Understanding Users

Jussi Kangasharju

University of Helsinki

Joint work with Lasse Nordgren, Jesse Lankila,

and Ossi Karkulahti

User-Generated Content (UGC)

Popular sites: Flickr, YouTube, Wikipedia, etc.

Significant fraction of Internet content

How is UCG different from commercial content?

In particular: Any difference for the network?

What can we learn about users?

Goals of Our Study

Get an idea of UCG on popular sites

Understand how user activity shows in network

 How to make network adapt better?

How is the real world visible on the net?

 Can we predict future needs?

Data Collection Methodologies

Monitor RSS-feeds from selected sources

- Real RSS feeds
- Wikipedia edits
- Discussion forum posts
- Twitter

Use API provided by system

- Flickr
- YouTube
- Twitter

Amount of Data Collected

Site	Data collected since
BBC	March 2009
Helsingin Sanomat	March 2009
Flyertalk	April 2009
Wikipedia (7 different languages)	May/August 2009
Flickr	July 2009
YouTube	July 2009
Twitter	August 2009

Commercial vs. Users

Is there a difference between commercially generated content and user-generated content?

Difference, such as when is content created?

Comparison: Two news sites and several Wikipedias

- BBC and Helsingin sanomat
- Wikipedia FI, SE, DK, NO, KR, AR, Simple
- On Wikipedia, only edits count

BBC (top)
HS (bottom)

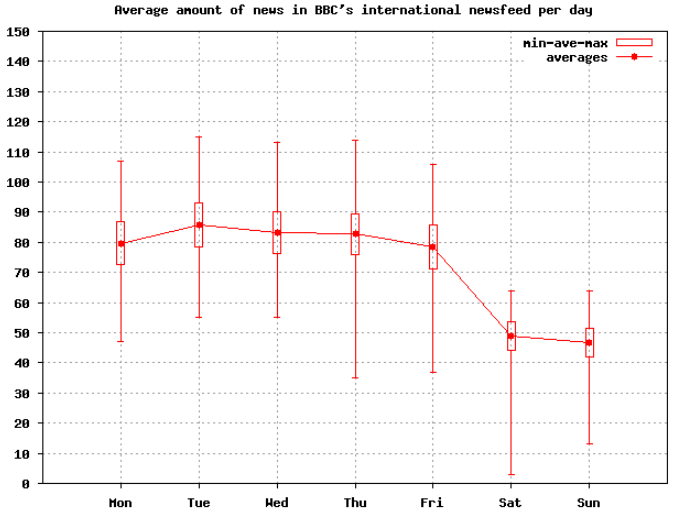
Left weekly activity

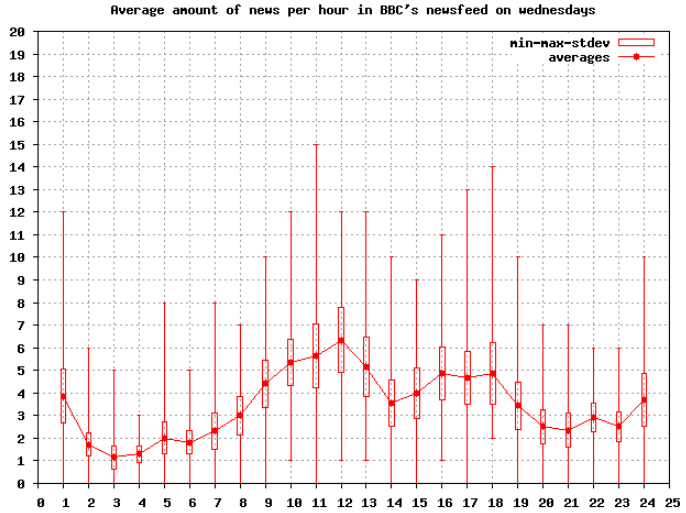
Right Wednesdays

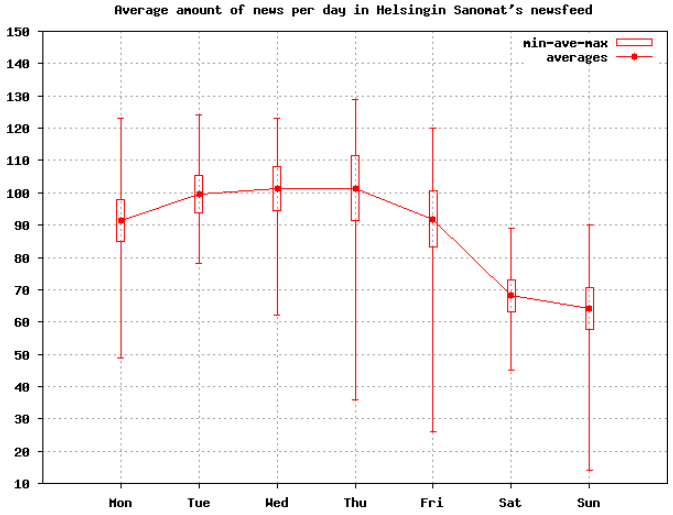
Both averaged over whole data set

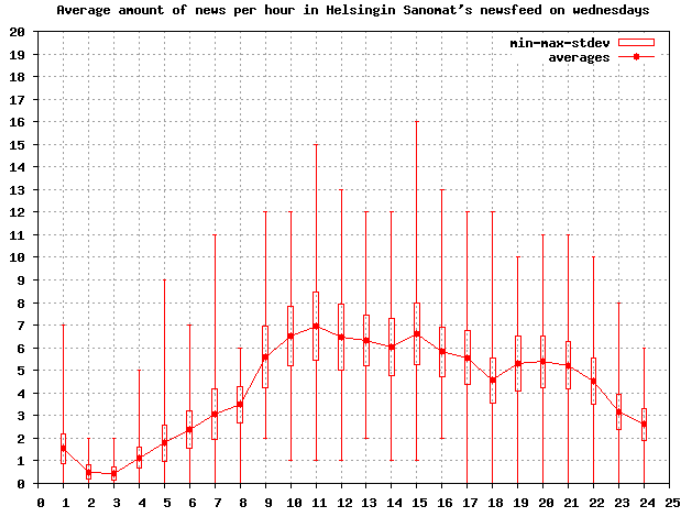
No major difference between weekdays

Weekends slightly different

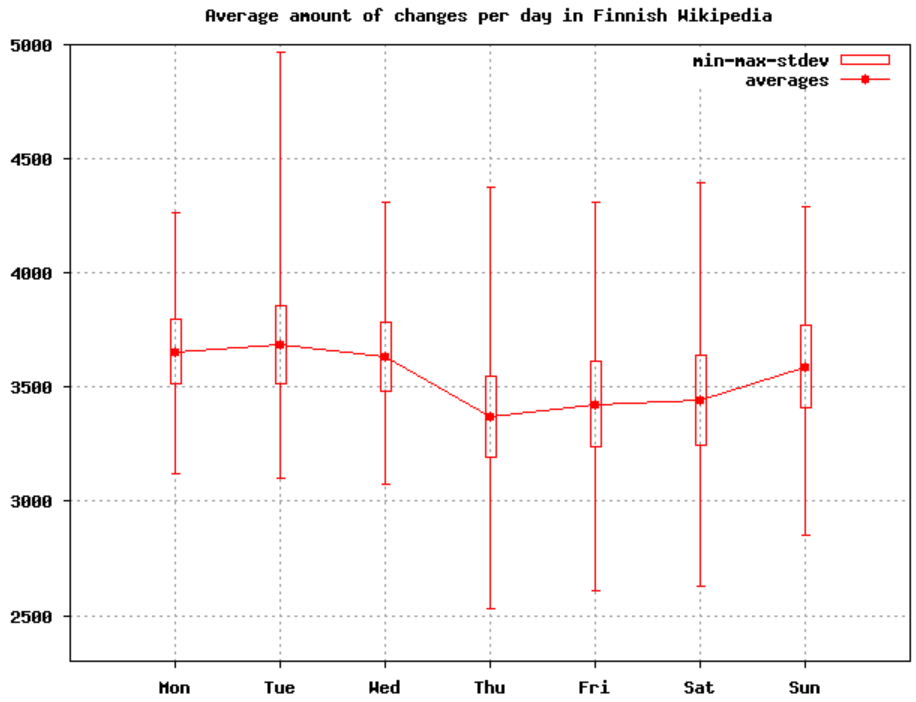




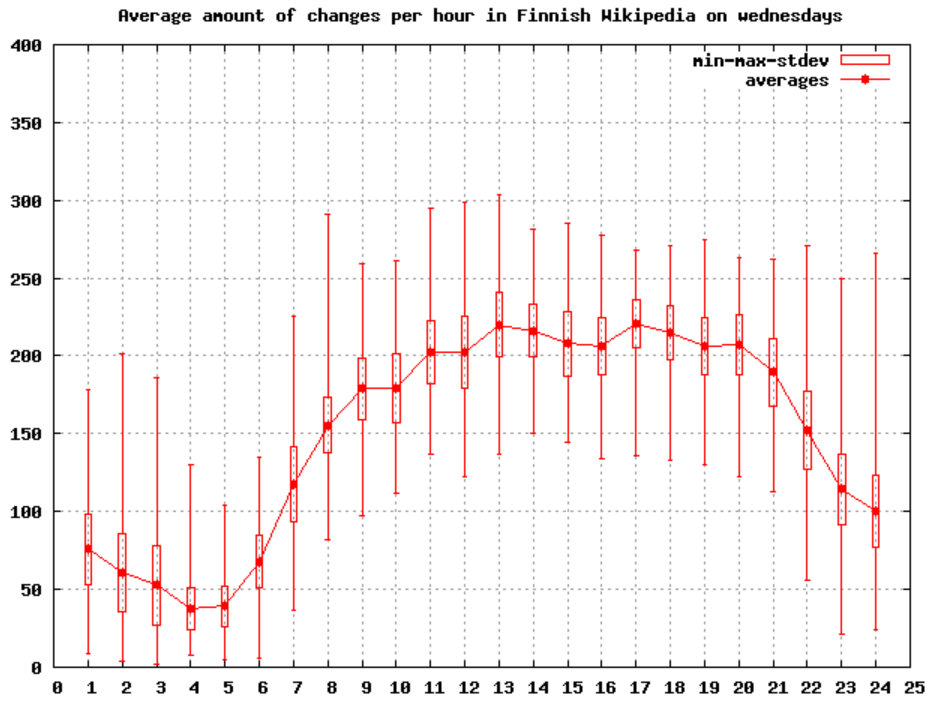




Wikipedia Edits: Finland



Weekly activity



Wednesday activity

Wikipedia: Cultural Differences

Are there differences between different cultures?

Selected for study:

- Sweden, Norway, Denmark (similar to Finland?)
- Arabic and Korean (larger spread?)
- Simple English (purely a hobby?)

Sweden (top)
Norway (middle)
Denmark (bottom)

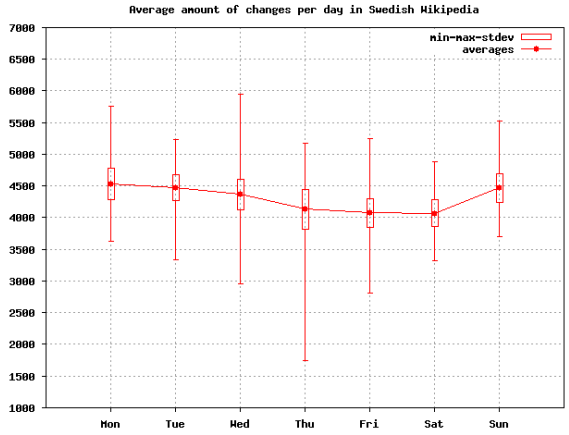
Note: Y-axis not same

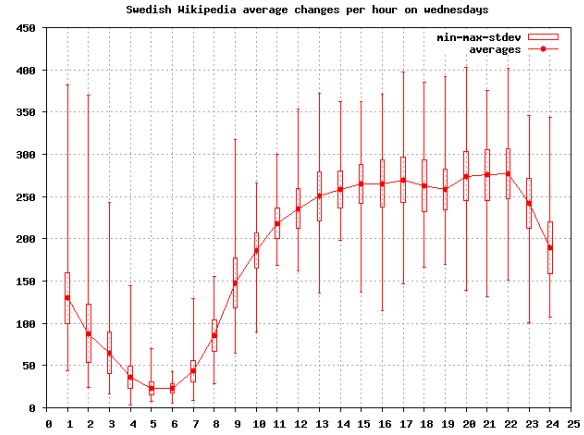
Very similar to Finland

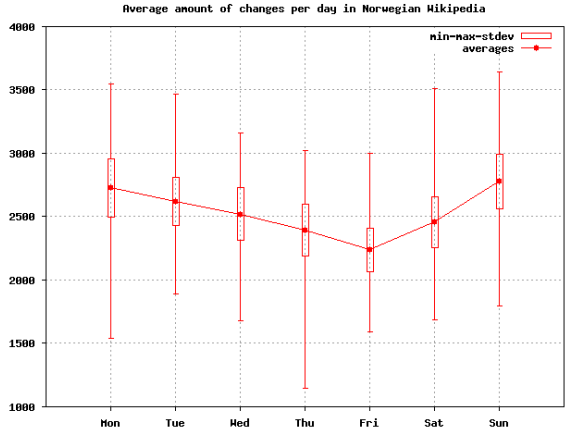
Larger weekly variation in Norway

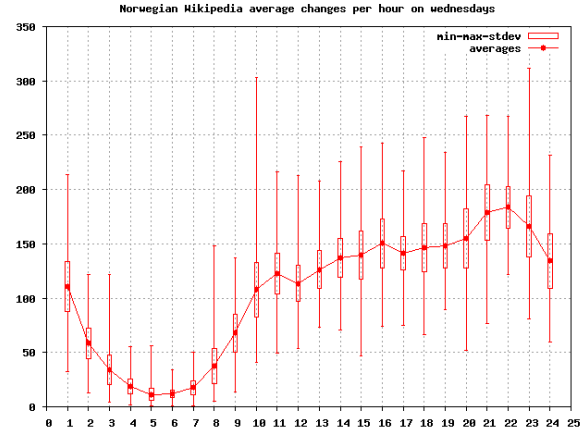
Sweden mostly constant

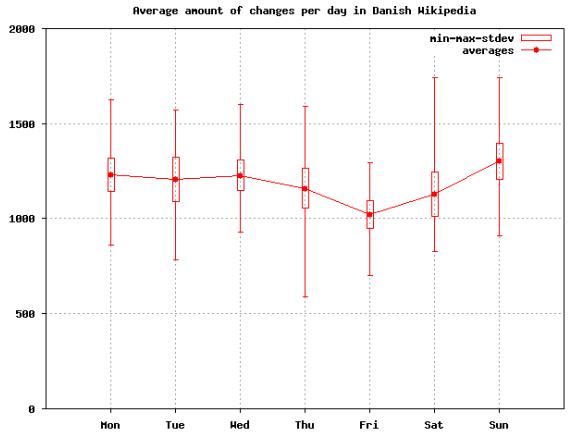
Edits during day and evening

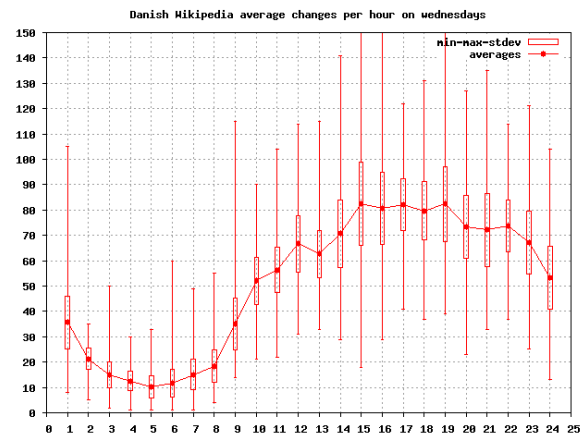












Kangasharju: U

www.h

Arabic (top)
Korea (bottom)

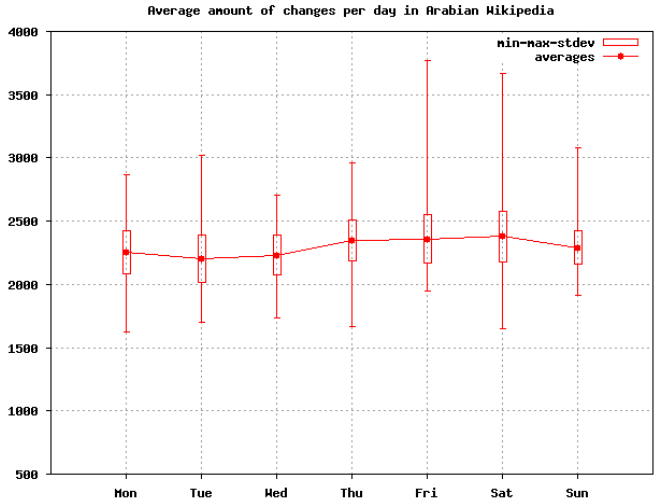
Arabic = UTC
Korea = KT

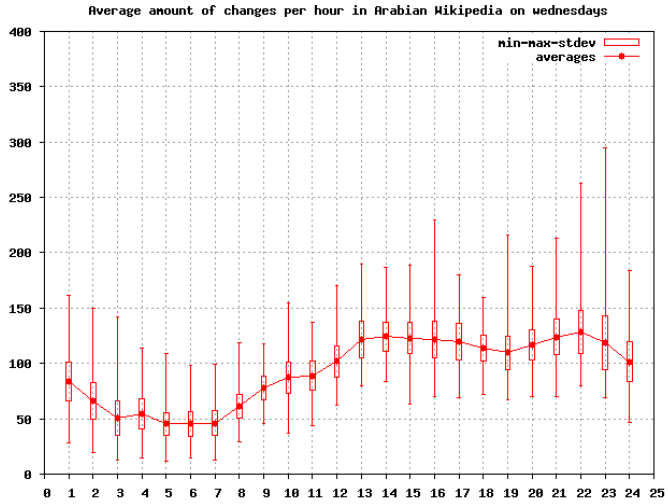
Weekly activity almost constant in both cases

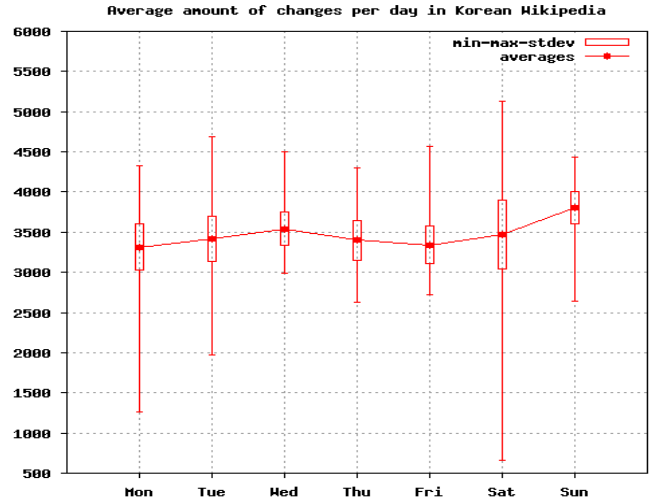
Less daily variation in Arabic, as expected

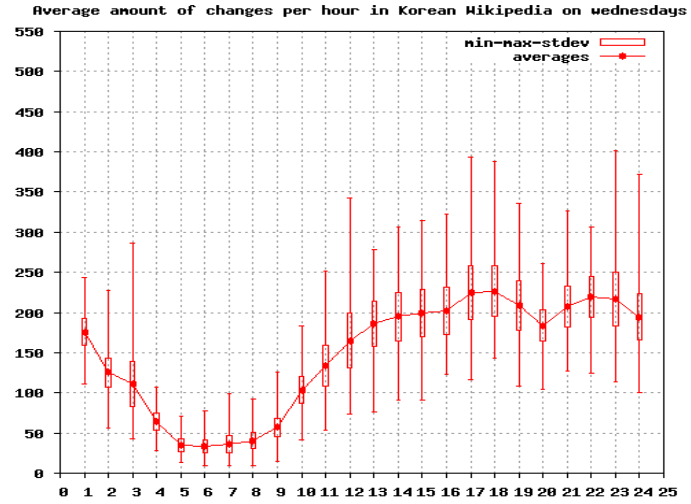
Strange dip in Korean at 8pm

Dip is visible every day, also weekends!

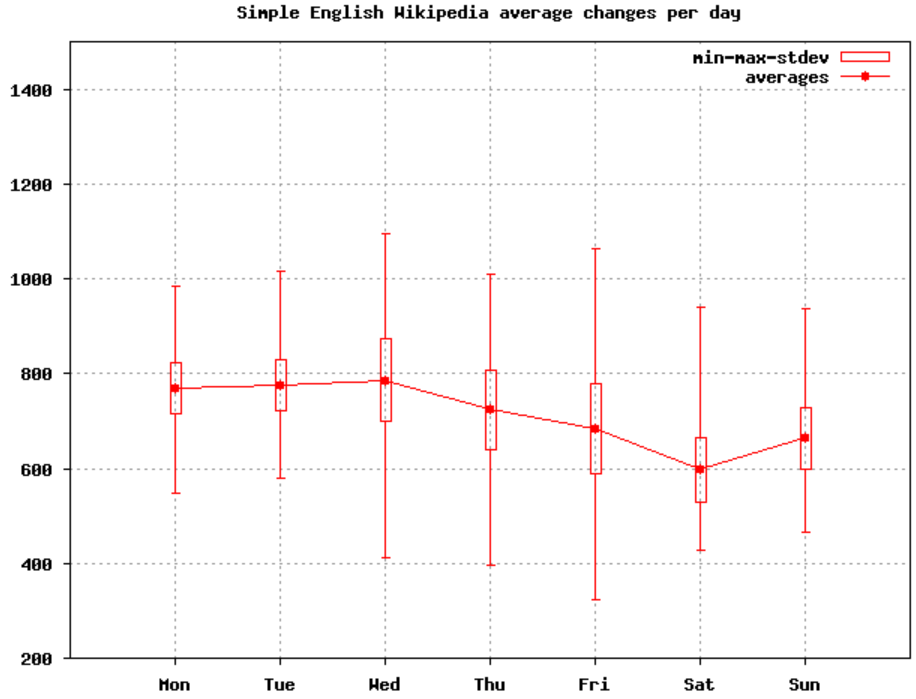


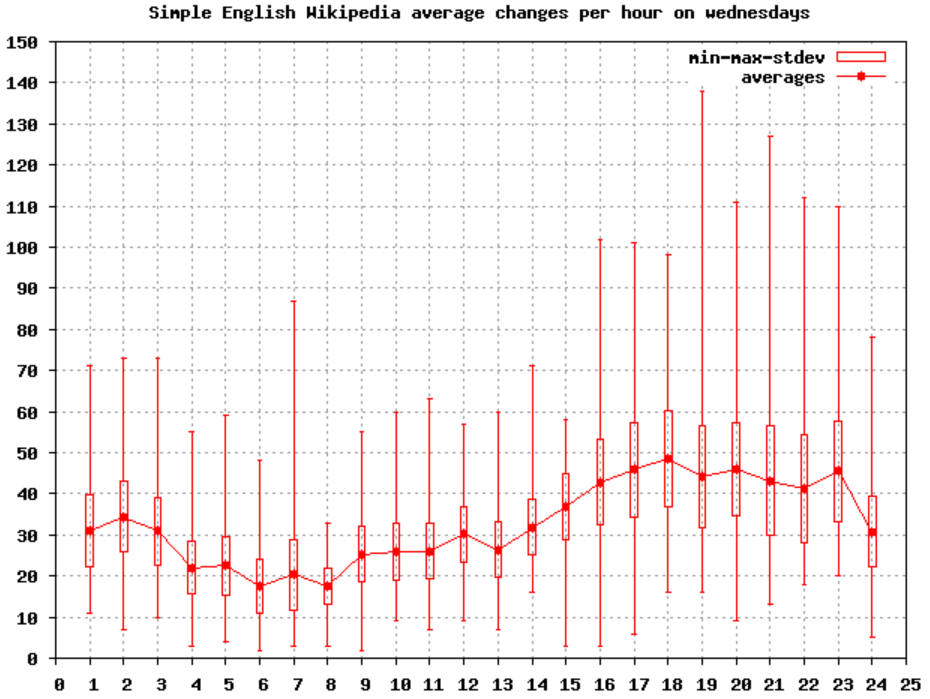






Wikipedia: Simple English





Activity spreads over day more evenly

Weekly activity focused on early working week

Activity mostly from the US? (evening in UTC)

Measuring Flickr & YouTube

Both provide a well-documented API

- Practical limit of 1 query/second

Two methods of data collection

1. Collect a larger set as a snapshot
2. Follow a small set of photos over time

Flickr: Large Set

Set of 129,000 photos, all 30 days old

- View count power-law distributed
- Same applies to favoriting and comments

What else can we find out about users?

- Tags
- Titles and descriptions
- Geographical data

Flickr: Large Set, Numbers

39% of photos have no views at all

- Similar number holds generally for all sets we saw
- View = Someone else clicks to see large photo
- Thumbnails and views by owner do not count

	Title	Description	Tags
Fraction	73%	20%	39%

Geodata: 16 levels of accuracy

- 2434 photos at highest level, other levels very rare (except no data)

Flickr: Large Set, Tags

Almost 80k unique tags in 55k photos with tags

Lot of photos have more than 1 tag

About 8000 photos have more than 10 tags

Most popular tags on right

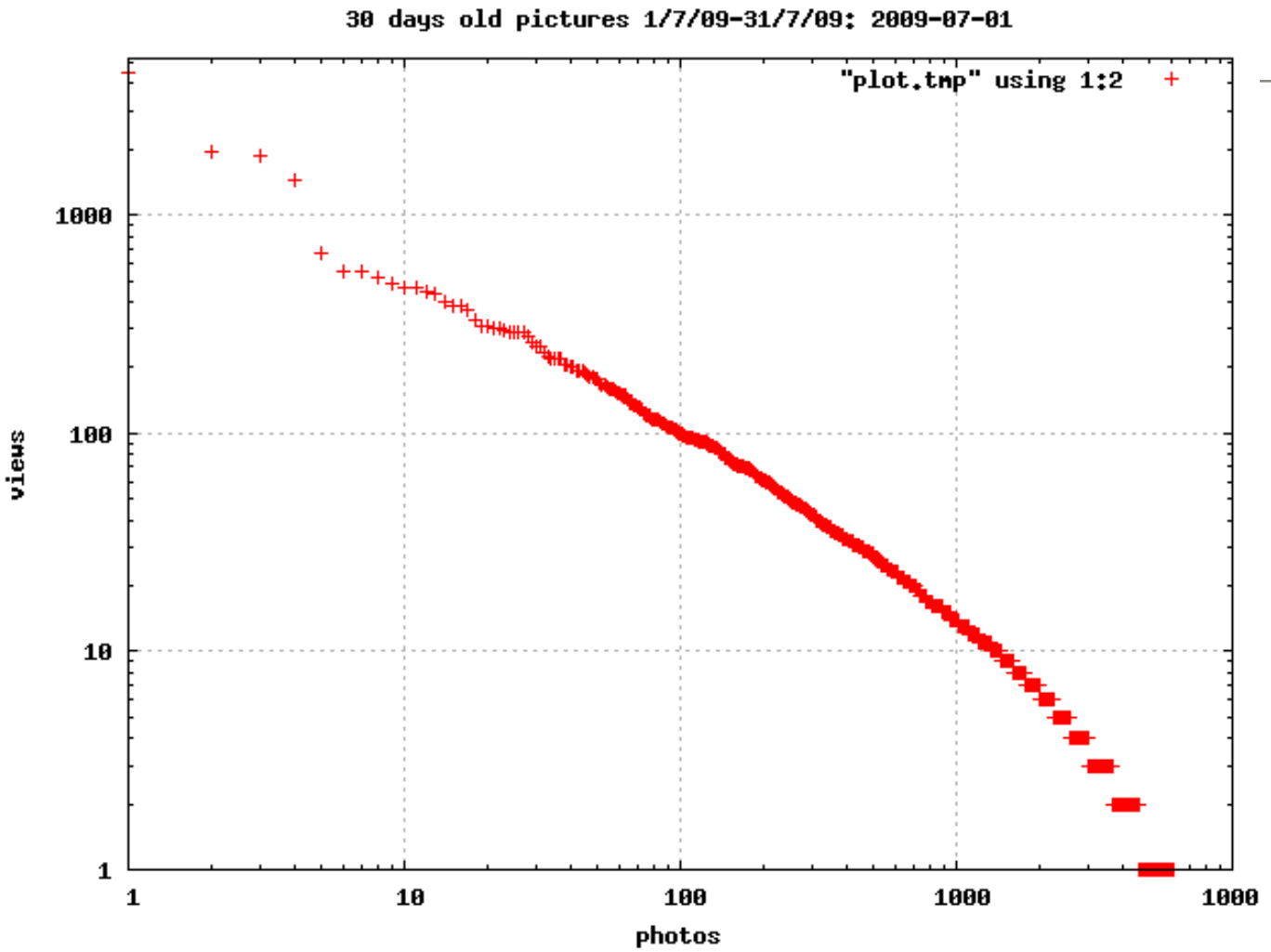
1. 2009 4105
2. nikon 862
3. canon 729
4. wedding 672
5. nature 661
6. london 660
7. usa 588
8. music 588
9. art 580

Flickr: Following Photos

We followed 10,000 photos over 30 days

	1 day	15 days	30 days
Photos left	9985	-255	-102
No views	4251	-866	-276
Has Title	7156	+137	+50
Has Description	1854	+38	+0
Has Tags	3536	+41	+15
Geodata/16	160	+18	+2
Unique tags	6165	+356	+71

Flickr: View Count Evolution



YouTube

Expected to be similar to Flickr in most respects

Extensive studies of YouTube already exist

Our goal: Compare to Flickr, look at some new stuff

View on YouTube = User watches until the end?

YouTube: Basic Stats

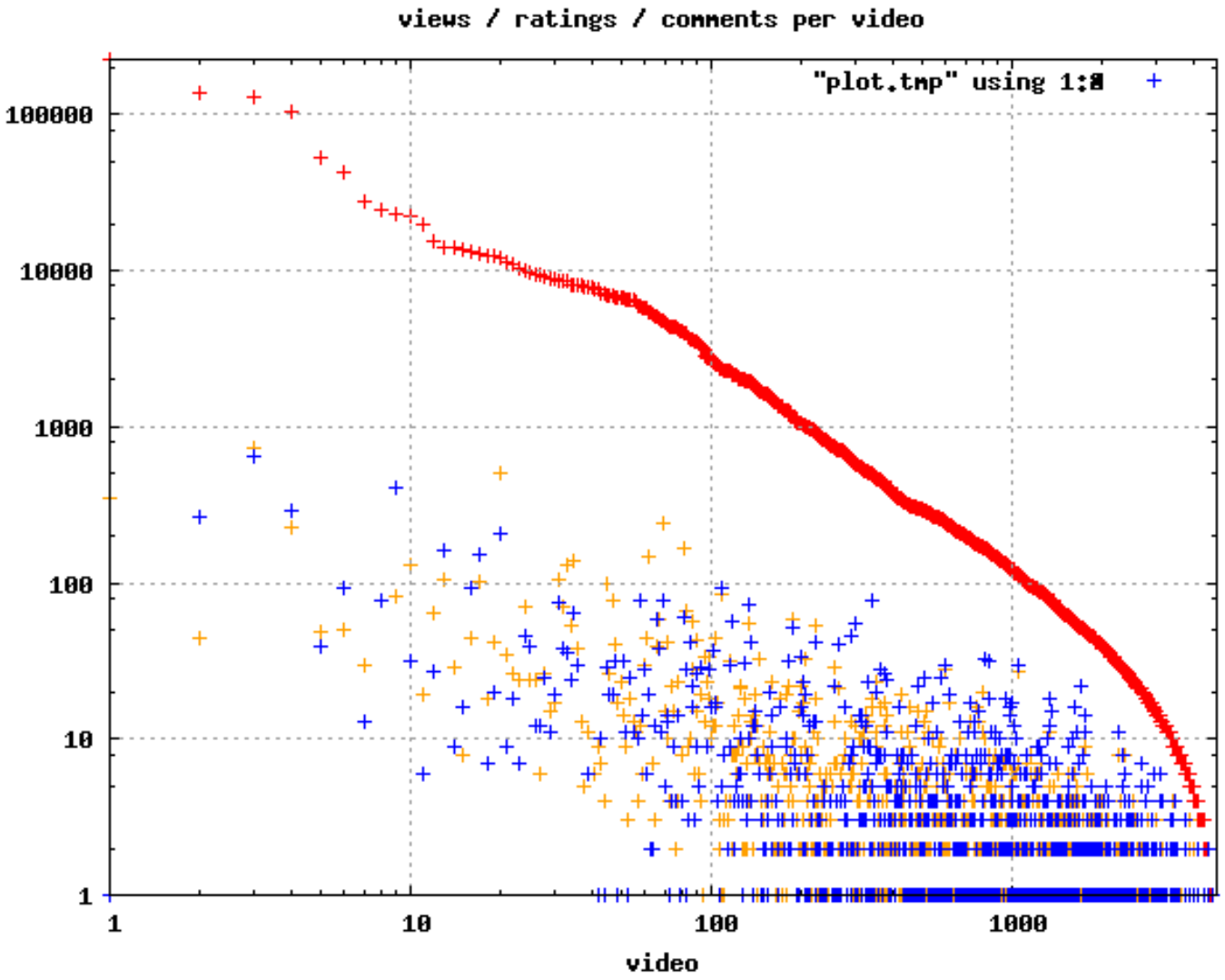
Views (red)
Ratings (orange)
Comments (blue)

Power-law rules as expected/known

Almost all videos have been viewed

Rather strong correlation between ratings and comments

Both correlated with views as well



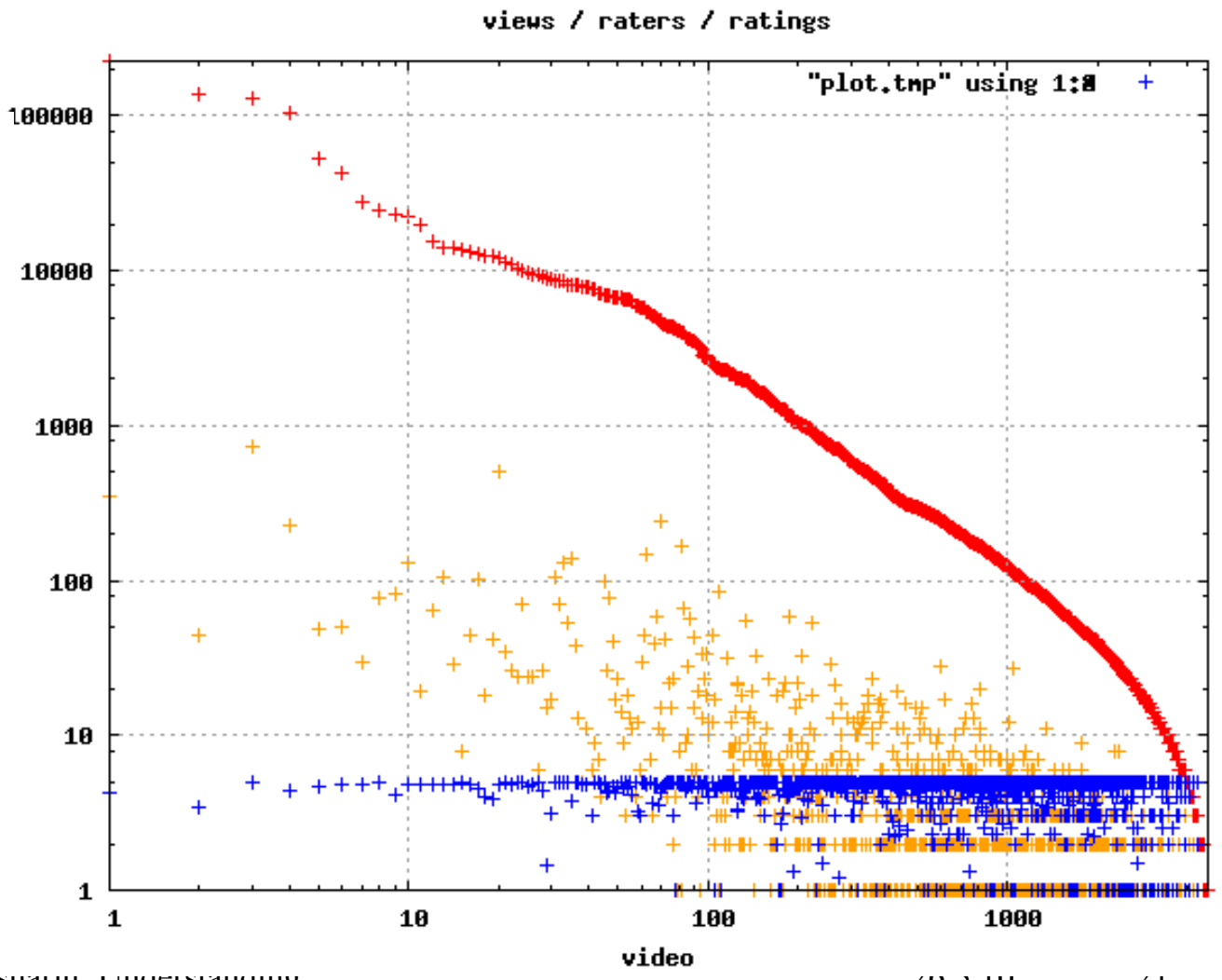
YouTube: Ratings in More Detail

Views (red)
# of ratings (orange)
Rating value (blue)

Rating is mostly 5

Independent of views or number of ratings

Actual rating of a video not a useful metric for much anything



YouTube: Following Videos

We followed 6000 videos over 30 days

	1 day	15 days	30 days
Videos left	5859	-767	-295
No views	455	-305	-28
No rating	4818	-725	-285
Highest views	>20000	>100000	>100000

Almost 20% of videos have disappeared in 30 days

Almost all videos have been viewed at least once

View counts grow fastest in first two weeks

YouTube: Comments

How is the language in YouTube comments?

We analyzed words in comments after 30 days

Removed most common words of English

Surprisingly “clean”…

1. like 390
2. are 312
3. your 286
4. lol 283
5. good 269
6. love 268
7. no 253
8. be 248
9. u 242
10. que 237

Lessons Learned

Wikipedia is edited during the working day and early evening

Wikipedia editing behavior independent of culture

Photos and videos consumed differently by users

Future Work

Continue collecting data

Improve existing workload models

Better distribution architectures

World Cup

See Ossi's talk for more measurement work on Twitter

Thank You!

Email: Jussi.Kangasharju@cs.helsinki.fi

URL: http://www.cs.helsinki.fi/u/jakangas/