Network Dynamics

Dynamics and behaviour in online conversations

Andreas Kaltenbrunner

andreas.kaltenbrunner@upf.edu

Based on slides from Vicenç Gómez

Table of contents

Outline

Introduction

Temporal patterns and popularity prediction

Modeling conversation threads

Motivation

Dynamics of online conversations

- Temporal dimension
  - What temporal patterns govern these social phenomena?
  - Can we predict popularity of news?
- Structural dimension
  - Can we model how conversation trees evolve in time?
  - Can we characterise user behaviour in terms of this model?

Motivation

General methodology

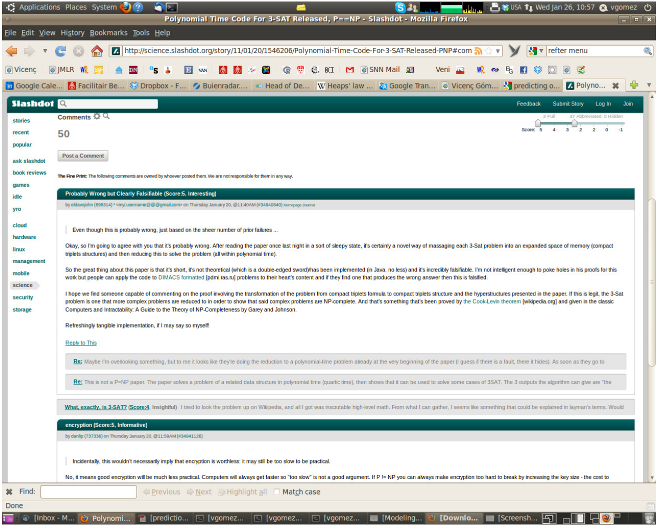
- Parsimonious data-driven approach
  - Few parameters that are interpretable
  - Simple optimisation problems
- Role of the content
  - Explain as much as possible without considering content
- Analysis at the population level
  - Single-user data is too noisy
  - Aggregate analysis averages out the noise

Motivation

Example of conversation in Slashdot (post):

Motivation

Example of conversation in Slashdot (comments):



Motivation

Example of conversation threads 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

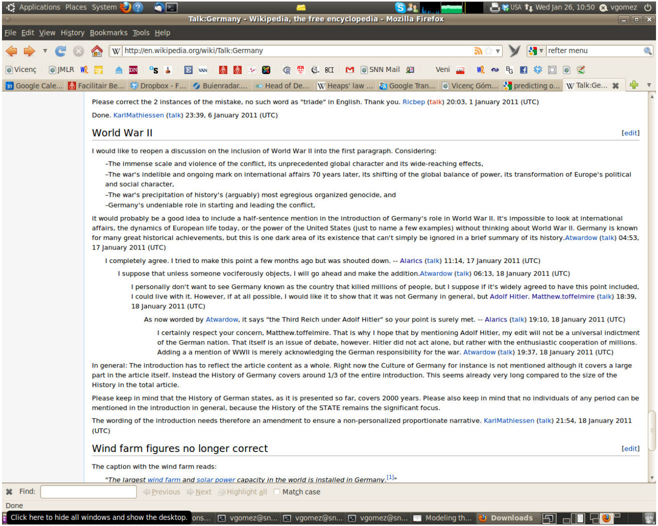
#6 #1 en manos de los responsables del mayor recorte social de la democracia, el único gobierno que ha aprobado un estado de emergencia contra na huelga (salvaje, pero huelga) de trabajadores , los que han promocionado una ley a los dictados de USA que cercena la libertad de los y el ciudadano en general...

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 conversation in Wikipedia:


Talk:Germany - Wikipedia, the free encyclopedia - Mozilla Firefox

http://en.wikipedia.org/wiki/Talk:Germany

Please correct the 2 instances of the mistake, no such word as "triade" in English. Thank you. Ricbep (talk) 20:03, 1 January 2011 (UTC)

Done. KarlMathiessen (talk) 23:39, 6 January 2011 (UTC)

World War II [edit]

I would like to reopen a discussion on the inclusion of World War II into the first paragraph. Considering:

-The immense scale and violence of the conflict, its unprecedented global character and its wide-reaching effects,

-The war's indelible and ongoing mark on international affairs 70 years later, its shifting of the global balance of power, its transformation of Europe's political and social character,

-The war's precipitation of history's (arguably) most egregious organized genocide, and

-Germany's undeniable role in starting and leading the conflict,

it would probably be a good idea to include a half-sentence mention in the introduction of Germany's role in World War II. It's impossible to look at international affairs, the dynamics of European life today, or the power of the United States (just to name a few examples) without thinking about World War II. Germany is known for many great historical achievements, but this is one dark area of its existence that can't simply be ignored in a brief summary of its history.Atwardow (talk) 04:53, 17 January 2011 (UTC)

I completely agree. I tried to make this point a few months ago but was shouted down. -- Alarics (talk) 11:14, 17 January 2011 (UTC)

I suppose that unless someone vociferously objects, I will go ahead and make the addition.Atwardow (talk) 06:13, 18 January 2011 (UTC)

I personally don't want to see Germany known as the country that killed millions of people, but I suppose if it's widely agreed to have this point included, I could live with it. However, if at all possible, I would like it to show that it was not Germany in general, but Adolf Hitler. Matthew.toffelmire (talk) 18:39, 18 January 2011 (UTC)

As now worded by Atwardow, it says "the Third Reich under Adolf Hitler" so your point is surely met. -- Alarics (talk) 19:10, 18 January 2011 (UTC)

I certainly respect your concern, Matthew.toffelmire. That is why I hope that by mentioning Adolf Hitler, my edit will not be a universal indictment of the German nation. That itself is an issue of debate, however. Hitler did not act alone, but rather with the enthusiastic cooperation of millions. Adding a a mention of WWII is merely acknowledging the German responsibility for the war. Atwardow (talk) 19:37, 18 January 2011 (UTC)

In general: The introduction has to reflect the article content as a whole. Right now the Culture of Germany for instance is not mentioned although it covers a large part in the article itself. Instead the History of Germany covers around 1/3 of the entire introduction. This seems already very long compared to the size of the History in the total article.

Please keep in mind that the History of German states, as it is presented so far, covers 2000 years. Please also keep in mind that no individuals of any period can be mentioned in the introduction in general, because the History of the STATE remains the significant focus.

The wording of the introduction needs therefore an amendment to ensure a non-personalized proportionate narrative. KarlMathiessen (talk) 21:54, 18 January 2011 (UTC)

Wind farm figures no longer correct [edit]

The caption with the wind farm reads:

"*The largest wind farm and solar power capacity in the world is installed in Germany.*[1]"

Outline

Introduction

Temporal patterns and popularity prediction

Modeling conversation threads

Motivation

Scientific questions

Temporal patterns in news aggregators

[Kaltenbrunner et al, 2007]

- What are the temporal patterns governing these responses?
- Is there a mathematical law that describes this patterns?
- Can we use this law to predict number of votes (popularity) in the long term?

Structure and evolution of conversation threads

[Gómez et al, 2013]

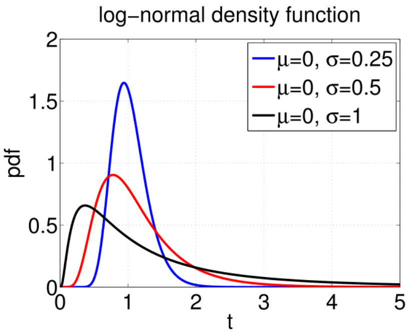
- What are the structural patterns governing these responses?
- Is there a generative model that captures their statistical properties?
- Can we use the model parameters to characterize websites, user behaviour, conversations?

Preliminaries

The Log-Normal distribution

- Continuous probability distribution of a random variable whose logarithm is normally distributed

$$f_{\text{LN}}(t;\mu,\sigma) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(\frac{-(\ln(t)-\mu)^2}{2\sigma^2}\right)$$

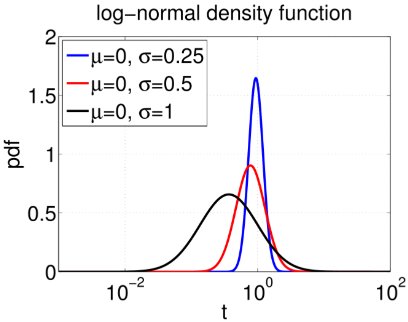


Preliminaries

The Log-Normal distribution

- Continuous probability distribution of a random variable whose logarithm is normally distributed

$$f_{\mathrm{LN}}(t;\mu,\sigma) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(\frac{-(\ln(t)-\mu)^2}{2\sigma^2}\right)$$

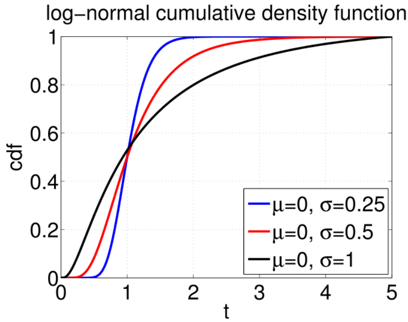


Preliminaries

The Log-Normal distribution

- Continuous probability distribution of a random variable whose logarithm is normally distributed

$$f_{\mathrm{LN}}(t;\mu,\sigma) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(\frac{-(\ln(t)-\mu)^2}{2\sigma^2}\right)$$

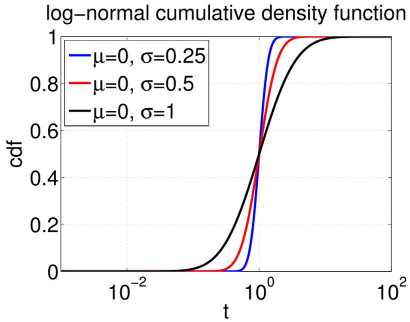


Motivation

The Log-Normal distribution

- Continuous probability distribution of a random variable whose logarithm is normally distributed

$$f_{\mathrm{LN}}(t;\mu,\sigma) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(\frac{-(\ln(t)-\mu)^2}{2\sigma^2}\right)$$



Preliminaries

Fitting log-normal distributions

- A dataset of points is given $\mathbf{t} = t_1, \ldots, t_n$
- Maximum likelihood

$$\mathcal{L}(\mathbf{t}; \mu, \sigma) = \prod_{i=1}^{n} \left(\frac{1}{t_i}\right) \mathcal{N}(\ln t_i; \mu, \sigma)$$

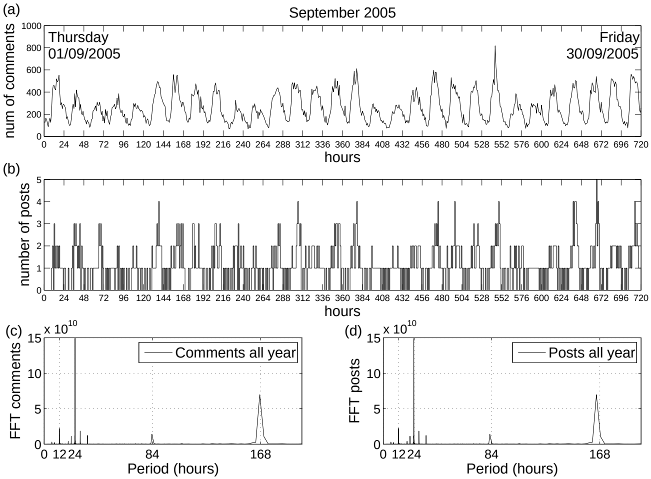
- Closed form

$$\hat{\mu} = \frac{\sum_i \ln t_i}{n}$$
$$\hat{\sigma}^2 = \frac{\sum_i (\ln t_i - \hat{\mu})^2)}{n}$$

- Alternatively: using `fminsearch` in Matlab or similar tools

Temporal patterns of Slashdot

Time series of total number of comments



- "Sustained" activity coupled with the circadian rythm.

Temporal patterns of Slashdot

Statistical approach for analyzing reaction times

- Guess a candidate probability distribution F for reaction times
- Kolmogorov-Smirnov (KS) test
- Following hypothesis
  - H_0: The reaction time is a sample of distribution F
  - H_1: The hypothesis H_0 is not true
- Compute point-wise maximal difference between the CDF of the data and the approximation (KS statistic)
- Calculate the p-value: probability of obtaining a result as different as F as the data
- The greater the p-value, the better the fit
- For a chosen level of significance α_0, the hypothesis H_0 is accepted

Temporal patterns of Slashdot

Log-normal model and circadian cycle

- Incorporating the circadian cycle in the log-normal model

$$f_{\mathsf{LN \times C}}(t; \mu, \sigma, C(\cdot)) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(\frac{-(\ln(t) - \mu)^2}{2\sigma^2}\right) C(t)$$

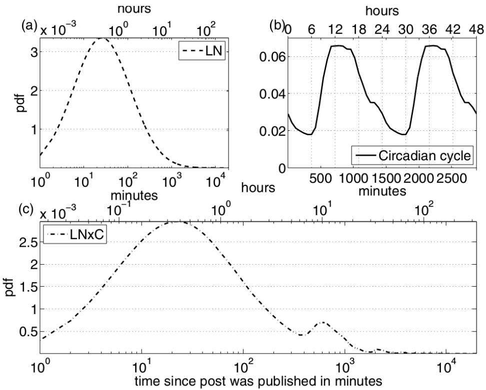
The function $C(\cdot)$ is computed from the data

Temporal patterns of Slashdot

Log-normal model and circadian cycle

- Incorporating the circadian cycle in the log-normal model

$$f_{\mathrm{LN{\times}C}}(t;\mu,\sigma,C(\cdot)) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(\frac{-(\ln(t)-\mu)^2}{2\sigma^2}\right) C(t)$$



Temporal patterns of Slashdot

A mixture of two log-normals

- A more flexible model
- Linear combination of two log-normals

$$f_{\mathrm{DLN}}(t;\theta) = k f_{\mathrm{LN}}(t;\mu_1,\sigma_1) + (1-k) f_{\mathrm{LN}}(t;\mu_2,\sigma_2)$$

- Parameters $\theta = (k, \mu_1, \sigma_1, \mu_2, \sigma_2)$

A mixture of two log-normals with circadian cycle

- Incorporating the circadian cycle in the mixture log-normal model

$$f_{\mathrm{DLNxC}}(t;\theta) = \left(k f_{\mathrm{LN}}(t;\mu_1,\sigma_1) + (1-k) f_{\mathrm{LN}}(t;\mu_2,\sigma_2)\right) C(t)$$

- Parameters $\theta = (k, \mu_1, \sigma_1, \mu_2, \sigma_2, C(\cdot))$

Temporal patterns of Slashdot

Summary of models

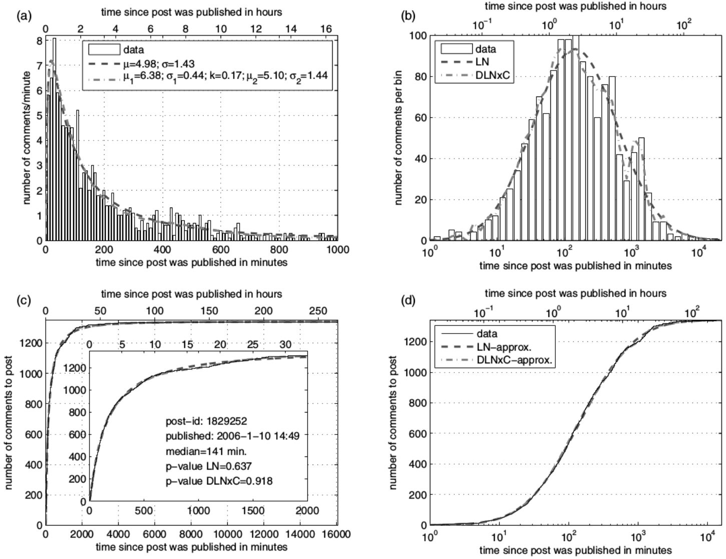
- (LN) Single log-normal model
- (LNxC) Single log-normal model with circadian cycle
- (DLN) Double log-normal model
- (DLNxC) Double log-normal model with circadian cycle

Tasks

- Model comparison
- Which model is better? How much? Why?
- Can we interpret the parameters?

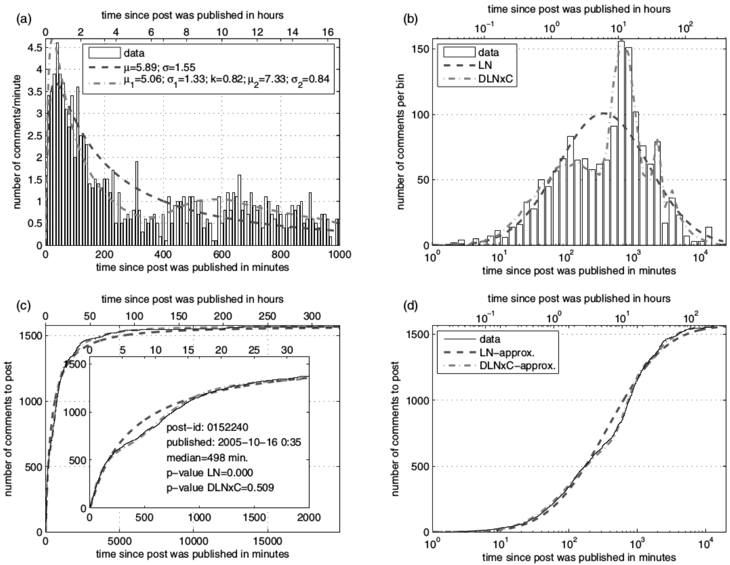
Temporal patterns of Slashdot

Single-post analysis (post published in the afternoon)



Temporal patterns of Slashdot

Single-post analysis (post published during night)



Temporal patterns of Slashdot

Some conclusions

- All posts show a stereotyped behavior
- Accurate fitting using models based on log-normal distributions
- LN model performs well for post published in daylight
- DLNxC model outperforms LN for post published during night

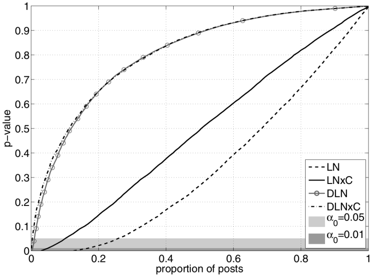
Temporal patterns of Slashdot

Approximating all posts

- Analysis of distribution of KS statistic and p-values

α_0	0.01	0.05
LN	16.68%	25.62%
LNxC	4.80%	9.88%
DLN	0.44%	0.96%
DLNxC	0.11%	0.33%

Table 1. Percentage of rejected 0-Hypotheses

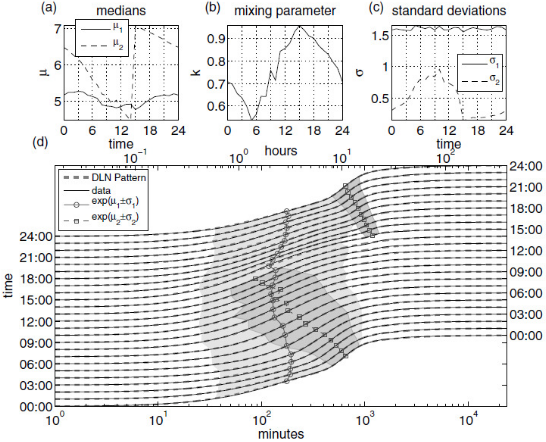


- LN model explains 83% of the posts
- Incorporating cycle in LN improves significantly
- DLNxC and DLN account form more than 99% of the data
- DLN accounts for the main part of variation caused by the circadian rhythm

Temporal patterns of Slashdot

Qualitative explanation: Two waves of activity

- First wave: locked to the post publication
- Second wave: depends on the publication hour
- Only the first wave is necessary in a short interval.



Temporal patterns of Slashdot

Popularity prediction

- At time t we want to **predict the number of comments** in the next s minutes of a post published x minutes ago and has received until now N comments
- Use available data window $[t-x, t]$ and predict the number of comments M in the prediction window $(t, t+s]$.

Challenges

- Large variability between posts
- Transient behaviour (sharp initial raise)
- Heavy tails: difficult to simply extrapolate based on evidence
- Limited information (no content)

Temporal patterns of Slashdot

Popularity prediction

- At time t we want to **predict the number of comments** in the next s minutes of a post published x minutes ago and has received until now N comments
- Use available data window $[t - x, t]$ and predict the number of comments M in the prediction window $(t, t + s]$.

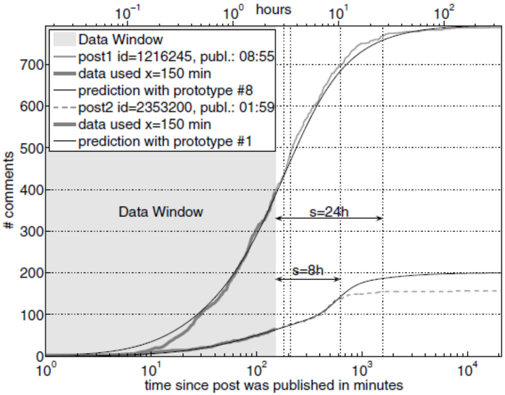
Methodology

- Compute DLN prototypes, one for every hour of the day
- Prediction is made by rescaling the corresponding prototype given the limited data window
- Use older posts (first months of data) as *training* set
- Error measure (relative):

$$\epsilon = |(M_{\text{predicted}} - M_{\text{real}})/M_{\text{real}}|$$

Temporal patterns of Slashdot

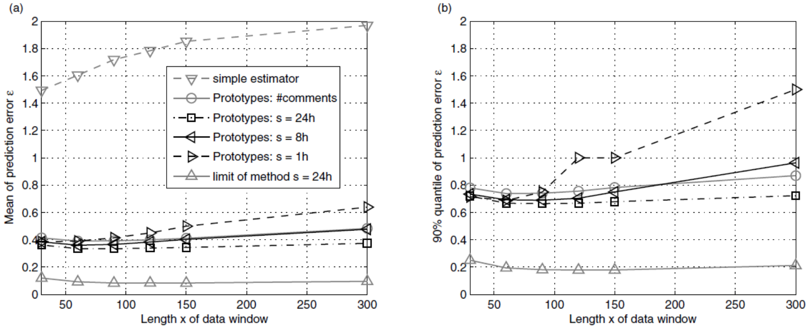
Popularity prediction: two illustrative examples



- Prediction of post1 is satisfactory at all times
- Prediction of post2 is satisfactory until 8 hours and overestimated afterward

Temporal patterns of Slashdot

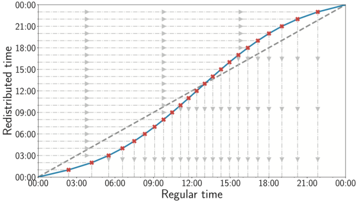
Popularity prediction: results



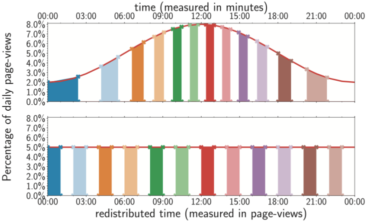
- Best results are obtained for a 24 hour prediction
- Num. comments more relevant than data window length
- Error increases in the tail: large number of posts with a very low number of comments in the prediction window

Alternative way to deal with Activity Cycles

Rescale Time



(b) Mapping time in minutes (t) to time in page-views (t^*). The gray arrows indicate the direction of the mapping.

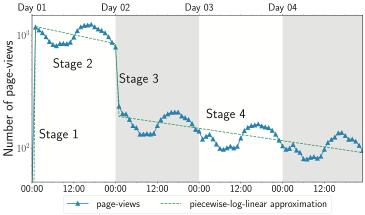


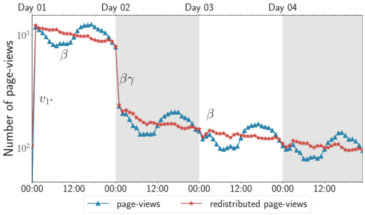
(c) Visualization of the effects of time redistribution.

- Image from (ten Thij et al., 2019)
- measure time in activity not in minutes

Alternative way to deal with Activity Cycles

Rescale Time





- Image from [ten Thij et al., 2019]
- Show regular decay of interest in new Items on Wikipedia's Featured Articles

Temporal patterns of Slashdot

Conclusions

- A parsimonious approach that disregards content is valid
- DLN distributions provide an excellent explanation for the reaction times
- Parameters have a nice interpretation: two waves of activity, each corresponding to a LN
- In some cases, this approach allows for reliable prediction based on limited amounts of data

Outline

Motivation

Scientific questions

Temporal patterns in news aggregators

[Kaltenbrunner et al, 2007]

- What are the temporal patterns governing these responses?
- Is there a mathematical law that describes this patterns?
- Can we use this law to predict number of votes (popularity) in the long term?

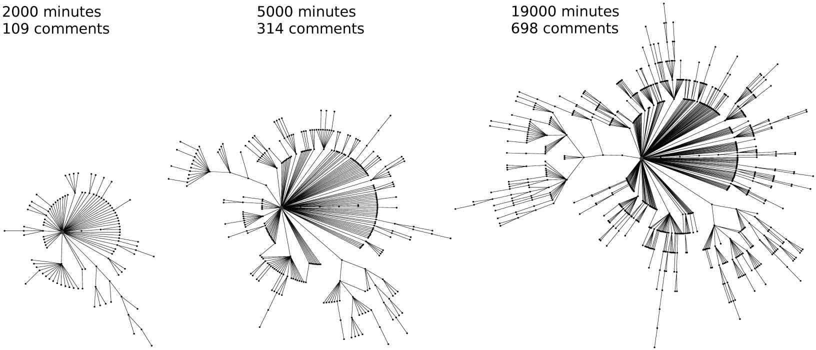
Structure and evolution of conversation threads

[Gómez et al, 2013]

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

Modeling conversation threads

Example of online conversation



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

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

Modeling conversation threads

Datasets:

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

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

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

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

Modeling conversation threads

General approach

- Suggest features based on prior empirical analysis
- Propose a generative model
- Learn the model parameters based on data
- Interpret, understand, predict the real system based on the learned parameters

Bottom-up

- Simple models are preferable (only a few features are relevant)
- First approach
  - Discard content, discard user network
  - Assume threads size is known

Modeling conversation threads

General approach:

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

Empirical facts

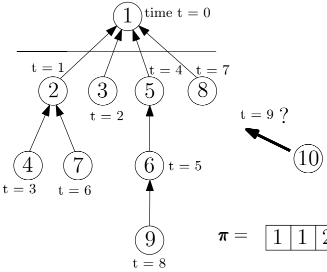
- Popular comments receive more replies: *preferential attachment*
- New comments are more *attractive* than old ones
- Replies to the post behave different than replies to comments

Modeling conversation threads

Representation of a conversation

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

$$\boldsymbol{\pi}_0 = ()$$
$$\boldsymbol{\pi}_1 = (1)$$
$$\vdots$$



$\boldsymbol{\pi} =$

1	1	2	1	5	2	1	6	?

Modeling conversation threads

Parameters of the model: popularity

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

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

- $d_{k,t}$ is weighted by α

Modeling conversation threads

Parameters of the model: novelty

- At time t, the **novelty** of node k is

$$n_{k,t} = \tau^{t-k+1}, \quad \tau \in [0,1]$$

- Captures an exponential decay of novelty

Modeling conversation threads

Parameters of the model: root bias

- The bias of a node k is is either zero or β for the root:

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

- Captures the different law governing post replies and replies to comments

Modeling conversation threads

Model definition

- We define a model by means of its associated *atractiveness* 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

- Full model (**FM**)

$$\phi(k) = \alpha d_{k,t} + b_k + \tau^{t-k+1}$$

- Parameters $\{\alpha, \tau, \beta\}$

Modeling conversation threads

Model definition

- We define a model by means of its associated *atractiveness* 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 without popularity model (**NO**-α)

$$\phi(k) = b_k + \tau^{t-k+1}$$

- Parameters $\{\tau, \beta\}$, $\alpha = 0$

Modeling conversation threads

Model definition

- We define a model by means of its associated *atractiveness* 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 without novelty (**NO**-τ)

$$\phi(k) = \alpha d_{k,t} + b_k + 1$$

- Parameters $\{\alpha, \beta\}$, $\tau = 1$

Modeling conversation threads

Model definition

- We define a model by means of its associated *atractiveness* 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 without bias (**NO-bias**)

$$\phi(k) = \alpha d_{k,t} + \tau^{t-k+1}$$

- Parameters $\{\alpha, \tau\}$, $\beta = 0$

Modeling conversation threads

Parameter estimation

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

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

Modeling conversation threads

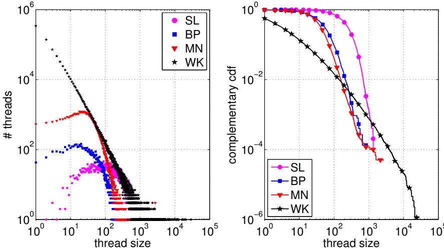
Parameter estimation

- Minimization problem

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

Modeling conversation threads

Global analysis of the data



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

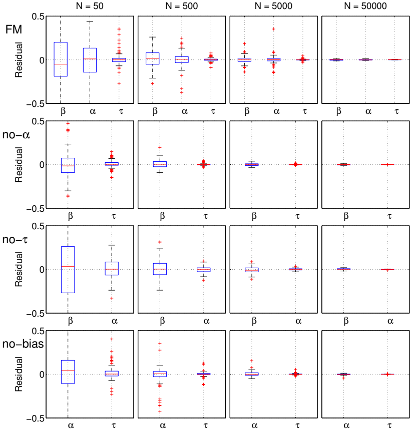
Parameter estimation

Validation

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

Modeling conversation threads

Validation



- Estimation is unbiased
- Good estimates can be obtained using $N = 50$

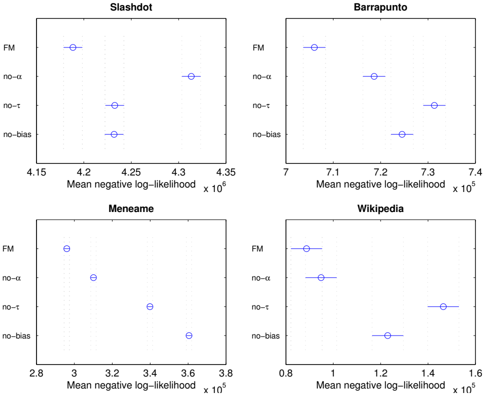
Modeling conversation threads

Model Comparison

For each dataset:

- Select N threads randomly with replacement
- Find estimates $\hat{\theta}$.
- Compute likelihoods

- Model comparison based on likelihoods for each dataset

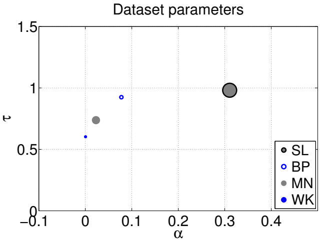


Modeling conversation threads

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



Modeling conversation threads

Validation of the model

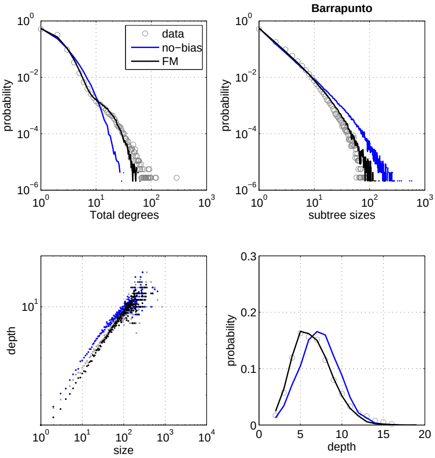
- Original data versus synthetic threads produced by the model
  - Degrees distribution
  - Subtree sizes distribution
  - Mean node depth versus size
  - Node depths distribution

Generating threads

- Threads sizes are drawn from the empirical distribution
- We use model **NO-BIAS** for comparison

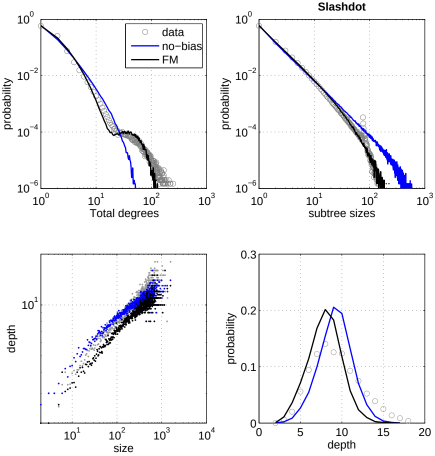
Modeling conversation threads

Barrapunto dataset



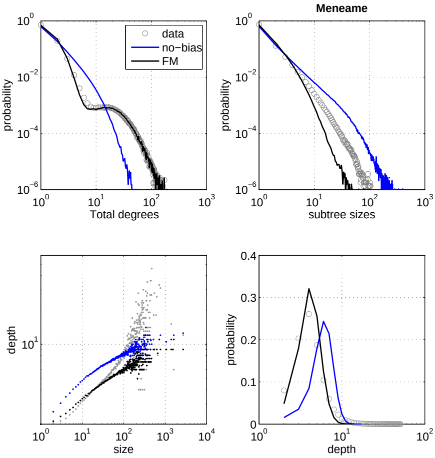
Modeling conversation threads

Slashdot dataset



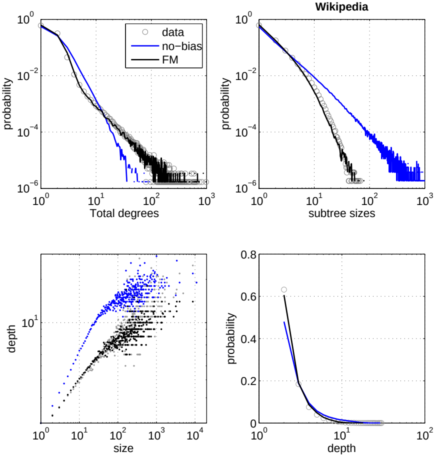
Modeling conversation threads

Meneame dataset



Modeling conversation threads

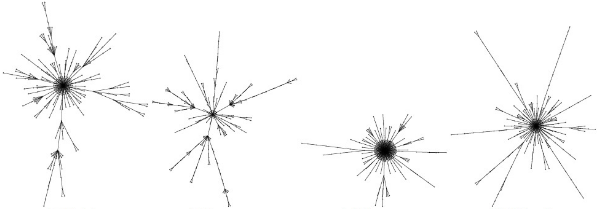
Wikipedia dataset



Growing tree model for conversation threads

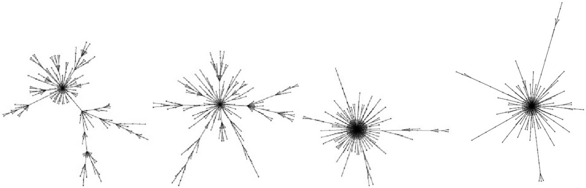
Comparison between real and synthetic threads

Real threads:



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

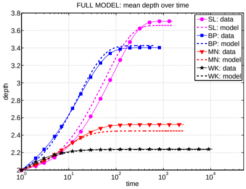
Synthetic threads:

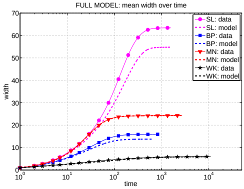


(a) Slashdot (b) Barrapunto (c) Meneame (d) Wikipedia

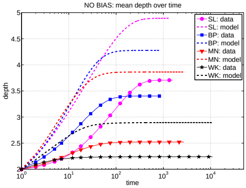
Evolution of mean depths and mean widths

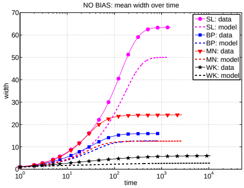
FULL MODEL:





NO-BIAS model:





Modeling conversation threads

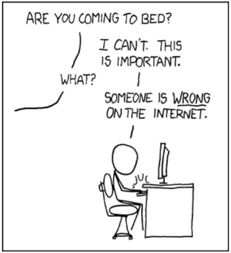
Adding authorship

Extending the model

- Main interest: understanding user behavior
- Is the author relevant to determine the structure of the discussion?
- Can we extend minimally the model to incorporate authorship?

Design choices

- User $\rightarrow$ Discussion?
- Discussion $\rightarrow$ User?
- Empirical observation: Reciprocity
  - User A tends to reply user B who previously replied to A



Modeling conversation threads

Adding authorship

Extending the model

- Two coupled processes
- Growing authorship vector $a_{1:t} = (a_1, a_2, \ldots, a_t)$
- In addition to $\pi_{1:t} = (\pi_1, \pi_2, \ldots, \pi_t)$
- At time $t + 1$
  - A new author is created with p_{new}
  - An existing author ν is chosen, otherwise
- If existing author ν, chosen according to the number of replies to ν in the thread, r_ν

$$p(a_{t+1} = v | a_{1:t}, \pi_{1:t}) = \begin{cases} p_{new}, & \text{for } v = U + 1 \\ \frac{(1-p_{new})2^{r_v}}{\sum_{i=1}^{U} 2^{r_i}}, & \text{for } v \in 1, \ldots, U \end{cases}$$

Modeling conversation threads

Adding authorship

Extending the model

- New reciprocity parameter κ, $\theta' = (\alpha, \tau, \beta, \kappa)$
- Extended attractiveness function $\phi'_j(\cdot)$

$$\phi'_j(\pi_{1:t}, a_{1:t}; \theta') := \phi_j(\pi_{1:t}; \theta) + \kappa\delta_{a_{\pi_j}, a_{t+1}}$$

- Leads to the extended full model

$$p'(\pi_{t+1} = j | \pi_{1:t}, a_{1:t}; \theta') \propto \phi'_j(\pi_{1:t}, a_{1:t}; \theta')$$

- Only when $a_{\pi_j} = a_{t+1}$, κ-term

  $\kappa = 0$: the new feature will play no role

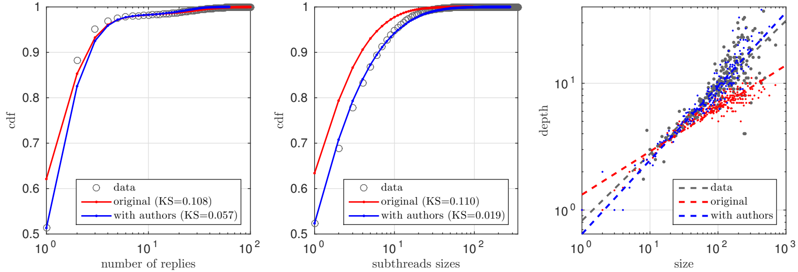
  $\kappa \gg 0$: all comments reciprocal

- Optimization of θ' using maximum likelihood

Modeling conversation threads

Adding authorship

Model comparison (degrees, subthread sizes, depth vs size)

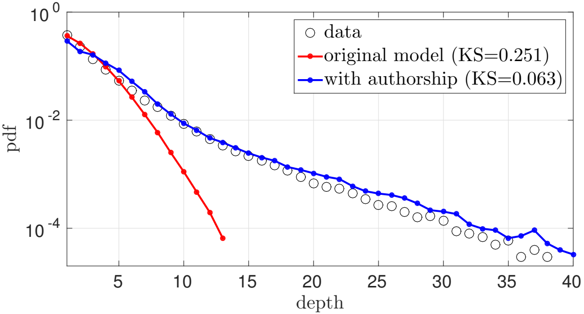


- Features are reproduced better thanks to the authorship model and the reciprocity feature

Modeling conversation threads

Adding authorship

Model comparison (thread depths)



- original model is FULL model
- Extended model reproduces the long tail created by reciprocal message chains accurately

Conclusions and current directions

Conclusions

- Framework which allows to re-create conversations with similar structural features as real instances
- Model captures the large heterogeneity of the data
- Parameters allow to characterize audience and platform:
  - Same platform : differences between SL and BP
  - Influence of the interface: MN (flat) characterized by bias
  - Main difference between news media and WK: popularity
- A minimal increase in complexity (authorship and reciprocity) greatly improves the overall descriptive power of the model

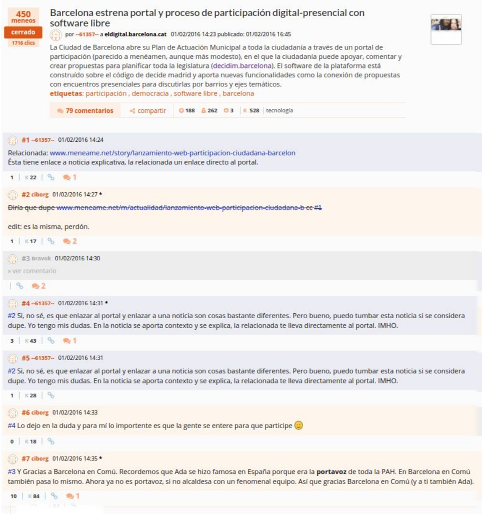
Application : Evaluation of platform design

- Can be used to assess the impact of a given design element on the user interaction patterns on a platform.
- Shows the interdependency between user interaction patterns and platform design elements.
- Can be exploited to help site owners and community managers to create a positive and constructive environment for large scale online discussions.

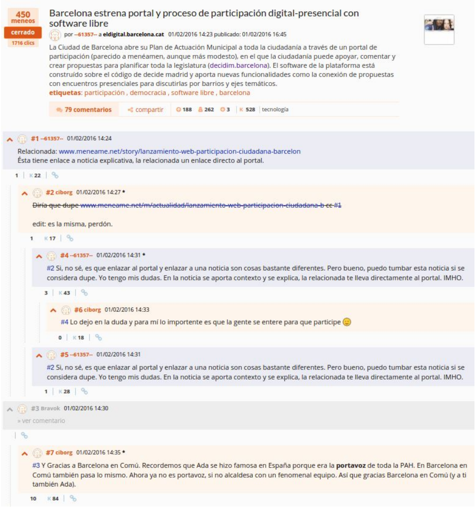
Application: Evaluation of platform design

Example: Change of how conversation threads are presented

- **Aragón et al. [2017]** analyze the impact of threaded vs. non-threaded conversation views

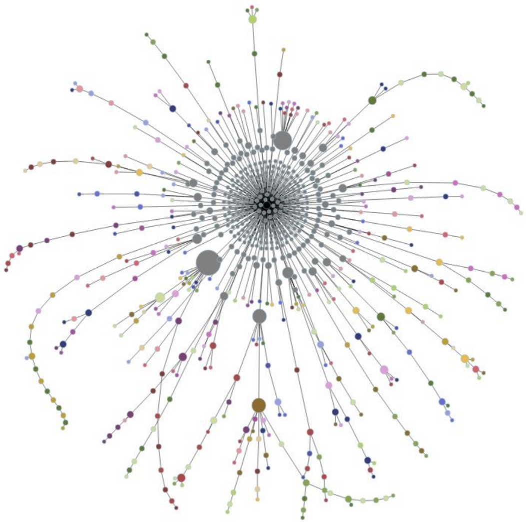


Aragón P., Gómez V., Kaltenbrunner A. (2017) To Thread or Not to Thread: The Impact of Conversation Threading on Online Discussion, ICWSM-17, Montreal, Canada.

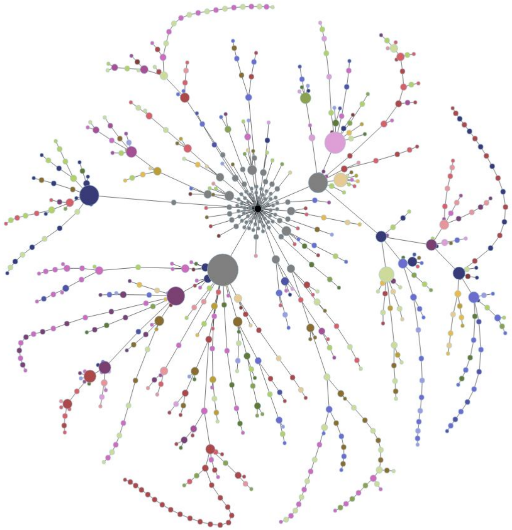


Application: Evaluation of platform design

Aragón et al. [2017] Visual differences visible



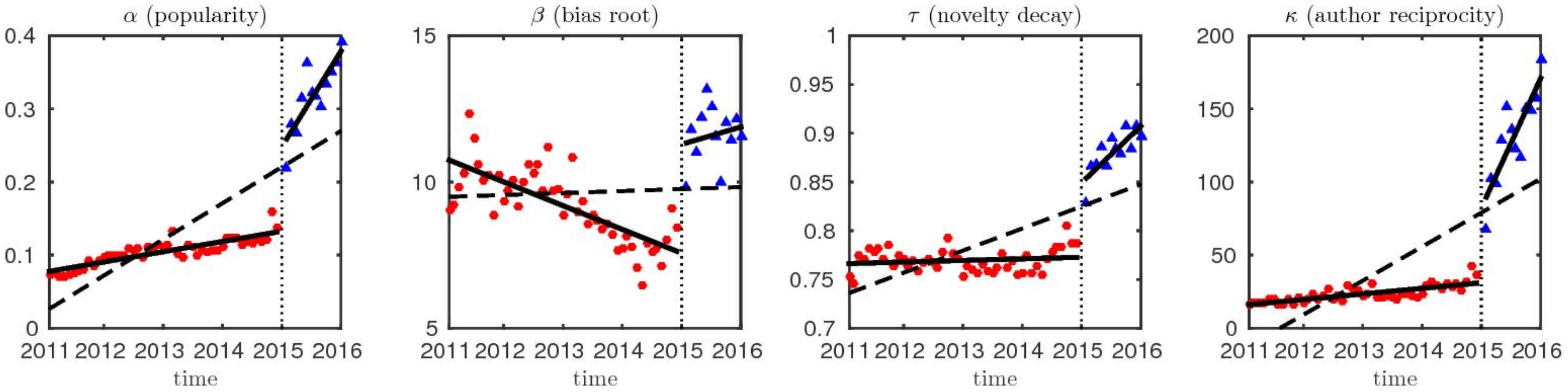
Thread in 2013
(linear view)



Thread in 2015
(hierarchical view)

Application: Evaluation of platform design

- **Aragón et al. [2017]** Behavioural features of a generative model undergo an notable increase when conversation threading is released (Jan 2015)
- Change in design can be detected with Regression Discontinuity Design applied on model parameters



Open challenges

- Competition between discussion threads
- Impact of sub-communities
- The role of content
- Influencing user activity

A related Tutorial

- Generative models of online discussion threads

  https://www.upf.edu/web/ai-ml/tutorial-ICWSM

- Related code (in R)

  https://github.com/alumbreras/discussion-threads

Bibliography I

- P. Aragón, V. Gómez, A. Kaltenbrunner.
  To Thread or Not to Thread: The Impact of Conversation Threading on Online Discussion.
  International AAAI Conference on Web and Social Media (ICWSM), 2017.
- V. Gómez, H. J. Kappen, N. Litvak & A. Kaltenbrunner.
  A likelihood-based framework for the analysis of discussion threads.
  World Wide Web, vol. 16, no. 5-6, pages 645-675, 2013.
- A. Kaltenbrunner, V. Gómez & V. López.
  Description and Prediction of Slashdot Activity.
  In Proceedings of the 5th LA-WEB 2007, IEEE Computer Society.
- M. ten Thij, A. Kaltenbrunner,D. Laniado, Y. Volkovich.
  Collective attention patterns under controlled conditions.
  Online Social Networks and Media (OSNEM), 13, 100047, 2019.

Bibliography II

Additional references

- P. Aragón, V. Gómez, D. García & A. Kaltenbrunner.
  Generative models of online discussion threads: state of the art and research challenges.
  Journal of Internet Services and Applications, 8(1) 15 (2017).
- P. Aragón, V. Gómez & A. Kaltenbrunner.
  To Thread or Not to Thread: The Impact of Conversation Threading on Online Discussion.
  ICWSM-17, Montreal, Canada (2017).
- A. N. Medvedev, R. Lambiotte, JC. Delvenne.
  The Anatomy of Reddit: An Overview of Academic Research
  Dynamics On and Of Complex Networks III pp 183-204 (2017).
- J. Bollenbacher, D. Pacheco, P. Hui, Y. Ahn, A. Flammini & F. Menczer
  On the challenges of predicting microscopic dynamics of online conversations
  Applied Network Science volume 6, Article number: 12 (2021).