-08-04T08:27:38.000Z|323|jeppeli123
- q00j-9vwAEA|Games|2009-08-15T19:36:50.000Z|64|GMLEGENDAZTEK
- Q00J-XhwEqA|People|2009-04-23T22:56:54.000Z|72|sjohnsgeo
- Q00j-9h8g0k|Games|2010-10-14T11:44:13.000Z|29|bebelulu91
- q00k-mgp9ak|Music|2008-02-12T16:51:02.000Z|169|grizzly9587
- Q00K-TZ53IY|People|2009-02-17T23:58:46.000Z|535|83diogosampaio
- q00K-VR6xT0|Comedy|2011-02-13T18:04:26.000Z|71|WhatsUpTay
- Q00L-OsxpfM|Comedy|2008-04-11T00:46:39.000Z|94|feergi
- Q00m-hFq_0Y|Music|2010-01-02T02:15:10.000Z|212|BakhtiyarHajiyev
- q00m-44nU7o|Sports|2007-07-23T21:17:16.000Z|27|smashingSurfer
- Q00m-Qha_nE|People|2009-11-29T03:54:40.000Z|29|swaggaqueens
- Q00N-LAzRgI|Entertainment|2010-12-12T03:03:20.000Z|321|BNMASS

Network sampling

sampling graphs

random sampling (uniform & independent)	crawling
▸ vertex sampling	▸ BFS sampling
	
▸ edge sampling	▸ random walk sampling
	

Course Project

YouTube Data API v3.0

Get Started

- Google Account
  - access the Google Developers Console, request an API key, and register your application
- Create a project
  - Google Developers Console and obtain authorization credentials so your application can submit API requests.
- Add YouTube Data API to your Project services
- Obtain a key like this
  - AIzaSyCTNWZ26RDrleu_aNMp9U34NkpYkzJppOc

YouTube Data API v3.0

Sample API Requests

- Retrieve and manipulate YouTube resources, including
  - videos,
  - channels,
  - playlists,
  - and etc
- More on tutorials online. Just name a few here.
  - Video 1
  - Video 2
  - Video 3
  - Find more in Google Search & YouTube.
- Note that API v2.0 is no longer maintained.
- https://support.google.com/youtube/answer/6098135?hl=en

YouTube Data API v3.0 Examples

Sample API Requests

- An individual Video
- https://www.googleapis.com/youtube/v3/videos?id=Im69kzhpR3I&key=AIzaSyCTNWZ26RDrIeu_aNMp9U34NkpYkzJppOc&part=snippet

- A prefix search
- https://www.googleapis.com/youtube/v3/search?part=snippet&q=%22watch?v=f6tz%22&type=video&key=AIzaSyCTNWZ26RDrIeu_aNMp9U34NkpYkzJppOc

YouTube Data API v3.0 Examples

Sample API Requests

- A prefix search
- Base URL: https://www.googleapis.com/youtube/v3/
- Function: Search?part=snippet
- Keyword: &q=%22watch?v=f6tz%22
- Type: &type=video
- Auth Key: &key=AIzaSyCTNWZ26RDrleu_aNMp9U34NkpYkzJppOc

For more configuration settings, please refer to YouTube Data API v3.0

For sample code in Python, Java, etc, please refer to Sample Code for YouTube Data API

Sampled Data

- Q00I-y9iePw|Tech|2008-08-19T02:52:52.000Z|23|blessingsolarenergy
- q00i--f2s4s|Entertainment|2008-10-12T18:29:22.000Z|602|corester69
- q00j-Zrs730|Music|2009-08-04T08:27:38.000Z|323|jeppeli123
- q00j-9vwAEA|Games|2009-08-15T19:36:50.000Z|64|GMLEGENDAZTEK
- Q00J-XhwEqA|People|2009-04-23T22:56:54.000Z|72|sjohnsgeo
- Q00j-9h8g0k|Games|2010-10-14T11:44:13.000Z|29|bebelulu91
- q00k-mgp9ak|Music|2008-02-12T16:51:02.000Z|169|grizzly9587
- Q00K-TZ53IY|People|2009-02-17T23:58:46.000Z|535|83diogosampaio
- q00K-VR6xT0|Comedy|2011-02-13T18:04:26.000Z|71|WhatsUpTay
- Q00L-OsxpfM|Comedy|2008-04-11T00:46:39.000Z|94|feergi
- Q00m-hFq_0Y|Music|2010-01-02T02:15:10.000Z|212|BakhtiyarHajiyev
- q00m-44nU7o|Sports|2007-07-23T21:17:16.000Z|27|smashingSurfer
- Q00m-Qha_nE|People|2009-11-29T03:54:40.000Z|29|swaggaqueens
- Q00N-LAzRgI|Entertainment|2010-12-12T03:03:20.000Z|321|BNMASS

Project 1 directions

What is your project goal?

- What new story you want to tell?
- New contents to sample?
- New sampling methods via API?
- New statistics of YouTube, view count distribution, dynamics, or # uploaders/active users?
- Analysis on other websites, Twitter, Facebook, Foursquare, Yelp, with API interfaces

Broad impacts? (Keep in mind)

- How YouTube is evolving?
  - More business or personal videos? How to distinguish the two
  - How special events, e.g., NBA game, breaking news, affect the uploading, viewing behaviors
- Online Marketing, advertising?

Project 1

As a Team

- Project work
- Project Presentations
- Topic Presentations

Project I

- Teams

- Timeline and Evaluation
  - Proposal: Week 3, 9/8
  - Methodology: Week 4, 9/15
  - Empirical Results: Week 5, 9/22
  - Introduction: Week 6, 9/29
  - Conclusion, Abstract: Week 7, 9/29
  - Final Report and Presentation: Week 8, 10/13

- Discussions

Next Class: Data Preprocessing &Cleaning

- Do assigned readings before class
  - Be prepared, read and review required readings *on your own in advance!*
  - *Do literature survey: find and read related papers if any*
  - *Bring your questions to the class and look for answers during the class.*
- Submit reviews/critiques
  - In mywpi before class
  - Bring 2 hardcopies to the class
  - Hand in one copy, and keep one copy with you.

Review Writing:
http://users.wpi.edu/~yli15/courses/DS504Spring16/Critiques.html

- Attend in-class discussions
  - Please ask and answer questions in (and out of) class!
  - Let's try to make the class interactive and fun!