Modeling the structure and evolution of online discussion cascades

Vicenç Gómez[1] Hilbert J Kappen[1] Nelly Litvak[2]
Andreas Kaltenbrunner[3]

[1] Donders Institute for Brain, Cognition and Behaviour,
Radboud University, Nijmegen, The Netherlands

[2] Faculty of Electrical Engineering, Mathematics and Computer Sciences,
University of Twente, Enschede, The Netherlands

[3] Social Media Research Group, Barcelona Media,
Barcelona, Spain

ETC* July 26th, 2012

Outline

1. Introduction
   - Institution
   - Motivation
   - Datasets
2. Likelihood-based framework
   - Model definition
   - Parameter estimation
   - Validation
3. Conclusions

Agenda

Structure and evolution of online discussion cascades

Gómez V., Kappen H. J., Litvak N., and Kaltenbrunner, A. (2012).
A likelihood-based framework for the analysis of discussion threads.
World Wide Web Journal, 2012, pp 1-31.

Gómez V., Kappen H. J., and Kaltenbrunner, A. (2011).
Modelling the Structure and Evolution of Discussion Cascades.
In *HT2011 22nd ACM Conference on Hypertext and Hypermedia*, , Eindhoven, The Netherlands.

But first ...

- a brief presentation of Fundació Barcelona Media.

Outline

1. Introduction
   - Institution
   - Motivation
   - Datasets
2. Likelihood-based framework
   - Model definition
   - Parameter estimation
   - Validation
3. Conclusions

Barcelona Media

What is Barcelona Media?

- A private non-profit organisation for the research and the innovation in the communication industry.
- A Technology Centre collaborating with companies and institutions to foster the competitiveness of the sector.
- Close relations with Universitat Pompeu Fabra.
- Participation in 41 European projects (12 as a coordinator).

Research Lines

- Audio
- Image
- Voice and Language
- Perception and Cognition
- Information retrieval (Yahoo! Labs Barcelona)
- Social Media

The Social Media Research Group at Barcelona Media

Mission

- **Combine qualitative and quantitative methods** to generate knowledge about the interplay of social behaviour and “Social Media”.

Main research lines

- Survey Research
- Sentiment analysis
- Social Network Analysis
- Study of online conversation

Main data sources

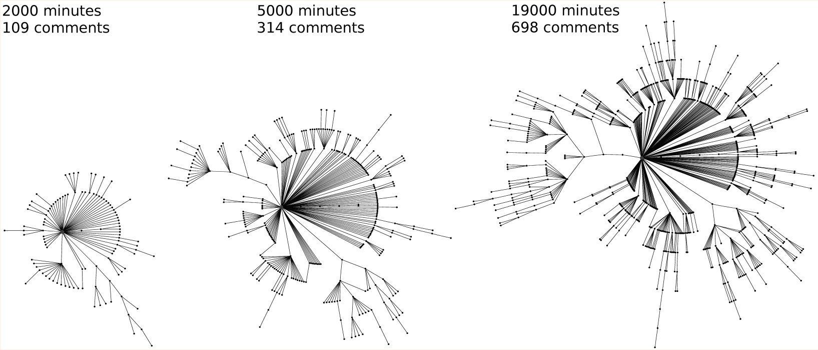
- Twitter
- Wikipedia
- Online Forums
- Online Social Networks

Outline

1. Introduction
   - Institution
   - Motivation
   - Datasets
2. Likelihood-based framework
   - Model definition
   - Parameter estimation
   - Validation
3. Conclusions

Motivation

Example of online discussion (from Slashdot)



Title: *"Can Ordinary PC Users Ditch Windows for Linux?*.

- Online conversations as networks: **nodes** correspond to comments, **edges** represent a reply action.

Motivation - Online discussion threads

Scientific questions

- What are the structural patterns governing these responses?
- What determines the growth of a conversation?
- Is there a generative model that captures their statistical properties?
- Can we use the model parameters to characterize websites, user behaviour, discussions?

Implications / Applications

- Understanding communication in large webspaces that comprise many-to-many interaction.
- Understanding diffusion of news and opinion in social networks.
- Community management, forum design/maintenance, ...

Outline

1. Introduction
   - Institution
   - Motivation
   - Datasets
2. Likelihood-based framework
   - Model definition
   - Parameter estimation
   - Validation
3. Conclusions

Online discussion threads

Datasets

We collected data from the following sources:

Slashdot (SL) : Technological news aggregator.
$473,065$ discussions, $2 \cdot 10^6$ comments, $93 \cdot 10^3$ users

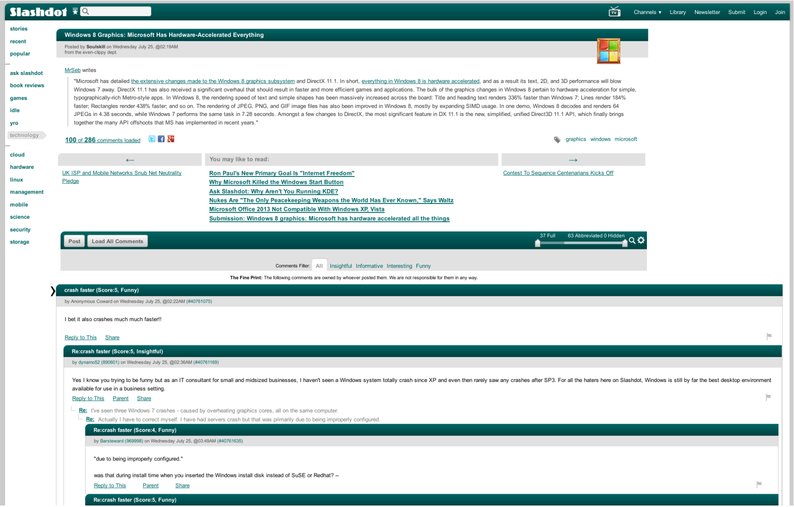
Barrapunto (BP) : Spanish version of Slashdot.
$44,208$ discussions, $4 \cdot 10^5$ comments, $50 \cdot 10^3$ users

Meneame (MN) : Spanish Digg clone (general news aggregator)
$58,613$ discussions, $2.1 \cdot 10^6$ comments, $5,4 \cdot 10^4$ users.

Wikipedia (WK) : discussion pages related to every article.
$871,485$ discussions, $\approx 10^7$ comments, $3.5 \cdot 10^5$ users.

Motivation

Example of discussion in Slashdot (post):



Windows 8 Graphics: Microsoft Has Hardware-Accelerated Everything

Posted by **Soulskill** on Wednesday July 25, @02:19AM
from the even-clippy dept.

MrSeb writes

"Microsoft has detailed the extensive changes made to the Windows 8 graphics subsystem and DirectX 11.1. In short, everything in Windows 8 is hardware accelerated, and as a result its text, 2D, and 3D performance will blow Windows 7 away. DirectX 11.1 has also received a significant overhaul that should result in faster and more efficient games and applications. The bulk of the graphics changes in Windows 8 pertain to hardware acceleration for simple, typographically-rich Metro-style apps. In Windows 8, the rendering speed of text and simple shapes has been massively increased across the board: Title and heading text renders 336% faster than Windows 7; Lines render 184% faster; Rectangles render 438% faster; and so on. The rendering of JPEG, PNG, and GIF image files has also been improved in Windows 8, mostly by expanding SIMD usage. In one demo, Windows 8 decodes and renders 64 JPEGs in 4.38 seconds, while Windows 7 performs the same task in 7.28 seconds. Amongst a few changes to DirectX, the most significant feature in DX 11.1 is the new, simplified, unified Direct3D 11.1 API, which finally brings together the many API offshoots that MS has implemented in recent years."

100 of 286 comments loaded

graphics windows microsoft

← UK ISP and Mobile Networks Snub Net Neutrality Pledge

You may like to read:

- Ron Paul's New Primary Goal Is "Internet Freedom"
- Why Microsoft Killed the Windows Start Button
- Ask Slashdot: Why Aren't You Running KDE?
- Nukes Are "The Only Peacekeeping Weapons the World Has Ever Known," Says Waltz
- Microsoft Office 2013 Not Compatible With Windows XP, Vista
- Submission: Windows 8 graphics: Microsoft has hardware accelerated all the things

→ Contest To Sequence Centenarians Kicks Off

The Fine Print: The following comments are owned by whoever posted them. We are not responsible for them in any way.

crash faster (Score:5, Funny)

by Anonymous Coward on Wednesday July 25, @02:22AM (#40761075)

I bet it also crashes much much faster!!

Reply to This Share

Re:crash faster (Score:5, Insightful)

by dynamo52 (890601) on Wednesday July 25, @02:36AM (#40761169)

Yes I know you trying to be funny but as an IT consultant for small and midsized businesses, I haven't seen a Windows system totally crash since XP and even then rarely saw any crashes after SP3. For all the haters here on Slashdot, Windows is still by far the best desktop environment available for use in a business setting.

Reply to This Parent Share

Re: I've seen three Windows 7 crashes - caused by overheating graphics cores, all on the same computer.

Re: Actually I have to correct myself. I have had servers crash but that was primarily due to being improperly configured.

Re:crash faster (Score:4, Funny)

by Barsteward (969998) on Wednesday July 25, @03:49AM (#40761635)

"due to being improperly configured."

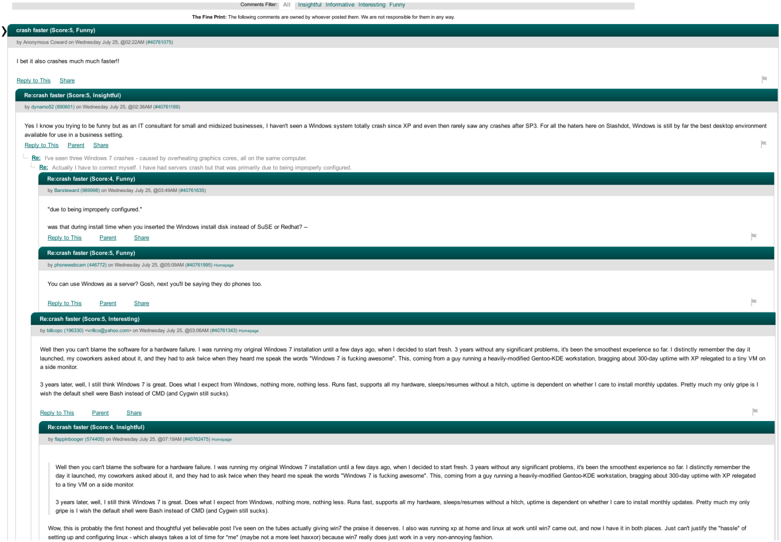
was that during install time when you inserted the Windows install disk instead of SuSE or Redhat? --

Reply to This Parent Share

Re:crash faster (Score:5, Funny)

Motivation

Example of discussion in Slashdot (comments):



Comments Filter: All Insightful Informative Interesting Funny

The Fine Print: The following comments are owned by whoever posted them. We are not responsible for them in any way.

crash faster (Score:5, Funny)

by Anonymous Coward on Wednesday July 25, @02:22AM (#40761075)

I bet it also crashes much much faster!!

Reply to This Share

Re:crash faster (Score:5, Insightful)

by dynamo52 (890601) on Wednesday July 25, @02:36AM (#40761169)

Yes I know you trying to be funny but as an IT consultant for small and midsized businesses, I haven't seen a Windows system totally crash since XP and even then rarely saw any crashes after SP3. For all the haters here on Slashdot, Windows is still by far the best desktop environment available for use in a business setting.

Reply to This Parent Share

Re: I've seen three Windows 7 crashes - caused by overheating graphics cores, all on the same computer.

Re: Actually I have to correct myself. I have had servers crash but that was primarily due to being improperly configured.

Re:crash faster (Score:4, Funny)

by Barsteward (969998) on Wednesday July 25, @03:49AM (#40761635)

"due to being improperly configured."

was that during install time when you inserted the Windows install disk instead of SuSE or Redhat? --

Reply to This Parent Share

Re:crash faster (Score:5, Funny)

by phonewebcam (446772) on Wednesday July 25, @05:09AM (#40761995) Homepage

You can use Windows as a server? Gosh, next you'll be saying they do phones too.

Reply to This Parent Share

Re:crash faster (Score:5, Interesting)

by billcopc (196330) <vrillco@yahoo.com> on Wednesday July 25, @03:06AM (#40761343) Homepage

Well then you can't blame the software for a hardware failure. I was running my original Windows 7 installation until a few days ago, when I decided to start fresh. 3 years without any significant problems, it's been the smoothest experience so far. I distinctly remember the day it launched, my coworkers asked about it, and they had to ask twice when they heard me speak the words "Windows 7 is fucking awesome". This, coming from a guy running a heavily-modified Gentoo-KDE workstation, bragging about 300-day uptime with XP relegated to a tiny VM on a side monitor.

3 years later, well, I still think Windows 7 is great. Does what I expect from Windows, nothing more, nothing less. Runs fast, supports all my hardware, sleeps/resumes without a hitch, uptime is dependent on whether I care to install monthly updates. Pretty much my only gripe is I wish the default shell were Bash instead of CMD (and Cygwin still sucks).

Reply to This Parent Share

Re:crash faster (Score:4, Insightful)

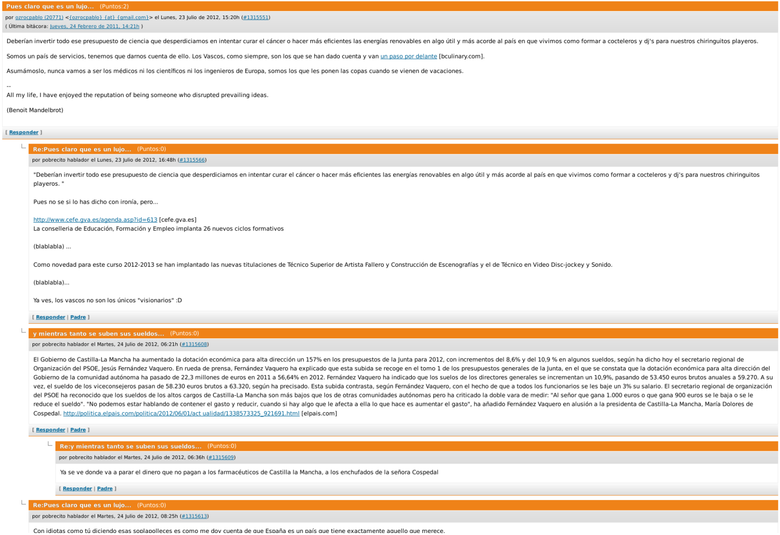
by flappinbooger (574405) on Wednesday July 25, @07:19AM (#40762475) Homepage

> Well then you can't blame the software for a hardware failure. I was running my original Windows 7 installation until a few days ago, when I decided to start fresh. 3 years without any significant problems, it's been the smoothest experience so far. I distinctly remember the day it launched, my coworkers asked about it, and they had to ask twice when they heard me speak the words "Windows 7 is fucking awesome". This, coming from a guy running a heavily-modified Gentoo-KDE workstation, bragging about 300-day uptime with XP relegated to a tiny VM on a side monitor.
>
> 3 years later, well, I still think Windows 7 is great. Does what I expect from Windows, nothing more, nothing less. Runs fast, supports all my hardware, sleeps/resumes without a hitch, uptime is dependent on whether I care to install monthly updates. Pretty much my only gripe is I wish the default shell were Bash instead of CMD (and Cygwin still sucks).

Wow, this is probably the first honest and thoughtful yet believable post I've seen on the tubes actually giving win7 the praise it deserves. I also was running xp at home and linux at work until win7 came out, and now I have it in both places. Just can't justify the "hassle" of setting up and configuring linux - which always takes a lot of time for *me* (maybe not a more leet haxxor) because win7 really does just work in a very non-annoying fashion.

Motivation

Example of discussion in Barrapunto (comments):



Pues claro que es un lujo... (Puntos:2)

por ozrocpablo (20771) <{ozrocpablo} {at} {gmail.com}> el Lunes, 23 Julio de 2012, 15:20h (#1315551)
(Última bitácora: Jueves, 24 Febrero de 2011, 14:21h)

Deberían invertir todo ese presupuesto de ciencia que desperdiciamos en intentar curar el cáncer o hacer más eficientes las energías renovables en algo útil y más acorde al país en que vivimos como formar a cocteleros y dj's para nuestros chiringuitos playeros.

Somos un país de servicios, tenemos que darnos cuenta de ello. Los Vascos, como siempre, son los que se han dado cuenta y van un paso por delante [bculinary.com].

Asumámoslo, nunca vamos a ser los médicos ni los científicos ni los ingenieros de Europa, somos los que les ponen las copas cuando se vienen de vacaciones.

--
All my life, I have enjoyed the reputation of being someone who disrupted prevailing ideas.

(Benoit Mandelbrot)

[Responder]

Re:Pues claro que es un lujo... (Puntos:0)

por pobrecito hablador el Lunes, 23 Julio de 2012, 16:48h (#1315566)

"Deberían invertir todo ese presupuesto de ciencia que desperdiciamos en intentar curar el cáncer o hacer más eficientes las energías renovables en algo útil y más acorde al país en que vivimos como formar a cocteleros y dj's para nuestros chiringuitos playeros. "

Pues no se si lo has dicho con ironía, pero...

http://www.cefe.gva.es/agenda.asp?id=613 [cefe.gva.es]
La conselleria de Educación, Formación y Empleo implanta 26 nuevos ciclos formativos

(blablabla) ...

Como novedad para este curso 2012-2013 se han implantado las nuevas titulaciones de Técnico Superior de Artista Fallero y Construcción de Escenografías y el de Técnico en Video Disc-jockey y Sonido.

(blablabla)...

Ya ves, los vascos no son los únicos "visionarios" :D

[Responder | Padre]

y mientras tanto se suben sus sueldos... (Puntos:0)

por pobrecito hablador el Martes, 24 Julio de 2012, 06:21h (#1315608)

El Gobierno de Castilla-La Mancha ha aumentado la dotación económica para alta dirección un 157% en los presupuestos de la Junta para 2012, con incrementos del 8,6% y del 10,9 % en algunos sueldos, según ha dicho hoy el secretario regional de Organización del PSOE, Jesús Fernández Vaquero. En rueda de prensa, Fernández Vaquero ha explicado que esta subida se recoge en el tomo 1 de los presupuestos generales de la Junta, en el que se constata que la dotación económica para alta dirección del Gobierno de la comunidad autónoma ha pasado de 22,3 millones de euros en 2011 a 56,64% en 2012. Fernández Vaquero ha indicado que los suelos de los directores generales se incrementan un 10,9%, pasando de 53.450 euros brutos anuales a 59.270. A su vez, el sueldo de los viceconsejeros pasan de 58.230 euros brutos a 63.320, según ha precisado. Esta subida contrasta, según Fernández Vaquero, con el hecho de que a todos los funcionarios se les baje un 3% su salario. El secretario regional de organización del PSOE ha reconocido que los sueldos de los altos cargos de Castilla-La Mancha son más bajos que los de otras comunidades autónomas pero ha criticado la doble vara de medir: "Al señor que gana 1.000 euros o que gana 900 euros se le baja o se le reduce el sueldo". "No podemos estar hablando de contener el gasto y reducir, cuando si hay algo que le afecta a ella lo que hace es aumentar el gasto", ha añadido Fernández Vaquero en alusión a la presidenta de Castilla-La Mancha, María Dolores de Cospedal. http://politica.elpais.com/politica/2012/06/01/act ualidad/1338573325_921691.html [elpais.com]

[Responder | Padre]

Re:y mientras tanto se suben sus sueldos... (Puntos:0)

por pobrecito hablador el Martes, 24 Julio de 2012, 06:36h (#1315609)

Ya se ve donde va a parar el dinero que no pagan a los farmacéuticos de Castilla la Mancha, a los enchufados de la señora Cospedal

[Responder | Padre]

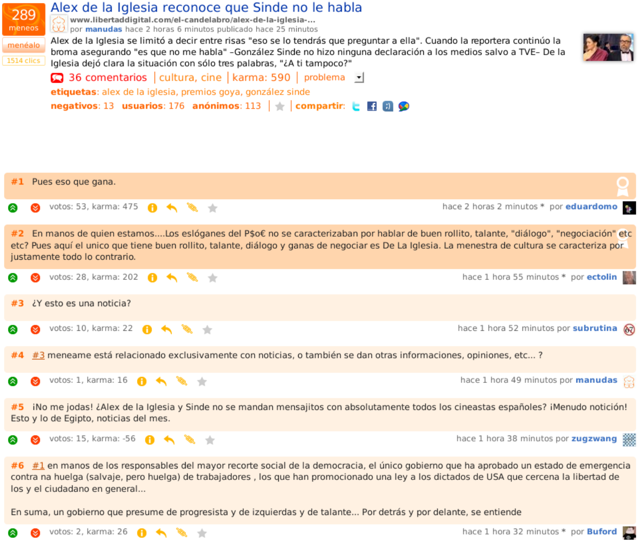
Re:Pues claro que es un lujo... (Puntos:0)

por pobrecito hablador el Martes, 24 Julio de 2012, 08:25h (#1315613)

Con idiotas como tú diciendo esas soplapolleces es como me doy cuenta de que España es un país que tiene exactamente aquello que merece.

Motivation

Example of discussion in Meneame:



289 meneos

menéalo

1514 clics

Alex de la Iglesia reconoce que Sinde no le habla

www.libertaddigital.com/el-candelabro/alex-de-la-iglesia-...

por **manudas** hace 2 horas 6 minutos publicado hace 25 minutos

Alex de la Iglesia se limitó a decir entre risas "eso se lo tendrás que preguntar a ella". Cuando la reportera continúo la broma asegurando "es que no me habla" –González Sinde no hizo ninguna declaración a los medios salvo a TVE– De la Iglesia dejó clara la situación con sólo tres palabras, "¿A ti tampoco?"

36 comentarios | cultura, cine | karma: 590 | problema

etiquetas: alex de la iglesia, premios goya, gonzález sinde

negativos: 13 **usuarios**: 176 **anónimos**: 113 | ★ | **compartir**:

#1 Pues eso que gana.

votos: 53, karma: 475 — hace 2 horas 2 minutos * por **eduardomo**

#2 En manos de quien estamos....Los eslóganes del P$o€ no se caracterizaban por hablar de buen rollito, talante, "diálogo", "negociación" etc etc? Pues aquí el unico que tiene buen rollito, talante, diálogo y ganas de negociar es De La Iglesia. La menestra de cultura se caracteriza por justamente todo lo contrario.

votos: 28, karma: 202 — hace 1 hora 55 minutos * por **ectolin**

#3 ¿Y esto es una noticia?

votos: 10, karma: 22 — hace 1 hora 52 minutos por **subrutina**

#4 #3 meneame está relacionado exclusivamente con noticias, o también se dan otras informaciones, opiniones, etc... ?

votos: 1, karma: 16 — hace 1 hora 49 minutos por **manudas**

#5 ¡No me jodas! ¿Alex de la Iglesia y Sinde no se mandan mensajitos con absolutamente todos los cineastas españoles? ¡Menudo notición! Esto y lo de Egipto, noticias del mes.

votos: 15, karma: -56 — hace 1 hora 38 minutos por **zugzwang**

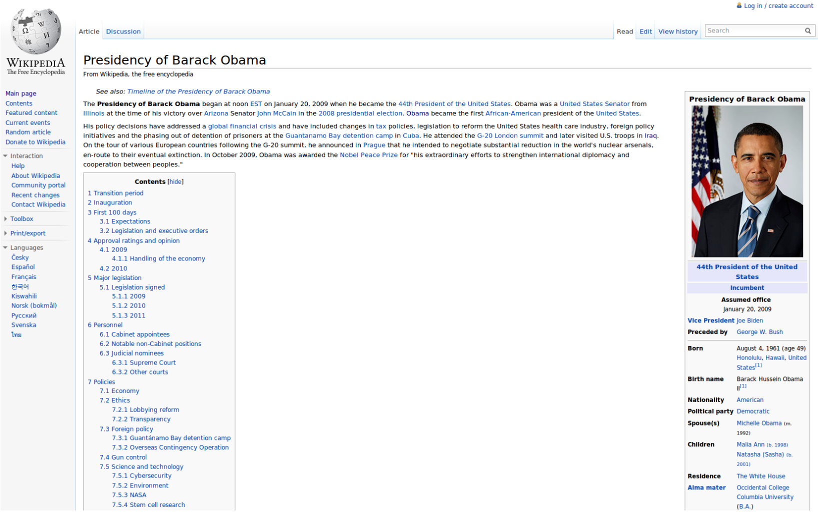
#6 #1 en manos de los responsables del mayor recorte social de la democracia, el único gobierno que ha aprobado un estado de emergencia contra na huelga (salvaje, pero huelga) de trabajadores , los que han promocionado una ley a los dictados de USA que cercena la libertad de los y el ciudadano en general...

En suma, un gobierno que presume de progresista y de izquierdas y de talante... Por detrás y por delante, se entiende

votos: 2, karma: 26 — hace 1 hora 32 minutos * por **Buford**

Motivation

Example of discussion in Wikipedia (I)



Presidency of Barack Obama

From Wikipedia, the free encyclopedia

See also: Timeline of the Presidency of Barack Obama

The **Presidency of Barack Obama** began at noon EST on January 20, 2009 when he became the 44th President of the United States. Obama was a United States Senator from Illinois at the time of his victory over Arizona Senator John McCain in the 2008 presidential election. Obama became the first African-American president of the United States.

His policy decisions have addressed a global financial crisis and have included changes in tax policies, legislation to reform the United States health care industry, foreign policy initiatives and the phasing out of detention of prisoners at the Guantanamo Bay detention camp in Cuba. He attended the G-20 London summit and later visited U.S. troops in Iraq. On the tour of various European countries following the G-20 summit, he announced in Prague that he intended to negotiate substantial reduction in the world's nuclear arsenals, en-route to their eventual extinction. In October 2009, Obama was awarded the Nobel Peace Prize for "his extraordinary efforts to strengthen international diplomacy and cooperation between peoples."

Contents [hide]

Presidency of Barack Obama

44th President of the United States

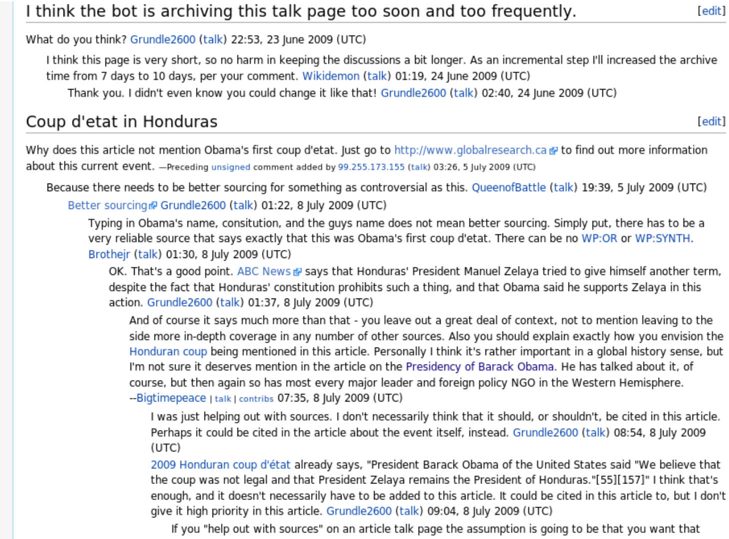
Incumbent

Assumed office
January 20, 2009

Vice President	Joe Biden
Preceded by	George W. Bush
Born	August 4, 1961 (age 49) Honolulu, Hawaii, United States[1]
Birth name	Barack Hussein Obama II[1]
Nationality	American
Political party	Democratic
Spouse(s)	Michelle Obama (m. 1992)
Children	Malia Ann (b. 1998) Natasha (Sasha) (b. 2001)
Residence	The White House
Alma mater	Occidental College Columbia University (B.A.)

Motivation

Example of discussion in Wikipedia (II)



I think the bot is archiving this talk page too soon and too frequently. [edit]

What do you think? Grundle2600 (talk) 22:53, 23 June 2009 (UTC)

I think this page is very short, so no harm in keeping the discussions a bit longer. As an incremental step I'll increased the archive time from 7 days to 10 days, per your comment. Wikidemon (talk) 01:19, 24 June 2009 (UTC)

Thank you. I didn't even know you could change it like that! Grundle2600 (talk) 02:40, 24 June 2009 (UTC)

Coup d'etat in Honduras [edit]

Why does this article not mention Obama's first coup d'etat. Just go to http://www.globalresearch.ca to find out more information about this current event. —Preceding unsigned comment added by 99.255.173.155 (talk) 03:26, 5 July 2009 (UTC)

Because there needs to be better sourcing for something as controversial as this. QueenofBattle (talk) 19:39, 5 July 2009 (UTC)

Better sourcing Grundle2600 (talk) 01:22, 8 July 2009 (UTC)

Typing in Obama's name, consitution, and the guys name does not mean better sourcing. Simply put, there has to be a very reliable source that says exactly that this was Obama's first coup d'etat. There can be no WP:OR or WP:SYNTH. Brothejr (talk) 01:30, 8 July 2009 (UTC)

OK. That's a good point. ABC News says that Honduras' President Manuel Zelaya tried to give himself another term, despite the fact that Honduras' constitution prohibits such a thing, and that Obama said he supports Zelaya in this action. Grundle2600 (talk) 01:37, 8 July 2009 (UTC)

And of course it says much more than that - you leave out a great deal of context, not to mention leaving to the side more in-depth coverage in any number of other sources. Also you should explain exactly how you envision the Honduran coup being mentioned in this article. Personally I think it's rather important in a global history sense, but I'm not sure it deserves mention in the article on the Presidency of Barack Obama. He has talked about it, of course, but then again so has most every major leader and foreign policy NGO in the Western Hemisphere. --Bigtimepeace | talk | contribs 07:35, 8 July 2009 (UTC)

I was just helping out with sources. I don't necessarily think that it should, or shouldn't, be cited in this article. Perhaps it could be cited in the article about the event itself, instead. Grundle2600 (talk) 08:54, 8 July 2009 (UTC)

2009 Honduran coup d'état already says, "President Barack Obama of the United States said "We believe that the coup was not legal and that President Zelaya remains the President of Honduras."[55][157]" I think that's enough, and it doesn't necessarily have to be added to this article. It could be cited in this article to, but I don't give it high priority in this article. Grundle2600 (talk) 09:04, 8 July 2009 (UTC)

If you "help out with sources" on an article talk page the assumption is going to be that you want that

Most discussed Wikipedia articles

Top 20 articles ordered by number of chains in the discussion [Laniado et al. 2011]

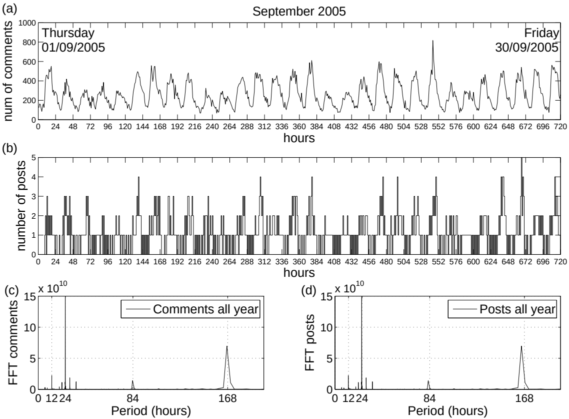
#	Title	chains	comments	users	h-index	max. depth	edits
1	Intelligent design	2413	22454 (3)	954 (13)	16 (20)	20 (358)	9179 (53)
2	Gaza War	2358	17961 (6)	607 (47)	19 (2)	27 (28)	11499 (29)
3	Barack Obama	2301	22756 (2)	2360 (2)	18 (6)	21 (245)	17453 (6)
4	Sarah Palin	2182	19634 (4)	1221 (9)	17 (10)	25 (56)	12093 (24)
5	Global warming	2178	19138 (5)	1382 (5)	17 (10)	20 (358)	14074 (15)
6	Main Page	2065	32664 (1)	5969 (1)	15 (34)	22 (169)	4003 (674)
7	Chiropractic	1772	13684 (13)	243 (389)	18 (6)	29 (17)	6190 (204)
8	Race and intelligence	1764	13790 (12)	410 (126)	17 (10)	24 (74)	7615 (100)
9	Anarchism	1589	14385 (9)	496 (76)	20 (1)	28 (22)	12589 (19)
10	British Isles	1556	12044 (16)	576 (56)	17 (10)	23 (113)	4047 (658)
11	CRU[1] hacking incident	1551	11536 (17)	474 (88)	17 (10)	20 (358)	2346 (2364)
12	Jesus	1397	17916 (7)	1239 (7)	13 (119)	16 (1383)	17081 (7)
13	Circumcision	1356	10469 (21)	436 (113)	17 (10)	26 (42)	7354 (117)
14	Homeopathy	1323	13509 (14)	516 (68)	17 (10)	25 (56)	6902 (151)
15	George W. Bush	1281	15257 (8)	1969 (3)	14 (65)	18 (676)	32314 (1)
16	September 11 attacks	1250	13830 (11)	1244 (6)	16 (20)	26 (42)	11086 (30)
17	Evolution	1165	13404 (15)	942 (16)	13 (119)	23 (113)	9780 (44)
18	Catholic Church	1162	14104 (10)	620 (43)	15 (34)	18 (676)	14082 (14)
19	Cold fusion	1098	8354 (29)	359 (174)	15 (34)	20 (358)	4320 (557)
20	2008 South Ossetia war	1075	10596 (20)	853 (20)	17 (10)	23 (113)	9930 (43)

In parenthesis: rank according to the corresponding variable

[1] Climatic Research Unit

Temporal patterns. From Kaltenbrunner et al (LAWEB 2007)

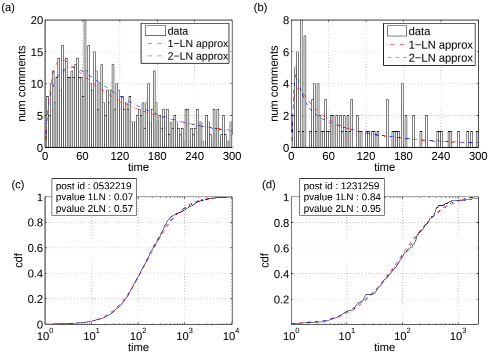
Time series of total number of comments



1. "Sustained" activity coupled with the circadian rhythm.

Temporal patterns. From Kaltenbrunner et al (LAWEB 2007)

Single post level analysis

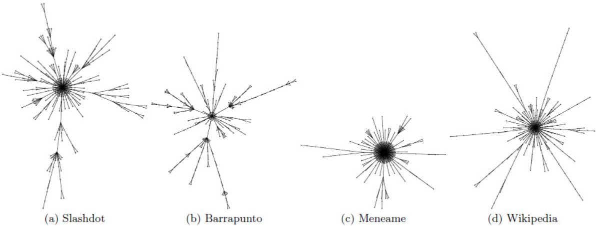


- Posts create cascades of comments which propagate over the network.
- All posts show a stereotyped behaviour.
- Response times can be described using a log-normal distribution.

Online discussion threads

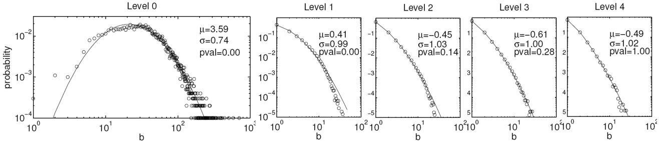
Examples of real discussions

Typical cascades for each website:



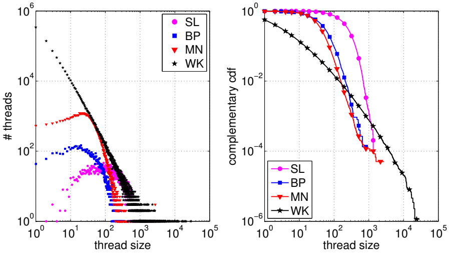
(a) Slashdot (b) Barrapunto (c) Meneame (d) Wikipedia

Degrees Slashdot:



Online discussion threads

Global analysis



- SL, BP and MN present a distribution with a defined scale.
- Cascade sizes in Wikipedia "seem to be" scale-free.

Outline

1 Introduction
- Institution
- Motivation
- Datasets

2 Likelihood-based framework
- Model definition
- Parameter estimation
- Validation

3 Conclusions

Model definition

Our approach:

- The model must reproduce:
  - The statistical structure of threads.
  - Their evolution.
- No content involved.
- No authorship.
- Essentially *"Which comment is going to be replied next?"*

Empirical facts:

- Popular comments receive more replies: *preferential attachment*.
- New comments are more *attractive* than old ones.
- Replies to the post behave different than replies to comments.

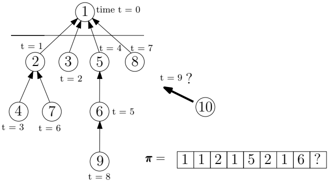
Model definition

- Thread **representation**: vector of parent nodes $\boldsymbol{\pi}$, where π_t denotes the parent of the node with id $t+1$ added at time-step t.

$$\boldsymbol{\pi}_0 = ()$$

$$\boldsymbol{\pi}_1 = (1)$$

$$\dots$$



Parameters of the model

- At time t, the **popularity** of node k is its degree:

$$d_{k,t}(\boldsymbol{\pi}_{(1:t-1)}) = \begin{cases} 1 + \sum_{m=2}^{t-1} \delta_{k\pi_m} & \text{for } k \in \{1,\dots,t\} \\ 0 & \text{otherwise,} \end{cases} \qquad (d_{k,t} \text{ is weighted by } \alpha)$$

- At time t, the **novelty** of node k is $n_{k,t} = \tau^{t-k+1}, \quad \tau \in [0,1]$.
- **Root bias**: The bias of a node k is is either zero or β for the root:

$$b_k = \beta, \qquad \text{for } k = 1, \text{ and } 0 \text{ otherwise.}$$

Model definition

- We define a model by means of its associated *attractiveness* function $\phi(\cdot)$, which is defined for each of the nodes.
- At time $t+1$, a new node is linked to node k with probability:

$$p(\pi_t = k | \boldsymbol{\pi}_{(1:t-1)}) = \frac{\phi(k)}{Z_t}, \qquad Z_t = \sum_{l=1}^{t} \phi(l),$$

- Different model variants:

Model	Attractiveness funct. $\phi(\cdot)$	Parameters $\boldsymbol{\theta}$	Constraint
Full model (**FM**)	$\alpha d_{k,t} + b_k + \tau^{t-k+1}$	$\{\alpha, \tau, \beta\}$	
Model without popularity (**NO-**α)	$b_k + \tau^{t-k+1}$	$\{\tau, \beta\}$	$\alpha = 0$
Model without novelty (**NO-**τ)	$\alpha d_{k,t} + b_k + 1$	$\{\alpha, \beta\}$	$\tau = 1$
Model without bias (**NO-bias**)	$\alpha d_{k,t} + \tau^{t-k+1}$	$\{\alpha, \tau\}$	$\beta = 0$

Outline

1. Introduction
   - Institution
   - Motivation
   - Datasets
2. Likelihood-based framework
   - Model definition
   - Parameter estimation
   - Validation
3. Conclusions

Parameter estimation

Maximum likelihood

We can compute the likelihood of the full model

- The likelihood of a set $\Pi := \{\boldsymbol{\pi}_1, \ldots \boldsymbol{\pi}_N\}$ of N trees with respective sizes $|\boldsymbol{\pi}_i|$, $i \in \{1, \ldots N\}$, given the values of $\boldsymbol{\theta}$ can be written as:

$$\mathcal{L}(\boldsymbol{\Pi}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\boldsymbol{\pi}_i|\boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{t=2}^{|\boldsymbol{\pi}_i|} p(\pi_{t,i}|\boldsymbol{\pi}_{(1:t-1),i}, \boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{t=2}^{|\boldsymbol{\pi}_i|} \frac{\phi(\pi_{t,i})}{Z_{t,i}}$$

- We minimise the negative of the log-likelihood function:

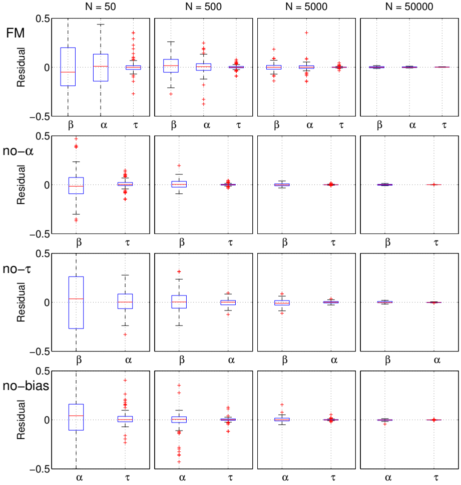
$$-\log \mathcal{L}(\Pi|\boldsymbol{\theta}) = -\sum_{i=1}^{N} \sum_{t=2}^{|\boldsymbol{\pi}_i|} \phi(\pi_{t,i}) - \log Z_{t,i}.$$

Parameter estimation

Validation

For each model:

- Choose $\boldsymbol{\theta}^*$ randomly.
- Generate N threads.
- Find estimates $\hat{\boldsymbol{\theta}}$.
- Compute residuals $\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}$.
- Repeat for 100 times.



- Estimation is unbiased.
- Good estimates can be obtained using $N = 500$.

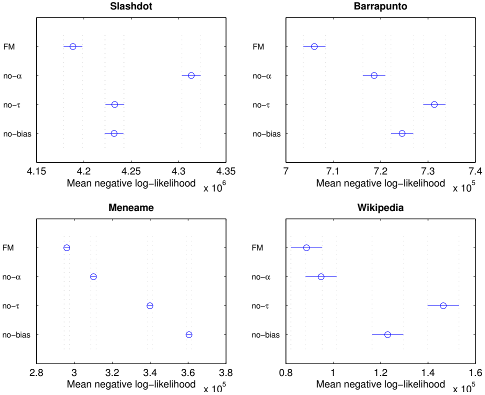
Parameter estimation

Model Comparison

For each dataset:

- Select $N = 5 \cdot 10^4$ threads randomly with replacement.
- Find estimates $\hat{\theta}$.
- Compute likelihoods.
- Repeat for 100 times.

- Model comparison based on likelihoods for each dataset.

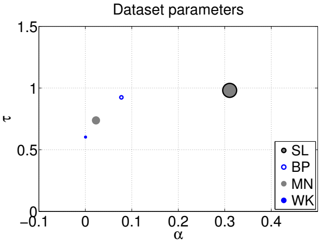


Parameter estimation

Parameter estimates for the different datasets

Dataset	$\log \beta$	α	τ
$N = 50$			
SL	2.39 (0.17)	0.31 (0.02)	0.98 (0.02)
BP	0.93 (0.12)	0.08 (0.04)	0.92 (0.00)
MN	1.66 (0.16)	0.03 (0.01)	0.72 (0.04)
WK	−0.21 (0.81)	0.00 (0.00)	0.40 (0.19)
$N = 5000$			
SL	**2.39** (0.01)	**0.31** (0.01)	**0.98** (0.00)
BP	**0.96** (0.02)	**0.08** (0.00)	**0.92** (0.00)
MN	**1.69** (0.03)	**0.02** (0.00)	**0.74** (0.01)
WK	**0.39** (0.22)	**0.00** (0.00)	**0.60** (0.01)

- Bootstrap with $N = 50$ threads already gives good estimates.



Outline

1. Introduction
   - Institution
   - Motivation
   - Datasets
2. Likelihood-based framework
   - Model definition
   - Parameter estimation
   - Validation
3. Conclusions

Growing tree model for discussion threads

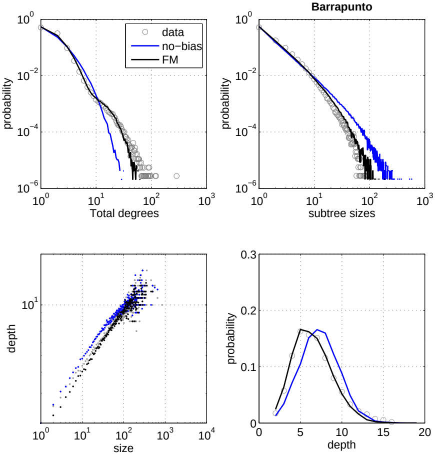
Validation of the model

We calculate the following quantities from the empirical data and from the synthetic threads produced by the model:

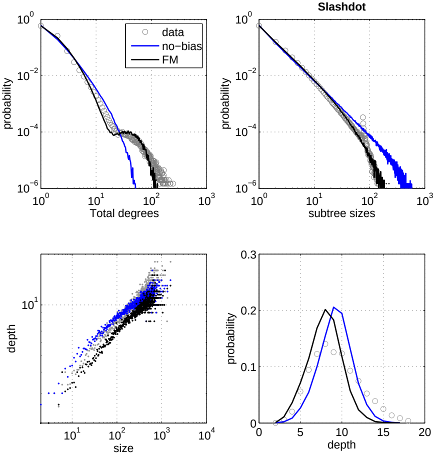
- Degrees distribution.
- Subtree sizes distribution.
- Mean node depth versus size.
- Node depths distribution.

- Size of the post N is drawn from the empirical distribution.
- We use model **NO-BIAS** for comparison [Kumar et al. 2010].

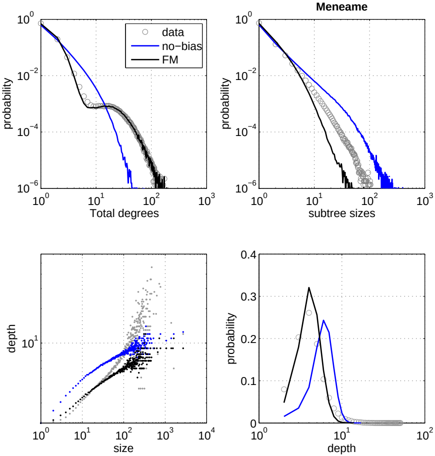
Barrapunto dataset



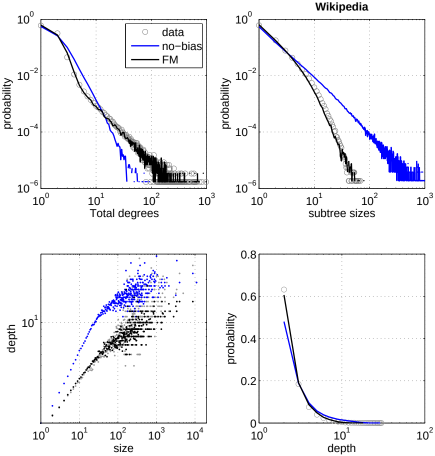
Slashdot dataset



Meneame dataset

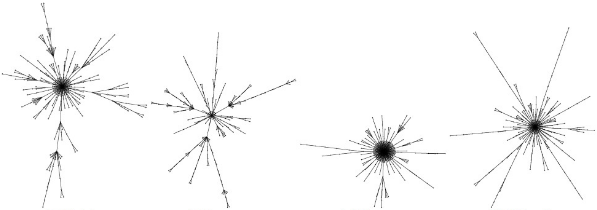


Wikipedia dataset



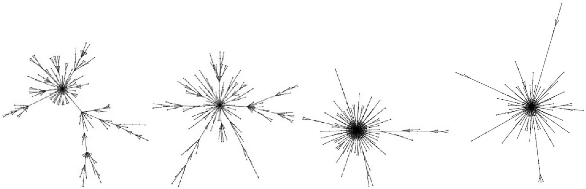
Growing tree model for discussion threads

Real cascades:



(a) Slashdot (b) Barrapunto (c) Meneame (d) Wikipedia

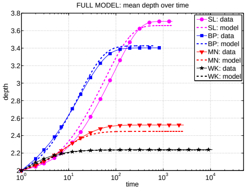
Synthetic cascades:

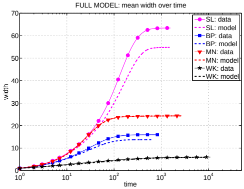


(a) Slashdot (b) Barrapunto (c) Meneame (d) Wikipedia

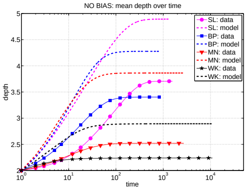
Evolution of mean depths and mean widths

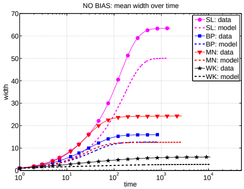
FULL MODEL:





NO-BIAS model:





Theoretical Result

Asymptotics for degree distribution in FM

It can be shown that ...

- the degree distribution follows (asymptotically) a power-law with exponent 3.
- The parameter τ does not affect the power-law exponent

but ...

- formally

$$c_1 x^{-2} \leq P(degree \geq x) \leq c_2 x^{-2}, 0 \leq c_1 \leq c_2$$

- with τ affecting c_2 which is bounded by $\exp(\frac{\tau}{1-\tau})$.
- Thus τ can affect the fraction of nodes with a degree larger than x by several orders of magnitude.

Related work

Galton-Watson branching process

Idea

- Tree grows level by level.
- Nodes at level i receive a random number of child-nodes at level $i+1$ (according to a probability distribution).

Pros

- The model is simple.
- Explains chain letter trees (combined with a selection bias) [Golub & Jackson 2010].

Cons

- Not a generative model.
- Does not capture the order of message creation.

Conclusions and future work

Conclusions

- Framework which allows to re-create discussions with similar structural features as real instances.
- Likelihood-based optimisation on the entire cascade evolution.
- Large datasets are not necessary.
- Parameters allow to characterize audience and platform:
  - Same platform : differences between SL and BP.
  - Influence of the interface: MN (flat) characterised by bias.
  - Main difference between news media and WK: popularity.

Future work

- Include prior authorship structure in model and analysis
- Application to other types of information cascades.

questions?
networks
friendship
users
reply
social
number
comments
links
friends
nodes
Meneame
behavioural
two
analysis
similar
larger
connections
replies
link
also
Section
degree
one
interactions
reciprocity
connected
directed
decomposition
Figure
comment
non-friends
case
probability
threshold
analyse
found
study
articulated
shell
votes
undirected
likely
value
Table
average
edges
relations
index
higher
relation
find
calculate
least
difference
stories
cases
original
large
use
section
intersection
reduced
core
present
explicit
slightly
friend

Bibliography I

V. Gómez, H. J. Kappen, N. Litvak & A. Kaltenbrunner.
A likelihood-based framework for the analysis of discussion threads.
World Wide Web Journal

V. Gómez, H. J. Kappen & A. Kaltenbrunner.
Modeling the structure and evolution of discussion cascades.
In HT '11, Eindhoven, The Netherlands, 2011. ACM.

B. Golub & M.O. Jackson.
Using selection bias to explain the observed structure of Internet diffusions.
Proceedings of the National Academy of Sciences, vol. 107, no. 24, page 10833, 2010.

A. Kaltenbrunner, V. Gómez & V. López.
Description and Prediction of Slashdot Activity.
In LA-WEB '07, Santiago de Chile, 2007. IEEE.

R. Kumar, M. Mahdian & M. McGlohon.
Dynamics of conversations.
In SIGKDD '10, pages 553–562, New York, USA, 2010. ACM.

D. Laniado, R. Tasso, Y. Volkovich & A. Kaltenbrunner.
When the Wikipedians talk: Network and tree structure of Wikipedia discussion pages.
In ICWSM-11 - 5th International AAAI Conference on Weblogs and Social Media. The AAAI Press, 2011.

Related work II

T-MODEL [Kumar et al. 2010]

Features

- Equivalent to model **NO-bias** with an extra parameter to model the death of a discussion.
- Model is illustrated on USENET.
- Authorship model (TI-model).
- Both are independent of the structure.
- Could be build on top of other structural models as well.
- T-model re-creates a power-law relation in the data between size and depth of the discussions, but is not the best model.