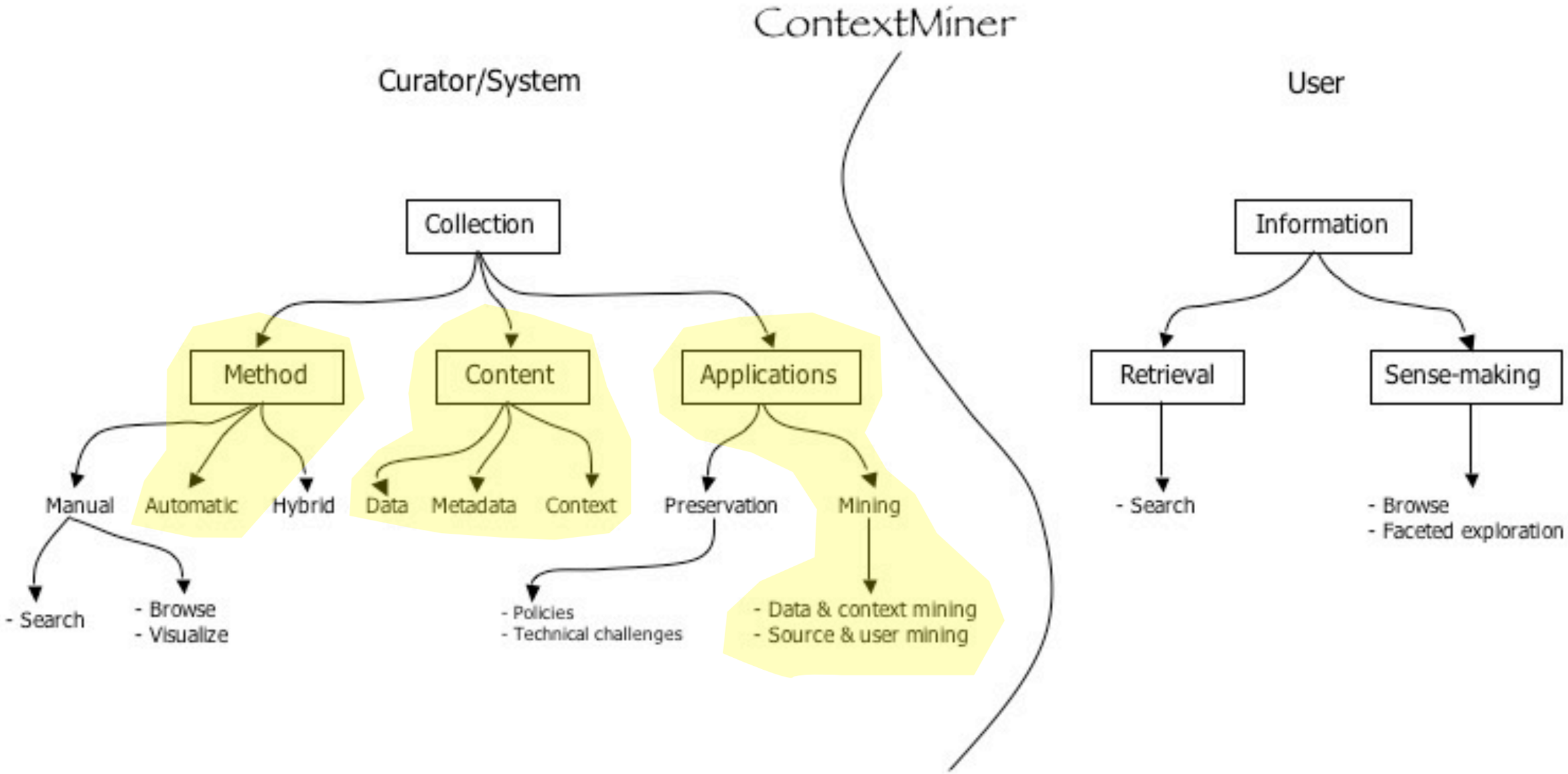
Hunting for Hip, Hipsters, and Happenings on YouTube - A *ContextMiner* Story

Chirag Shah
CRADLE Talk
SILS, UNC Chapel Hill
September 21, 2007

Flow of this story

- What are we trying to do?
- Why are we doing it?
- How did we do it?
- What came out of it?
- How to make sense of it?
- Where do we go from here?

ContextMiner - the big picture

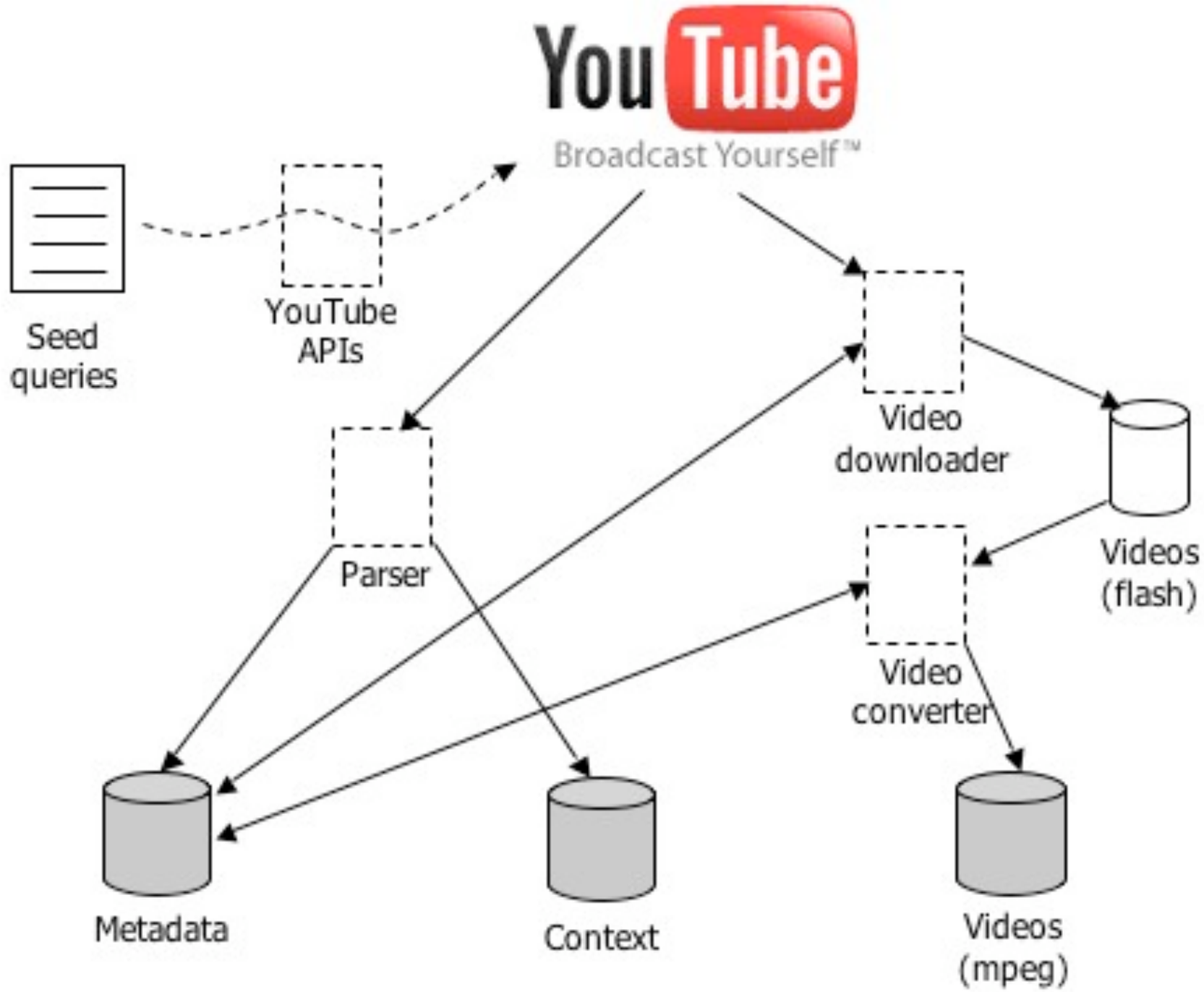


YouTube spread

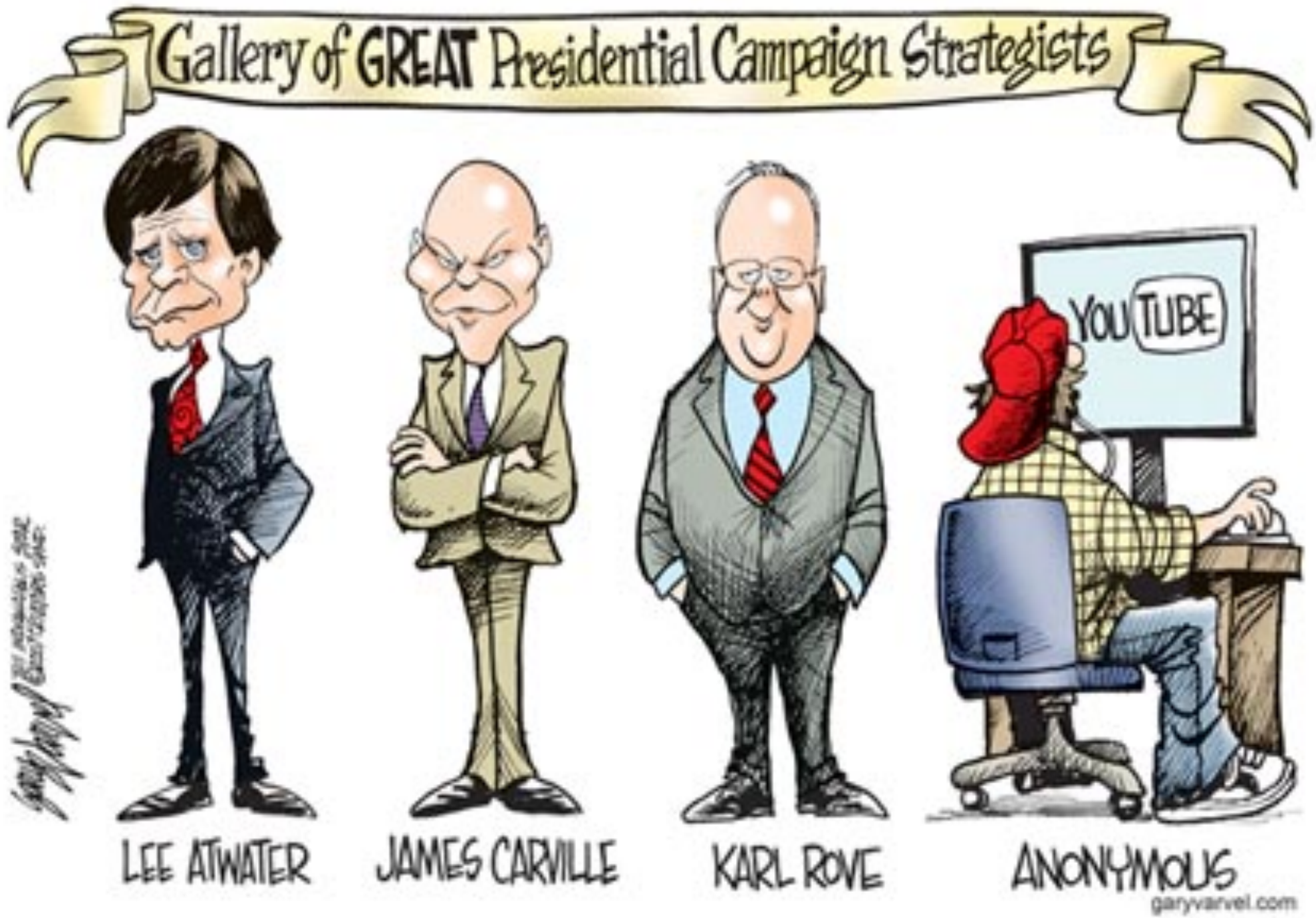
- YouTube's market share: 42.94% (Source: Hitwise.com)
- Average visit duration on YouTube: 13:20 mins (Source: Xeep.net)
- As of August, 2006: (Source: Wall Street Journal)
  - Videos > 6,000,000 (growth: 20% every month)
  - Storage > 45 TB
  - Views > 1,730,000,000
  - Time spent watching YouTube videos = 9,035 years
- Time taken for the spread among 50 million users (Source: Dean José-Marie Griffith's talk on 09/17/2007)
  - Radio: 17 years
  - TV: 13 years
  - Internet: 5 years
  - MySpace: 3 years
  - YouTube: 1 year

YouTube spread

Video harvesting from YouTube



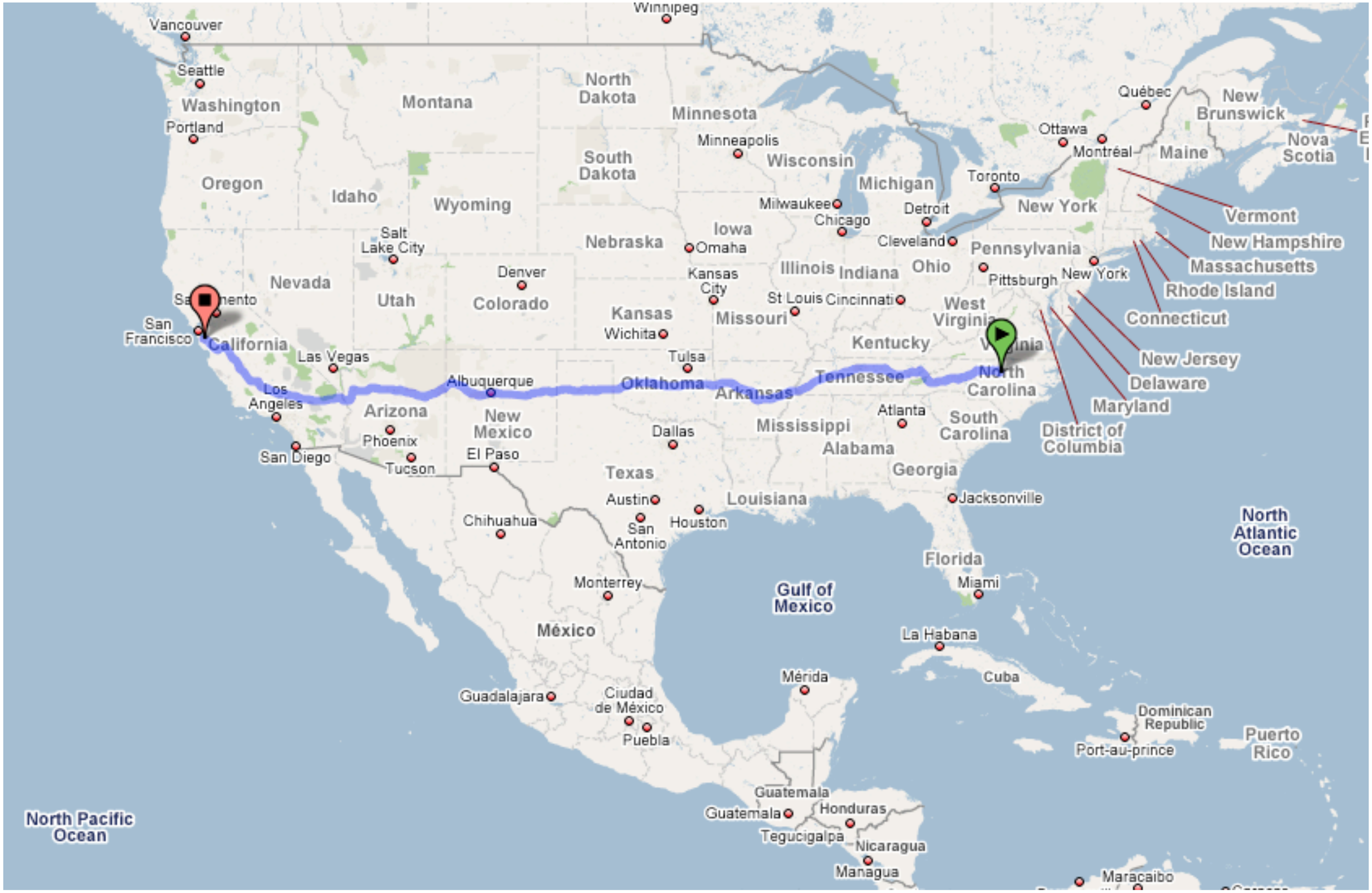
Video harvesting from YouTube for election 2008



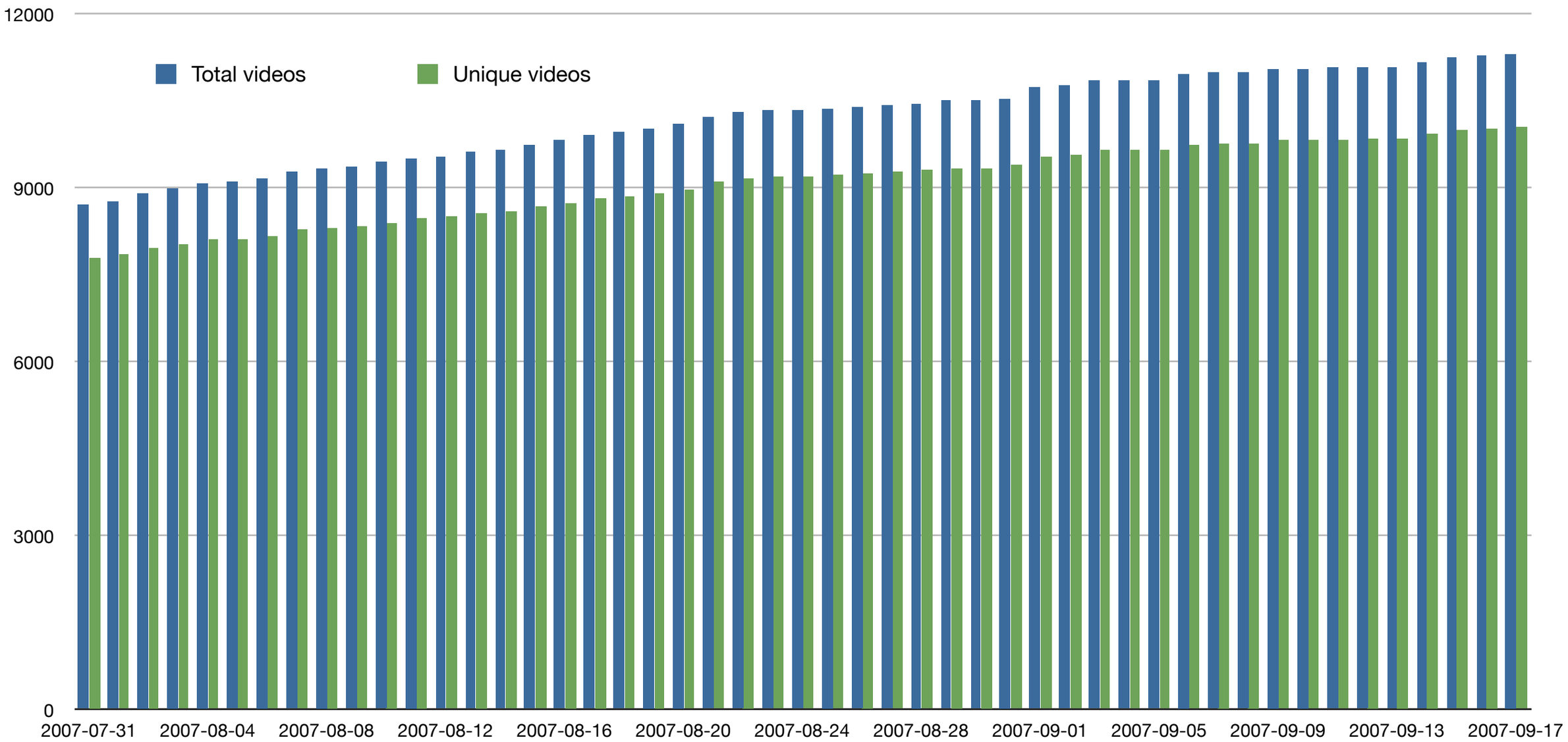
Video harvesting from YouTube for election 2008

- Number of seed queries: 56 (now 57)
  - 6 general queries and the rest are the names
- Crawl everyday (almost)
- Get top 100 results for each query
- Collect more than 20 attributes (including all the comments)
- Download flash videos
- Use YouTube APIs, screen scrapping, and other tools

ContextMiner's journey - from Carolina to California



Overview of the collection



Overview of the collection (as of 09/17/2007)

- Crawls = 100
- Unique videos > 10,000
- Video files > 100 GB
- Total honors > 200
- Total views > 125,000,000
- Total ratings > 800,000
- Total comments ~ 800,000

JCDL 2007 (Vancouver, BC, Canada)

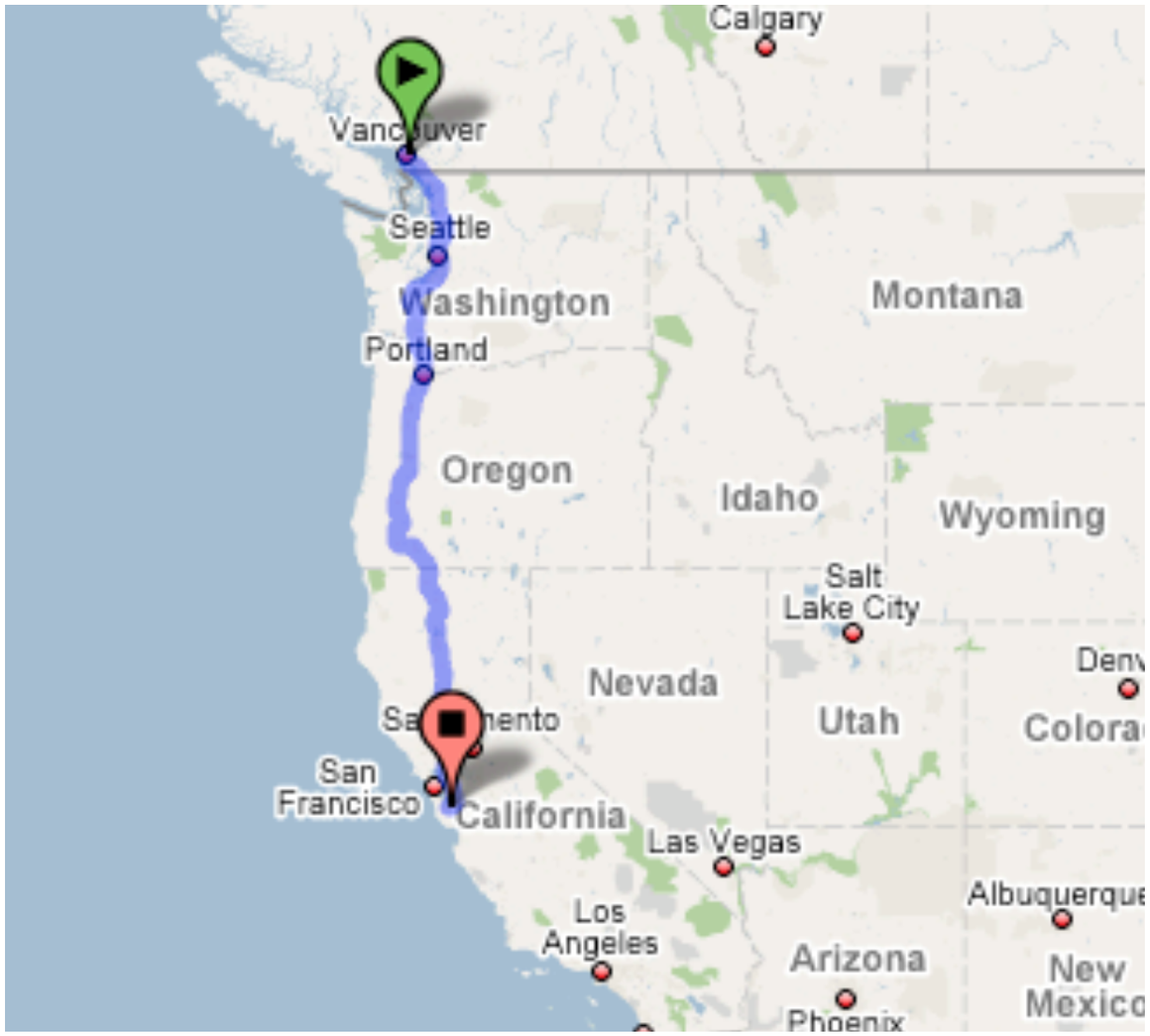
- Shah, Chirag and Marchionini, Gary. *Capturing Relevant Information for Digital Curation*. In IEEE ACM Joint Conference on Digital Libraries (JCDL 2007).

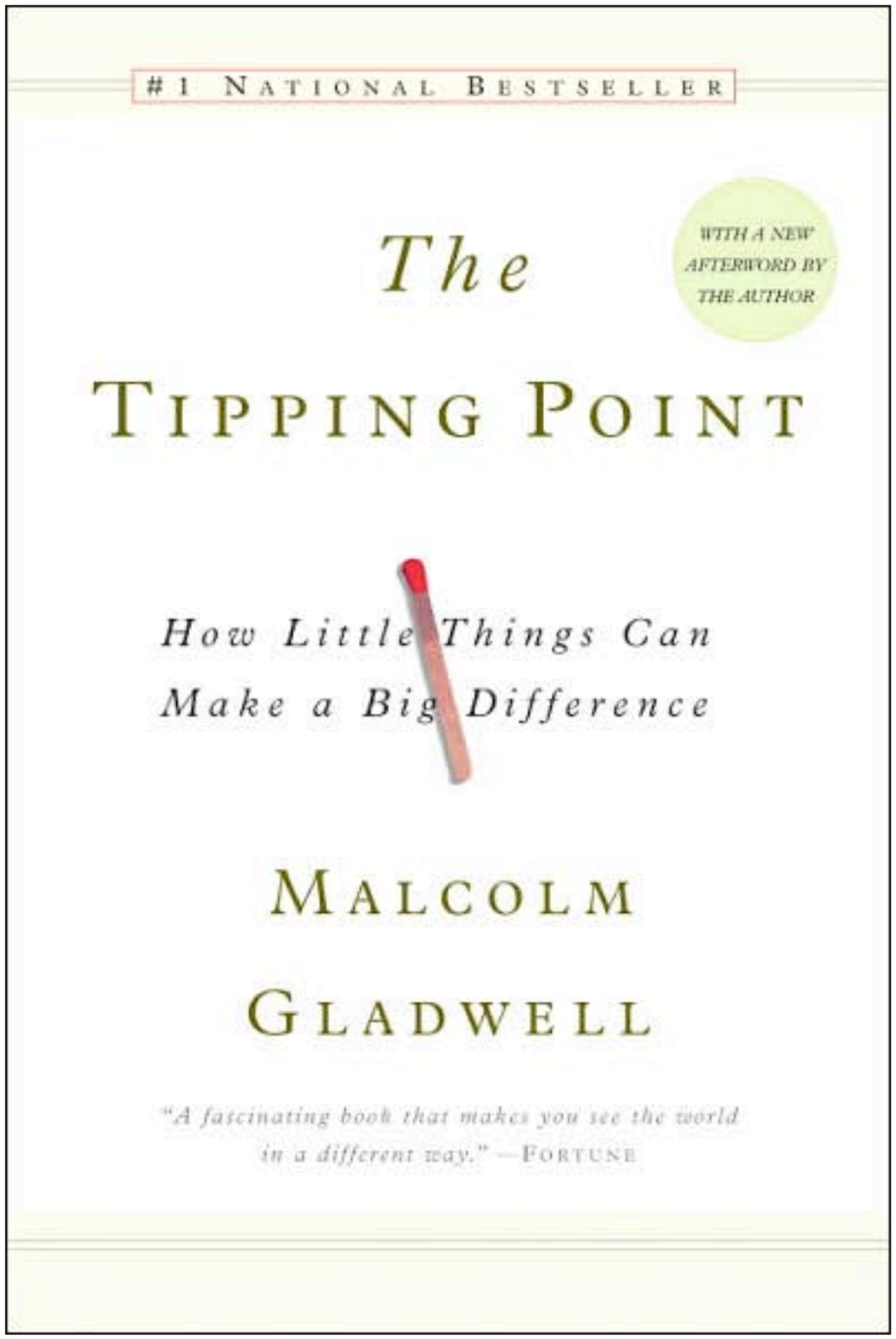
- Shah, Chirag and Marchionini, Gary. *ContextMiner: A Tool for Digital Curator.* In IEEE ACM Joint Conference on Digital Libraries (JCDL 2007).

- Shah, Chirag and Marchionini, Gary. *Preserving 2008 US Presidential Election Videos*. In the Proceedings of International Web Archiving Workshop (IWAW) 2007.


YOU KNOW WHAT I LIKE ABOUT SUMMER DAYS?
THEY'RE JUST MADE FOR *DOING* THINGS
...EVEN IF IT'S NOTHING.
ESPECIALLY IF IT'S NOTHING.
WATTERSON

Tipping Point



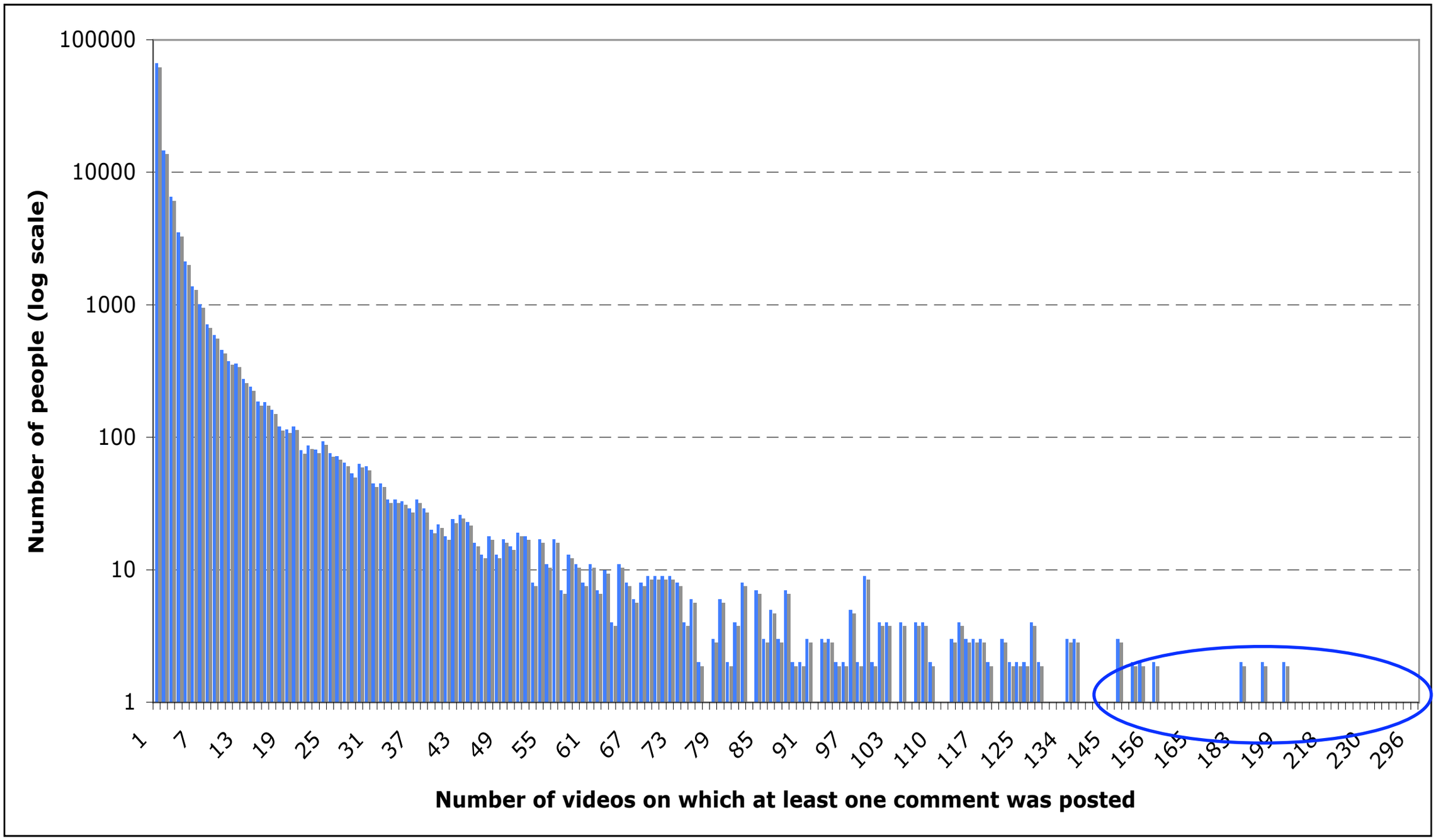


Looking for the “movers and shakers”

- Connectors
  - Know many people
  - Many acquaintances; may not be very strong
- Mavens
  - Know a lot about something
  - Special powers because of their in-depth knowledge
- Salesmen
  - Have skills to persuade people
  - May not have a lot of knowledge, but can convince someone

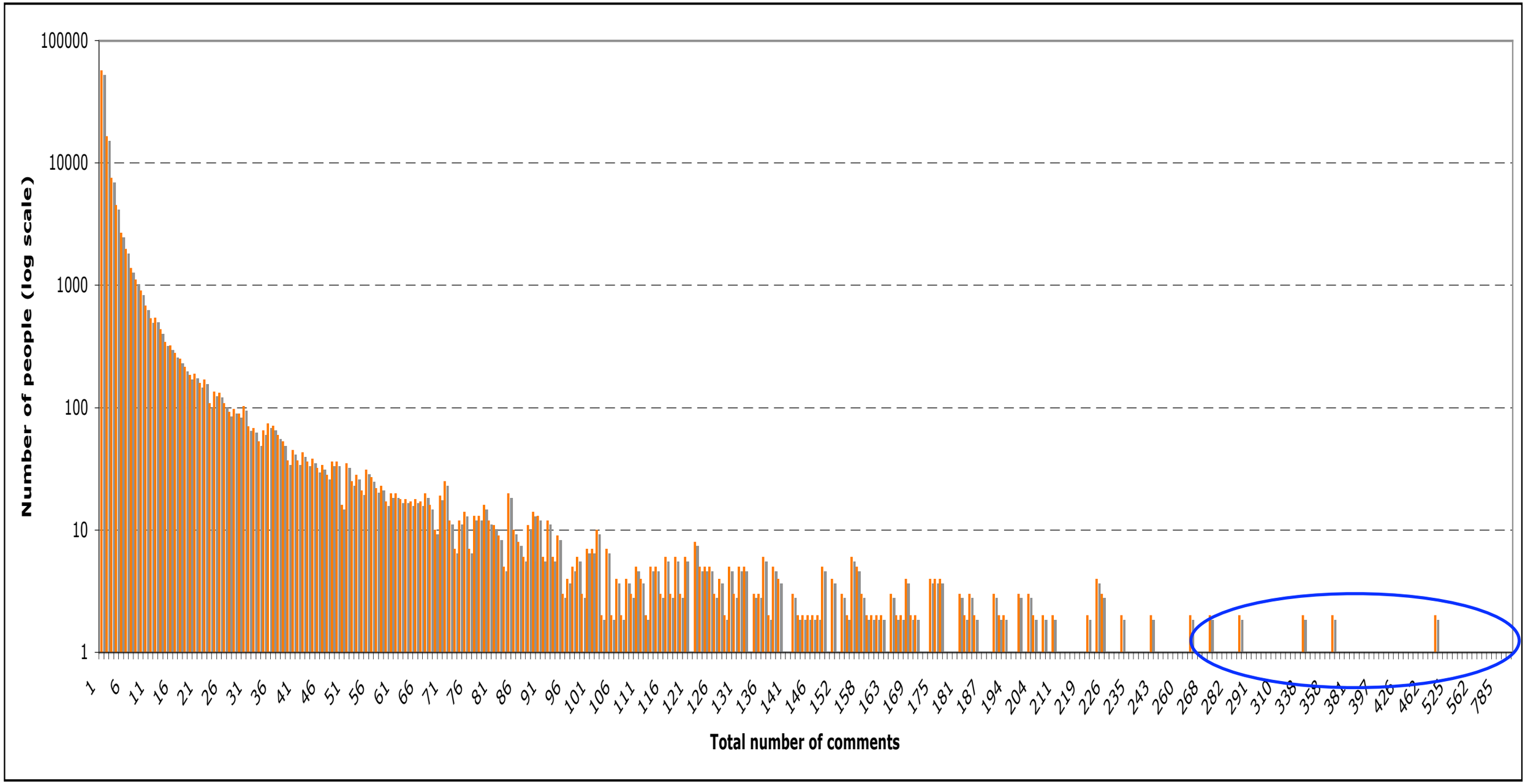
Election “connectors” on YouTube

People who posted at least one comment on many videos



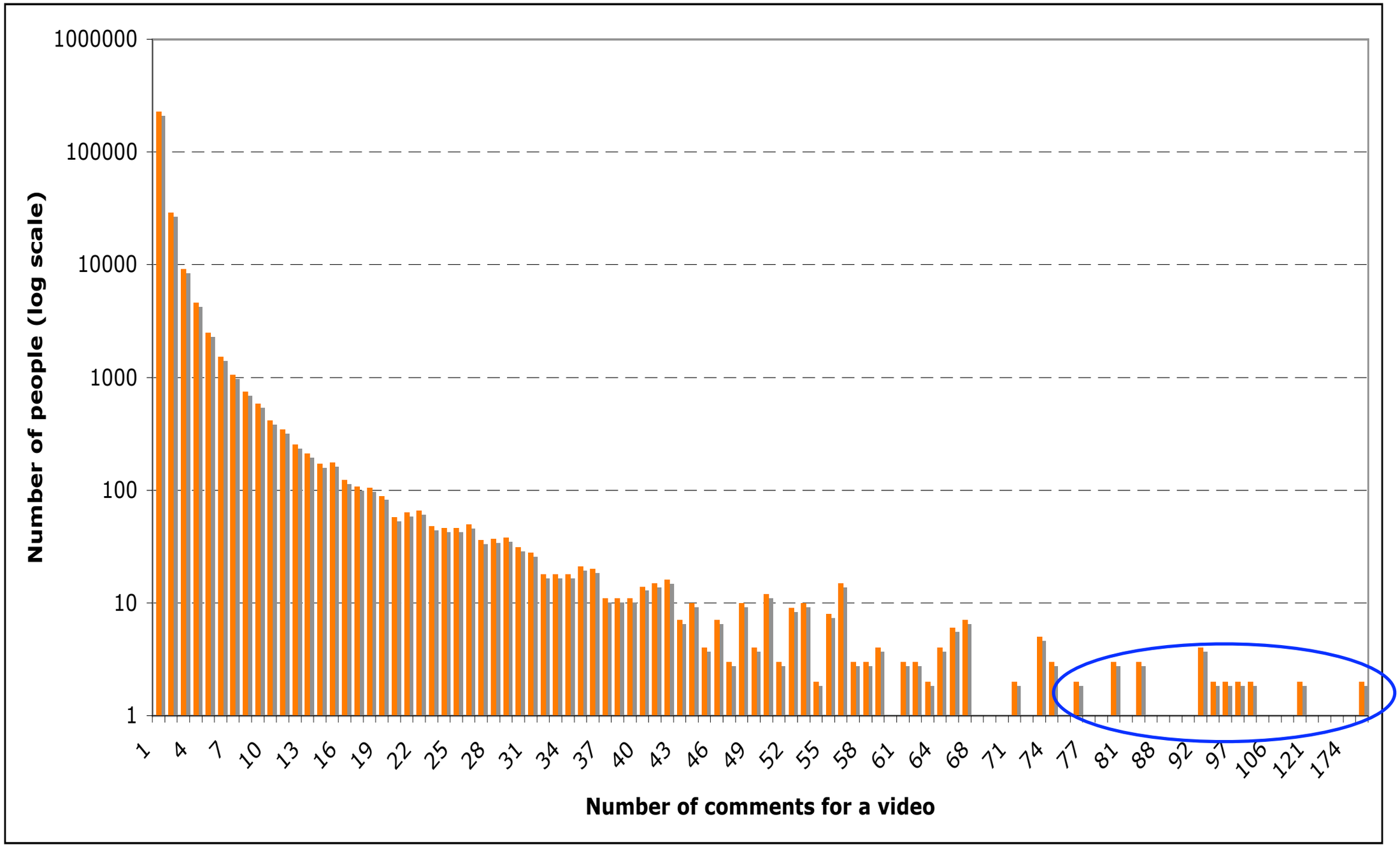
Election “mavens” on YouTube

People who commented a lot in the entire collection



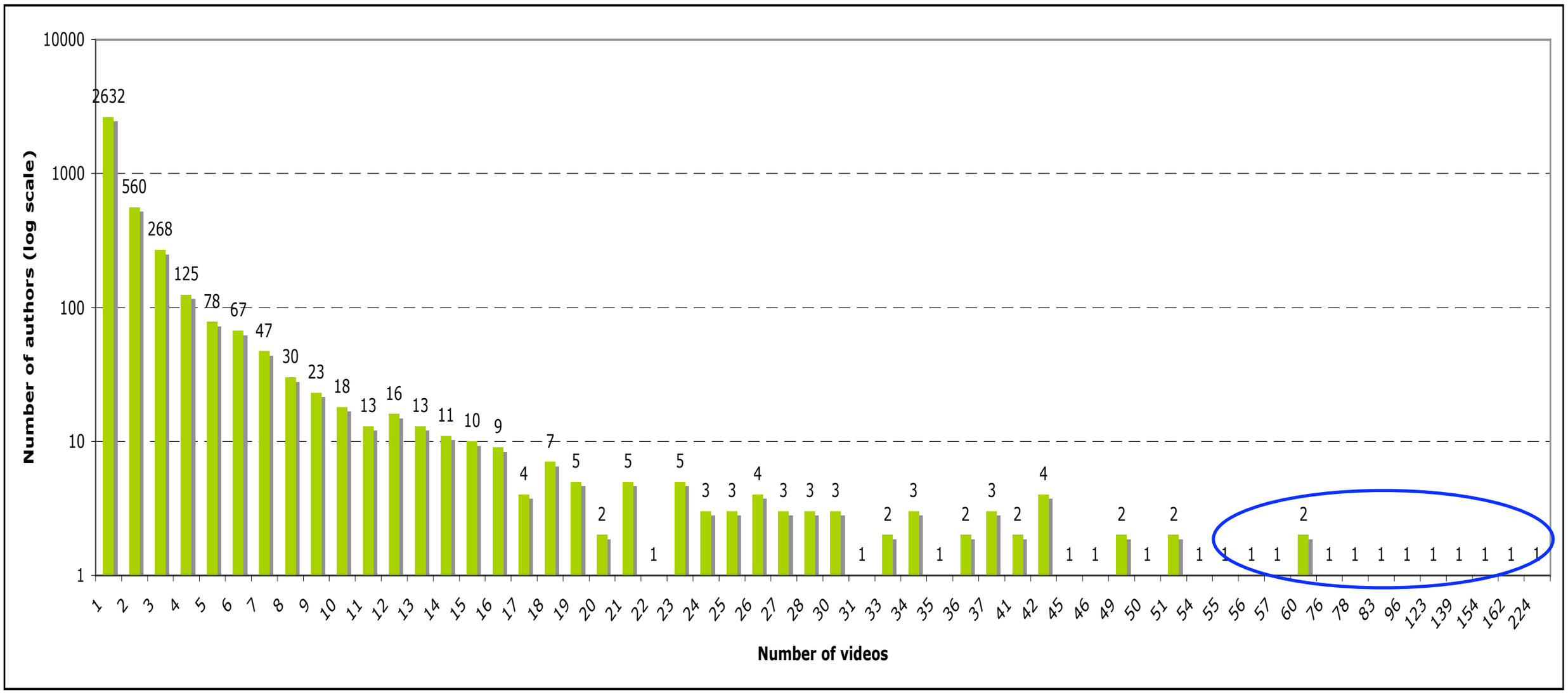
Election “mavens” on YouTube

People who commented a lot for any given video



Election “salesmen” on YouTube

People who posted a lot of videos



SIGIR 2007 (Amsterdam, The Netherlands)

- Shah, Chirag and Marchionini, Gary. *DiscoverInfo: A Tool for Discovering Information with Relevance and Novelty*. In ACM SIGIR 2007.

Detecting events from the collection

ϕ may not be a good representation, it its not a linear function.

We want to know how peoples participation ~~increase~~ change with changing number of visitors and this may not be linear. For instance, double number of visits may not lead to doubling number of comments.

We need to learn this function

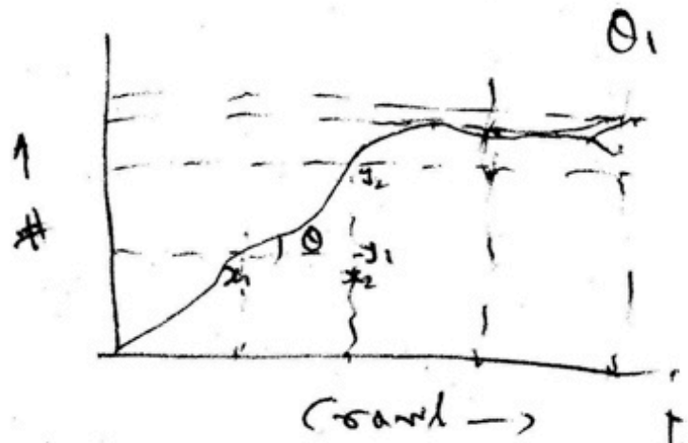
Evaluation: show that this learned function succeeds in predicting "right" thing when simple ϕ fails.

Local changes in ϕ can help us understand the change in level of participation

$\Delta\phi = 0$ No change

> 0 More participation

< 0 Less participation



$$\theta_1^i = \tan^{-1}\left(\frac{y_2 - y_1}{x_2 - x_1}\right)^i$$

$$\Delta\theta_1 = \theta_1^j - \theta_1^i$$

$$\Delta\theta_2 = \theta_2^j - \theta_2^i$$

$$\Delta\theta_3 = \theta_3^j - \theta_3^i$$

Model: (θ, ϕ)

$\Delta\theta = \theta_j - \theta_i$

$\phi = \frac{\text{participants}}{\text{visitors}}$

$\Delta\theta = 0$ No change

> 0 More interest

< 0 Less interest

Can we measure the strength/confidence of this measure?

θ_1 for # views

θ_2 " # comments

θ_3 " # ratings

$\phi_1 = \frac{\text{comments}}{\text{views}}$

$\phi_2 = \frac{\text{ratings}}{\text{views}}$

Model for detecting changes for a video

$$M = (\Theta, \Phi)$$

Changes in different parameters

Level of participation

$$\Theta = (\theta_1, \theta_2, \theta_3)$$

$$\theta_i = tan^{-1}\left(\frac{y_2^i - y_1^i}{x_2^i - x_1^i}\right)$$

$i = 1$: #views

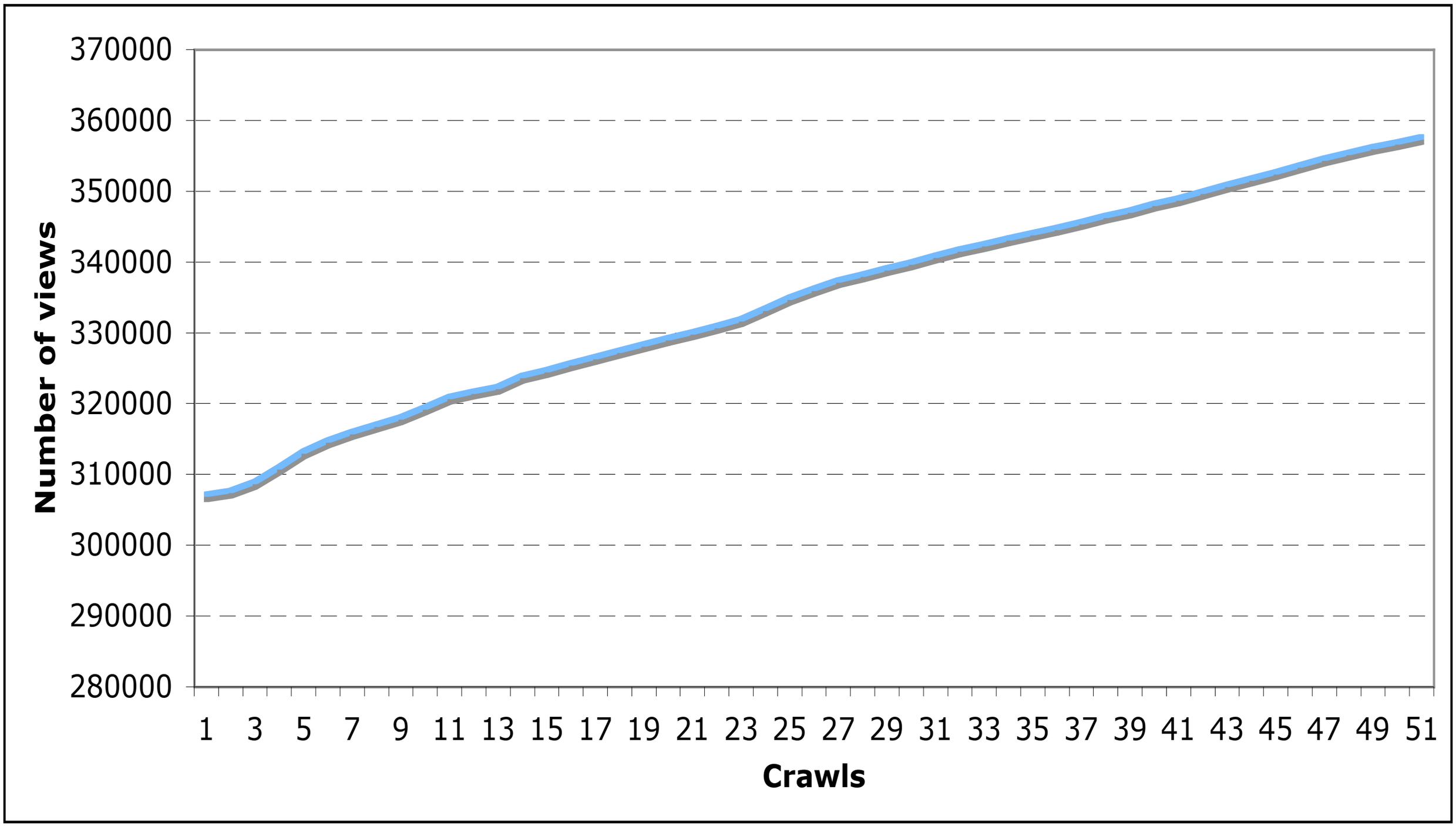
$i = 2$: #comments

$i = 3$: #ratings

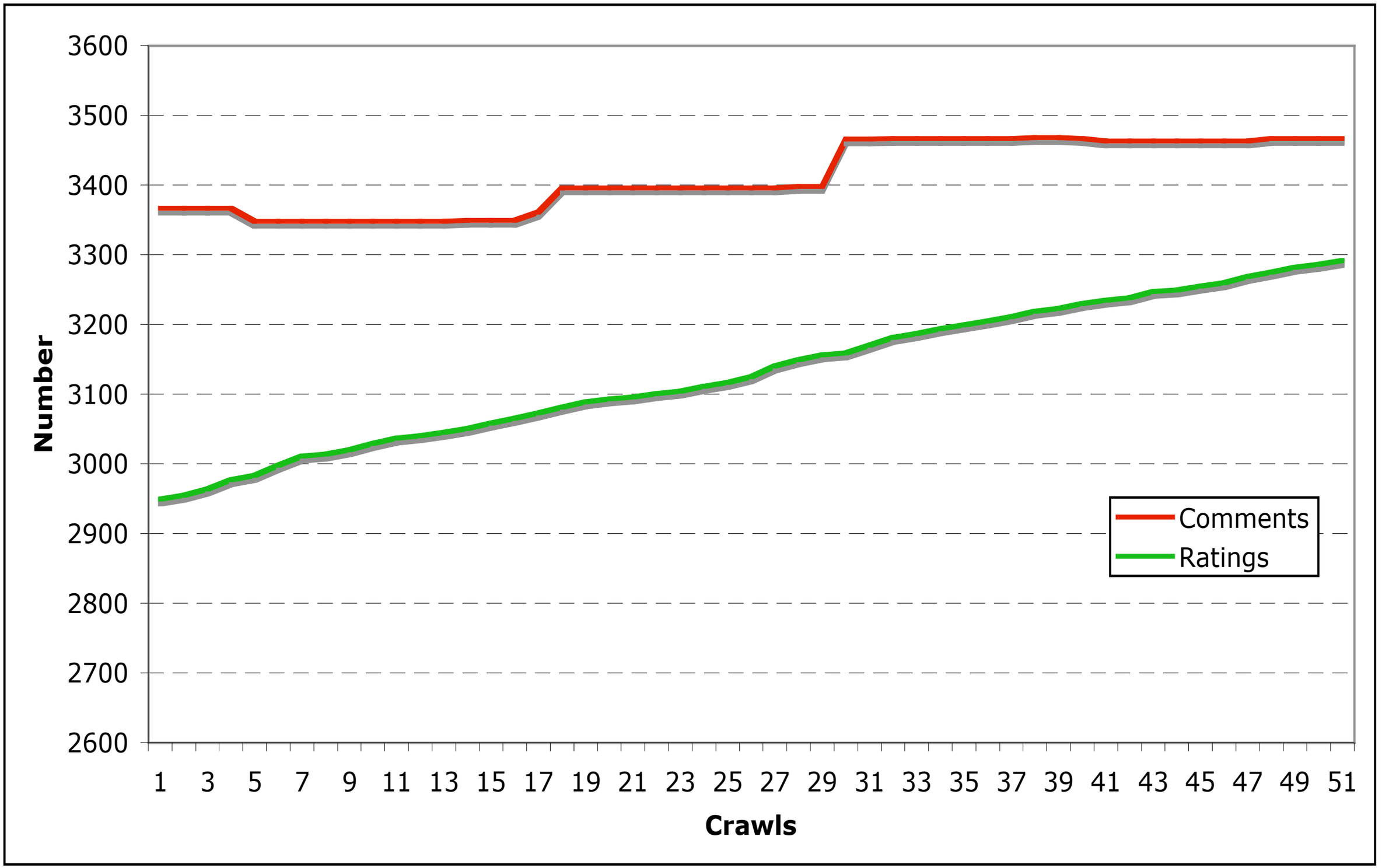
$$\Phi = (\phi_1, \phi_2)$$

$$\phi_1 = \frac{\#\text{comments}}{\#\text{views}} \qquad \phi_2 = \frac{\#\text{ratings}}{\#\text{views}}$$

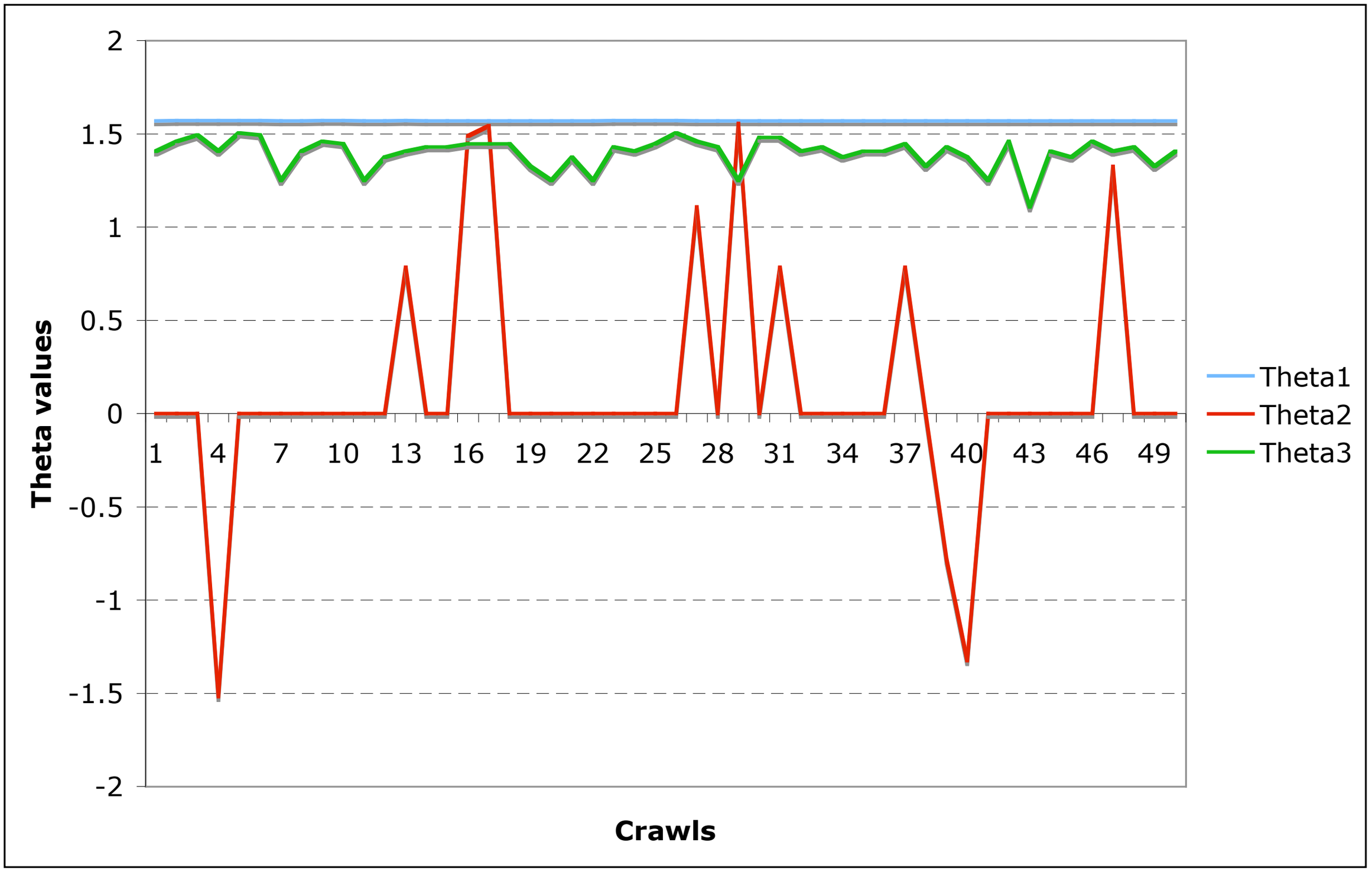
View counts for a video



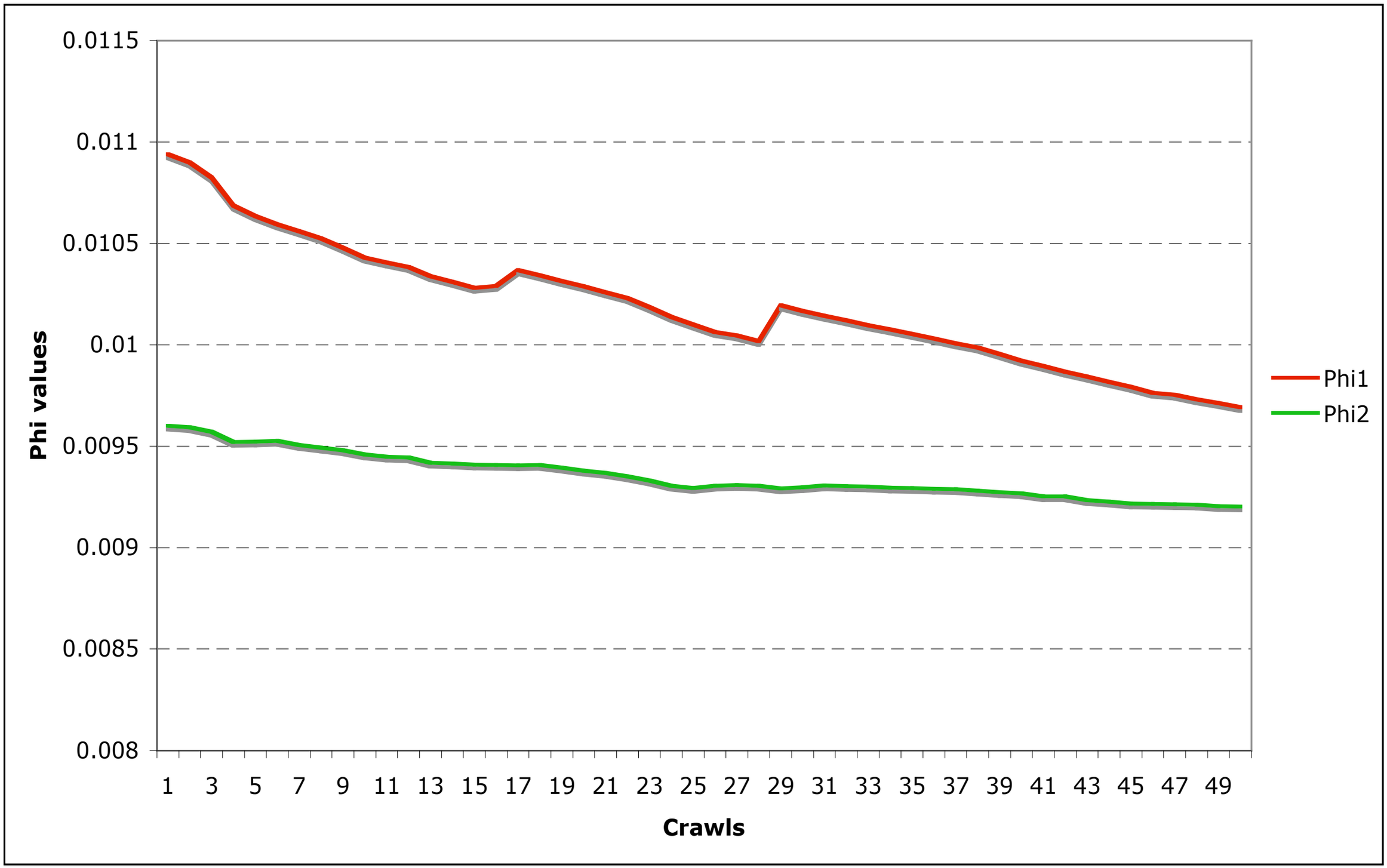
Comments and ratings counts for a video



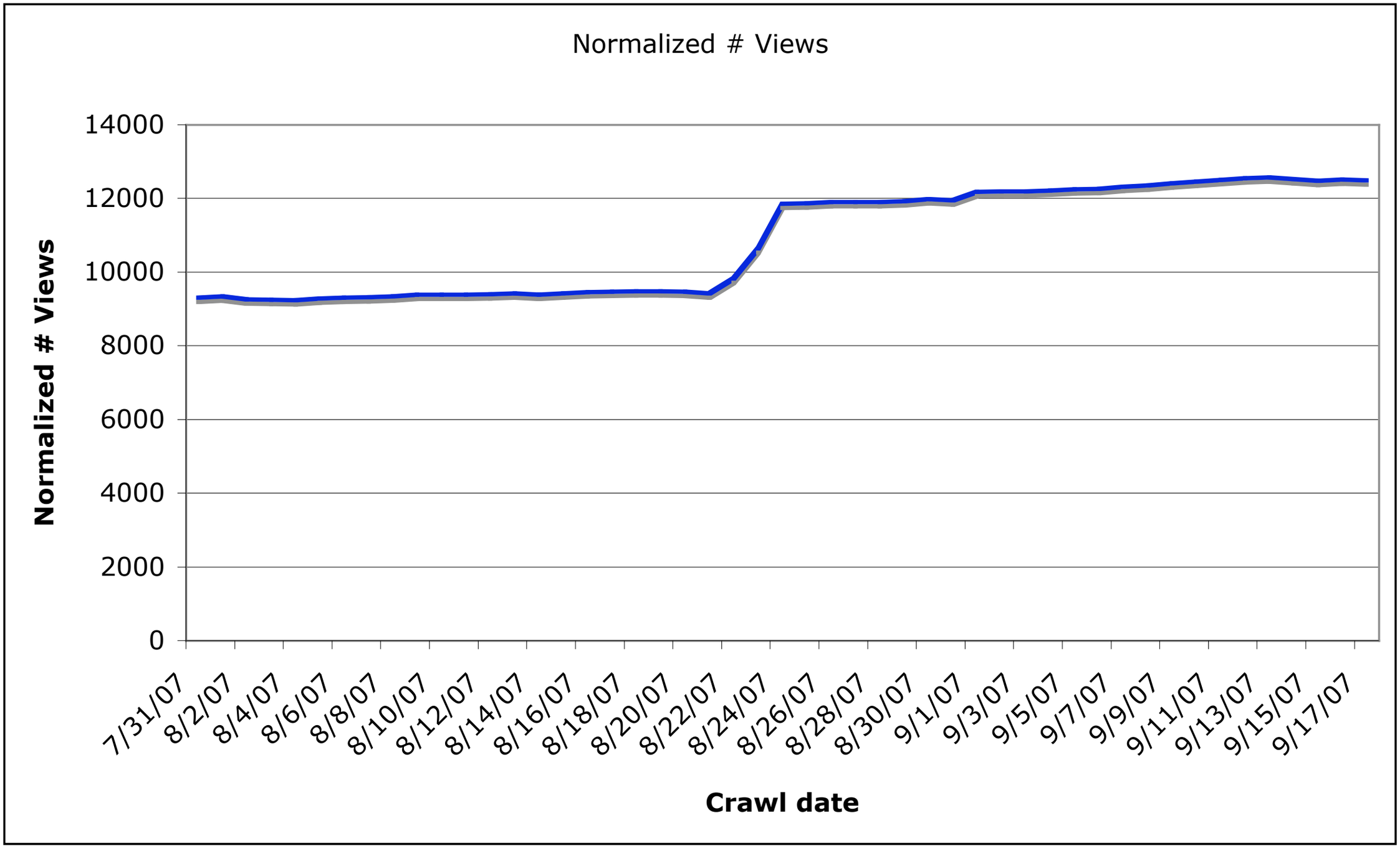
Theta values



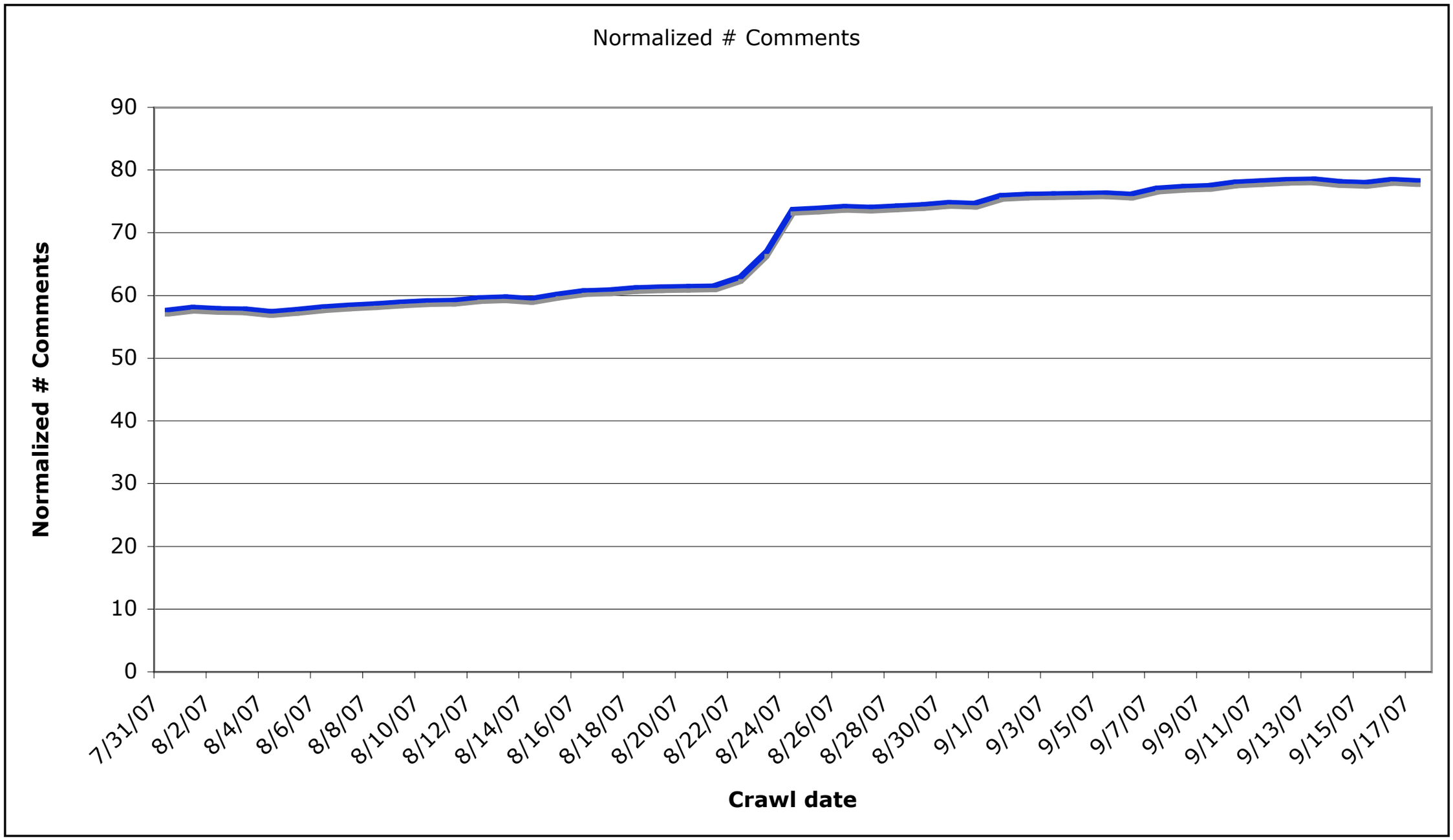
Phi values


0.0115
0.011
0.0105
0.01
0.0095
0.009
0.0085
0.008
Phi values
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
Crawls
Phi1
Phi2

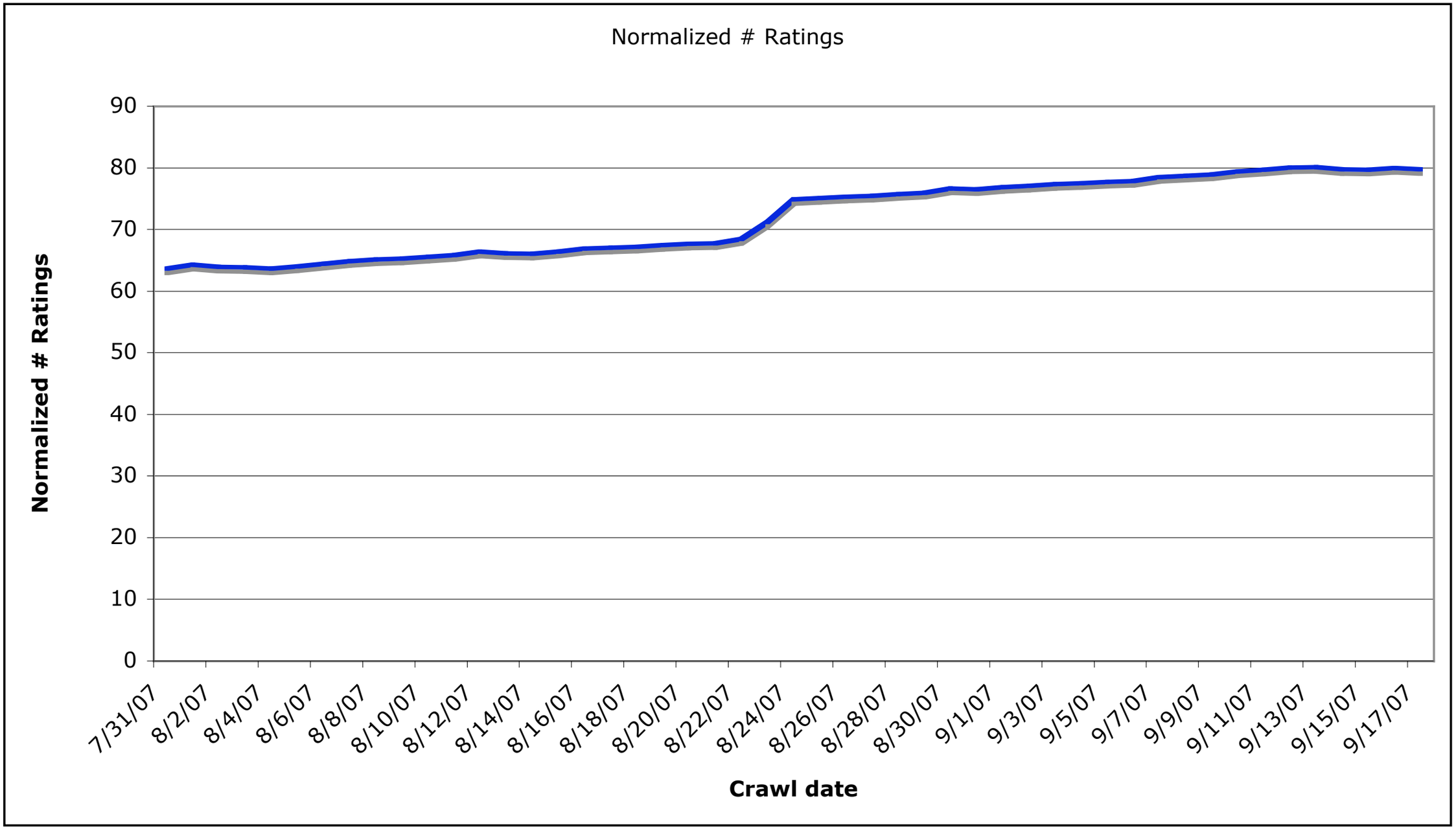
Detecting events from the collection



Detecting events from the collection



Detecting events from the collection

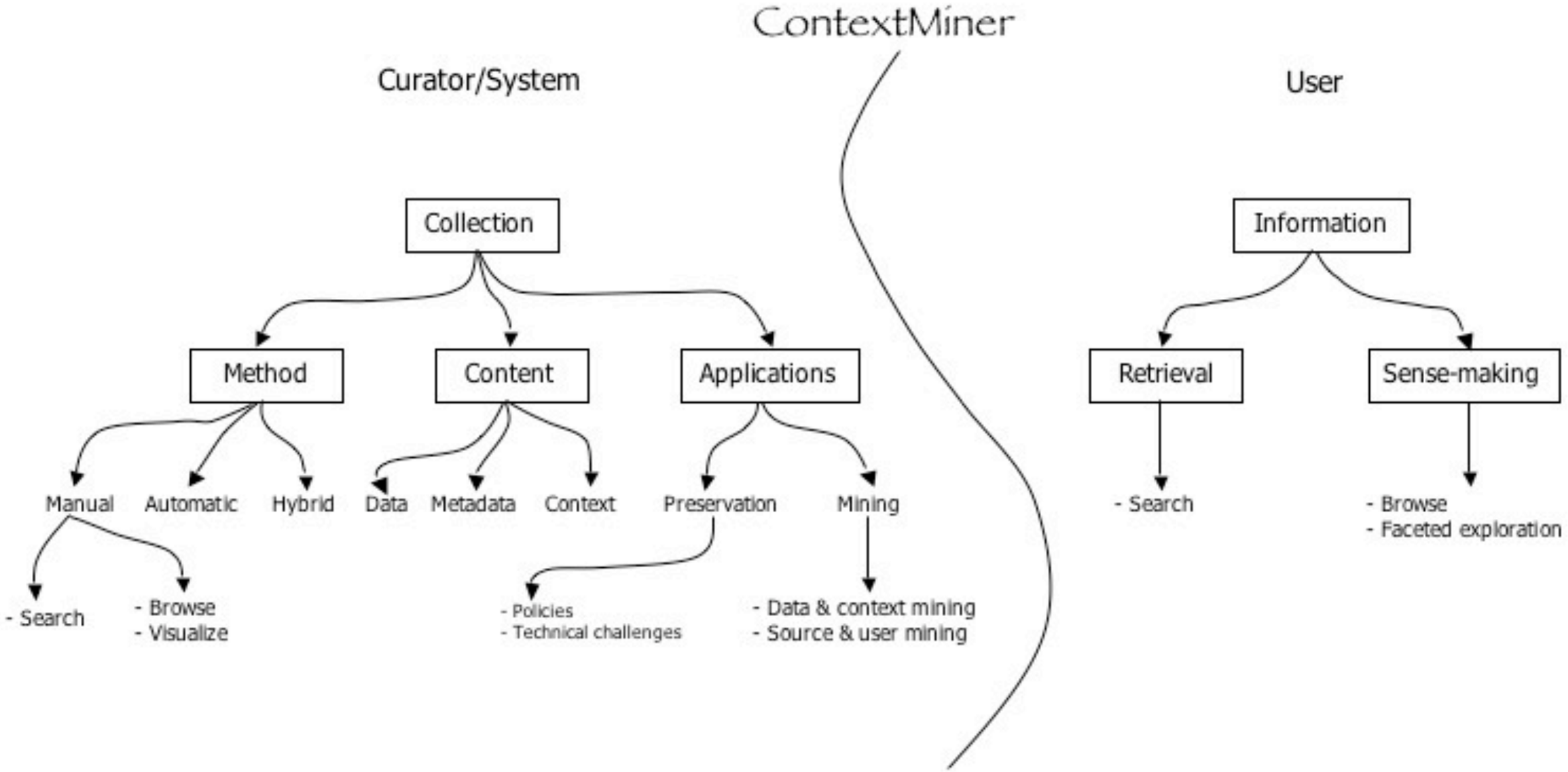


Interesting issues to investigate

- How to detect significant changes or events?
- How faithfully the changes in online activities reflect the real-life events?

- Who is responsible for these significant changes in online world?
- Can we detect opinions and sentiments of people by analyzing their online behavior and participation?

ContextMiner - the big picture



Becoming a part of this story

- Access the *ContextMiner* website
- Access the data using Developer APIs
- Build your own crawler using **TubeKit**
  A YouTube Crawling Toolkit

Ideas, ideas, ideas...

Idea: Create a relevance / retrieval function using title, descr, and the model for the video

Relevance ∝ Text similarity (title, descr, tags)
∝ Access
∝ participation } popularity

We don't have click through data

- Text matching
- Popular
- Active participation

} How to evaluate?

Idea: Measure "query performance". Watch a video's rank changing and compare it with its popularity and ~~partic~~ participation values.

What's the point?

Moral of the story

Research areas

Policy issues
Technical challenges
Identifying and capturing context
Collection visualization
Understanding online user behavior and participation
Event detection
User interface
Retrieval performance

Tools

ContextMiner
DiscoverInfo
DIToolkit
ContextMiner APIs
TubeKit
FEX

Websites

Author's homepage: http://www.unc.edu/~chirags
VidArch homepage: http://www.ils.unc.edu/vidarch/
ContextMiner: http://idl63.ils.unc.edu/chirag/ContextMiner/
DiscoverInfo: http://idl.ils.unc.edu/~chirag/DiscoverInfo/
DIToolkit: http://idl.ils.unc.edu/~chirag/DIToolkit
ContextMiner APIs: http://idl63.ils.unc.edu/chirag/ContextMiner/developer.php
TubeKit: http://idl.ils.unc.edu/~chirag/TubeKit/

Thanks!


WAP WAP WAP
CALVIN! WHAT ARE YOU DOING TO THE COFFEE TABLE?!?
IS THIS SOME SORT OF TRICK QUESTION, OR WHAT?