Statistical analysis of the social network & discussion threads in Slashdot

Vicenç Gómez Andreas Kaltenbrunner Vicente López

Barcelona Media Innovation Center (BM)
Barcelona, Spain

Department of Information and Comunication Technologies (DTIC)
Pompeu Fabra University (UPF), Barcelona, Spain

Outline

Motivation

Analyze social interaction in form of discussions

- **Message boards** are an excellent source of information.
- **Slashdot** is the most prominent example.

We study

- The social network generated by the discussions.
- The structure of these discussions.

Goals

- Find relevant patterns using statistical methods.
- Gain understanding on this type of social interaction.
- Derive useful metrics to rank and describe discussions.

Slashdot

A tech-news website (1997)

Post:

BitTorrent Use Up 24% Since November

Posted by Soulskill on Friday April 18, @02:13AM
from the **not-what-they-wanted-to-hear** dept.

dingalig writes

"It looks as though the MPAA's fight against The Pirate Bay and other BitTorrent sites isn't going very well. Ars Technica reports that BitTorrent traffic is up by 24% since before the holidays. 'BitTorrent traffic spiked over the December holidays. After a peaking at almost 12.5 million downloaders on the 200 most popular files, traffic dropped at the beginning of January — about the time that school started up again. But one figure that will prove alarming to the content creation industry is that the numbers are higher now than they used to be. "The baseline has been elevated," notes [BigChampagne CEO Eric] Garland. "Not only did the spike happen, but the bar was raised."'"

▶ internet, media, p2p, bittorrent *(tagging beta)*

Comments:

Victimless (Score:5, Insightful)
by Repton (60818) on Friday April 18, @02:26AM (#23114548) Homepage

"We need to highlight that [copyright infringement] is not a victimless crime and take appropriate actions."

Anyone know any victims? Artists or creators whose works are widely pirated but who struggle to make a living?

Reply to This

1 hidden comment

Re:Victimless (Score:5, Insightful)
by mollymoo (202721) * on Friday April 18, @03:27AM (#23114794) Journal

You seem to be implying that depriving someone of something doesn't make them a victim as long as it doesn't leave them struggling to survive. Which is of course complete and utter bullshit.

Reply to This Parent

3 hidden comments

Re:Victimless (Score:5, Insightful)
by evanbd (210358) on Friday April 18, @05:17AM (#23115174)

Ideas are owned by society. They are what make up our culture. Sometimes we, as a society, have seen fit to let their creator exercise some limited degree of control over them. That does not mean any one person can own an idea any more than they can own a sunset.

Reply to This Parent

Re:Victimless (Score:5, Insightful)
by Plutonite (999141) on Friday April 18, @04:20AM (#23114994)

Or maybe he's implying, correctly, that sharing digital information for free does not deprive anyone of anything, let alone make them penniless.

Reply to This Parent

- Users can comment to posts.
- Posts trigger easily hundreds of comments.
- Distributed moderation system.

Dataset [Aug '05, Aug '06]

$\sim 10^4$ news posts.
$\sim 2 \cdot 10^6$ comments.
$\sim 10^5$ different users.

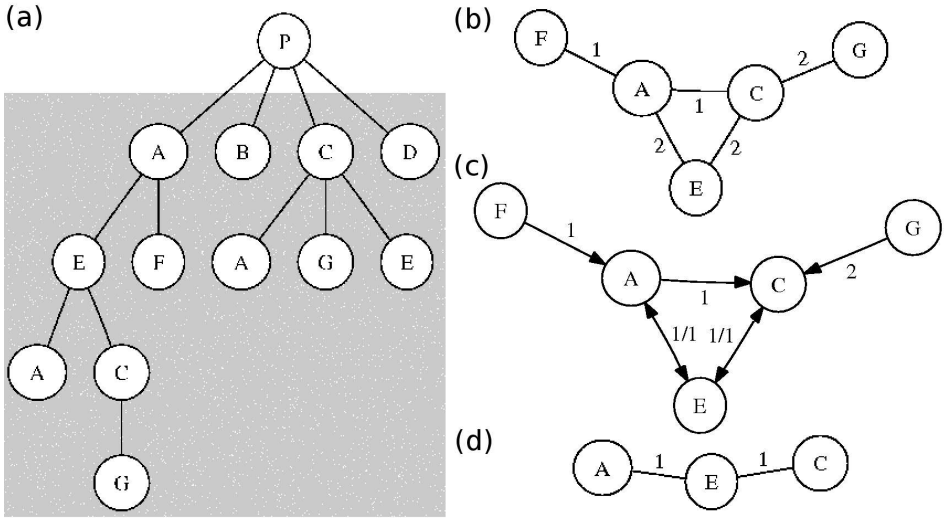
We consider:

- Id message
- type (post/comment)
- autor
- time
- score of a comment $\in [-1, 5]$
- nesting level of a comment

The social network of Slashdot

Network construction

- Users are connected according to their **posting activity**:



- Three interpretations of a link between two users:
  - (b) Undirected dense
  - (c) Directed
  - (d) Undirected sparse

Results in three **weighted** networks amenable to analyze.

The social network of Slashdot

Main Indicators

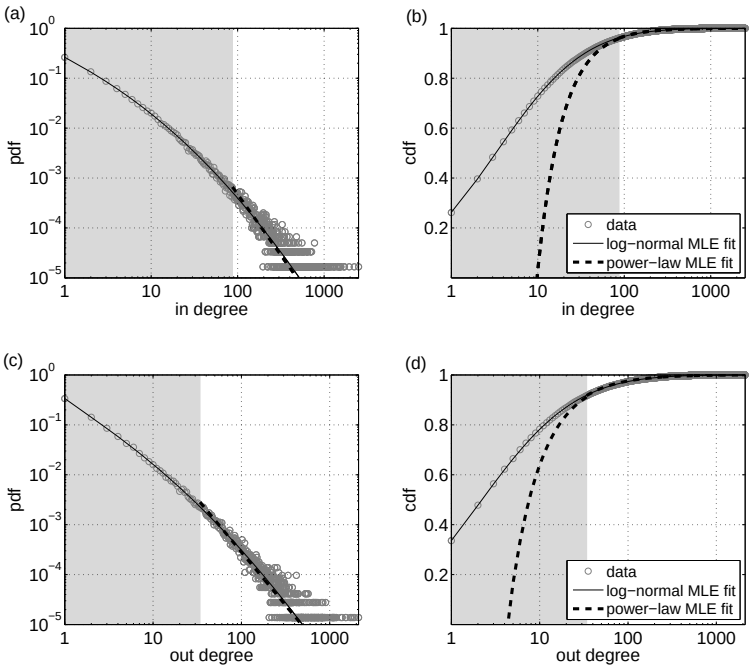
Indicator	Directed	Und.Dense	Und.Sparse
Number of nodes	80,962	80,962	37,087
Number of edges	1,052,395	905,003	294,784
Max.clust.size	73.12%	97.90%	97.15%
Av. degree	13(50.1/49.4)	22.36(79.3)	7.95(25.7)
Av. path length	3.62(0.7)	3.48(0.7)	4.02(0.8)
Av. path length (random)	4.38	3.62	5.05
Diameter	10	9	11
Clustering coef.	0.027(0.075)	0.046(0.12)	0.017(0.078)
Clustering coef. (weighted)	0.026(0.074)	0.047(0.12)	0.018(0.080)
Clustering coef. (random)	$1.67 \cdot 10^{-4}$	$2.88 \cdot 10^{-4}$	$2.27 \cdot 10^{-4}$
Assortativity by degree	−0.016	−0.039	−0.016
Reciprocity	0.28	–	–

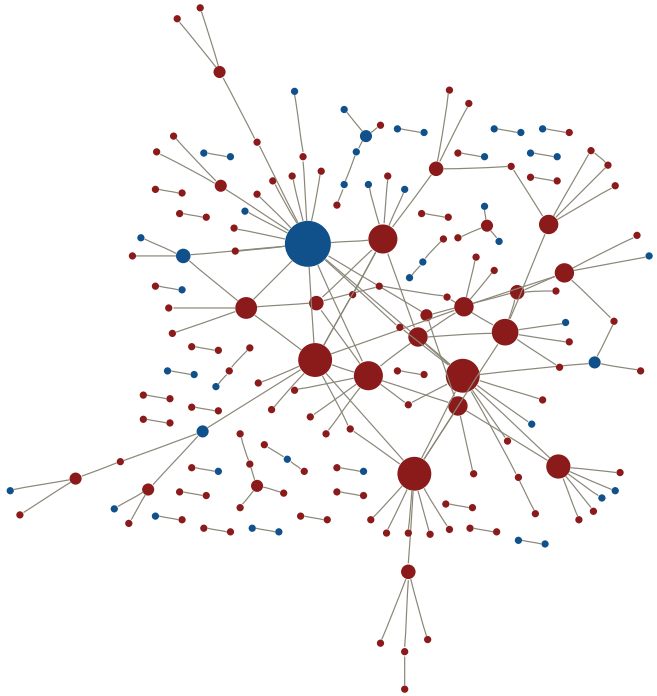
Comparison with traditional social networks

- **Similarities**: Giant component, small-world network, ...
- **Discrepancies**: Neutral assortativity, moderated reciprocity.

The social network of Slashdot

Degree Distributions





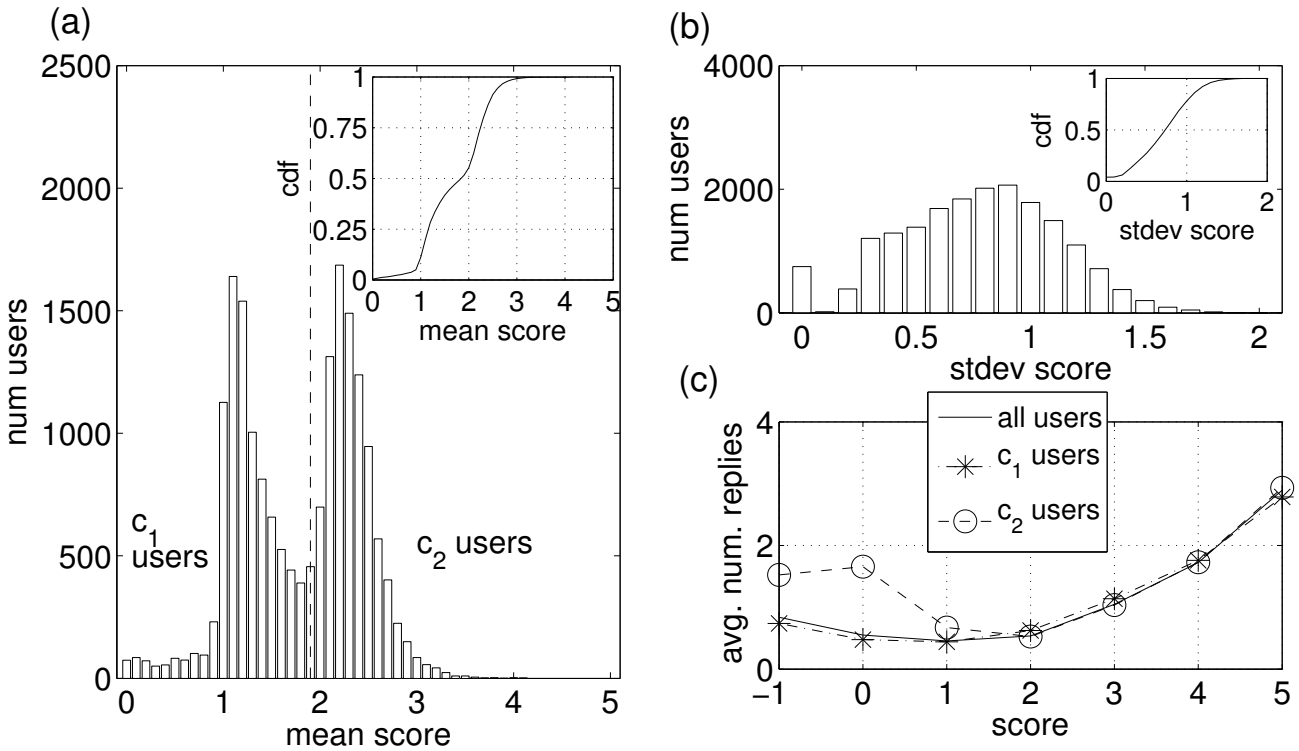
Statistical analysis (Maximum Likelihood & KS test)

- Rejects the Power-law hypothesis.
- A (truncated) log-normal fits the entire dataset.
- Similar In- and out-degree distributions.

The social network of Slashdot

Mixing patterns by score

- Users can be characterized by the mean score of their comments.
- 2 classes of users: **good** and **regular** commentators.
- Number of received comments correlates with the score.
- **Neutral mixing** by mean score, but c_2 users receive more replies for low-scored comments than c_1 users $\Rightarrow$ reputation $\sim$ score.



The social network of Slashdot

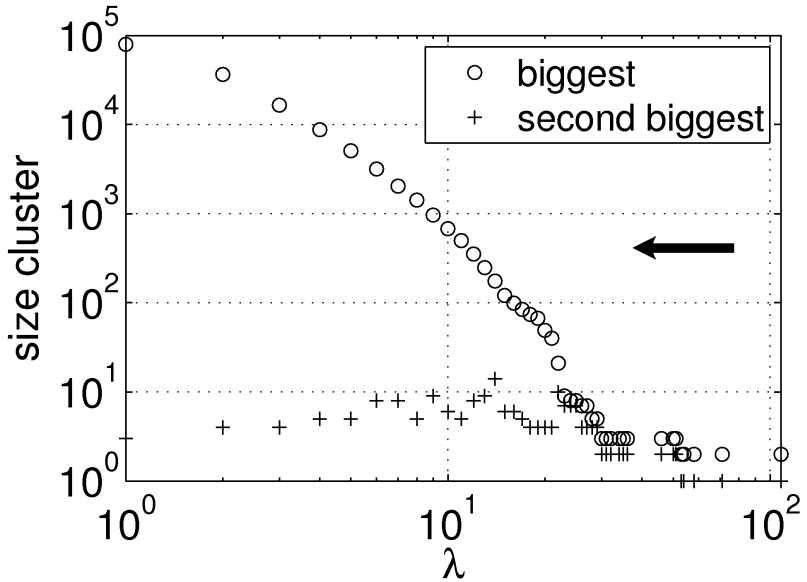
Community structure

Agglomerative clustering (dendrogram).
Only pairs i, j of users with weight $w_{ij} > \lambda$ are included.

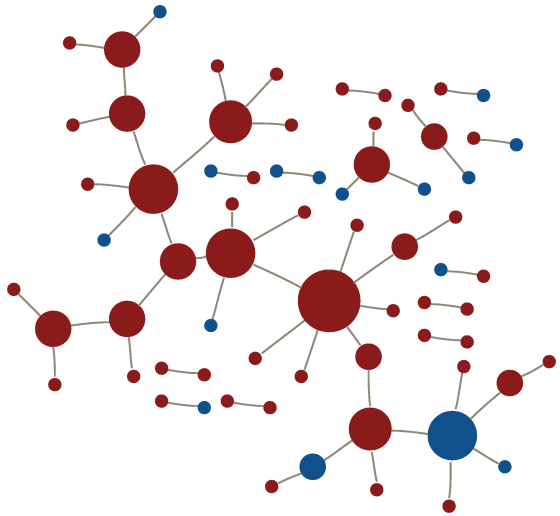
Result

- One giant component present in all scales.

Backbone is composed mainly of good writers.



$\lambda = 20$



The social network of Slashdot

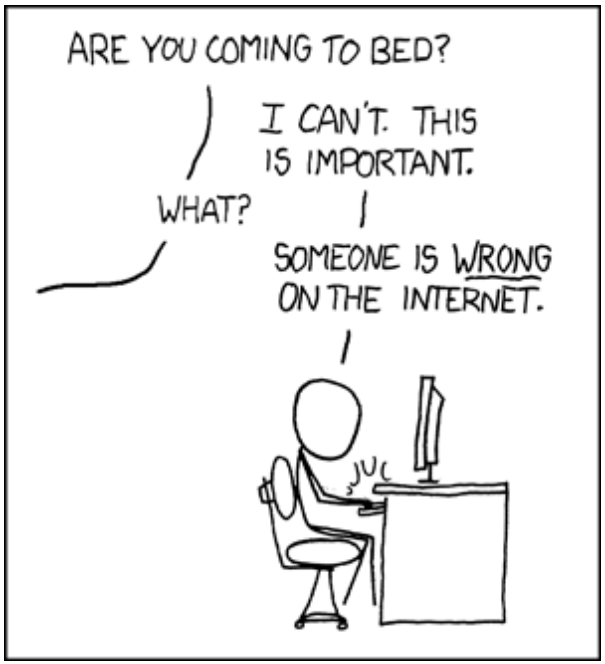
- Absence of a complex community structure.
- A small set of strongly connected users exist.
- First link occurs easily...

What induces a user to comment?

The social network of Slashdot

- Absence of a complex community structure.
- A small set of strongly connected users exist.
- First link occurs easily...

What induces a user to comment?



Taken from http://xkcd.com/386

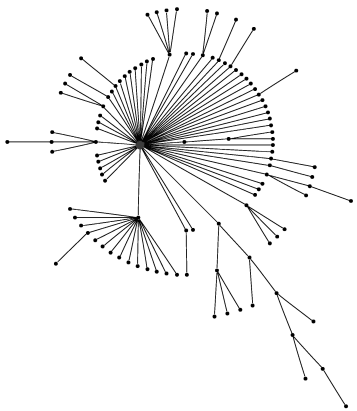
Discussion threads

Radial tree representation

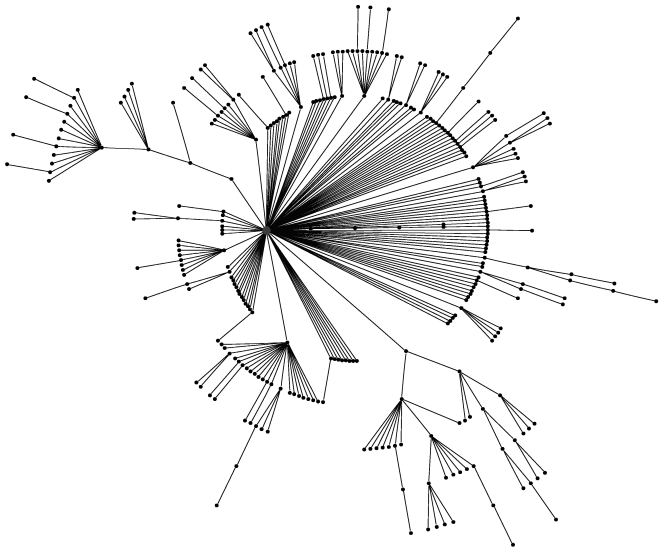
- Discussion threads have a radial tree structure.
- What are their statistical properties?

Example of evolution of a controversial post:

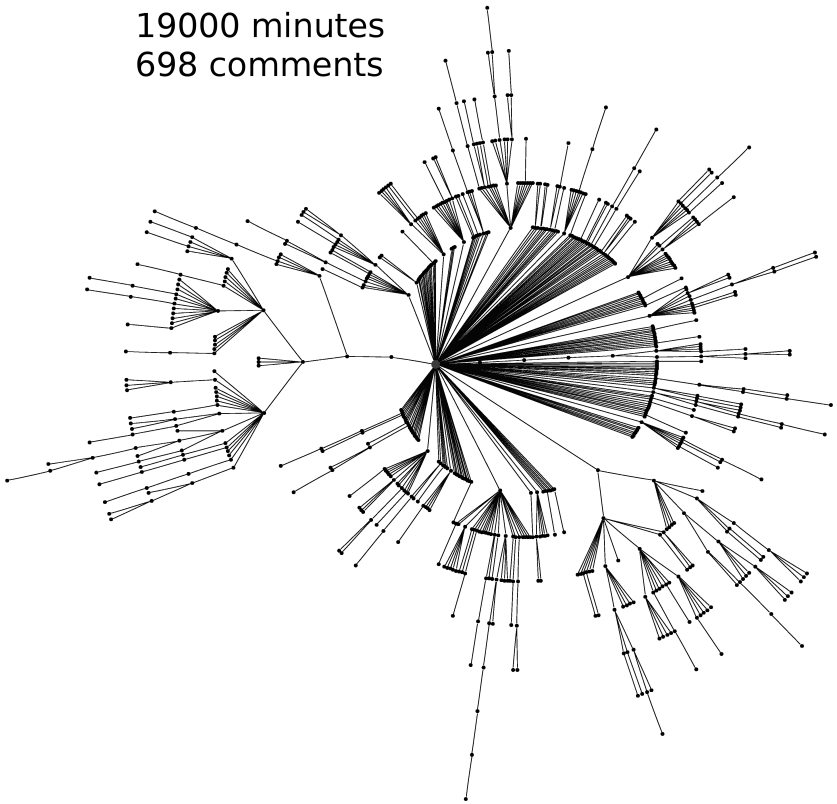
2000 minutes
109 comments



5000 minutes
314 comments



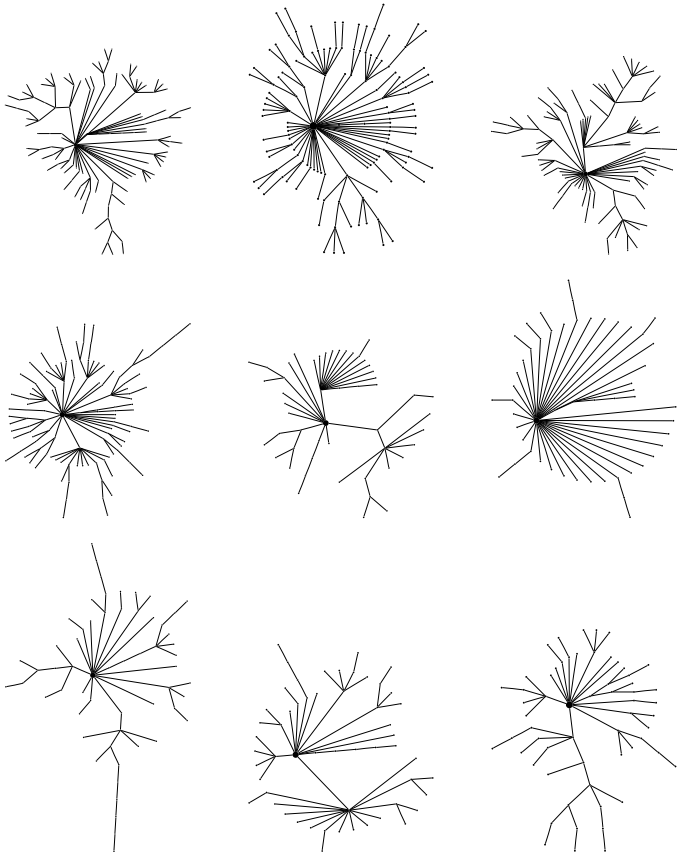
19000 minutes
698 comments

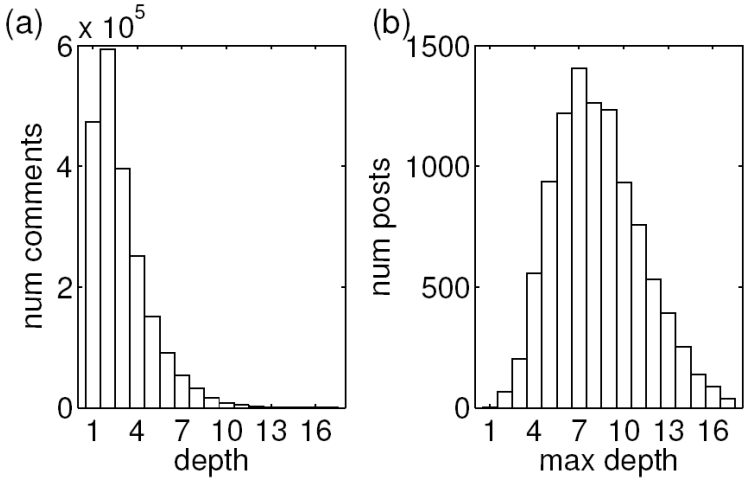


The discussion threads

Global characterization

Heterogeneity in radial trees:





(a) Distribution of comments throughout nesting levels.

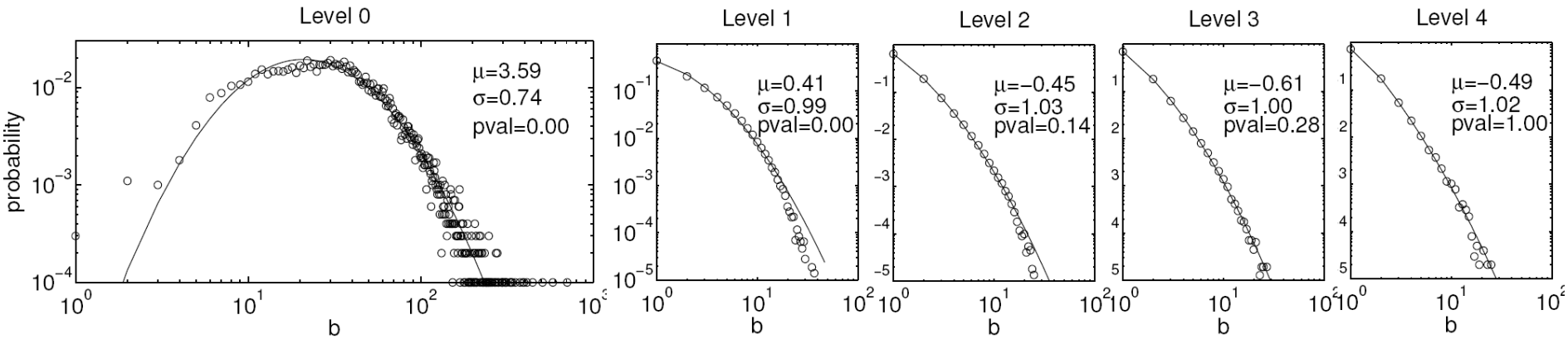
(b) Distribution of threads per maximum depth.

The discussion threads

Probability distribution of branching factors

Branching factors

- For each level: Distribution of number of replies.
- Direct answers to the post differ from comments to comments.
- Nesting levels $\Rightarrow$ **Depth-invariant** mechanism.



The discussion threads

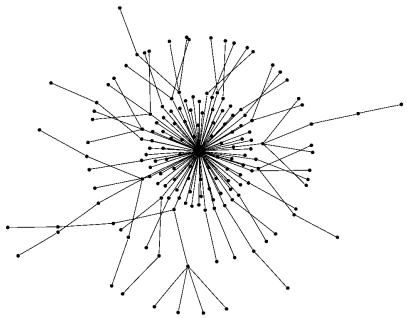
Measuring controversy

How can we measure controversy of a post?

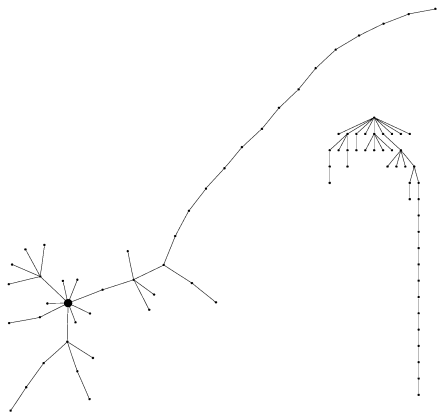
- Keep in mind that controversy is *subjective*.
- A simple and efficient procedure.
- Based on structural properties of the radial tree.

Number of comments or **maximum depth** are not enough:

- A thread can receive many messages but short discussions



- 2 users can increase the depth without general interest



The discussion threads

The h-index as a measure of scientific production

We propose a measure based on the **h-index**.

- Measures scientific impact of a researcher [Hirsch '05].

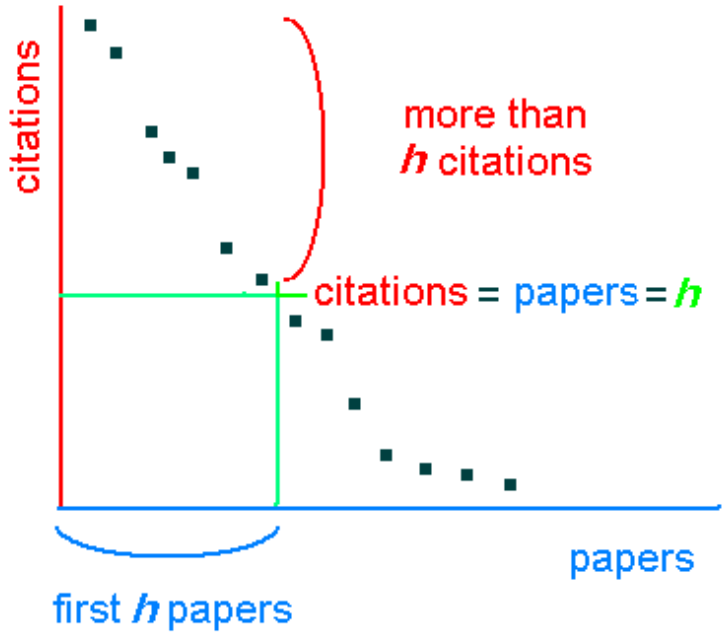


Figure taken from wikipedia.org

- **Maximum** rank-number for which the number of citations is **greater or equal** to the rank-number.

The discussion threads

The h-index as a measure of controversy

We propose an adapted version of the h-index

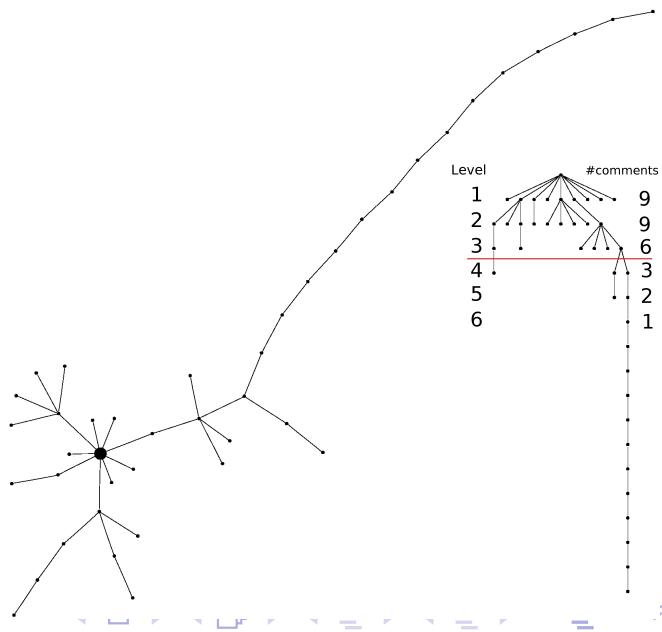
- The h-index of a post is h if $h+1$ is the first nesting level i which has less than i comments.
- Choose the thread with *less* comments to break ties.

The **controversy rank** of post i is:

$$\text{h-index}_i + \frac{1}{\text{num comments}_i}.$$

Example

Controversy is $3 + \frac{1}{41}$ $\Longrightarrow$



The discussion threads

The h-index as a measure of controversy

We propose an adapted version of the h-index

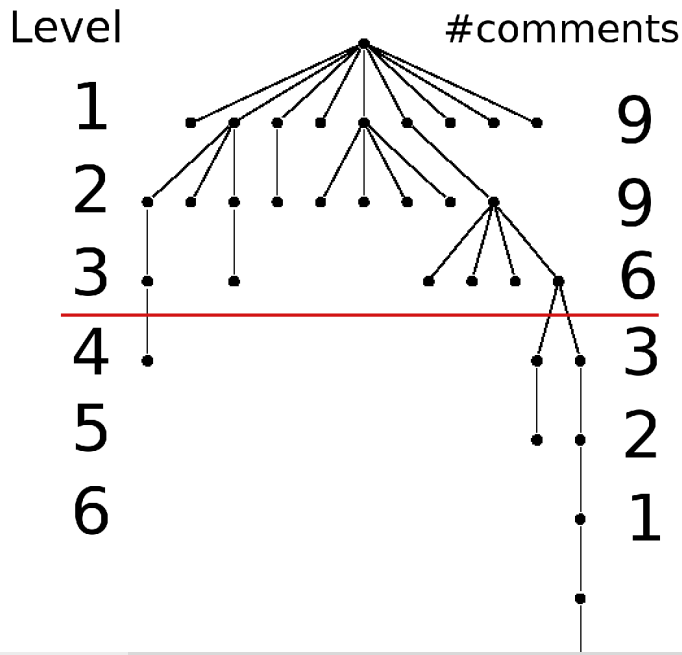
- The h-index of a post is h if $h+1$ is the first nesting level i which has less than i comments.
- Choose the thread with *less* comments to break ties.

The **controversy rank** of post i is:

$$\text{h-index}_i + \frac{1}{\text{num comments}_i}.$$

Example

Controversy is $3 + \frac{1}{41}$ $\Longrightarrow$



The discussion threads

The h-index as a measure of controversy

We propose an adapted version of the h-index

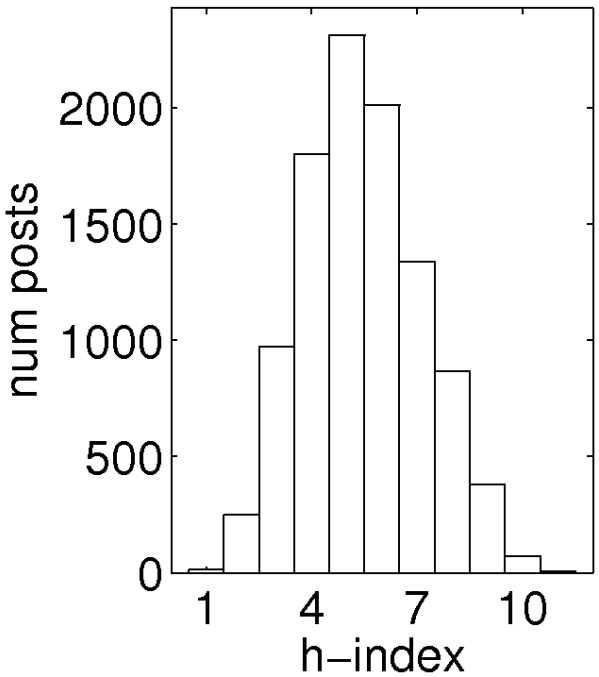
- The h-index of a post is h if $h+1$ is the first nesting level i which has less than i comments.
- Choose the thread with *less* comments to break ties.

The **controversy rank** of post i is:

$$\text{h-index}_i + \frac{1}{\text{num comments}_i}.$$

Example

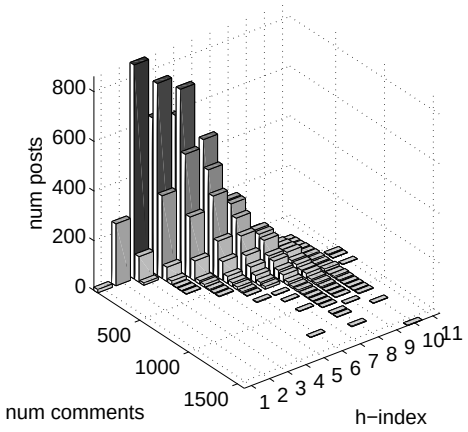
Controversy is $3 + \frac{1}{41}$ $\Longrightarrow$

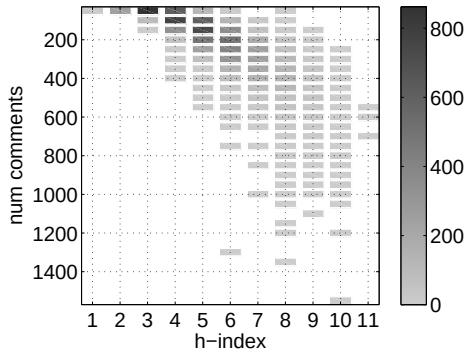


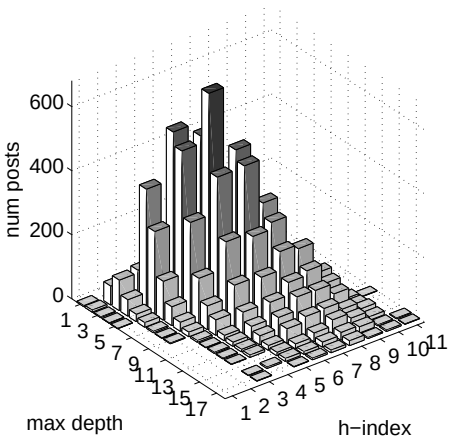
The discussion threads

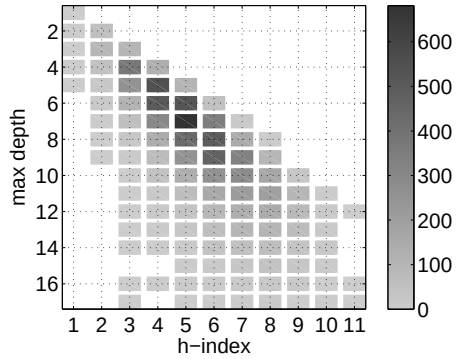
The h-index as a measure of controversy

- Relations with number of comments and maximum depth:









Global features of our proposed measure

- Considers total number of comments and the replies.
- A simple measure (efficient).
- The h-index is robust and monotonic (never decreases).

The discussion threads

The h-index as a measure of controversy

#	H	Num cmnts (#)		Depth (#)		Title
1	11	527	(401)	16	(113)	Violating A Patent As Moral Choice
2	11	529	(390)	12	(1374)	Human Genes Still Evolving
3	11	605	(208)	16	(120)	Powell Aide Says Case for War a 'Hoax'
4	11	693	(96)	17	(34)	US Releasing 9/11 Flight 77 Pentagon Crash Tape
5	10	243	(3287)	15	(159)	Apple Fires Five Employees for Downloading Leopard
6	10	288	(2431)	14	(356)	Linus Speaks Out On GPLv3
7	10	290	(2409)	11	(1774)	New Mammal Species Found in Borneo
8	10	309	(2078)	13	(698)	Biofuel Production to Cause Water Shortages?
9	10	315	(1999)	12	(1168)	Torvalds on the Microkernel Debate
10	10	355	(1511)	17	(17)	Well I'll Be A Monkey's Uncle
11	10	361	(1446)	13	(747)	Windows Vista Delayed Again
12	10	366	(1394)	14	(416)	NSA Had Domestic Call Monitoring Before 9/11?
13	10	367	(1379)	11	(1922)	Unleashing the Power of the Cell Broadband Engine
14	10	380	(1279)	12	(1238)	Making Ice Without Electricity
15	10	384	(1243)	14	(424)	Evidence of the Missing Link Found?

Table: Top-15 controversial posts according to our proposed measure and corresponding positions according to the number of comments and maximum depth rankings.

Conclusions

Conclusions

- Similarities and discrepancies between traditional social networks.
- Weak evidence of reputation influencing the connectivity.
- Depth invariant mechanism generating discussion threads.
- Simple and efficient measure to asses the controversy of a post.

Future work

- Compare results with other websites.
- Build a model to understand the process generating the discussions.
- Empirical evaluation of the validity of the h-index.
- Study temporal evolution of the h-index.

References

A. Kaltenbrunner, V. Gómez, V. López.
Description and Prediction of Slashdot Activity
In *Proceedings of the 5th Latin American Web Congress (LA-WEB 2007)*.

J. E. Hirsch.
An index to quantify an individual's scientific research output.
Proc. Natl. Acad. Sci. USA, 102(46):16569–16572, 2005.