World Wide Web
DOI 10.1007/s11280-012-0162-8

A likelihood-based framework for the analysis of discussion threads

Vicenç Gómez · Hilbert J. Kappen · Nelly Litvak · Andreas Kaltenbrunner

Received: 1 July 2011 / Revised: 16 March 2012 /
Accepted: 29 March 2012

Abstract Online discussion threads are conversational cascades in the form of posted messages that can be generally found in social systems that comprise many-to-many interaction such as blogs, news aggregators or bulletin board systems. We propose a framework based on generative models of growing trees to analyse the structure and evolution of discussion threads. We consider the growth of a discussion to be determined by an interplay between *popularity*, *novelty* and a trend (or *bias*) to reply to the thread originator. The relevance of these features is estimated using a full likelihood approach and allows to characterise the habits and communication patterns of a given platform and/or community. We apply the proposed framework on four popular websites: *Slashdot*, *Barrapunto* (a Spanish version of Slashdot), *Meneame* (a Spanish *Digg*-clone) and the article discussion pages of the English *Wikipedia*. Our results provide significant insight into understanding how discussion cascades grow and have potential applications in broader contexts such as community management or design of communication platforms.

Keywords discussion threads · online conversations · information cascades · preferential attachment · novelty · maximum likelihood · Slashdot · Wikipedia

V. Gómez (✉) · H. J. Kappen
Donders Institute for Brain Cognition and Behaviour,
Radboud University Nijmegen, 6525 EZ Nijmegen, The Netherlands
e-mail: v.gomez@science.ru.nl

H. J. Kappen
e-mail: bertk@science.ru.nl

N. Litvak
Faculty of Electrical Engineering, Mathematics and Computer Science, Department of Applied Mathematics, University of Twente, Enschede, The Netherlands
e-mail: n.litvak@ewi.utwente.nl

A. Kaltenbrunner
Social Media Research Group, Barcelona Media, Barcelona, Spain
e-mail: andreas.kaltenbrunner@barcelonamedia.org

Published online: 15 April 2012

1 Introduction

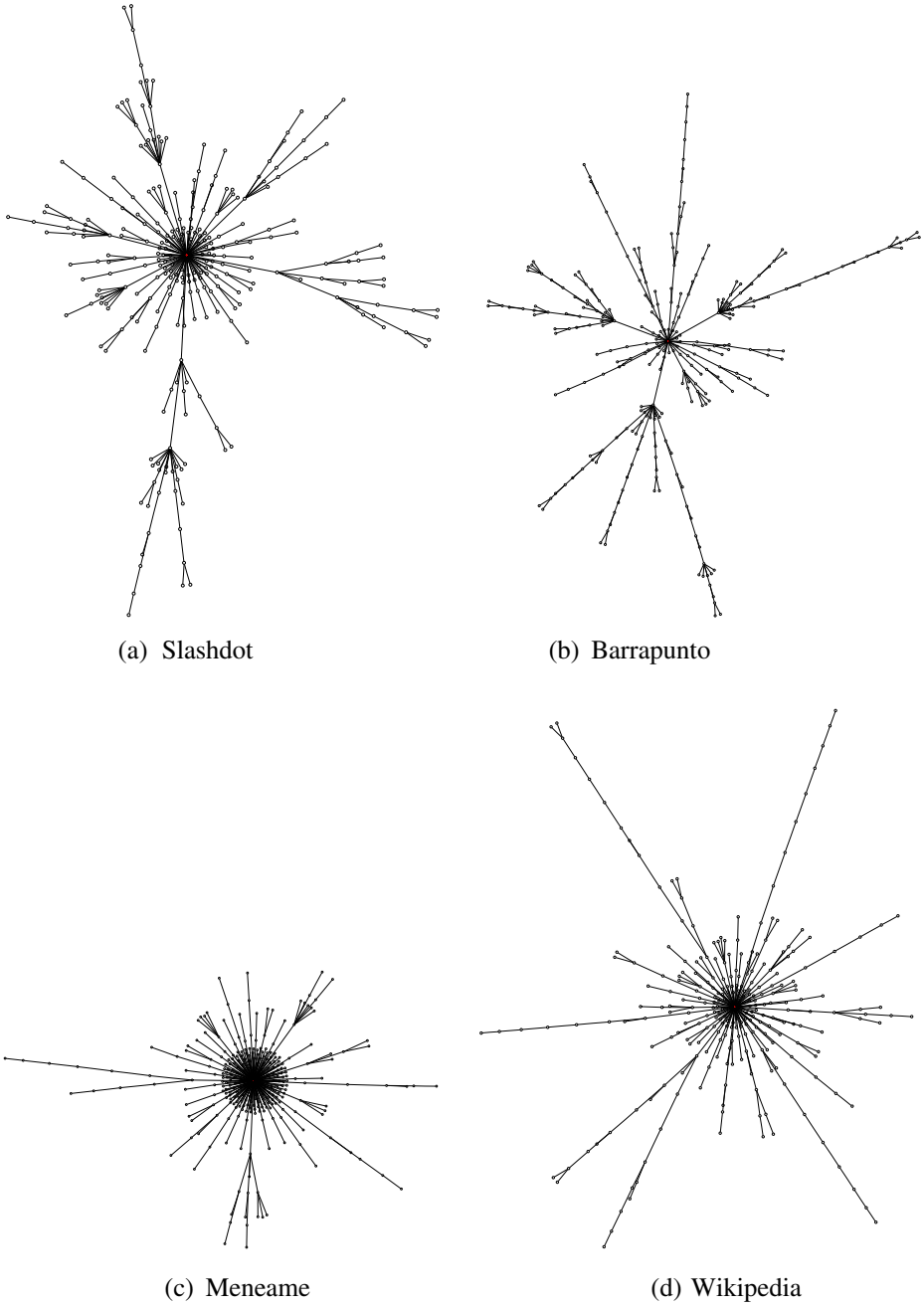
Nowadays, online platforms where users interchange messages about a topic of interest are ubiquitous on the Internet. Examples range from online message boards, blogs, newsgroups, or news aggregators to the discussion pages of the Wikipedia. A discussion typically starts with a broadcasted posting event that triggers a chain reaction involving some users who actively participate in the cascade.

Unlike other types of information cascades, such as those corresponding to massively circulated chain letters [36], Twitter [30], photo popularity on Flickr [10] or diffusion of pages on Facebook [49], where a small piece of information is just forwarded from one individual to another, discussion threads involve a more elaborated interaction between users, with uncertain (and possibly multiple) directions of information flow, more similar, for instance, to the cascades extracted from phone calls [41]. Since threaded discussions are in direct correspondence with the information flow in a social system, understanding their governing mechanisms and patterns plays a fundamental role in contexts like the spreading of technological innovations [44], diffusion of news and opinion [20, 35], viral marketing [34] or collective problem-solving [27].

What determines the growth of a discussion thread? How to predict which comment will elicit the next reply? Is there a simple mechanism that can capture the structure and evolution of online discussions? To answer these questions, it is usually believed that popular comments attract more replies, which in turn increases their popularity, so a *rich-get-richer* phenomenon seems to play an important role. On the other hand, given the transitory character of certain fads, our interest decays with time, and novelty also appears to be fundamental in determining our attention [57].

In this work, we propose a modelling framework which focuses on structural aspects of discussion threads and sheds light on the interplay between popularity and novelty. We introduce a parametric generative model which combines three basic features: *popularity*, *novelty* and a trend (or *bias*) to reply to the thread originator. We show that a model which combines these three ingredients is able to capture many of the statistical properties, as well as the thread evolution, in four popular and heterogeneous websites. We also use statistical tests to analyse the impact of neglecting one of these basic features on the explanatory power of the model. In this way, we are able to make statistical inferences that can assess, for instance, whether popularity is significantly more relevant than novelty in a given web-space.

To illustrate and validate the framework, we consider four popular websites: *Slashdot*, *Barrapunto* (a Spanish version of Slashdot), *Meneame* (a Spanish *Digg*-clone) and the article discussion pages of the English *Wikipedia*. These datasets are quite heterogeneous (see Figure 1 for a typical thread example of each dataset). For instance, whereas the first three websites can be classified as news aggregators, Wikipedia discussion pages represent a collaborative effort towards a well-defined goal: producing a free, reliable article. Also, at the interface level, while Slashdot and Barrapunto provide the same hierarchically threaded interface, Meneame provides a linear (flat) view which allows users to reply to other users via a tagging mechanism only. Using the same model for the four datasets, we can segregate the heterogeneities, which are captured via the corresponding parameter values.



(a) Slashdot (b) Barrapunto

(c) Meneame (d) Wikipedia

Figure 1 Examples of discussion threads. The illustrations represent the structure of the discussion after removing the content of the messages. The central node corresponds to the main post (news article) and the rest of the nodes to regular comments which are attached as replies to the main post or to other existing comments. Each figure corresponds to a discussion selected randomly from each of the four websites considered in this study.

1.1 Motivation and methodology

The aim of the present work is to propose a quantitative framework for the statistical analysis of online discussion threads. For that, we propose a model that can reproduce the structural and evolving patterns of the discussion threads of a particular website or platform. The model considers little semantic information. In particular, the discussion thread is treated as a growing network where nodes correspond to messages and links to reply actions. The growing networks are therefore the discussions themselves, and *not* subgraphs of an underlying social network. Identifying a valid generative model for this type of networks that disregards the content helps to find meaningful regularities which uncover universal patterns and provide a fundamental understanding of users' communication habits.

The model is a stochastic process that assumes that such patterns can be reproduced by means of three simple features: popularity, novelty and a trendiness to the thread initiator. Other aspects such as the dynamics of an underlying social network or the precise temporal timings (termination criteria) for the discussions are not included. These aspects could be built "on top" of the current framework a posteriori.

Associated to each feature, there is a parameter that captures its relative influence, and that depends on the particular website or platform under consideration. The framework includes a parameter estimation procedure that can be performed independently for each dataset (a collection of discussion threads). Parameter estimation is based on the likelihood of the entire evolution of each single thread of a given dataset, providing the parametrised model which globally fits the data best. This prevents over-fitting, that is, reproducing very accurately particular quantities such as the number of replies per comment (the degrees of the nodes in network terminology), at the cost of poor approximation of other quantities. We show that the estimation procedure is robust, in the sense that optimal parameters are not biased, and does not require very large datasets to be estimated.

Parameter estimates have descriptive power, since they allow the habits and communication patterns of a given dataset to be characterised in terms of the aforementioned features. They can also be used to establish differences between topics or users groups within a given website. The relevance of each of the features is determined by the framework via model comparison, that is, comparing the likelihood of the general model that includes all the three features with the three reduced variants of the model that omit one of them.

1.2 Related work

There is an overwhelming and vastly growing amount of literature related to online discussion threads. In this section, we review some of the most related papers.

1.2.1 Data analysis of threaded conversations

Data from threaded conversations has been used extensively to characterise human behaviour, for instance, to quantify how moderation affects the quality [31] and to detect social roles in Usenet [9, 13]. Usenet is considered the first message board and the precursor of Internet forums. The first large-scale empirical analysis of Usenet threads was developed in [55]. The authors reported significant heterogeneities in the levels of user participation and thread depths, and determined meaningful

correlations between certain indicators such as message sizes, thread depths and cross-posting between groups.

Typical user behaviours have been characterised, for example, behaviours that are dominated by responding to questions posed by other users ("answer persons") on Usenet [54] or lurking behaviour (the act of reading but rarely posting on forums) on MSN bulletin board communities [40, 42]. The factors that cause users who initially posted to an online group to again contribute to it were analysed for different platforms in [24]. In contexts of knowledge sharing, such as Yahoo answers, interesting patterns have been found which differentiate between discussion- and question-answer forums and their relation with the different levels of specialisation of the users [1]. These studies have important implications for cultivating valuable online communities.

From a social network perspective, discussion threads comprise user interaction from which a social network can be extracted and analysed [39]. The reply networks emerging from the comment activity (that link two users according to their interaction) have been analysed for bulletin board systems [14, 58], Slashdot [16], Digg [43] or Wikipedia [32]. Although global network features of these networks show only minor discrepancies to other social networks, e.g. friendship networks, a rigorous comparative analysis revealed fundamental differences in the practise of establishing reply and friendship links in the case of Meneame, a Digg-like website [26].

At the thread-level, visualisation techniques of the conversations have facilitated the understanding of the social and semantic structures [46, 48]. Statistical analysis of the threads has made it possible to identify the distinctive properties of online political conversations in relation to other types of discussions [18], or to derive measures that can improve the assessment of information diffusion [38], popularity prediction [25, 33, 50] or controversy [16].

1.2.2 Information cascades

Recently, the term *information cascades* has been adopted to describe similar phenomena. It has been introduced in the economic sciences for the analysis of herding behaviour, when an individual adopts/rejects a behaviour based on the decisions of other individuals [4, 7]. In the case of discussion threads, a user adopts a behaviour by actively participating in the conversation.

The increasing availability of electronic communication data has prompted extensive empirical work on information cascades. The diffusion patterns seem to depend on the nature of the cascades under consideration. For instance, while a Twitter study suggested that cascades spread very fast and are predominantly shallow and wide [30], photo popularity on Flickr seems to spread slowly and not widely [10]. Another study found that fan pages on Facebook are triggered typically by a substantial number of users and are not the result of single chain-reaction events [49].

Empirical analysis of email threads has been the subject of intense analysis and controversy. The diffusion patterns of two large-scale Internet chain-letters were analysed in [36]. The authors concluded that, rather than fanning out widely and reaching many people in a few steps, chain-letters propagate in a narrow and very deep tree-like pattern. This result seems to contradict the way a small-world network would operate, and several hypotheses have been proposed to account for this observation while preserving the small-world intuition. One of them is the *selection*

bias hypothesis, which states that the observed structures may not be typical instances of the processes that generated them, but instead exceptional realisations [15]. Recently, another study [52] reported that cascades composed of forwarded emails fan out widely and quickly die out. Despite the differences of the different studies, however, certain regularities are pervasive in all datasets, for instance, that the largest cascades occur with very small probability and affect a very small proportion of the whole population.

Theoretical model analysis to understand these phenomena usually considers the underlying networks where cascades originated. In [53] two cascading regimes which show rare but very large cascades are identified, depending on the network connectivity: for sufficiently sparse connectivity, cascade sizes follow a power-law distribution at a critical point, while for sufficiently dense connectivity, cascade sizes follows a bimodal distribution. The analysis also concludes that endogenous heterogeneities in the underlying network (high threshold or degree variability) has mixed effects on the likelihood of observing global cascades.

Attempts to find the underlying connectivity of associated networks using epidemic models have been made using data from blogs [2, 20, 29]. The conversation cascades of blogs have been considered in [35], with special emphasis on the scale-free character of related distributions such as cascade sizes or degree distributions. A simple, parameter-free model able to generate power-law distributed cascade sizes and temporal patterns resembling the real-world ones was proposed in [19]. However, the role the underlying social network plays in information diffusion also seems to be dependent on the particular domain. Whereas a study about social influence concluded that diffusion of content strongly depends on the network topology [3], email forwarding seems to be less dependent [52]. Despite the existing discrepancy about the role of network topology, it is believed that network topology strongly determines diffusion at a microscopic level, in the beginning of the cascade only. At a macroscopic scale, after a critical propagation threshold is reached, network topology does not seem to be much relevant.

If one disregards the underlying social network which generates the cascades, the simplest phenomenological model for cascades is a branching process, where a random number of descendants is generated at each time step (or generation), for each node, according to a fixed probability distribution which is equal for all nodes in the cascade. Galton-Watson processes are a particular type of branching processes, and have been suggested in [15] to support the selection bias hypothesis, in [47] to account for missing information in the cascades and in [52] as baseline models. However, branching processes are insufficient for our purposes, mainly because they assume each node (comment) to be independent, and therefore do not provide a basis for the evolution of the cascade. Instead, we are interested in the stochastic process governing the cascade growth.

1.2.3 Discussion threads as information cascades

As stated previously, the aforementioned cascades involve the forwarding of a piece of information from one individual to another one. Discussion threads involve a more elaborated interaction between users, with uncertain (and possibly multiple) directions of information flow. More recently, [28] proposed a model for conversation threads which combines popularity and novelty. The model improves on the simple branching process and qualitatively reproduces certain statistical properties of the

resulting threads and authorships, and is illustrated using data from three popular forums, with special emphasis on Usenet. Similarly, [17] showed that a growth model based on a modified preferential attachment which differentiates between the root of the thread and the rest of the nodes captures many statistical quantities associated with the structures and the evolution of the empirical threads.

We build on these previous works and compare extensions of both models, providing a unifying likelihood-based framework for their parameter estimation and validation. Our approach allows the interplay of the different parameters to be analysed and is validated in detail for four datasets.

1.3 Outline

In the next section we introduce our framework and present growth models of discussion threads and their parametrisation. Section 3 describes our likelihood approach for parameter estimation and its validation. In Section 4, we present the empirical results of this work. First, we provide a global description of the threading activity in the four datasets under study in Sections 4.1 and 4.2. Our analysis also highlights the importance of repetitive user participation in relation to other types of cascades and their impact on the entire social network. We compare the explanatory power of the different proposed models in Section 4.3. Validation of the structure and evolution of the model generated threads is analysed in detail in Sections 4.4 and 4.5 respectively. Finally, in Section 5 we discuss the results and implications of this work. In the appendices we provide an analytical deviation of the limit behaviour of the proposed models and details of the parameter estimation procedure.

2 Growing tree models for discussion threads

Before we introduce the formal model, we provide first the required mathematical terminology. We consider an abstract representation of a discussion thread as a graph, where nodes correspond to comments and links between nodes denote reply actions. The initial (root) node has a special role: it corresponds to the triggering event of the discussion (a news article, for instance) and we will refer to it in what follows often as the "post". We model the growth of such a graph, in which new nodes are added sequentially at discrete time-steps. We consider the case that comments

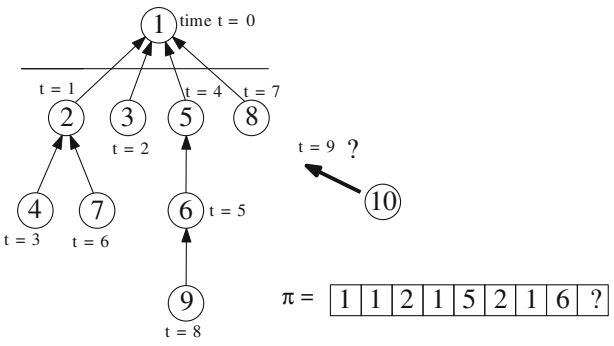


Figure 2 Small example of a discussion thread represented as a tree: at time-step $t = 9$, node (comment) number 10 is added to the thread. At the *bottom right* we show the corresponding vector of parents $\boldsymbol{\pi}$. Each node attracts the new comment with different probability according to the model under consideration (see text).

are single-parental, that is, the same comment cannot be a reply of more than one comment. In this way, the resulting graph is a tree (it does not contain cycles).

A compact way to represent trees consists of a vector of parent nodes that we denote by $\boldsymbol{\pi}$. We use the indices of the vector $\boldsymbol{\pi}$ as the identifiers of the comments and elements of $\boldsymbol{\pi}$ correspond to identifiers of the replied comments. In this way, π_t denotes the parent of the node with identifier $t+1$, which was added at time-step t. The total number of nodes, or size of the discussion, is denoted by $|\boldsymbol{\pi}|$. See Figure 2 for an illustration.

The growth of the tree is characterised by the probability that existing nodes attract new ones. Thus we are interested in the probability of node k being the parent π_t of node $t+1$ given the past history, that we denote as the vector $\boldsymbol{\pi}_{(1:t-1)}$. Such probability can be written as $p(\pi_t = k|\boldsymbol{\pi}_{(1:t-1)})$, for $t > 1$, $k \in \{1, \ldots, t\}$. The vector $\boldsymbol{\pi}$ at time-step $t = 1$ contains only the first reply to the root and is denoted as $\boldsymbol{\pi}_{(1)} = (1)$.[1] Note that by construction, $\pi_t \leq t, \forall t$.

We define a growing tree model by means of its associated *attractiveness* function $\phi(k)$ (to be defined later) for each of the nodes. Generally:

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

where for clarity we have omitted the dependency of $\phi(k)$ and Z_t on the thread history $\boldsymbol{\pi}_{(1:t-1)}$. The term Z_t is just a normalisation sum which ensures that at every time-step, the probability of receiving a reply is normalised and adds up to one.

Once we have introduced the stochastic process governing the thread evolution, we present the three features that determine the attractiveness of a node.

Popularity Comments receive new replies depending on how much replies they already have. This mechanism, known as preferential attachment (PA) or as *Mathew effect* in social sciences, has a long tradition to characterise many types of complex networks. Its origins date back to the early twentieth century [12]. More recently, PA became popularised in the model of Barabási [5] to explain the scale-free nature of degree distribution in complex networks.

At time t, we relate the *popularity* of a comment with its number of occurrences in the vector of parents. Mathematically, the degree of a node k is its number of links (degree $d_{k,t}$) before node $t+1$ is added:

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

where δ is the Kronecker delta function. In the following, we omit the explicit dependence on $\boldsymbol{\pi}_{(1:t-1)}$, so that $d_{k,t} \equiv d_{k,t}(\boldsymbol{\pi}_{(1:t-1)})$. Note that we consider an undirected graph, and every existing node has degree equal to one initially.

To parametrise the popularity, we introduce a weight α common to *all* the degrees. This factor captures the relevance of the popularity during the growth of the tree, so a value of α very close to zero would mean almost no influence of popularity and its relevance will be proportional to α. This model corresponds to a *linear* PA model.

[1] At time 0 we have $\boldsymbol{\pi}_0 = ()$ and for all trees, $p(\pi_1 = 1) = 1$ and 0 otherwise, i.e. $\boldsymbol{\pi}_1 = (1)$ always.

Novelty Either because of saturation or competition, old comments gradually become less attractive than new ones. We model the novelty of comment k as an exponentially decaying term:

$$n_{k,t} = \tau^{t-k+1}, \qquad \tau \in [0, 1]. \tag{3}$$

Note that empirical evidence exists that novelty in online spaces decays slower than exponentially [22, 57] and is strongly coupled with circadian rhythms [37]. Decay in novelty also depends on the data, e.g. news fade away rapidly compared with video popularity [50]. However, since we use comment arrivals as time units, these heterogeneities are alleviated and an exponential decay is justified [51]. In [28] the same mechanism is also proposed.

Root bias Finally, we explicitly distinguish between the root node of a thread and the regular comments. On many platforms users are more inclined to start a new sub-thread than to reply to an other comment. A convenient way to establish such a difference is to assume that the root node has an initial popularity, parametrised with β, which acts as a bias. The bias of a node k is either zero or β:

$$b_k = \begin{cases} \beta & \text{for } k = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

2.0.1 Attractiveness function

We define the attractiveness $\phi(k)$ of a comment k as the sum of the previous parametrised features. The interplay or relative importance between them is determined by the concrete values (to be estimated given the data) of the different parameters: α, τ or β.

We propose a model that combines all the features, and name it full model (FM). For comparison, we also consider three reduced variants which miss one of them. We denote the model without popularity as NO-α, the model without novelty as NO-τ and the model without bias as NO-bias. According to our formulation, the three reduced models are nested within the full model. Table 1 shows the four models we consider and their respective attractiveness function $\phi(k)$ together with the corresponding parameter set and constraint. Note that the NO-τ only requires the knowledge of node degrees and the distinction between the root and the rest of the nodes. Including the novelty term τ makes the process dependent on the full past history.

Other variants of this model are possible: in [17] popularity is modelled as a sub-linear PA process where the parameter α is exponentiating the degree and no novelty term exists. We found no significant differences between the FM model introduced here and a more general model with an extra parameter exponentiating the degrees. Thus the conclusions derived here are general and do not depend on whether a linear

Table 1 The four models considered in this work.

Model	Attractiveness function $\phi(k)$	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)	$\alpha d_{k,t} + \tau^{t-k+1}$	$\{\alpha, \tau\}$	$\beta = 0$

o sub-linear PA process is used to model popularity. The proposed formulation is more convenient mathematically, since the normalisation constant Z_t does not depend on the particular structure of the thread. For the FM, we have:

$$Z_t = \sum_{l=1}^{t} \alpha d_{l,t} + b_l + \tau^{t-l+1} = 2\alpha(t-1) + \beta + \frac{\tau(\tau^t - 1)}{\tau - 1}. \tag{5}$$

This allows to derive the asymptotic properties of certain quantities of interest, such as degree distributions. If one neglects the bias to the root node and considers instead a termination parameter γ independent of the thread structure, one recovers the T-MODEL proposed in [28], which is also based on a linear PA. The NO-bias model can thus be used to illustrate the T-MODEL in the datasets considered here.

3 Likelihood-based approach for parameter estimation

We explain here our approach to find parameter estimates given a set of data. In the following, a dataset denotes a generic collection of threads. It can include the entire set of conversations extracted from a particular website such as Wikipedia, but also conversations focused on particular topic domain, for instance, the domain *science* in Slashdot.

Typically, existing approaches for parameter estimation of evolving graph models require certain assumptions to be hold. For instance, the parameters of a PA process in large networks are usually measured by calculating the rate at which groups of nodes with identical connectivity form new links during a small time interval Δt [8, 23]. However, this approach is suitable only for networks with many nodes that are stationary in the sense that the number of nodes remain constant during the interval Δt. This is not a reasonable assumption in our data, which is often produced by a transient, highly non-stationary response.

Another approach for parameter estimation relies on fitting a measured property, for instance the degree distribution, for which an analytical form can be derived in the model under consideration. For the PA model, extensive results exist with emphasis precisely on the degree distributions [6, 45]. Following the standard heuristic (see e.g. [21, Chapter 8]), we obtain the following power law behaviour of the degree distribution in FM:

$$c_1 x^{-2} \leq P(\text{degree} \geq x) \leq c_2 x^{-2}, \quad 0 < c_1 < c_2, \tag{6}$$

where c_2 depends strongly on τ. The derivation of this result is provided in Appendix A. We see that the power law exponent of the cumulative distribution function equals 2 and does not depend on the model parameters. Furthermore, we see from the derivation that the difference between c_1 and c_2 can be several orders of magnitude. Thus, our results, obtained by existing analytical methods, are too rough to enable a statistical evaluation of τ. Finally, the parameter β does not affect (6), but we will see from the experiments that this parameter defines the shape of the distribution for lower values of the degrees. We note that the analytical derivation for the NO-τ model leads to the power law exponent $2 + 1/\alpha$, which does depend on α but this dependence is not prominent enough to accurately evaluate α from the power law exponent estimation on the data.

Our approach considers instead the likelihood function corresponding to the *entire* generative process (instead of particular measures such as degree distributions or subtree sizes) introduced before. We can assign to each observation (each node arrival in each thread) a given probability using (1). The parameters for which the probability of the observed data is maximised are the ones that best explain the data given the model assumptions (see [56] for a similar approach for other network growth model).

Formally, we observe a set $\Pi := \{\boldsymbol{\pi}_1, \ldots \boldsymbol{\pi}_N\}$ of N trees with respective sizes $|\boldsymbol{\pi}_i|$, $i \in \{1, \ldots N\}$ and we want to obtain estimates $\hat{\boldsymbol{\theta}}$ which best explain the data Π. If we assume that the threads in the dataset are independent and identically distributed, the likelihood function can be written as:

$$\begin{aligned}\mathcal{L}(\Pi|\boldsymbol{\theta}) &= \prod_{i=1}^{N} p(\boldsymbol{\pi}_i|\boldsymbol{\theta}) \\ &= \prod_{i=1}^{N}\prod_{t=2}^{|\boldsymbol{\pi}_i|} p(\pi_{t,i}|\boldsymbol{\pi}_{(1:t-1),i}, \boldsymbol{\theta}) \\ &= \prod_{i=1}^{N}\prod_{t=2}^{|\boldsymbol{\pi}_i|} \frac{\phi(\pi_{t,i})}{Z_{t,i}} \quad , \end{aligned} \tag{7}$$

where $\boldsymbol{\pi}_{(1:t-1),i}$ is the vector of parents in the tree i after time $t-1$ and $Z_{t,i}$ is the normalisation constant Z_t for thread i. We can apply this approach to each of the model variants presented in the previous section by choosing the attractiveness function ϕ accordingly. Numerically, instead of maximising (7) directly, it is more convenient to use the log-likelihood function. We consider the following error function to be minimised:

$$-\log \mathcal{L}(\Pi|\boldsymbol{\theta}) = -\sum_{i=1}^{N}\sum_{t=2}^{|\boldsymbol{\pi}_i|} \phi(\pi_{t,i}) - \log Z_{t,i}. \tag{8}$$

3.1 Validation of the maximum likelihood estimation procedure

In this subsection we show numerically that the parameter estimates found using the previously described optimisation are correct, i.e. not biased. We proceed as follows: for a given model, we choose randomly real parameter values $\boldsymbol{\theta}^*$. We use them to generate a set of synthetic threads (the set Π previously described). Then, we calculate the estimated parameters $\hat{\boldsymbol{\theta}}$ using the synthetic set via minimisation of (8). We repeat this procedure 100 times for different sizes N of the synthetic set. If the residuals (defined as $\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}$) go to zero as a function of the number of threads generated N, our estimates are non-biased.

We also tested for different local minima using five different random initialisations for a given $\boldsymbol{\theta}^*$. In practise, we only experienced different optimal solutions (local minima) for small N, and every time this happened, each solution had a different likelihood, which shows evidence that the optimisation problem is well defined.

Figure 3 shows box plots of the residuals. Columns indicate size N of the synthetic dataset and each row corresponds to one of the models under consideration. We can see that all models asymptotically converge to the true values, since the residuals are practically zero for large enough N. Overall, outliers (red crosses) are most

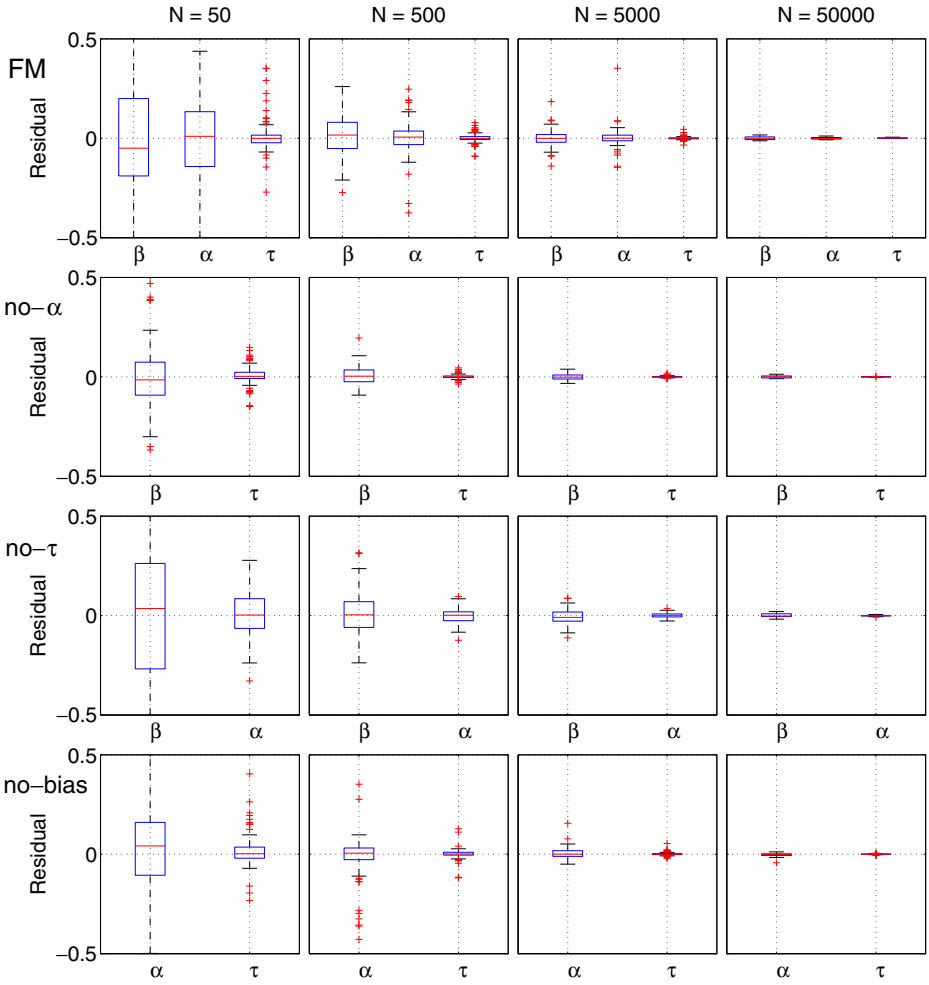


Figure 3 Validation of the maximum likelihood estimates: box plots of the residuals (differences between estimates $\hat{\boldsymbol{\theta}}$ and real values $\boldsymbol{\theta}^*$) for synthetic data. Columns indicate number of threads N and each row corresponds to one model (see Table 1). Data represents the outcome of 100 independent experiments with $\boldsymbol{\theta}^*$ selected randomly. Estimates were initialised using five different random initial conditions to test for multiple local minima. The selected solution was the one corresponding to the best likelihood (color online).

frequent in the FM estimates, which is the model with most number of parameters. In contrast, the model without novelty term, NO-τ, is the one which shows the most stable behaviour. This occurs because for the other models, estimating τ is difficult for small values ($\tau^* < 0.5$) and small N, since novelty decays exponentially. We will see later that for our four datasets this is not a problem. Interestingly, although for $N = 50$ the residuals are broadly distributed, their medians are centred at zero, which implies that even for small number of threads, one can get a fair estimate of the optimal values.

Table 2 Dataset statistics.

Dataset	#Threads	#Nodes	Total users
SL	9,820	2,028,518	93,638
BP	7,485	397,148	6,864
MN	58,613	2,220,714	53,877
WK	871,485	9,421,976	350,958

We therefore can conclude that the proposed maximum likelihood method is unbiased and that it is possible to obtain good parameter estimates using a few hundreds of threads.

4 Empirical results

In this section we first describe the datasets we consider and then give a brief overview about some general characteristics. The datasets which we consider contain complete information of the thread evolution and are therefore not prone to selection bias. A summary of the datasets statistics can be found in Table 2.

4.1 Description of the datasets

- **Slashdot (SL)** (http://slashdot.org/): Slashdot is a popular technology-news website created in 1997 that publishes frequently short news posts and allows its readers to comment on them. Slashdot has a community based moderation system that awards a score to every comment and upholds the quality of discussions [31]. The interface displays hierarchically the conversations, so users have direct access to the thread structure. A single news post triggers typically about 200 comments (most of them in a few hours) during the approximately 2 weeks it is open for discussion. Our dataset contains the entire amount of discussions generated at Slashdot during a year (from August 2005 to August 2006). See [16] for more details about this dataset.
- **Barrapunto (BP)** (http://barrapunto.com/): Barrapunto is a Spanish version of Slashdot created in 1999. It runs the same open source software as Slashdot, making the visual and functional appearance of the two sites very similar. Although Slashdot currently runs a more sophisticated interface than Barrapunto, they both shared the same interface at the time the data were retrieved. They differ in the language (audience) they use and the content of the news stories displayed, which normally does not overlap. The volume of activity on Barrapunto is significantly lower. A news story on Barrapunto triggers on average around 50 comments. Our dataset contains the activity on Barrapunto during three years (from January 2005 to December 2008).
- **Meneame (MN)** (http://www.meneame.net/) Meneame is the most successful Spanish news aggregator. The website is based on the idea of promoting user-submitted links to news (stories) according to user votes. It was launched in December of 2005 as a Spanish equivalent to `Digg`. The entry page of Meneame consists of a sequence of stories recently promoted to the front page, as well as a link to pages containing the most popular, and newly submitted stories. Registered users can, among other things: *(a)* publish links to relevant news which are retained in a queue until they collect a sufficient number of votes

to be promoted to the front page of Meneame, *(b)* comment on links sent by other users (or themselves), *(c)* vote (*menear*) comments and links published by other users. Contrary to both BP and SL, Meneame lacks an interface for nested comments, which are displayed as a list. However, the tag `#n` can be used to indicate a reply to the *n*-th comment in the comment list and to extract the tree structures we analyse in this study. To focus on the most representative cascades, we filter out stories that were not promoted, that is marked as discarded, abuse, etc. Our dataset contains the promoted stories and corresponding comments during the interval between Dec. 2005 and July 2009.

– **Wikipedia (WK)** (http://en.wikipedia.org) : The English Wikipedia is the largest language version of Wikipedia. Every article in Wikipedia has its corresponding *article talk page* where users can discuss on improving the article. For our analysis we used a dump of the English Wikipedia of March 2010 which contained data of about 3.2 million articles, out of which about 870,000 articles had a corresponding discussion page with at least one comment. In total these article discussion pages contained about 9.4 million comments. Note that the comments are never deleted, so this number reflects the totality of comments ever made about the articles in the dump. The oldest comments date back to as early as 2001. Comments who are considered a reply to a previous comment are indented, which allows to extract the tree structure of the discussions. Note that Wikipedia discussion pages contain, in addition to comments, structural elements such as subpages, headlines, etc. which help to organise large discussions. We eliminate all this elements and just concentrate our analysis on the remaining pure discussion trees. More details about the dataset and the corresponding data preparation can be found in [32]. For our experiments we selected a random subset of 50,000 articles from the entire dataset. Results did not vary significantly when using different random subsets of the data.

4.2 Global analysis

To globally characterise the threads, we analyse some properties related to the sizes of the threads (number of comments they receive) and the authorships.

Figure 4 shows histograms of the thread sizes (left) and their complementary cumulative distributions (right). As expected, all distributions are positively skewed, showing a high concentration of relatively short threads and a long tail with large threads. However, although all distributions are heavy tailed, we clearly see a different pattern between the three news aggregators and the Wikipedia. Whereas SL, BP and MN present a distribution with a defined scale, the distribution of thread sizes of Wikipedia is closer to a scale-free distribution, in line with the threads found in weblogs [35] and Usenet [28]. We remark that, even in the Wikipedia case, the power-law hypothesis for the tail of this distribution is not plausible via rigorous test analysis: we obtain an exponent of 2.17 at the cost of discarding 97% of the data.

We also observe a progressive deviation from websites with a well defined scale such as Slashdot, which could be described using a log-normal probability distribution, toward websites with less defined scale such as Meneame, which may show a power-law behaviour for thread sizes > 50. Barrapunto falls in the middle and, interestingly, is more similar to Meneame than to Slashdot.

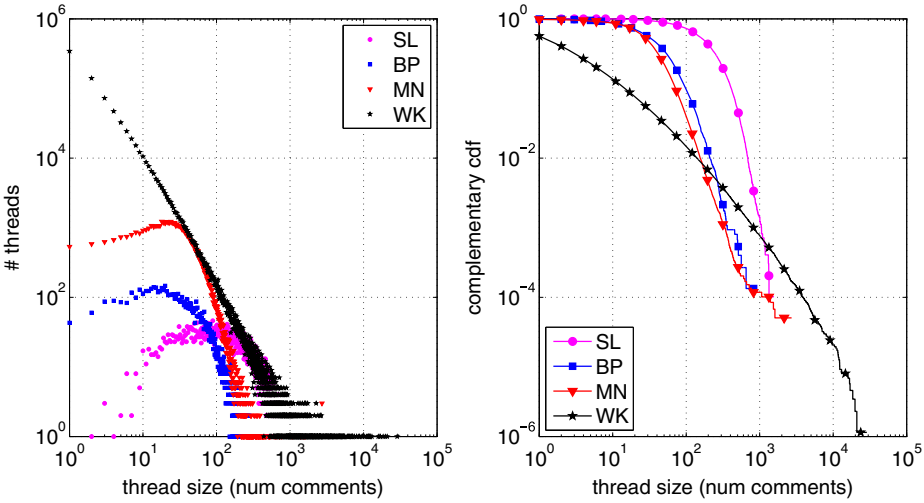


Figure 4 Thread sizes for the different datasets. (*left*) Histogram of the sizes. (*right*) Complementary cumulative distribution of the sizes (color online).

The previous considerations imply that, in general, a new post in Slashdot can hardly stay unnoticed and will propagate almost surely over several users. Conversely, most of the news in Meneame will only provoke a small reaction and reach, if they do, a small group of users. We can say that, according to the behaviour of the thread sizes, Meneame is the news aggregator that shares most similarities with Wikipedia.

A characteristic feature of discussion threads, unlike other form of information cascades, is the repeated user participation. To end this section, we briefly mention some properties related to the authors of the comments. Figure 5 (left) shows the distribution of the number of comments per user and thread (participations) for the four datasets. Although the proportion of participations with only one comment is large, a significant number involve at least two or more comments. The proportion of these participations lies between 15% for Meneame and 31% in the case of Wikipedia. Occasionally, some users react react ten, hundreds or even thousands of times in the same discussion thread. It is also interesting to analyse the relation between the size of the threads and the authorships. This is depicted in figure Figure 5 (right). Although we observe a close linear relationship in all datasets (log-log scale) as reported for Usenet in [28], we can differentiate a small decay in the gradient present in MN and WK only for large threads, indicating that the proportion of users that comment at least twice in the same thread becomes larger in larger threads, something that does not seem to happen for SL at all. We observe that the frequency of participation on the WK talk pages is significantly higher than the rest.

Figure 1 illustrates the different types of threads which we found. We plot representative threads with similar sizes selected randomly from each of the four datasets. For Slashdot we can see that the chain reaction is located mainly on the initiator event (direct reactions), but some nodes also have high degree, resulting in bursty disseminations. We could say that after a news article is posted, the collective attention is constantly drifting from the main post to some new comments which

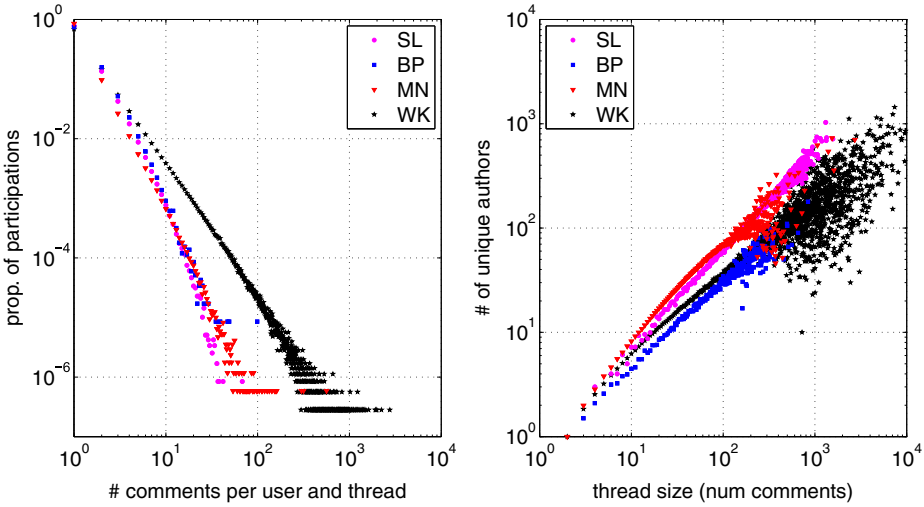


Figure 5 Authorship: (*Left*) distribution of the number of comments per user and thread in the different datasets. (*Right*) relation between sizes and average number of distinct authors per thread (color online).

become more popular. In Barrapunto we observe similar structures, although their persistence is less noticeable. On the contrary, Meneame is characterised by having high concentration of nodes at the first level together with rare but long chains of thin threads. This represents a pattern where only a few comments receive multiple replies, but that sporadically can trigger a long dialogue between a few users. We note that this phenomenon might be caused by the fact that the thread tree and, more importantly, the number of replies a comment receives are hidden in the interface of Meneame. Finally, the case of Wikipedia is very similar to Meneame, but with even longer, more frequent and finer threads of nodes with very low degree.

4.3 Comparison of models and interplay of the features

In this section we compare the explanatory power of the different proposed models using the datasets previously described. These results allow to characterise the interplay between novelty, popularity and root bias for a particular website. In order to compare models, we perform two types of statistical tests based on the likelihoods.

We first check whether the full model is significantly better than any of the reduced models by means of a likelihood ratio test. Results show that for all datasets except for the Wikipedia, the full model is preferable over any of the three reduced models. For the Wikipedia discussions, the full model, although is better than NO-τ and NO-bias, it is not significantly different when compared against the model without popularity (NO-α). This result is important, since it highlights the main difference between the three news websites and the Wikipedia discussion pages, namely, whereas popularity plays a role in the news aggregators, it does not in the Wikipedia.

To compare the reduced models, we can say that the model with better likelihood is preferable only if the hypothesis that the likelihoods are significantly different

holds. To test whether the likelihood between the different groups (models) differ significantly, we perform a one-sided ANOVA test and subsequently, a Tukey's range test for multiple comparisons. The results are shown on Figure 6 for each dataset.

For Slashdot, we observe that the models NO-bias and NO-τ are not significantly different. More importantly, the model NO-α is the one which performs significantly worst. This indicates that neglecting the preferential attachment mechanism has the strongest impact. We can therefore conclude that popularity is the most relevant feature of Slashdot: users tend to write to popular comments more than to novel ones, for instance.

Interestingly, the PA mechanism seems to be crucial only for Slashdot. Although one would expect very similar characterisation for Barrapunto, the relevance of novelty and popularity differs. For Barrapunto, we observe that all the four models

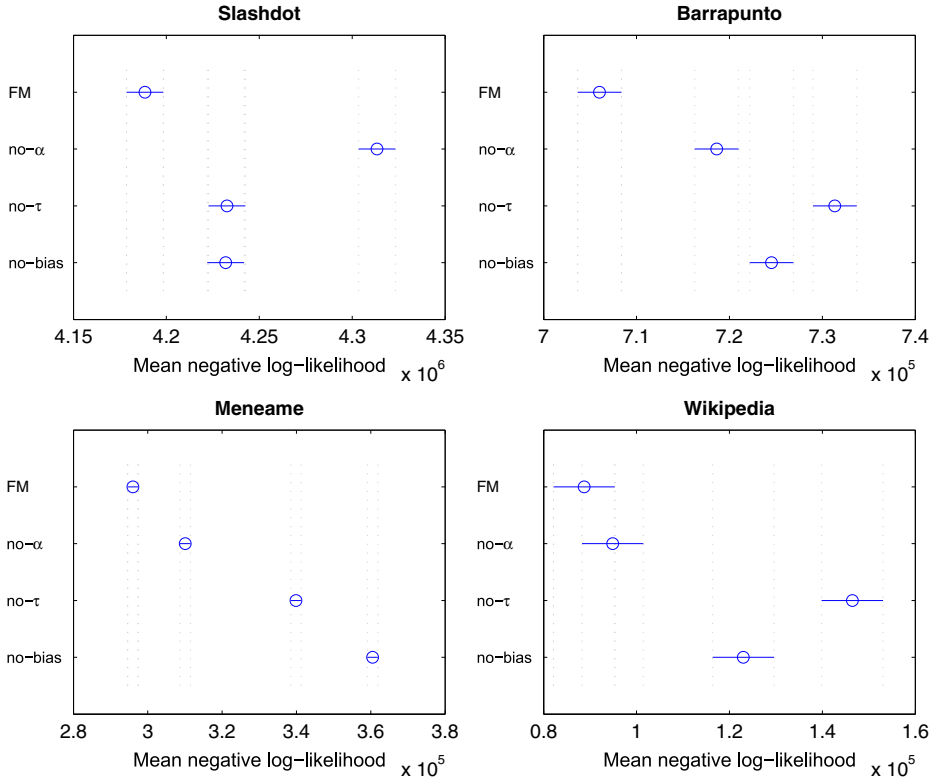


Figure 6 Model comparison for each dataset: *horizontal axis* shows negative log-likelihood (left is better) and *vertical axis* indicates the four different models we consider. Two models are not statistically different if their range plots overlap, for instance, models NO-τ and NO-bias in Slashdot. Conversely, model A is preferable than model B if their range plots do not overlap and A is positioned on the left of B. The full model FM outperforms any of the reduced models except NO-α in Wikipedia. The best reduced model depends on the dataset. Range plots are computed via one-sided ANOVA and Tukey's range test on the mean of Likelihoods across 100 (bootstrap) different random subsets. The number of sampled threads is $N = 5 \cdot 10^4$ for all datasets.

Table 3 Average parameter estimates for the full model over 100 different subsets and two different subset sizes N. Values within parenthesis indicate the standard deviation of the estimated parameter.

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)

are statistically different. In decreasing order of accuracy, we have FM, NO-α, NO-bias and finally NO-τ. The impact of removing the novelty term τ is therefore larger than removing the root-bias, and both features are more relevant than the popularity. We can conclude that the novelty is the most relevant of the three features in Barrapunto. In contrast to Slashdot users, Barrapunto users tend to write preferably based on how new a comment is than how popular a comment is.

The case of Meneame is also different. The results show that the key feature to describe Meneame is the difference between the process of writing to the post and the process of writing to regular comments. After this distinction is made, we can also say that novelty is more relevant than popularity, thus in Meneame, as in Barrapunto, users write preferably to new comments than to popular ones.

Finally, for the Wikipedia discussion pages, as noted before, we see that models FM and NO-α are not differentiable (in accordance with the likelihood ratio test) and second, that novelty is more relevant than differentiating between the article and the comments.

In the following we will contrast these conclusions with an analysis of the parameters of the full model. In Table 3 we can find their values for two different subset sizes. We observe that for small subsets ($N = 50$ threads), we already obtain a reliable estimation. Only the bias to the root term (β) for Wikipedia shows larger fluctuations. Using larger subsets of 5000 threads does not change the mean parameter estimates significantly, except again for the case of β in Wikipedia. Thus we can conclude that the estimated parameter values are stable using different, sufficiently large random subsets of the data.[2]

If we compare the actual parameter values among the different datasets we observe results that confirm the previous conclusions like an only minor influence of the age of a comment (novelty τ close to 1, indicating a very slow decay) but a large impact of popularity (α) in SL and, on the contrary, a zero value for α and the biggest dependency on novelty in WK. Furthermore, if we look at the parameter β we find that it is largest for SL and smaller for BP and WK. This bias to the root parameter is most important at the beginning of a discussion where it determines whether new comments go mainly to the root node or are replies to already existing comments. We would expect thus to have initially broader trees for SL (and to a

[2]Note that this also indicates that a cross-validation (train-test) procedure would yield to very similar parameter estimates in train and test set.

minor degree for MN) while BP and WK should experience a faster initial growth. We will analyse this point further in section in Section 4.5.

4.4 Model validation: structure of the threads

To compare the real and the synthetic threads, we focus on the following structural properties, which are calculated for both types of threads:

- *Degree probability distribution:* we consider the probability distribution the degrees, which is equivalent to distribution of the number of direct replies minus one. For this calculation we use all nodes, including root and non-root nodes.
- *Subtree sizes distribution:* for each non-root node, we compute the probability distribution of the total number of its descendants, i.e. the size of the conversation triggered by a comment. We discard the root node because its associated distribution is precisely the one we use to generate the thread sizes.
- *Relation between the sizes and depths:* we analyse the thread depths as a function of the thread size by taking the average depth of all threads with a certain size.

Figures 7, 8, 9 and 10 show the three previous properties for each dataset independently in log-log axis. For clarity, they are illustrated for the best (FM, black lines) and the worst (red lines) model only.

Overall, the full model is able to capture reasonably well the relevant quantities in all datasets. In particular, the degree distributions are very accurately reproduced, even though each dataset exhibits a different profile (see left plot of the figures). The effect of using a bias term is clearly manifested in the bi-modality of that distribution, with a first peak dominated by the comments' replies and a second peak dominated by the direct replies to the root. This effect is strongest in Meneame (Figure 9 left) and less pronounced in Barrapunto or Wikipedia, in agreement with the analysis of the previous section, since the bias term is fundamental in the former and less relevant in the latter datasets.

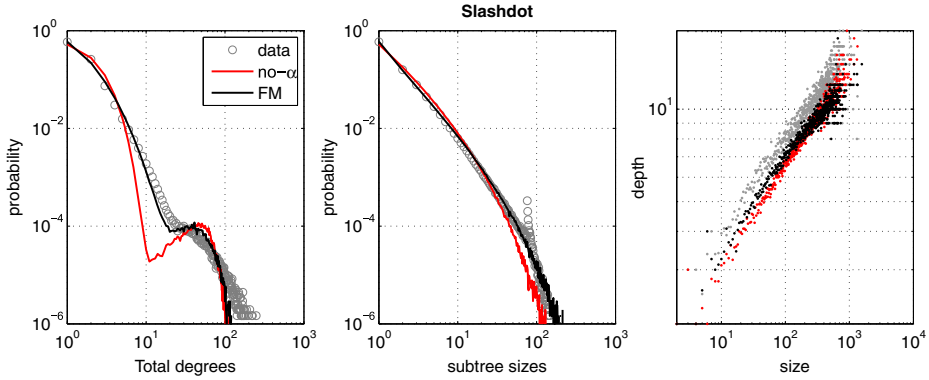


Figure 7 Comparison of the degree distribution (*left*), subtree sizes (*centre*) and correlation between depth and number of comments (*right*) for the original discussion trees (*grey circles*) from the Slashdot dataset and synthetic trees generated with the full model (*black curves and dots*) or a model without popularity term (*red curves and dots*) (color online).

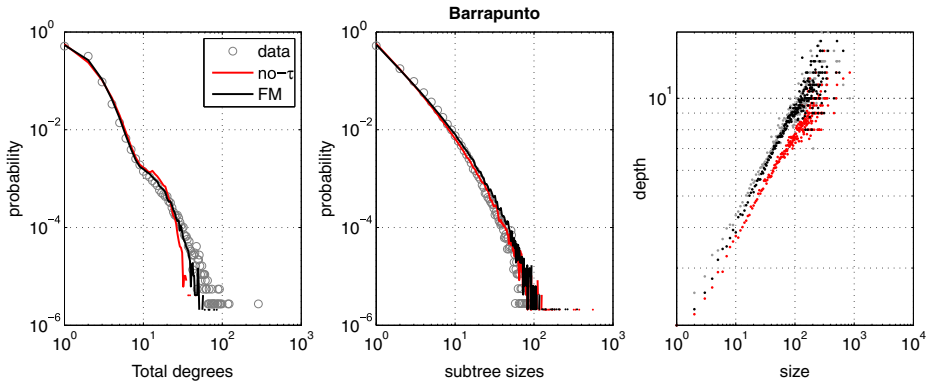


Figure 8 Comparison of the degree distribution (*left*), subtree sizes (*centre*) and correlation between depth and number of comments (*right*) for the original discussion trees (*grey circles*) from the Barrapunto dataset and synthetic trees generated with the full model (*black curves and dots*) or a model not using the novelty term (*red curves and dots*) (color online).

The effect of neglecting the root-bias term is that the weights of popularity and novelty are increased and decreased respectively with respect to the weights obtained for the FM. This effect strengthens the PA process and results in degree and subtree sizes distributions that are too skewed for the non-root nodes. This issue seems to be the main limitation when the global (direct reactions to the root) and the localised replies are not differentiated, such as in the T-MODEL [28] and is closely related with the so-called stage dependency found in [52] and modelled using a branching process.

The full model also generates correct subtree sizes of the non-root nodes in all datasets, with the exception of Meneame, which we postulate is caused by the particularities of the platform. With the exception of NO-bias, all models reproduce adequately this quantity. For the Wikipedia, however, we also observe that the

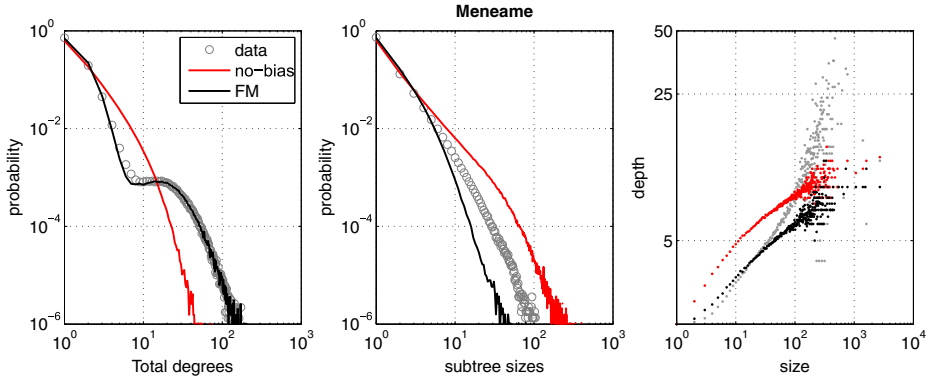


Figure 9 Comparison of the degree distribution (*left*), subtree sizes (*centre*) and correlation between depth and number of comments (right) for original discussion trees (*grey circles*) from the Meneame dataset and synthetic trees generated with the full model (*black curves and dots*) or a model not using the bias term (*red curves and dots*) (color online).

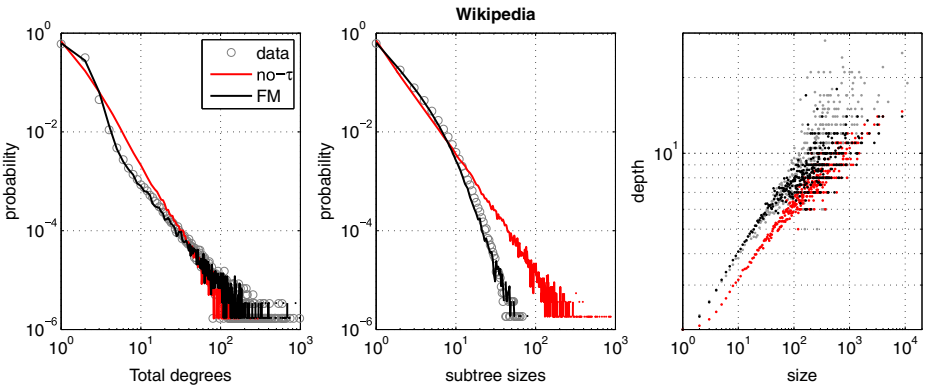


Figure 10 Comparison of the degree distribution (*left*), subtree sizes (*centre*) and correlation between depth and number of comments (*right*) for original discussion trees (*grey circles*) from the Wikipedia dataset and synthetic trees generated with the full model (*black curves and dots*) or a model not using the novelty term (*red curves and dots*) (color online).

model without novelty NO-τ also produces longer tails (Figure 10, middle). Since, as we have shown before, Wikipedia can be characterised using bias and novelty only irrespective of popularity, neglecting any of these two crucial features affects dramatically the approximation quality of the model.

The third quantity we compare is the average depth as a function of the size of the threads. We can see that the full model reproduces very accurately this quantity for Wikipedia and Barrapunto, but only qualitative agreement is reported for Slashdot. For Meneame, it clearly differs from the data, especially for large threads.

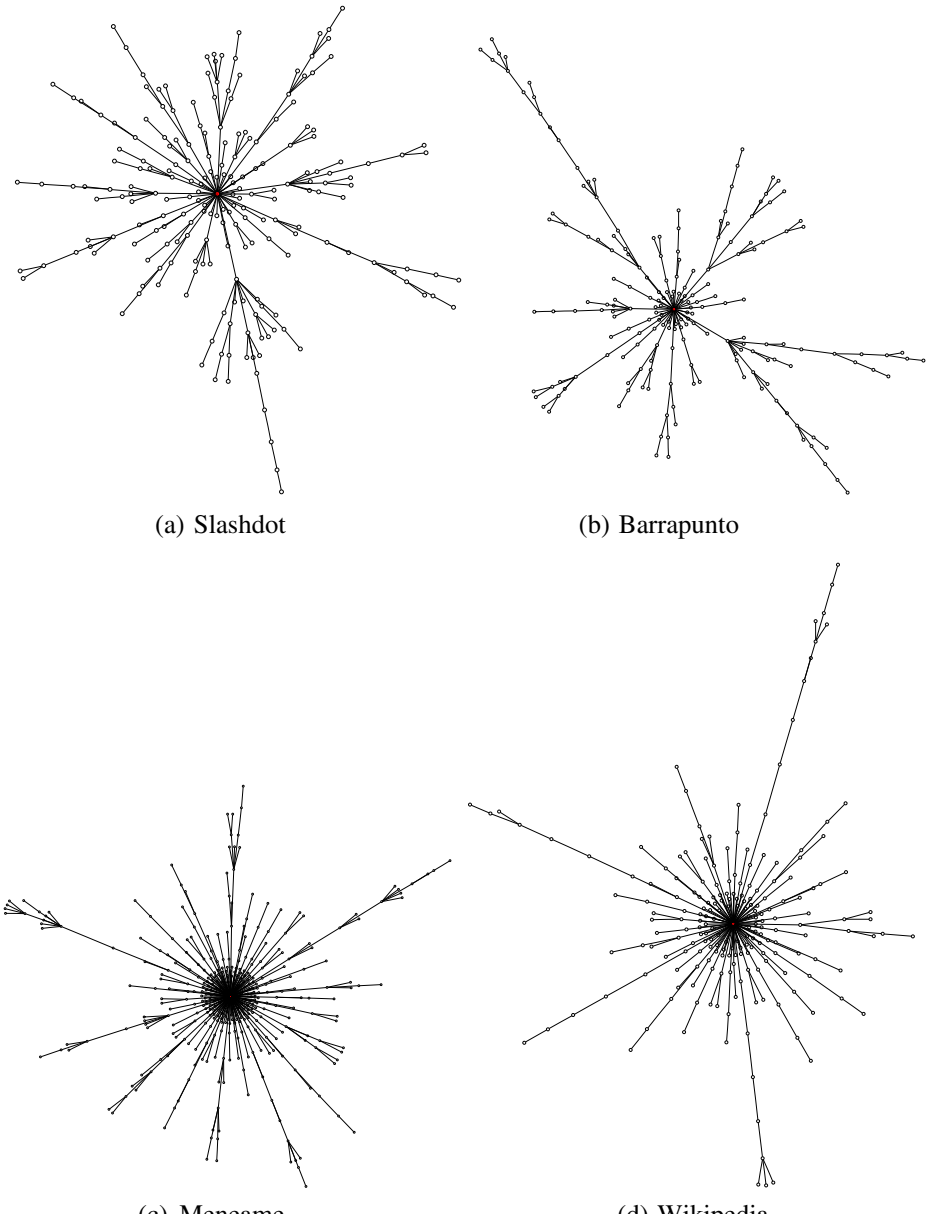
Capturing the thread depths correctly is difficult, as pointed out in [17], where it was shown that a model without novelty was unable to reproduce accurately the mean depth distribution in any of the datasets we consider here. Indeed, Figures 8 and 10 (red curves) show that the model NO-τ, which is comparable to the one of [17], clearly underestimates the depths. We see that the novelty term incorporated in the full model substantially the depths to be very close to the empirical observations in all datasets.

It is interesting to compare the observed distributions with the ones reported in other studies. The discussion threads analysed in this work are very similar to the conversations of Usenet [28], although the relevance of the bias term seems to be smaller for Usenet than for any of our datasets. Compared to chain-letters, discussion threads are much shallower than the chain-letters analysed in [36] (median depths are of around 500 levels), but much deeper than the trees extracted from forwarded email in [52] (max depth found was four). We have to keep in mind that the type of interaction considered in [52] and [36] differs substantially from ours.

One could consider whether the power-law hypothesis is a valid explanation for the relation between thread sizes and thread depths as suggested for Usenet [28], or for the degree distributions as advocated for blogs data [35], for instance. In our datasets, we observe that rigorous statistical tests systematically rejects the power-law hypothesis for either degrees or subtree sizes, or does not reject it at the cost of discarding almost all of the data. Further, average depth does not cover more than

two orders of magnitude and are very noisy in the tail (very few posts exist with large depths), complicating a phenomenological explanation using power laws.

To conclude this section, we show in Figure 11 the synthetic counterpart of Figure 1, where we plot representative threads with similar sizes selected randomly



(a) Slashdot (b) Barrapunto

(c) Meneame (d) Wikipedia

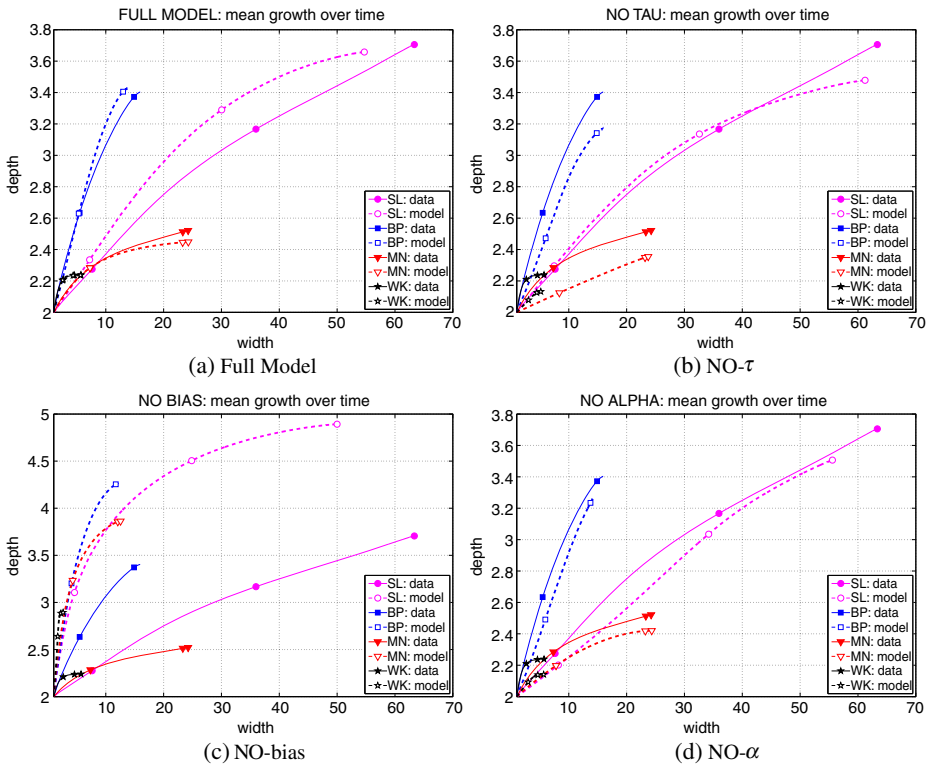
Figure 11 Examples of synthetic discussion threads.

from each of the four synthetic datasets. We can see that the generated threads present a strong resemblance with the real ones.

4.5 Model validation: evolution of the threads

After having compared the main structural properties of the synthetic trees with the real ones, we now investigate whether the models considered in this study are also able to reproduce the growth process of the threads. In other words, if we take intermediate snapshots of the threads during their evolution, how close match the synthetic trees their archetypes?

To this end we record two quantities: the **width** (maximum over the number of nodes per level) and the **mean depth** of the trees every time a new node is added. The evolution of the relationship between the averages of these two quantities is depicted in Figure 12 for all datasets. The marker symbols indicate the size of the discussions after 10, 100 and 1000 comments, respectively. We compare the original



(a) Full Model (b) NO-τ

(c) NO-bias (d) NO-α

Figure 12 Evolution of the discussions on a width-depth plane. *Full lines* correspond to the discussion obtain from our datasets while the *dashed lines* the growth of synthetic discussion generated with one of the four different model types analysed in this study. The markers indicate the size of the discussions after 10, 100 or 1000 comments. With the exception of Slashdot the full model is the one who best approximates the growth of the discussions (color online).

threads (continuous lines with symbols) with the different model variants (dashed lines) and observe a good coincidence between the full model and the data in the evolution of the width of the discussions (Figure 12a), for three of the four datasets. Wikipedia shows a nearly perfect coincidence while both, Meneame and Barrapunto initially follow the growth process of the synthetic threads but later slightly sub-estimate the mean depth of the discussions. However, in the case of Slashdot the full model generates threads which are deeper but also thinner than the ones observed in our dataset.

Interestingly, while for all other datasets all three reduced models reproduce less accurately the growth process of the trees (being NO-α the second best choice), for Slashdot the NO-τ model is initially closer to the evolution of the real threads that the full model. How is it possible then that the FM still has a better log-likelihood than this model? We can observe that towards the end of the discussions the full-model produces trees that have nearly the same average depth as the real instances (albeit in a thinner tree), while the NO-τ model produces trees with similar width but shallower sub-threads. This in the end leads to a better log-likelihood. And although the actual parameter value of τ is very close to one (compare with Table 3), it still has an influence in large discussions where it finally leads to greater depths, by favouring more recent replies. It seems that in the specific case of this dataset a two step model would be a better choice where the effect of τ would set in only when the discussions are already of a certain size.

We also observe in coincidence with the reflection on the parameter β at the end of Section 4.3 that Barrapunto and Wikipedia have the largest initial growth (largest initial gradient between width and depth), while Meneame (and even more Slashdot) cause broader discussions with more comments to the root node. The nearly identical initial gradient in the NO-bias model, visible in Figure 12c, confirms this further.

5 Discussion

We have presented a statistical analysis and comparison of the structure and evolution of the different discussion threads associated to three popular news media websites and the discussion pages of the English Wikipedia. Without examining the thread content, our analysis already highlights the heterogeneities between datasets, which could be conditioned mainly by two factors, namely, the page design, or platform, and the community of users. Despite these heterogeneities, we have provided evidence that our proposed mathematical model can capture many of the structural properties and evolution profiles of the real threads accounting for the particularities of each dataset. To the best of our knowledge, this is the first study that analyses four large-scale datasets in such a level of detail.

The contrast found between Slashdot and Barrapunto (the Spanish version of Slashdot) is a clear example of how a diverse community of users can differently shape the activity of a website. Since both websites share the same platform, the contrasting activity patterns governing their threads should be mainly attributed to social factors related to the different user communities. As we show in Section 4.3, the proposed framework provides a rigorous quantitative basis for this effect, which

is succinctly represented in the degree of relevance of novelty and popularity in both spaces.

Another interesting aspect concerns the ability of the proposed framework to quantify to a certain extent how the thread patterns are influenced by the interface and/or platform. In Meneame, our analysis identifies the distinction between two processes, post and comment reply (what we call root-bias), as the most fundamental feature among novelty and popularity. Meneame is also the only website that provides a flat view of the conversations, in contrast to Slashdot, Barrapunto and Wikipedia, where threads are displayed hierarchically. We believe that the strongest relevance of the root-bias in Meneame is mainly caused by the flat interface, which effectively increases the attractiveness of the post, especially during the beginning of the discussion. We postulate that this result is general and could be present in other forums which provide a flat view of the conversations as well.

The different motivation between Wikipedia discussion pages and typical conversations on the news media websites manifests in the role of popularity in the Wikipedia. Whereas popularity seems to be important in the three news media websites, it is irrelevant in the growth process of Wikipedia discussions. This conclusion arises naturally from our framework when we compare the likelihoods of the data between the full model and the model without popularity, since the differences in the average likelihoods between the two models are not statistically significant.

The implications of the work presented in this manuscript are diverse. At a fundamental level, we have identified a common model that captures the heterogeneities found in different web-spaces, suggesting that human conversations are built following a kind of universal law. This has implications in the basic understanding of the communication patterns in large web-spaces that comprise many-to-many interaction.

On an application side, since the parameter estimates of the model allow for a figurative description of the communication habits of a website, one could model the structural differences between conversation threads associated with different communities which share the same platform. This would lead to results similar to the one reported for Barrapunto and Slashdot, but for pre-defined target user communities. For instance, in terms of political orientation, how would the parametrisations differ between left/right -wing oriented forums? What is the impact of popularity in a model of threads from a particular political tendency? The proposed model would be particularly useful to answer such questions, because of its robustness and its lack of content interpretation, since semantic analysis may introduce an additional source of error.

Since novelty and popularity are relevant features in forum design or maintenance in collaborative applications, it is easy to devise potential applications in these contexts. The proposed approach does not require extensive amounts of data for parameter estimation, which makes it adequate to track the evolution of the parametrisation over time. The associated dynamic patterns could be used to characterise user community evolution and adapt the design of a forum accordingly. One could further change the model parameters and predict how the structural patterns described in Sections 4.4 and 4.5 would change in a particular web-space.

We would like to address several points that deserve further investigation. First the proposed framework is simple and general. It could be easily applied to other types of cascades that have similar tree-like structure, such as chain-letters or

forwarded emails.[3] We also note that the popularity measure proposed here (the degree or number of replies) can be replaced by other measures such as votes or other types of scores. It would be also interesting to analyse the impact of using alternative measures of popularity.

As a consequence of our bottom-up approach, we have focused solely on structural aspects of the conversational threads and deliberately discarded other factors such as the precise timing of each node arrival or the cascade length. A direction of further research would include the incorporation of the size of the cascade in the model. To what extent does the structure of the discussions depend on their size? Do short and long discussions exhibit the same global pattern at different temporal scales, or do large cascades fan out deeper and narrower while small cascades follow a more shallow pattern? Our work represents a step towards answering these questions and suggests that the platform or domain, rather than its size, determines more the structure of the discussions.

While finishing the final version of this manuscript we became aware of [51], a recent paper which aims to explain the actual sizes of the discussion cascades, given the user's temporal behaviour patterns (the waiting time distribution between two consecutive comments) and the distributions of exposure times of the cascades (i.e. the time the cascades are exposed on a privileged place such as the front page of the hosting website).

To explain the in-degree distribution of the cascades, the authors of [51] propose a PA process. The structural part of their model represents a special case of the model presented here. A promising direction of future research would be to combine our structural model with the temporal behaviour modelling of [51] to generate a more powerful integrated model for online conversations.

Finally, as suggested in [11], decay of novelty can be due to competition (limited resources) or habituation. One would expect competition to play a more important role within news media which are more sensitive to fads and habituation to be more typical in Wikipedia. It would be interesting to analyse refinements of the novelty term which incorporate these principles.

Appendix A: Asymptotics for the degree distribution in FM.

Below we provide the heuristic derivation of the power law distribution in FM. We use notations as defined in Section 2. The exact derivation is possible in the same lines as in [21, Chapter 8]. In the full model we have

$$Z_t = 2\alpha t + \beta - 2\alpha + \tau(1-\tau^t)/(1-\tau).$$

Theorem 1 *For large t and k we have*

$$E[d_{k,t}] = \Theta\left(\left(\frac{t}{k}\right)^{1/2}\right).$$

[3]The datasets considered here and the source code for parameter estimation and processing the threads are available on request.

Proof For $k > 1$ we get

$$E[d_{k,t+1}] = E[d_{k,t}] + E[E[d_{k,t+1} - d_{k,t}|\pi_{(1:t-1)}]] = E[d_{k,t}] + \frac{\phi(k)}{Z_t}$$
$$= E[d_{k,t}] + \frac{\alpha E[d_{k,t}] + \tau^{t-k+1}}{2\alpha t + \beta - 2\alpha + \tau(1-\tau^t)/(1-\tau)}. \tag{9}$$

Next, we construct the sequences $E[\underline{d}_{k,t}]$ and $E[\bar{d}_{k,t}]$ such that the average degree can be bounded as follows:

$$E[\underline{d}_{k,t}] \le E[d_{k,t}] \le E[\bar{d}_{k,t}]. \tag{10}$$

To this end, we define

$$\underline{d}_{k,k} = d_{k,k} = \bar{d}_{k,k} = 1, \tag{11}$$

and for $E[\underline{d}_{k,t}]$ and $E[\bar{d}_{k,t}]$ we derive the recursive equations, which constitute, respectively, the lower and the upper bound for the recursion (9). The lower-bound recursion is constructed similarly as in [28]:

$$E[\underline{d}_{k,t+1}] = E[\underline{d}_{k,t}] + \frac{\alpha E[\underline{d}_{k,t}]}{2\alpha t + \beta - 2\alpha + \tau/(1-\tau)} = E[\underline{d}_{k,t}] \frac{2\alpha t + \beta - \alpha + \tau/(1-\tau)}{2\alpha t + \beta - 2\alpha + \tau/(1-\tau)} \tag{12}$$

For the upper bound recursion note that the right-most expression in (9) is bounded from above by

$$E[d_{k,t}] + \frac{(\alpha + \tau^{t-k+1}) E[d_{k,t}]}{2\alpha t + \beta - 2\alpha} = E[d_{k,t}] \frac{2\alpha t + \beta - \alpha}{2\alpha t + \beta - 2\alpha} \left(1 + \frac{\tau^{t-k+1}}{2\alpha t + \beta - 2\alpha}\right)$$
$$\le E[d_{k,t}] \frac{2\alpha t + \beta - \alpha}{2\alpha t + \beta - 2\alpha} e^{\frac{\tau^{t-k+1}}{2\alpha t + \beta - 2\alpha}}.$$

Thus, we define $E[\bar{d}_{k,t}]$ as

$$E[\bar{d}_{k,t+1}] = E[\bar{d}_{k,t}] \frac{2\alpha t + \beta - \alpha}{2\alpha t + \beta - 2\alpha} e^{\frac{\tau^{t-k+1}}{2\alpha t + \beta - 2\alpha}}. \tag{13}$$

With $E[\underline{d}_{k,t}]$ and $E[\bar{d}_{k,t}]$ defined by (11), (12) and (13) the inequalities (10) clearly hold.

Iterating (12) and applying the Stirling's approximation, for large t and k we obtain

$$E[\underline{d}_{k,t}] = \prod_{s=k}^{t} \frac{2\alpha s + \beta - \alpha + \tau/(1-\tau)}{2\alpha s + \beta - 2\alpha + \tau/(1-\tau)} = \prod_{s=k}^{t} \frac{s + \frac{\beta - \alpha + \tau/(1-\tau)}{2\alpha}}{s + \frac{\beta - 2\alpha + \tau/(1-\tau)}{2\alpha}}$$
$$= \frac{\Gamma\left(t + 1 + \frac{\beta - \alpha + \tau/(1-\tau)}{2\alpha}\right) \Gamma\left(k + \frac{\beta - 2\alpha + \tau/(1-\tau)}{2\alpha}\right)}{\Gamma\left(k + \frac{\beta - \alpha + \tau/(1-\tau)}{2\alpha}\right) \Gamma\left(t + 1 + \frac{\beta - 2\alpha + \tau/(1-\tau)}{2\alpha}\right)} \sim \left(\frac{t}{k}\right)^{\frac{1}{2}},$$

where $a \sim b$ denoted an asymptotic equivalence of a and b. Analogously, for (13), using that

$$C := \prod_{s=k}^{t} e^{\frac{\tau^{t-k+1}}{2\alpha t + \beta - 2\alpha}} \le e^{\frac{\tau(1-\tau^t)}{1-\tau}} \le e^{\frac{\tau}{1-\tau}} < \infty$$

we get that

$$E[\bar{d}_{k,t}] \sim C\left(\frac{t}{k}\right)^{\frac{1}{2}},$$

which, together with (10), proofs the result. □

Let us now count the number $\bar{N}_{\geq x}$ of values of k satisfying $E[\bar{d}_{k,t}] \geq x$, $x > 0$.

$$\bar{N}_{\geq x} = \sum_{k=1}^{t} \mathbf{1}_{\{E[\bar{d}_{k,t}] \geq x\}} \sim \sum_{k=1}^{t} \mathbf{1}_{\left\{C\left(\frac{t}{k}\right)^{\frac{1}{2}} \geq x\right\}} = \sum_{k=1}^{t} \mathbf{1}_{\{k \leq tC^2x^{-2}\}} = tC^2x^{-2}.$$

Similarly, for the number $\underline{N}_{\geq x}$ of values of k satisfying $E[\underline{d}_{k,t}] \geq x$ we get

$$\underline{N}_{\geq x} = \sum_{k=1}^{t} \mathbf{1}_{\{E[\underline{d}_{k,t}] \geq x\}} \sim \sum_{k=1}^{t} \mathbf{1}_{\left\{\left(\frac{t}{k}\right)^{\frac{1}{2}} \geq x\right\}} = \sum_{k=1}^{t} \mathbf{1}_{\{k \leq tx^{-2}\}} = tx^{-2}.$$

Finally, if $N_{\geq x}$ of values of k satisfying $E[d_{k,t}] \geq x$ then $\underline{N}_{\geq x} \leq N_{\geq x} \leq \bar{N}_{\geq x}$. Heuristically, $N_{\geq x}/t$ gives the asymptotic fraction of nodes with degree at least x at time t, thus we obtain the result (6) and conclude that the degree distribution follows a power law with exponent 2 for the cumulative distribution function, or 3 for the probability distribution function. The formal proof is more involved but will lead to the same result because of the concentration of the martingale probability measure around its mean.

Note that C is bounded by $\exp\left(\frac{\tau}{1-\tau}\right)$, which ranges from one to infinity, in particular for $\tau = 0.9$ the value is $e^9 \approx 8.1 \times 10^3$ and for $\tau = 0.99$ it becomes $\approx 9.89 \times 10^{42}$. Thus, although τ does not affect the power law exponent, it can change the fraction of nodes with degree at least x by several orders of magnitude.

Appendix B: Parameter estimation and model validation

We describe here our procedure to obtain parameter estimates for the different datasets and to validate the model.

For each dataset, we select *with replacement* a subset of N threads that we use to learn the parameters. We repeat this procedure for 100 different random subsets of size N. This bootstrap-based approach has several advantages: first, it reduces the over-fitting when we are dealing with a small dataset. Note, however, that this is not a concern in our case since the number of threads is much larger than the number of parameters (features) in all datasets under consideration. Second, it reduces the computational cost and makes the parameter estimation problem more tractable. In our case, for Wikipedia, which contains almost one million of threads, optimisation on the full dataset was too computationally demanding, but became feasible for a reduced subset of $N = 50 \cdot 10^3$ threads. As we will see and we already suggested in Section 3, estimates are already stable for subset sizes of that order.

The results presented in Section 4 are based on outcomes of this estimation procedure. In particular, we show results for $N = 50$ and $N = 5 \cdot 10^3$ and averaged likelihoods and parameter estimates over the 100 random realizations. To validate the model, we analyse the structure and evolution of the threads generated with

respect to the real ones. We generate as many threads as the number of threads for each dataset with sizes pre-determined drawing a pseudo-random number from the empirical distribution of cascade sizes (see Figure 4).

Acknowledgements We wish to thank David Laniado and Riccardo Tasso for providing the pre-processed Wikipedia dataset and Meneame.net for allowing to access an anonymised dump of their database. We also thank Mohammad Gheshlaghi, Wim Wiegerinck and Alberto Llera for useful discussions.

References

1. Adamic, L.A., Zhang, J., Bakshy, E., Ackerman, M.S.: Knowledge sharing and Yahoo answers: everyone knows something. In: Proceedings of the 17th International Conference on World Wide Web, WWW '08, pp. 665–674. ACM, New York, NY, USA (2008)
2. Adar, E., Adamic, L.A.: Tracking information epidemics in blogspace. In: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, WI '05, pp. 207–214. IEEE Computer Society, Washington, DC, USA (2005)
3. Bakshy, E., Karrer, B., Adamic, L.A.: Social influence and the diffusion of user-created content. In: Proceedings of the 10th ACM conference on Electronic Commerce, EC '09, pp. 325–334. ACM, New York, NY, USA (2009)
4. Banerjee, A.V.: A simple model of herd behavior. Q. J. Econ. **107**(3), 797–818 (1992)
5. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**(5439), 509–512 (1999)
6. Ben-Naim, E., Krapivsky, P.L.: Stratification in the preferential attachment network. J. Phys. A: Math. Theor. **42**(47), 475,001 (2009)
7. Bikhchandani, S., Hirshleifer, D., Welch, I.: A theory of fads, fashion, custom, and cultural change as informational cascades. J. Polit. Econ. **100**(5), 992–1026 (1992)
8. Blasio, B.F., Svensson, A., Liljeros, F.: Preferential attachment in sexual networks. Proc. Natl. Acad. Sci. **104**(26), 10,762–10,767 (2007)
9. Brush, A.B., Wang, X., Turner, T.C., Smith, M.A.: Assessing differential usage of Usenet social accounting meta-data. In: Proc. SIGCHI '05, pp. 889–898. ACM, New York, USA (2005)
10. Cha, M., Mislove, A., Gummadi, K.P.: A measurement-driven analysis of information propagation in the Flickr social network. In: Proceedings of the 18th International Conference on World Wide Web, WWW '09, pp. 721–730. ACM, New York, USA (2009)
11. D'Souza, R.M., Borgs, C., Chayes, J.T., Berger, N., Kleinberg, R.D.: Emergence of tempered preferential attachment from optimization. Proc. Natl. Acad. Sci. **104**(15), 6,112–6,117 (2007)
12. Eggenberger, F., Pólya, G.: Über die Statistik verketteter Vorgänge. Z. Angew. Math. Mech. **3**, 279–289 (1923)
13. Fisher, D., Smith, M., Welser, H.T.: You are who you talk to: detecting roles in Usenet newsgroups. In: Proc. HICSS '06. IEEE CS, Washington, USA (2006)
14. Goh, K.I., Eom, Y.H., Jeong, H., Kahng, B., Kim, D.: Structure and evolution of online social relationships: heterogeneity in unrestricted discussions. Phys. Rev. E **73**(6), 66,123 (2006)
15. Golub, B., Jackson, M.O.: Using selection bias to explain the observed structure of internet diffusions. Proc. Natl. Acad. Sci. **107**(24), 10,833–10,836 (2010)
16. Gómez, V., Kaltenbrunner, A., López, V.: Statistical analysis of the social network and discussion threads in Slashdot. In: Proceedings of the 17th international conference on World Wide Web, WWW '08, pp. 645–654. ACM, New York, NY, USA (2008)
17. Gómez, V., Kappen, H.J., Kaltenbrunner, A.: Modeling the structure and evolution of discussion cascades. In: Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia, HT '11, pp. 181–190. ACM, New York, NY, USA (2011)
18. Gonzalez-Bailon, S., Kaltenbrunner, A., Banchs, R.E.: The structure of political discussion networks: a model for the analysis of e-deliberation. J. Inf. Technol. **25**, 230–243 (2010)

19. Götz, M., Leskovec, J., McGlohon, M., Faloutsos, C.: Modeling blog dynamics. In: International Conference on Weblogs and Social Media, ICWSM '09 (2009)
20. Gruhl, D., Guha, R., Liben-Nowell, D., Tomkins, A.: Information diffusion through blogspace. In: Proceedings of the 13th International Conference on World Wide Web, WWW '06, pp. 491–501. ACM Press, New York, USA (2004)
21. van der Hofstad, R.: Random graphs and complex networks. Lecture notes, available online at: http://www.win.tue.nl/∼rhofstad/NotesRGCN2011.pdf (2011)
22. Iribarren, J.L., Moro, E.: Impact of human activity patterns on the dynamics of information diffusion. Phys. Rev. Lett. **103**(3), 38,702 (2009)
23. Jeong, H., Néda, Z., Barabási, A.L.: Measuring preferential attachment in evolving networks. Europhys. Lett. **61**(4), 567 (2003)
24. Joyce, E., Kraut, R.E.: Predicting continued participation in newsgroups. J. Comput-Mediat. Comm. **11**, 723–747 (2006)
25. Kaltenbrunner, A., Gómez, V., López, V.: Description and prediction of Slashdot activity. In: Proceedings of the 5th Latin American Web Congress, LA-WEB '07. IEEE Computer Society, Santiago de Chile (2007)
26. Kaltenbrunner, A., Gonzalez, G., Ruiz de Querol, R., Volkovich, Y.: Comparative analysis of articulated and behavioural social networks in a social news sharing website. New Rev. Hypermedia Multimed. **7**(3), 243–266 (2011)
27. Kearns, M., Suri, S., Montfort, N.: An experimental study of the coloring problem on human subject networks. Science **313**(5788), 824–827 (2006)
28. Kumar, R., Mahdian, M., McGlohon, M.: Dynamics of conversations. In: SIGKDD '10, pp. 553–562. ACM, New York, USA (2010)
29. Kumar, R., Novak, J., Raghavan, P., Tomkins, A.: On the bursty evolution of blogspace. World Wide Web **8**(2), 159–178 (2005)
30. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web, WWW '10, pp. 591–600. ACM, New York, USA (2010)
31. Lampe, C., Johnston, E.: Follow the (slash) dot: effects of feedback on new members in an online community. In: Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work, pp. 11–20. ACM, New York, USA (2005)
32. Laniado, D., Tasso, R., Volkovich, Y., Kaltenbrunner, A.: When the Wikipedians talk: network and tree structure of Wikipedia discussion pages. In: International Conference on Weblogs and Social Media, ICWSM '11. The AAAI Press (2011)
33. Lerman, K., Hogg, T.: Using a model of social dynamics to predict popularity of news. In: Proceedings of the 19th International Conference on World Wide Web, WWW '10, pp. 621–630. ACM, New York, NY, USA (2010)
34. Leskovec, J., Adamic, L.A., Huberman, B.A.: The dynamics of viral marketing. ACM Trans. Web. **1**(1) (2007)
35. Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N., Hurst, M.: Cascading behavior in large blog graphs: patterns and a model. In: SDM '07 (2007)
36. Liben-Nowell, D., Kleinberg, J.: Tracing information flow on a global scale using internet chain-letter data. Proc. Natl. Acad. Sci. **105**(12), 4633–4638 (2008)
37. Malmgren, R.D., Stouffer, D.B., Motter, A.E., Amaral, L.A.N.: A Poissonian explanation for heavy tails in e-mail communication. Proc. Natl. Acad. Sci. **47**(105), 18,135–18,158 (2008)
38. Mcglohon, M., Hurst, M.: Community structure and information flow in Usenet: improving analysis with a thread ownership model. In: International Conference on Weblogs and Social Media, ICWSM '09 (2009)
39. Musiał, K., Kazienko, P.: Social networks on the internet. World Wide Web 1–42 (2012). doi:10.1007/s11280-011-0155-z
40. Nonnecke, B., Andrews, D., Preece, J.: Non-public and public online community participation: needs, attitudes and behavior. Electron. Commerce Res. **1**(6), 7–20 (2006)
41. Peruani, F., Tabourier, L.: Directedness of information flow in mobile phone communication networks. PLoS ONE **6**(12), e28,860 (2011)
42. Preece, J., Nonnecke, B., Andrews, D.: The top five reasons for lurking: improving community experiences for everyone. Comput. Hum. Behav. **2**(20), 201–223 (2004)
43. Rangwala, H., Jamali, S.: Defining a coparticipation network using comments on Digg. IEEE Intell. Syst. **25**, 36–45 (2010)
44. Rogers, E.M.: Diffusion of Innovations, 5th edn. Free Press, New York (2003)

45. Rudas, A., Tóth, B., Valkó, B.: Random trees and general branching processes. Random Struct. Algorithms **31**, 186–202 (2007)
46. Sack, W.: Discourse diagrams: Interface design for very large-scale conversations. In: Proc. HICSS '00. vol. 3, p. 3034. IEEE CS, Washington, DC, USA (2000)
47. Sadikov, E., Medina, M., Leskovec, J., Garcia-Molina, H.: Correcting for missing data in information cascades. In: Proceedings of the 4th ACM International Conference on Web Search and Data Mining, WSDM '11, pp. 55–64. ACM, New York, NY, USA (2011)
48. Smith, M.: Tools for navigating large social cyberspaces. Commun. ACM **45**(4), 51–55 (2002)
49. Sun, E., Rosenn, I., Marlow, C., Lento, T.M.: Gesundheit! Modeling contagion through Facebook news feed. In: International Conference on Weblogs and Social Media, ICWSM '09. The AAAI Press (2009)
50. Szabo, G., Huberman, B.A.: Predicting the popularity of online content. Commun. ACM **53**, 80–88 (2010)
51. Wang, C., Ye, M., Huberman, B.A.: From User Comments to On-line Conversations. SSRN eLibrary (2012). doi:10.2139/ssrn.2012183
52. Wang, D., Wen, Z., Tong, H., Lin, C.Y., Song, C., Barabási, A.L.: Information spreading in context. In: Proceedings of the 20th International Conference on World Wide Web (2011)
53. Watts, D.J.: A simple model of global cascades on random networks. Proc. Natl. Acad. Sci. **99**, 5766–5771 (2002)
54. Welser, H.T., Gleave, E., Fisher, D., Smith, M.: Visualizing the signatures of social roles in online discussion groups. Journal of Social Structure (JoSS) **8**(2), 1–32 (2007)
55. Whittaker, S., Terveen, L., Hill, W., Cherny, L.: The dynamics of mass interaction. In: Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work, CSCW '98, pp. 257–264. New York, USA (1998)
56. Wiuf, C., Brameier, M., Hagberg, O., Stumpf, M.P.H.: A likelihood approach to analysis of network data. Proc. Natl. Acad. Sci. **103**(20), 7,566–7,570 (2006)
57. Wu, F., Huberman, B.: Novelty and collective attention. Proc. Natl. Acad. Sci. **104**(45), 17,599–17,601 (2007)
58. Zhongbao, K., Changshui, Z.: Reply networks on a bulletin board system. Phys. Rev. E **67**(3), 36,117 (2003)