Modeling the structure and evolution of discussion cascades

Vicenç Gómez
Donders Institute for Brain Cognition and Behaviour
Radboud University Nijmegen
v.gomez@science.ru.nl

Hilbert J. Kappen
Donders Institute for Brain Cognition and Behaviour
Radboud University Nijmegen
bertk@science.ru.nl

Andreas Kaltenbrunner
Information, Technology and Society Research Group
Barcelona Media
kaltenbrunner@gmail.com

ABSTRACT

We analyze the structure and evolution of discussion cascades in four popular websites: *Slashdot*, *Barrapunto*, *Meneame* and *Wikipedia*. Despite the big heterogeneities between these sites, a simple preferential attachment (PA) model with bias to the root can capture the temporal evolution of the observed trees and many of their statistical properties, namely, probability distributions of the branching factors (degrees), subtree sizes and certain correlations. The parameters of the model are learned efficiently using a novel maximum likelihood estimation scheme for PA and provide a figurative interpretation about the communication habits and the resulting discussion cascades on the four different websites.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences—*Sociology*; G.2.2 [**Mathematics of Computing**]: Graph Theory—*Network problems, Trees*

General Terms

measurement, algorithms, human factors

Keywords

discussion cascades, preferential attachment, maximum likelihood, slashdot, wikipedia

1. INTRODUCTION

Human communication patterns on the Internet are characterized by transient responses to social events. Examples of such phenomena are the discussion threads generated in news aggregators, the propagation of massively circulated Internet chain letters, or the synthesis of articles in collaborative web-based spaces such as Wikipedia.

These responses can be regarded as tree-like cascades of activity generated from an underlying social network. Typically, a trigger event, or a small set of initiators, generate a chain reaction which may catch the attention of other users who end up participating in the cascade (see Figure 1 for examples).

Since these cascades are in direct correspondence with the information flow in a social system, understanding the mechanisms and patterns which govern them plays a fundamental role in contexts like spreading of technological innovations [21], diffusion of news and opinion [9, 18], viral marketing [17] or collective problem-solving [13]. Although information cascades have been extensively analyzed for particular domains, such as blogs [9, 18], chain letters [19], Flickr [5], Twitter [14] or page diffusion on Facebook [22], the cascades under consideration in those studies rarely involve elaborated discussions or complex interchange of opinions: generally, a small piece of information is just forwarded from an individual to its direct neighbors. It remains as a open question whether the spread of information in discussion-based cascades follows similar patterns and is governed by the same mechanisms.

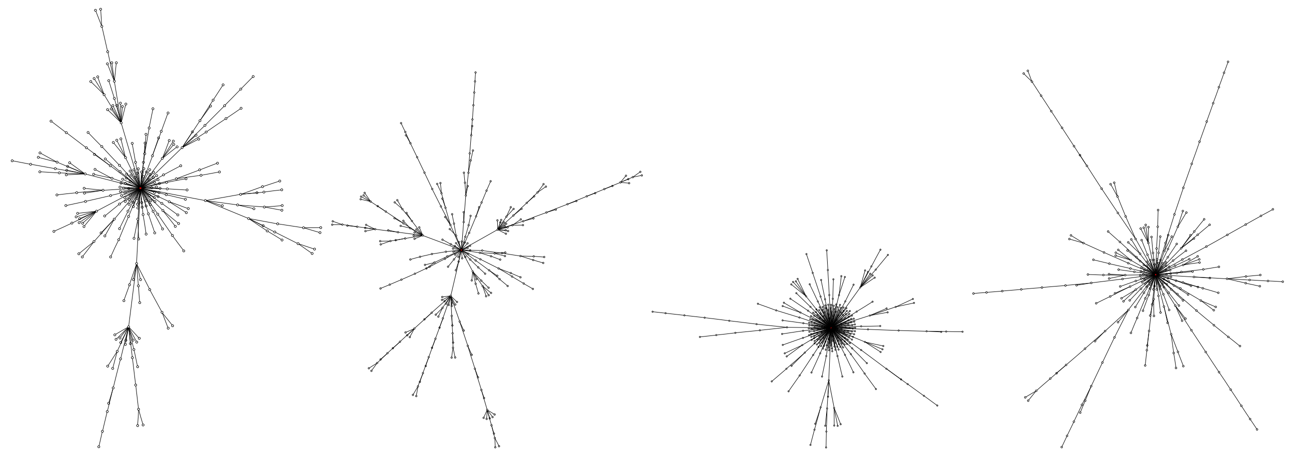
Here we consider several websites where the associated (discussion) cascades contain high level of interaction. We analyze the cascades of popular news aggregators such as *Slashdot*, *Barrapunto* (a Spanish version of Slashdot) and *Meneame* (a Spanish *Digg*-clone) and the English *Wikipedia*. As the reader may notice, these datasets are quite heterogeneous. For instance, although posts from both Slashdot and Meneame correspond to popular news which rely on broadcasted events, Slashdot contains rich and very extensive comments, which are less frequent in Meneame. The cascades found in Wikipedia, on the other hand, represent collaborative effort towards a well defined goal: produce a free, reliable article.

In this study we address the following questions: what are the statistical patterns that determine the structure of such cascades and their evolution? Can these patterns be largely determined regardless of semantic information using a simple parametric model? Can the parameterization corresponding to a given website provide a global characterization for it?

We first provide a global analysis of the cascade behavior in the four mentioned websites. We identify spaces where the sizes of the cascades occur at a clear defined scale whereas for others the scale is not so well defined. 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 also present a growth model for discussion cascades which is validated in the four datasets. Our approach is

.



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

Figure 1: Examples of real discussion cascades.

based on a simple model of preferential attachment (PA) [1], where new contributions in the cascade tree are linked to existing contributions with a probability which depends on their popularity (degree).

Two key ingredients characterize our approach: First, we account for a certain bias favoring the root, or event initiator. In this way, we are able to capture the different processes governing the global (direct reactions) and the localized responses of the system. Second, we use a likelihood method particularly developed for this study which allows an efficient estimation of the model parameters. The method is applicable not only for the data considered here but for a more general class of growing graphs. It is important to emphasize that we do not explicitly model the network dynamics over which the diffusion happens (dynamics at the level of the user). We focus only on the stochastic process which generates the cascade.

This paper is organized in the following way. In the next Section, we explain the proposed model and how we estimate its parameters. Section 3 introduces the datasets and provides a global analysis about their main characteristics. In Section 4 we explain the main results and give an interpretation of the parameters of the model. Finally, in Section 5 we describe related work and discuss the results in Section 6. In the Appendix we explain some aspects of the likelihood approach which are important for the estimation of parameters.

2. GROWING TREE MODEL FOR DISCUSSION CASCADES

We model a discussion cascade as a growing network. A new node is added sequentially at discrete time-steps. Our model is based on the original PA model to which we add a bias to the first node. Since each new node adds only one new link to the existing graph, the resulting network is a tree. We also assume that the total number of nodes N is known. It is convenient to represent compactly the cascade as a vector of parent nodes $\boldsymbol{\pi}$, where π_t denotes the parent of the node added at time-step $t+1$.

We are interested in the probability of π_{t+1} given the past history $\boldsymbol{\pi}_{(1:t)}$, that is $p(\pi_{t+1} = k|\boldsymbol{\pi}_{(1:t)})$, for $k = [1, \dots, t+1]$ and initial vector $\boldsymbol{\pi}_1 = (1)$ [1]. Note that by construction, $\pi_t \leq t, \forall t$.

At a given time t, we relate the *popularity* of a node k with its number of links in the following way:

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

where δ is the Kronecker delta function and c is usually denoted as the *initial attractiveness* of a node. In the following, we assume $c = 1$ and omit the explicit dependence on $\boldsymbol{\pi}_{(1:t)}$, so that $d_{k,t} \equiv d_{k,t}(\boldsymbol{\pi}_{(1:t)})$.

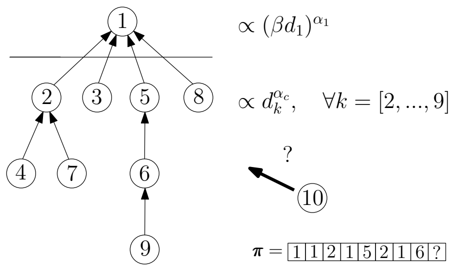


Figure 2: Small example: at time-step 9, node number 10 is added to the cascade. With probability proportional to $(\beta d_1)^{\alpha_1}$ it is added to the root node (initiator) and with probability proportional to $d_k^{\alpha_c}$ to one of the k non-root nodes. At the bottom right we show the corresponding parent vector $\boldsymbol{\pi}$ (see text).

At time $t+1$, the PA model attaches the new node to a node k with probability proportional to its popularity. See Figure 2 for an illustration. For completeness, we consider two models: the simplest model *without bias* to the root and the model which differentiates between the root node and the rest. In the general model the probability for attaching

[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.

the node $t+1$ to node k can be written as:

$$p(\pi_{t+1} = k|\boldsymbol{\pi}_{(1:t)}) = \frac{1}{Z_t}(\beta_k d_{k,t})^{\alpha_k}, \quad Z_t = \sum_{l=1}^{t} (\beta_l d_{l,t})^{\alpha_l}. \tag{2}$$

Model without bias: If we set $\alpha_k = \alpha$ and $\beta_k = 1$, for $k = [1, \ldots, t+1]$, we recover an important generalization of Barabasi's PA model, where the probability of attachment to a node goes as some general power α of the degree [15, 11]. For $\alpha = 1$, the linear preferential attachment is recovered. In this case, nodes have power-law distributed degrees. For $\alpha < 1$, or sublinear PA, the degrees are distributed according to a stretched exponential. For $\alpha > 1$ there is a "condensation" phenomenon, in which a single node gets a finite fraction of all the connections in the network [15].

Model with bias: Consider the following parameterization:

$$\alpha_k = \begin{cases} \alpha_1 & \text{for } k = 1 \\ \alpha_c & \text{for } k = [2, \ldots, t+1] \end{cases}$$

$$\beta_k = \begin{cases} \beta & \text{for } k = 1 \\ 1 & \text{for } k = [2, \ldots, t+1] \end{cases}. \tag{3}$$

In this case, α_1 and α_c are the exponents of the PA processes corresponding to the root and the non-root nodes respectively. β can be regarded as an additional degree of freedom weighting the root of the tree. In Section 4.4 we discuss about the interpretability of these parameters.

Note that, although we explicitly model the event which triggers the cascade as a root node, this representation does not limit the cascade to be originated from an individual event only. The root node can of course represent a group of initiators.

2.1 Maximum likelihood parameter estimation

Usually, PA in evolving networks is measured by calculating the rate at which groups of nodes with identical connectivity form new links during a small time interval Δt [11, 3]. 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 nonstationary, 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 [2]. However, two important aspects are worth to mention here. First, analytical results usually rely on assumptions like a continuum limit or on an infinite size network, which is also not the case of our data. Second, it is important to stress here that when parameters are learned for a particular observable for which an analytical form has been derived, the model may *overfit* on this measure, introducing a bias in other structural properties such as subtree sizes, average depths, or other correlations.

Our approach considers the likelihood function corresponding to the probabilistic model introduced before. We can assign to each observation (each node arrival in each cascade) a given probability using Equation (2). The parameters for which the likelihood is maximal are the ones that best explain the data given the model assumptions (see [23] 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 find the values of $\boldsymbol{\theta} := (\alpha_1, \alpha_c, \beta)$ which best explain the data.

The likelihood function can be written as:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\Pi}|\boldsymbol{\theta}) &= \prod_{i=1}^{N} p(\boldsymbol{\pi}_i|\boldsymbol{\theta}) \\ &= \prod_{i=1}^{N} \prod_{t=1}^{|\boldsymbol{\pi}_i|-1} p(\pi_{t+1,i}|\boldsymbol{\pi}_{(1:t),i}, \boldsymbol{\theta}) \\ &= \prod_{i=1}^{N} \prod_{t=1}^{|\boldsymbol{\pi}_i|-1} (\beta_x d_{x,t,i})^{\alpha_x} \left(\sum_{l=1}^{t} (\beta_l d_{l,t,i})^{\alpha_l}\right)^{-1}. \end{aligned} \tag{4}$$

where $\boldsymbol{\pi}_{(1:t),i}$ is vector of parents in the tree i until time t, $x := \pi_{t+1,i}$ is the parent of comment $t+1$ in the tree i, and $d_{x,t,i} := d_{x,t}(\boldsymbol{\pi}_{(1:t),i})$ denotes the degree of node x at time t also in the tree i. Instead of maximizing (4) directly, it is more convenient to minimize the negative of the log-likelihood function:

$$\log \mathcal{L}(\boldsymbol{\Pi}|\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{t=1}^{|\boldsymbol{\pi}_i|-1} \alpha_x(\log \beta_x + \log d_{x,t,i}) - \log Z_{t,i}(\boldsymbol{\pi}_i|\boldsymbol{\theta}), \tag{5}$$

where $Z_{t,i}(\boldsymbol{\pi}_i|\boldsymbol{\theta}) = \sum_{l=1}^{t} (\beta_l d_{l,t,i})^{\alpha_l}$.

For more details about the optimization see the Appendix.

3. DATASETS

We have analyzed the discussion cascades of four websites. The first three can be regarded as news aggregators and contain cascades where people discuss news or interesting links during some period. The fourth dataset corresponds to the discussion pages associated to the articles of the English Wikipedia. In the following paragraphs we give a more detailed description of the datasets and the corresponding websites. The detailed sizes of the datasets can be found in Table 1.

Slashdot (SL) : Slashdot[2] 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. The comments can be nested which allows us to extract the tree structure of the discussion. A single news post triggers typically about 200 comments (most of them in a few hours) during the approx. 2 weeks he is open for discussion. Our dataset contains the entire amount of discussions generated at Slashdot during a year (from August 2005 to August 2006) about 2 million comments to 10,000 different news post. More details about this dataset can be found in [7].

Barrapunto (BP) : Barrapunto[3] 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. They differ in the language they use and the content of the news stories displayed, which normally do not overlap. The volume of activity on Barrapunto is significantly lower than the one of its

[2] `http://slashdot.org/`
[3] `http://barrapunto.com/`

Table 1: Dataset statistics

dataset	#cascades	#nodes	max. size	max. users	total users	repeated user
SL	9,820	2,028,518	1,567	1,031	93,638	>1 99%
BP	7,485	397,148	1,040	180	6,864	>1 85%
MN	58,613	2,220,714	2,718	1,021	53,877	>1 34% >5 70%
WK	871,485	9,421,976	32,664	5,969	350,958	>1 34% >5 96%

English counterpart. A news story on Barrapunto triggers on average around 50 comments. Our dataset contains the activity on Barrapunto during 3 years (from January 2005 to December 2008), about 7,500 posts which receive a total of approx. 400,000 comments.

Meneame (MN) : Meneame[4] 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, comments are displayed in from of a list so that the tree structure is hidden to the user. However, users use `#n` in the text of their comments to indicate a reply to the n-th comment in the comment list. This can be used to extract the tree structures we analyze in this study. We focus only on the comments under the promoted stories in this study. Our dataset contains all promoted stories (nearly 60,000) and the corresponding comments (more than 2.2 millions) from Meneame during the interval between its first story in December 2005 until July 2009.

Wikipedia (WK) : The English Wikipedia[5] 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 organize 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 and cleaning process will be published elsewhere [16].

[4] `http://www.meneame.net/`

[5] `http://en.wikipedia.org`

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.

3.1 Global analysis

In this section we give a brief overview about some general characteristics of the four datasets. Several indicators are shown in Table 1. As the columns 4 and 5 show, the largest cascades in the datasets can be composed of hundreds of nodes and propagate across hundreds of users. The largest cascade corresponds to the Wikipedia involving 5,969 users and 32,664 nodes. The largest cascade of Barrapunto affects 180 users only and has 1,040 nodes.

It is interesting to consider this quantity relative to the size of the underlying social network (compare columns 5 and 6, where we indicate the total number of users during the crawled period). We see a remarkable fact: the percentage of users affected by the largest cascade is very small. In particular, it varies from a 1.1% for Slashdot and 2.6% in Barrapunto, the dataset which we saw that presented the smallest cascade in absolute terms. Globally, these results show that even the largest cascades only affect a very small portion of the entire underlying social network.

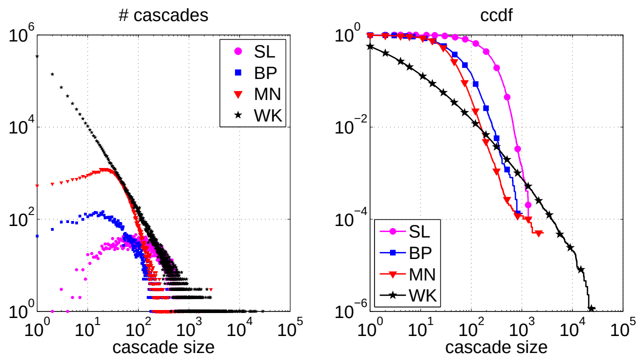
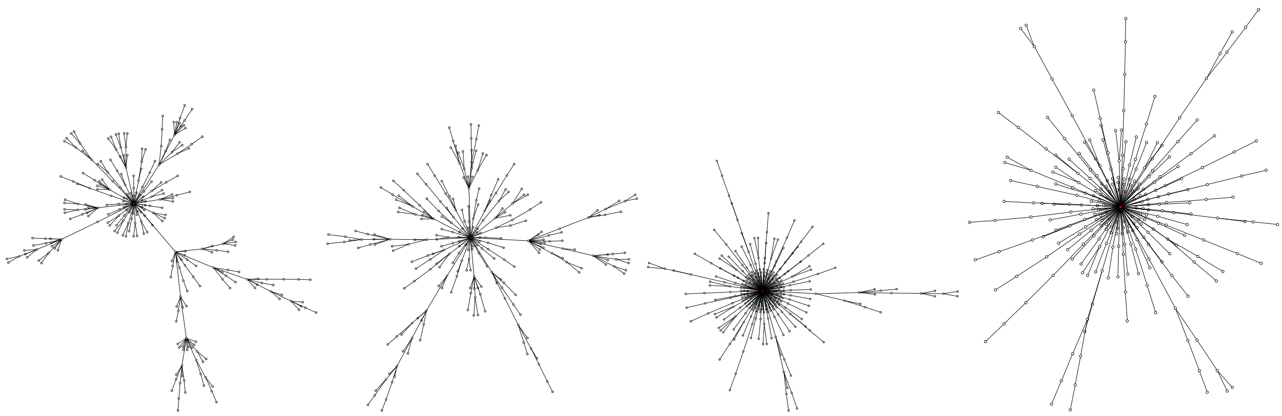


Figure 3: Cascade sizes for the different datasets

A characteristic feature of discussion cascades is the high frequency of user participation. Evidence for this idea is provided in column 7, where we show the percentage of cascades in which at least one user is involved more than once for cascades with more than two nodes (for MN and WK, we also show the percentage for cascades with more than five nodes). With the exception of Meneame, all datasets show very high values. In Slashdot, practically all posts contained at least one user who commented more than once. [6] An important consequence of this fact is that information diffusion may not be properly explained using epidemic models

[6] We only considered registered users here.



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

Figure 4: Examples of synthetic discussion cascades.

such as SIR (susceptible-infected-recovered) models unlike in other scenarios like photo popularity [4] or fanning pages [22].

Figure 3 shows the distribution of the cascade sizes of the four datasets. As expected, all distributions are positively skewed, showing a high concentration of relatively short cascades and a long tail with large cascades.

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 cascade sizes of Wikipedia is closer to a scale-free distribution, in line with the cascades found in weblogs [18]. However, 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 behavior for cascade sizes > 50. Barrapunto falls in the middle and ,interestingly, is more similar to Meneame than to Slashdot.

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. Compared with Wikipedia, we can say that Meneame is the news aggregator which has most similarities with it.

Figure 1 illustrates the different types of cascades which we found. We plot representative cascades 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 are highly scored and become more popular. In Barrapunto we observe similar structures, although their persistence is less noticeable. On the contrary, Meneame is characterized by having high concentration of nodes at the first level together with rare but long chains of thin threads. This seems to represent a pattern where most people just leave a comment which often is not replied by other users, but which sporadically can trigger a long dialog between a few users. We note that this phenomenon might be caused by the fact that the cascade tree is 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. RESULTS

In this section we validate the proposed model by comparing the real cascades to the synthetic ones generated using the model.

4.1 Model validation description

We use the cascades from the four datasets to validate the proposed PA model with bias. The parameters are optimized for each dataset independently and we generate the same number of synthetic cascades as the number of real cascades extracted from each dataset. The size of each synthetic cascade is pre-determined drawing a pseudo-random number from the empirical distribution of cascade sizes (see Figure 3). We calculate the following quantities from the empirical data and from the synthetic cascades produced by the model:

Root node degree probability distribution: Each cascade has a root degree, which is the number of *direct* contributions to the root.

Non-root nodes degree distribution: We also consider all non-root nodes and compute the probability distribution of their degrees or branching factors, i.e. how many times a comment is replied.

Subtree sizes distribution: For each non-root node, we compute the probability distribution of the total number of its descendants.

Mean node depth: Each non-root node belongs to one level of the cascade. We compute the mean over all the levels of all the nodes.

Size - Proportion of direct reactions : We compute the

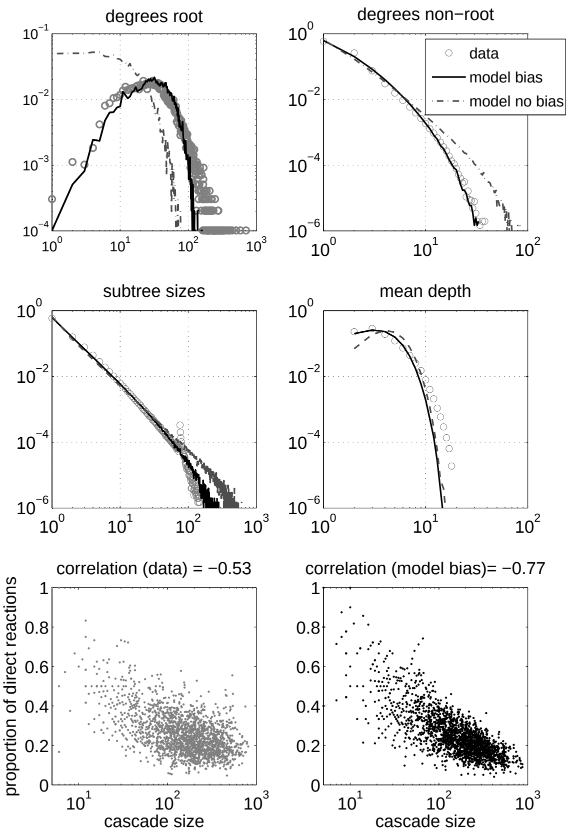


Figure 5: Slashdot

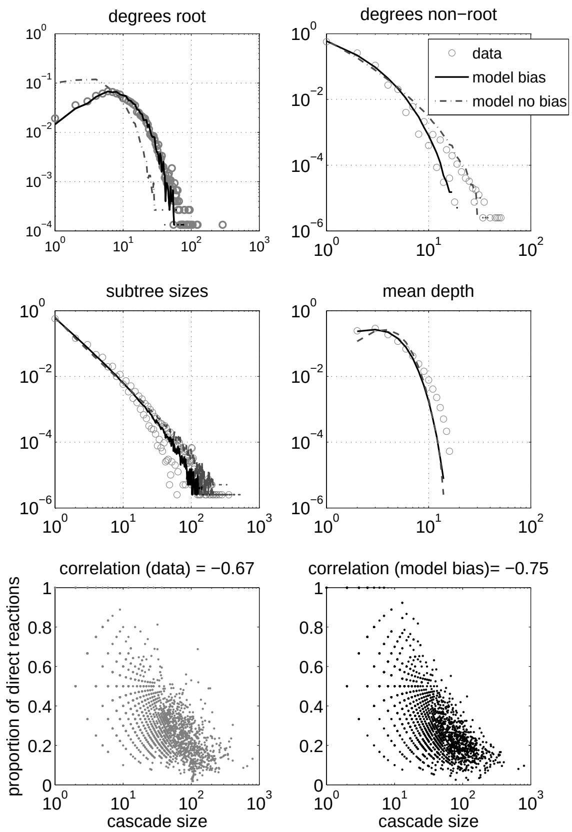


Figure 6: Barrapunto

relation between the size of a cascade and the proportion of direct reactions to the root and analyze if they are correlated.

4.2 Structure of the cascades

Figures 5, 6, 7 and 8 show plots of the previous quantities for each dataset and the outcomes of both PA models with and without bias to the root.

Overall, the model with bias is able to capture reasonably well all the measured properties, except the mean depth. In particular the degree distributions of the root nodes are very accurately reproduced, even though each dataset exhibits a different profile (see top-left plot of the figures). For this quantity, the difference between using or not a bias term is clearly manifested. A similar behavior is observed in the correlations between the size of the cascade and the proportion of direct reactions (bottom plots of the figures). Although the scatter plots differ substantially across datasets, the model with bias is able to reproduce them qualitatively. The correlations of the model without bias are very poor (data not shown).

The model with bias also generates correct subtree sizes in general, with the exception of Meneame, which we postulate is caused by the particularities of the platform (see Section 6 for details). On the contrary, the model without bias systematically produces longer tails than the real ones. We also see a slight deviation on the tails of the non-root degree distributions, more noticeable in Meneame and the Wikipedia, for which the distributions are marginally less skewed. Nevertheless, the model with bias produces much better approximations than the model without bias, which again systematically produces longer tails than the real ones.

Both models tend to produce shorter tails for the mean depth distribution in all datasets. This seems to be a current limitation of the model. Although for Slashdot and Barrapunto this deviation is not very severe, for the other two datasets we observe clear discrepancies at the tail of the distributions. Notice that in this case, the model without bias is unable even to reproduce the probability mass corresponding to the first values of the distribution. We will return to this point in Section 6.

To conclude this section, we show in Figure 4 the synthetic counterpart of Figure 1, where we plot representative cascades with similar sizes selected randomly from each of the four synthetic datasets. We can see that the generated cascades present a strong resemblance with the real ones.

4.3 Evolution of the cascades

After having compared the main structural properties of the synthetic trees generated by our model with the real ones, we will now investigate whether the PA model with bias is also able to reproduce the growth process of the cascades. In other words, if we take intermediate snapshots of the cascades during their evolution, how close match the synthetic trees their archetypes?

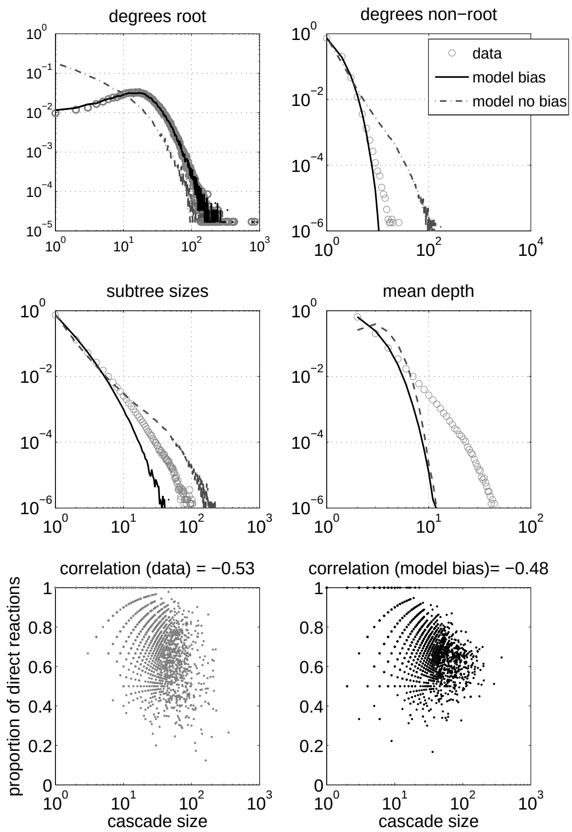


Figure 7: Meneame

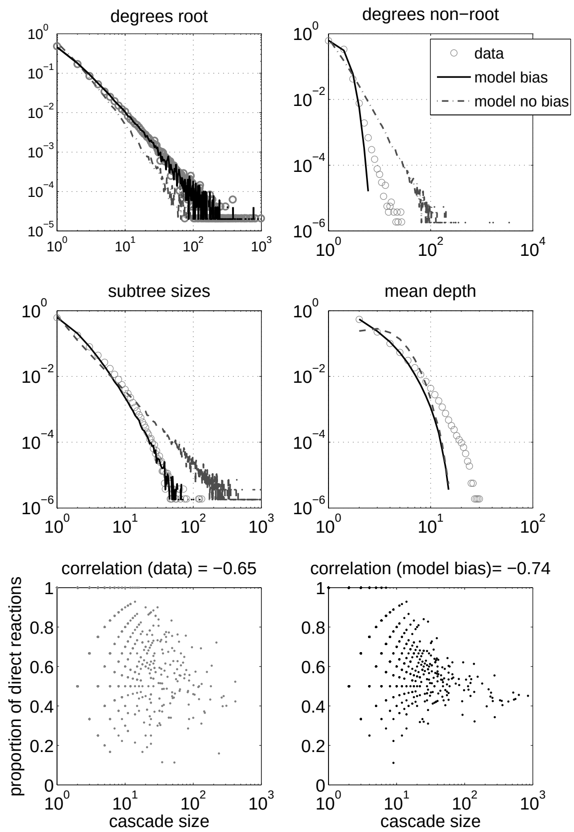


Figure 8: Wikipedia

To this end we record two quantities: the **width**[7] and the **mean depth** of the trees every time a new node is added (at every timestep). Note that the timesteps in the model do not coincide with the actual time differences between the comments. They just reflect the sequence of the comments attaching to the cascade. In reality, information spreading is conditioned to the large heterogeneity present in human activity, for instance induced by circadian cycles, which results in information transmission speeds governed by subexponential distributions, i.e. log-normals or power-laws [10, 12, 20]. Capturing the growth process of the real cascades is therefore a challenging task for our model.

The average over all these with- and depth evolution curves is presented in Figures 9 and 10 for all four datasets comparing the original cascades (continuous lines with symbols) as well as the biased model (dashed lines). We observe a nearly perfect coincidence between our model and the data in the evolution of the width of the discussions (Figure 9), for three of the four datasets. Only in the case of Slashdot the PA model with bias underestimates the width of the tree, although it still reproduces the same curve if normalized by the final depth.

The picture in the case of the mean depth (Figure 10) of the cascades is less favorable, but still shows a reasonable coincidence of our model with the data. In the case of Wikipedia, although the model underestimates the mean depth, it reproduces a rescaled version of it. The other three datasets show a similar profile. Initially, the trees produced by the model are too deep and the mean depth is overestimated. This deviation is corrected at some point and afterward, the opposite effect takes place: when the depth of the synthetic trees saturates, the depth of the real ones still grows. The initial deviation is specially severe in Slashdot, for which remarkably the final mean depth eventually is very close to the one of the real cascades. A possible way to overcome this problem is discussed in Section 6.

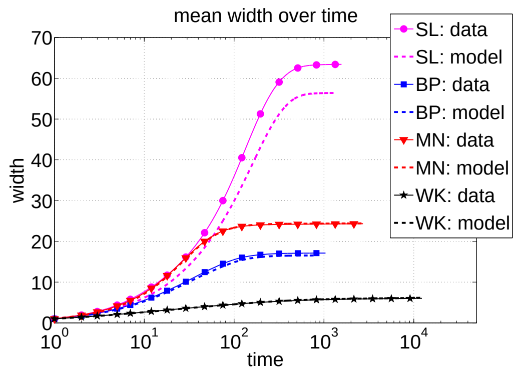


Figure 9: Evolution of the width

[7]The width of the tree is the maximum over the number of nodes per level.

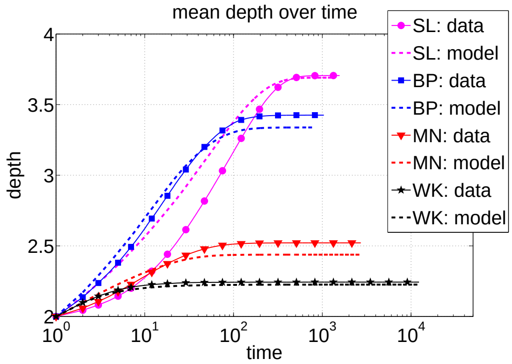


Figure 10: Evolution of the mean depth

4.4 Interpretation of parameters

Can we derive conclusions about the communication habits which characterize each website based on the obtained parameters which best fit each model? Figure 11 shows the optimal parameter values for each dataset in a three dimensional plot, where the horizontal and vertical axis correspond to α_1 and α_c respectively and the size of the marker to the value β. Table 2 shows the same values numerically.

The role of the exponents α_1 and α_c in the model is to quantify the degree of preferential attachment of the root node and the non-root nodes respectively. The higher the values they take, the more relevant is our popularity measure to determine the attractiveness to new nodes in the cascade. For instance, values very close to zero imply a random cascade where new nodes are linked to existing ones with uniform probability. We can use the established theoretical results described in Section 2 to characterize the websites under study.

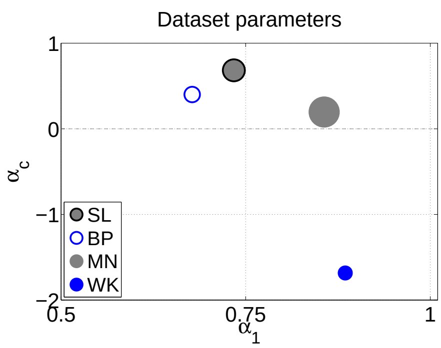


Figure 11: Parameters obtained for the different datasets for initial attractiveness $c = 1$.

The different exponents α_1 are all sublinear (< 1), relatively high, and very similar in all datasets, indicating a strong PA in the root process of all cascades. According to this quantity, Slashdot and Barrapunto appear clustered, in agreement with the known similarities of both websites. On the other hand, Meneame and Wikipedia present a higher and almost identical value, suggesting a very similar role of the root nodes in the PA mechanism of both websites.

A clear segregation between the group of three news media websites and the Wikipedia is manifested on Figure 11 in

Table 2: Optimal parameters

dataset	α_1	α_c	β
SL	0.734	0.683	1.302
BP	0.678	0.400	0.950
MN	0.856	0.196	1.588
WK	0.884	−1.684	0.794

the value of α_c. The first group, in order of increasing value of α_c, consists of Meneame, Barrapunto and Slashdot. The smaller value of ≈ 0.2 found for Meneame indicates that the diffusion of the news comments in this website follows predominantly a random process. This can be again influenced by the lack of explicit information about the popularity a given comment. Slashdot in this case, has the highest value, $\alpha_2 \approx 0.68$, even higher than α_1 for Barrapunto. It is also very similar to α_1 for the same dataset. This similarity may capture the special quality of the Slashdot comments. In a sense, good comments may behave like posts and may act eventually as effective sources of information diffusion.

The case of Wikipedia is very special. It presents the largest (in absolute value) and negative α_c. What are the implications of this result? An interpretation is that the process of cascade growth in Wikipedia is actually an *inverse* PA process. The suggested process here is that once an article has been originated, it will derive in a collaborative reciprocal chain between a very reduced group of contributors. So once a node has received a reply it will be less likely to receive another one than the replying new comment itself. In other words, nodes with degree equal to one (leaf nodes) are much more likely to be linked with a new node than nodes which have higher degrees.

Finally, the parameter β acts as a weight which expresses how important is the root of the cascade in relation with the non-root nodes. The parameter β is especially important when the degree of the root node is low, i. e. in the beginning of the cascade. It acts as initial weight of the root node and determines whether initially many nodes attach to the root or rather to one of the first comments. We observe thus that Meneame shows the largest initial predominance of direct reactions, while Wikipedia gives higher probability mass to the comments, allowing thus large chains already with a small number of nodes. The values for Slashdot and Barrapunto lie in between indicating an intermediate initial preference for the root node, showing Barrapunto a higher probability for early reply-chains than Slashdot.

What would be the scenario if one tries to explain the cascades using the simple model without bias to the root? In this case, the exponents are higher. Barrapunto (0.820) and Slashdot (0.975) are still sublinear whereas the values for Meneame (1.360) and Wikipedia (1.161) are both larger than one. In the latter models, one node of the cascade (not necessarily the post) acts as the "gel" node and attracts many comments. However, this mechanism, as we have established in the previous section, does not provide valid explanation for the observed cascades.

Summarizing, the optimal parameters allow an interpretation of the communication habits of each social space. This representation also leads to different classification as a function of the parameters. For instance, α_1 separates Slashdot and Barrapunto from Wikipedia and Meneame, and α_c splits up the Wikipedia from the three news media.

5. RELATED WORK

Due to the increasing availability of empirical data on cascades, extensive work is appearing with focus on how information cascades are propagated in a social network.

At a statistical description level, information cascades have been analyzed in detail for particular social spaces. In a Twitter analysis [14], the authors report cascades (called retweet trees) which are predominantly shallow and wide (maximum depth is 11). Flickr [5] shows the remarkable phenomena that popular photos spread slowly and not widely. This is in harmony with our findings which report that even the largest realizations reach a very small proportion of the social network.

Blog cascades, which conceptually resembles most similarities with the discussion cascades of this work, have been analyzed in [18]. Interestingly, although one would expect blog cascades to share more similarities with the discussion cascades existing in Slashdot or Meneame, it is the Wikipedia the dataset which shows most similar patterns to the blog cascades (see Figure 3). In [8], a model of both blogger (user) and cascades was presented which reproduces global temporal and structural aspects of the blogosphere. We note that the motivation of our work is rather different. Whereas [18, 8] aims for finding the simplest, parameter-free model able to describe both user network and cascade behavior, we look for a parameterized model from which we can describe communication habits which characterize a particular website (see Section 4.4).

The most important difference between our approach and the aforementioned ones is that in our datasets a new comment can in principle choose its parent between the entire set of comments. Further, (with the exception of Meneame) the complete information of the cascade is available to it which makes it a perfect medium for testing models which assume such accessibility, such as the one considered here. In this sense, our data avoids selection bias which strongly influences the estimation of these processes [6].

6. DISCUSSION

We have presented for the first time a thoughtful analysis and comparison of the different discussion cascades to characterize three popular news media websites and the English pages of the Wikipedia. Our analysis highlights the heterogeneities between the discussion cascades, which can be conditioned from two factors, namely, the page design, or platform, and the audience. Despite this, we have given evidence that a simple model can capture most of the structural properties and the evolution profiles of the real cascades with the particularities of each dataset. Furthermore, its parameters, which can be efficiently learned, allow for a figurative description characterize the communication habits of a website.

For some datasets, the presented model tends to produce too shallow cascades. We postulate that this limitation appears from the inability of the model to reproduce the situation where older nodes gradually become less attractive than newer ones. This occurs especially in mature discussions, where interaction in the end only happens between a few individuals who start to reply mutually to each other and increase the mean depth of the cascades considerably. An extension of the model which could account for this deviation considers the inclusion of an age term. In such a model, the attractiveness of a node is generally determined from a fitness function, which consists of a combination between its novelty (age term) and popularity.

Our findings open several possibilities for extension. Here we have analyzed only four websites, but the method is robust and can be easily adapted to any website with cascades of similar characteristics. The resulting parameters of the model would for example allow to draw a conversation map of different websites, where websites with similar dynamics would lay close by. This would allow an easy and fast characterization of conversation habits on a website. Since more and more websites implement nested commenting features such a map will become increasingly populated in the near future. One could even apply the same methodology to different subgroups of discussion cascades within the same site, e.g, to differentiate the conversation behavior according to content categories or user communities.

6.1 Acknowledgments

We wish to thank David Laniado and Riccardo Tasso for providing the preprocessed Wikipedia dataset and `Meneame.net` for allowing to access an anonymised dump of their database. We also thank Mohammad Gheshlaghi and Alberto Llera for useful discussions.

APPENDIX

Log-likelihood function

In this appendix we describe some considerations related to the log likelihood function (5) we want to minimize. Briefly, we show that the PA model can be formulated as a probability distribution which belongs to the exponential family. Consequently, the optimization problem is convex, e.g. has the convenient property that any local minimum is global.

Without loss of generality we can assume that parameter β_k is of the following form:

$$\beta_k^{\alpha_k} := \exp\left(\beta'_k\right).$$

We can rewrite the PA model defined in Equation (2) as:

$$p(\pi_{t+1} = k|\boldsymbol{\pi}_{(1:t)}) = \frac{1}{Z'_t}\exp\left(\beta'_k + \alpha_k \log d_{k,t}\right),$$

where $Z'_t = \sum_{l=1}^{t}\exp\left(\beta'_l + \alpha_l \log d_{l,t}\right)$. This probability distribution is equivalent to that of the Equation (2), but expressed in terms of the exponential family. The log-likelihood function (4) can be rewritten as:

$$\log \mathcal{L}'(\boldsymbol{\Pi}|\boldsymbol{\theta}) = \sum_{i=1}^{N}\sum_{t=1}^{|\boldsymbol{\pi}_i|-1} \beta'_x + \alpha_x \log d_{x,t,i} - \log Z'_{t,i}(\boldsymbol{\pi}_i|\boldsymbol{\theta}),$$

where $Z'_{t,i}(\boldsymbol{\pi}_i|\boldsymbol{\theta}) = \sum_{l=1}^{t}\exp\left(\beta'_l + \alpha_l \log d_{l,t,i}\right)$. The Hessian of this function (matrix of second order partial derivatives) is always positive semi-definite.

The presented method can therefore be applied to any set of observations which can be expressed as a collection of parent vectors $\boldsymbol{\Pi}$ from which the degrees of each node at each time-step can be obtained. Once the minimization is performed, We can recover the original parameter β_k using:

$$\beta_k = \exp\left(\frac{\beta'_k}{\alpha_k}\right).$$

The basic PA model is the special case where $\alpha = \alpha_1 = \alpha_c$. Note that the bias to the root node can be introduced in several ways:

(A) Using two alphas α_1, α_c but no β ($\beta = 0$).
(B) Using one alpha $\alpha = \alpha_1 = \alpha_c$ and β.
(C) Using two alphas $\alpha = \alpha_1 = \alpha_c$ and β (the approach presented in this manuscript).

As expected, since model **(C)** uses more parameters than **(A)** and **(B)**, the resulting likelihoods and fits are better. In particular, the impact of adding β as a parameter is notable in the approximated measures related to the root node, for instance the root degree distributions.

Notice that the convexity condition does not imply a unique global minimum. It could happen that the same minimum is attained for a range of parameter values. In practice, we used as an optimization procedure the Nelder-Mead simplex algorithm (implemented as the function `fminsearch` in Matlab) which is an unconstrained non-linear direct search method that does not use numerical or analytic gradients. Starting from many different random initial conditions, we did not experience multiple local minima problems in any of the datasets, so we can conclude that the presented optimal values for each dataset are unique.

Finally, we would like to mention that, although the optimal parameters depend on the choice of c (the initial attractiveness), the resulting models are equivalent for a large set of c (we experimented with $c \in \{1, 5, 10\}$. For consistency with the existent literature [15] we consider $c = 1$.

A. REFERENCES

[1] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[2] E Ben-Naim and P L Krapivsky. Stratification in the preferential attachment network. *Journal of Physics A: Mathematical and Theoretical*, 42(47):475001, 2009.

[3] B. F. Blasio, A. Svensson, and F. Liljeros. Preferential attachment in sexual networks. *Proceedings Of The National Academy Of Science USA*, 104(26):10762–10767, 2007.

[4] M. Cha, A. Mislove, B. Adams, and K. P. Gummadi. Characterizing social cascades in flickr. In *WOSP '08: Proceedings of the first workshop on Online social networks*, pages 13–18, New York, NY, USA, 2008. ACM.

[5] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on WWW*, pages 721–730, New York, NY, USA, 2009. ACM.

[6] B. Golub and M. O. Jackson. Using selection bias to explain the observed structure of internet diffusions. *Proceedings of the National Academy of Sciences*, 107(24):10833–6, 2010.

[7] V. Gómez, A. Kaltenbrunner, and V. López. Statistical analysis of the social network and discussion threads in slashdot. In *Proceeding of the 17th international conference on WWW*, pages 645–654. ACM, 2008.

[8] M. Götz, J. Leskovec, M. McGlohon, and C. Faloutsos. Modeling blog dynamics. In *ICWSM*, 2009.

[9] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th international conference on WWW*, pages 491–501, New York, NY, USA, 2004. ACM.

[10] J. L. Iribarren and E. Moro. Impact of human activity patterns on the dynamics of information diffusion. *Phys. Rev. Lett.*, 103(3):038702, Jul 2009.

[11] H. Jeong, Z. Néda, and A. L. Barabási. Measuring preferential attachment in evolving networks. *EPL (Europhysics Letters)*, 61(4):567, 2003.

[12] A. Kaltenbrunner, V. Gómez, and V. López. Description and prediction of Slashdot activity. In *Proc. 5th Latin American Web Congress (LA-WEB 2007)*, Santiago de Chile, 2007. IEEE CS.

[13] M. Kearns, S. Suri, and N. Montfort. An experimental study of the coloring problem on human subject networks. *Science*, 313(5788):824–827, 2006.

[14] H. Kwak, Ch. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on WWW*, pages 591–600, New York, NY, USA, 2010. ACM.

[15] Krapivsky P. L., Redner S., and F. Leyvraz. Connectivity of growing random networks. *Phys. Rev. Lett.*, 85(21):4629–4632, Nov 2000.

[16] D. Laniado, R. Tasso, Y. Volkovich, and A. Kaltenbrunner. When the Wikipedians talk: Network and tree structure of Wikipedia discussion pages. under review.

[17] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. In *EC '06: Proceedings of the 7th ACM conference on Electronic commerce*, pages 228–237, New York, NY, USA, 2006. ACM.

[18] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs: Patterns and a model. In *Society of Applied and Industrial Mathematics: Data Mining (SDM07)*, 2007.

[19] D. Liben-Nowell and J. Kleinberg. Tracing information flow on a global scale using Internet chain-letter data. *Proceedings of the National Academy of Sciences*, 105(12):4633–4638, 2008.

[20] R. D. Malmgren, J. M. Hofman, L. A.N. Amaral, and D. J. Watts. Characterizing individual communication patterns. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 607–616, New York, NY, USA, 2009. ACM.

[21] Everett M. Rogers. *Diffusion of innovations*. Free Press, New York, NY [u.a.], 5th edition, 2003.

[22] E. Sun, I. Rosenn, C. Marlow, and T. M. Lento. Gesundheit! modeling contagion through facebook news feed. In *ICWSM*. The AAAI Press, 2009.

[23] C. Wiuf, M. Brameier, O. Hagberg, and M. P. H. Stumpf. A likelihood approach to analysis of network data. *Proceedings Of The National Academy Of Science USA*, 103(20):7566–7570, 2006.