HOMOGENEOUS TEMPORAL ACTIVITY PATTERNS IN A LARGE ONLINE COMMUNICATION SPACE

Andreas Kaltenbrunner[*,‡,§], Vicenç Gómez[*,‡], Ayman Moghnieh[*], Rodrigo Meza[*,‡],
Josep Blat[*,‡] and Vicente López[*,‡]

[*] *Universitat Pompeu Fabra, Departament de Tecnologia, Passeig de Circumval·lació 8, 08003 Barcelona, Spain*

[‡] *Barcelona Media Centre d'Innovació, Ocata 1, 08003 Barcelona, Spain*

ABSTRACT

The many-to-many social communication activity on the popular technology-news website Slashdot has been studied. We have concentrated on the dynamics of message production without considering semantic relations and have found regular temporal patterns in the reaction time of the community to a news-post as well as in single user behavior. The statistics of these activities follow log-normal distributions. Daily and weekly oscillatory cycles, which cause slight variations of this simple behavior, are identified. The findings are remarkable since the distribution of the number of comments per users, which is also analyzed, indicates a great amount of heterogeneity in the community. The reader may find surprising that only two parameters, those of the log-normal law, allow a detailed description, or even prediction, of social many-to-many information exchange in this kind of popular public spaces.

KEYWORDS

Social interaction, information diffusion, log-normal activity, heavy tails, Slashdot

1. INTRODUCTION

Nowadays, an important part of human activity leaves electronic traces in form of server logs, e-mails, loan registers, credit card transactions, blogs, etc. This huge amount of generated data allows to observe human behavior and communication patterns at nearly no cost on a scale and dimension which would have been impossible some decades ago. A considerable number of studies have emerged in recent years using some part of these data to investigate the time patterns of human activity. The studied temporal events are rather diverse and reach from directory listings and file transfers (FTP requests) (Paxson and Floyd 1995), job submissions on a supercomputer (Kleban and Clearwater 2003), arrival times of consecutive printing-job submissions

[§] Corresponding author: e-mail: andreas.kaltenbrunner@upf.edu; Fax: +34 93 542 2517

(Harder and Paczuski 2006) over trades in bond (Mainardi et al. 2000) or currency futures (Masoliver et al. 2003) to messages in Internet chat systems (Dewes et al. 2003), online games (Henderson and Bhatti 2001), page downloads on a news site (Dezso et al. 2006) and e-mails (Johansen 2004). A common characteristic of these studies is that the observed probability distributions for the waiting or inter-event times are heavy tailed. In other words, if the response time ever exceeds a large value, then it is likely to exceed any larger value as well (Sigman 1999). A recent study (Barabási 2005) tries to explain this behavior under the assumption that these heavy tailed distributions can be well approximated by a power-law or at least by a power-law with an exponential cut-off (Newman 2005). The cited study presents a model which seems to explain the distribution of e-mail response times and has been used later to account for the inter-event times of web-browsing, library loans, trade transactions and correspondence patterns of letters (Vázquez et al. 2006). However, the hypothesis of a power-law distribution is not generally accepted, at least in case of e-mail response times. Stouffer et al. (2006) claim that the data can be much better fitted with a log-normal distribution (Limpert et al. 2001). This debate has been repeated across many areas of science for decades, as noticed by Mitzenbacher (2004).

To the authors' knowledge no study of this type has been performed on systems where social interaction occurs in a more complex manner than just person to person (one-to-one) communication. We think it is valuable to analyze the temporal patterns of the many-to-many social interaction on a technology-related news-website which supports user participation. We have chosen Slashdot[1], a popular website for people interested in reading and discussing about technology and its ramifications. It gave name to the “Slashdot effect” (Adler 1999), a huge influx of traffic to a hosted link during a short period of time, causing it to slow down or even to temporarily collapse.

Slashdot was created at the end of 1997 and has ever since metamorphosed into a website that hosts a large interactive community capable of influencing public perceptions and awareness on the topics addressed. Its role can be metaphorically compared to that of commercial malls in developed markets, or hubs in intricate large networks. The site's interaction consists of short-story **posts** that often carry fresh news and links to sources of information with more details. These posts incite many readers to **comment** on them and provoke discussions that may trail for hours or even days. Most of the commentators register and comment under their nicknames, although a considerable amount participates anonymously.

Although Slashdot allows users to express their opinion freely, moderation and meta-moderation mechanisms are employed to judge comments and enable readers to filter them by quality. The moderation system was analyzed by Lampe and Resnick (2004) who concluded that it upholds the quality of discussions by discouraging spam and offending comments, marking a difference between Slashdot and regular discussion forums. This high quality social interaction has prompted several socio-analytical studies about Slashdot. Poor (2005) and Baoill (2000) have both conducted independent inquiries on the extent to which the site represents an online public sphere as defined by Habermas (1962/1989).

Given that a great amount of users with different interests and motivations participate in the discussions, one would expect to observe a high degree of heterogeneity on a site like Slashdot. However, what if the posts and comments were analyzed just as imprints of an occurring information exchange, with no regard to semantic aspects? Is there a homogeneous behavior pattern underlying heterogeneity? To answer these and related questions we collected and studied one

[1] http://www.slashdot.org

year's worth of interchanged messages along with the associated metadata from Slashdot. We show here that the temporal patterns of the comments provoked by a post are very similar, indicating that homogeneity is the rule not the exception. The temporal patterns of the social activity fit accurately log-normal distributions, thus giving empirical evidence of our hypothesis and establishing a link with previous studies where social interaction occurs in a simpler way.

Finally, our analysis allows more insight into questions such as: is there a time-scale common to all discussions, or are they scale-free? What does incite a user to write a comment, is it the relevance of the topic, or maybe just the hour of the day? Can we predict the amount of activity triggered by a post already some minutes after it has been written? Which type of applications can we devise on the basis of using these conclusions?

The rest of the article is organized as follows: In section 2 we briefly explain the process of data acquisition. We then present the results in section 3 providing first an overview of the global activity and then explaining our analysis in detail. We finish the paper with section 4 where we discuss the results.

2. METHODS

In this section we explain the methods used to crawl and analyze Slashdot. The crawled[2] data correspond to posts and comments published between August 26th, 2005 and August 31th, 2006. We divided the crawling process into two stages. The first stage included crawling the main HTML (posts) and first level comments and the second stage covered all additional comment pages. Crawling all the data took 4.5 days and produced approximately 4.54 GB of data. Post-processing caused by the presence of duplicated comments was necessary (due to an error of representation on the website). Although a high amount of information was extracted from the raw HTML (sub-domains, title, topics, hierarchical relations between comments) we concentrated only on a minimal amount of information: **type** of contribution (either post or comment), its **identifier**, **author**'s identifier and **time-stamp** or date of publishing. The selected information was extracted to XML-files and imported into Matlab where the statistical analysis was performed. Table 1 shows the main quantities of the crawling and the extracted data.

Table 1. Main quantities of crawling and retrieved data.

Period covered	26-8-05 – 31-9-06
Time needed for crawling	4.5 days
Amount of data mined	4.54 GB
Posts	10016
Comments	2075085
Commentators	93636
Anonymous comments	18.6%

The time-stamps of post and comments can be obtained from Slashdot with minute-precision and corresponded to the EDT time zone (= GMT$-$4 hours). They allow to calculate the follow-

[2] Software used: wget, Perl scripts, and Tidy on a GNU/Linux, Ubuntu 6.0.6 OS.

ing two quantities:

The **Post-Comment-Interval (PCI)** stands for the difference between the time-stamps of a comment and its corresponding post.

The **Inter-Comment-Interval (ICI)** refers to the difference between the time-stamps of two consecutive comments of the same user (no matter what post he/she comments on).

3. RESULTS

In this section we first give an overview of the global activity looking at the data on different temporal scales and analyzing some relations between variables of interest. We then focus on the activity provoked by single posts and analyze the behavior of single users, concentrating on the most active ones.

3.1 Global cyclic activity

As previously explained, comments can be considered as reactions triggered by the publishing of posts. This difference in nature between both types of contributions justifies a separate analysis of their dynamics.

Figure 1 shows (normalized) mean activity and standard deviations of both posts and comments. It illustrates patterns in agreement with the social activity outside the public sphere. Figure 1a shows regular, steady activity during working days which slows down during weekends. This weekly cycle is interleaved by daily oscillations illustrated in Figure 1b. The daily activity cycle reaches its maximum at 1pm approximately and its minimum during the night between 3am and 4am. Although Slashdot is open to public access around the world, we see that its activity profile is clearly biased towards the American time-schedule.

Interestingly, although post activity shows more fluctuations and higher standard deviations than comment activity, there is little discrepancy between their mean temporal profiles. This difference in the deviations is not surprising given the greater number of comments (see Table 1). We notice that the standard deviations of the daily post- and commenting activities also show

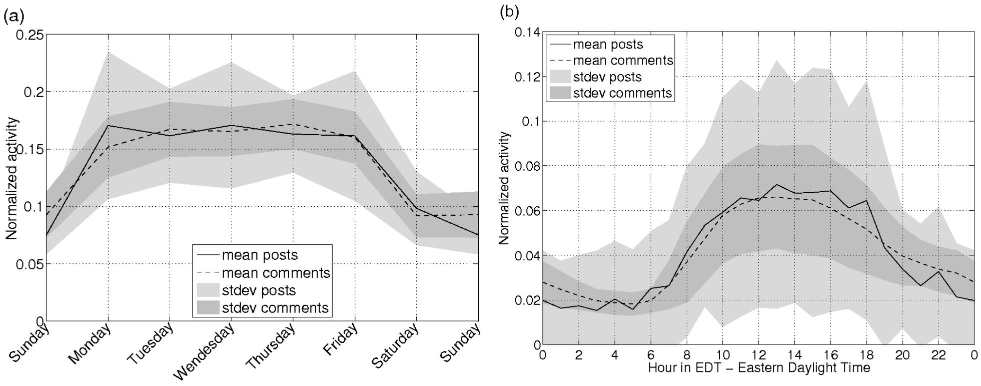

Figure 1. **(a)** Weekly and **(b)** daily activity cycles.

similar cyclic behavior (Figure 1b).

3.2 Post-induced activity

In this section we analyze the activity (comments) a post induces on the site. The histogram of Figure 2 gives an idea of the number of comments the posts receive. Note that half of the posts provoke more than 160 comments and some of them even trigger more than 1000. To analyze the time-distribution of these comments we study their post-comment intervals (PCIs).

3.2.1 Analysis of the activity generated by a single post

We are especially interested in the resulting probability distribution of all the PCIs of a certain post. This distribution reveals us the probability for a post to receive a comment t minutes after it has been published. Figures 3a and 3b show this distribution for a post which provoked 1341 comments. Although there are some important fluctuations, the characteristic shape of the probability density function (pdf) resembles a log-normal distribution. This becomes even clearer if the cumulative probability distribution (cdf) is observed, since there the fluctuations of the pdf are averaged out. Figures 3c and 3d show a good fit of the PCI-cdf of the data with the cdf of the log-normal distribution.

To classify the quality of the fit we have used a normalized error measure ε based on the ℓ^1-norm (see Appendix A). For the post shown in Figure 3 we obtain $\varepsilon = 0.007$, meaning that the average error is below 1%.

The PCI-cdf of three more posts can be observed in Figure 4. The top two sub-figures show good fits, indicating that the PCI is well approximated even for a small number of comments. However, the fit is not that accurate for all posts. For example, the comments of the post shown in Figure 4 (bottom) start to show considerable different behavior from the expected log-normal approximation about 3 hours after its publication. The activity is lower than the predicted one, but starts to increase again at about 6am in the morning the next day. At around at 8:30pm it increases further to recover the lost activity during the night. More such increases and decreases of activity can be observed during the following days. The time-spans of variations in activity coincide quite exactly with the average daily activity cycle shown in Figure 1b. We analyze this coincidence further in the next section.

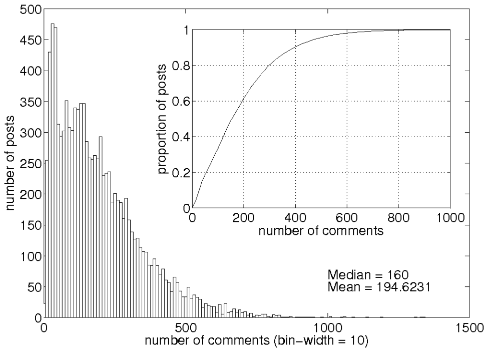

Figure 2. Histogram of the number of comments per post (inset shows the corresponding cdf).

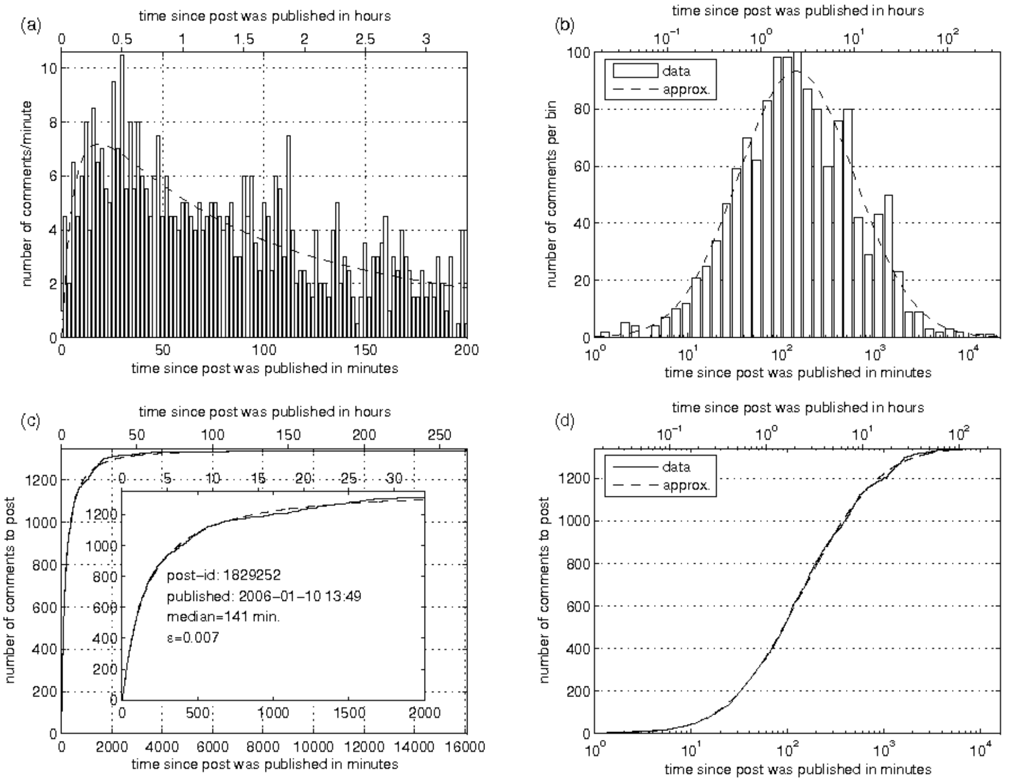

Figure 3. Log-normal approximation (dashed lines) of the PCI-distribution (solid lines and bars) of a post which received 1341 comments. **(a)** Comments per minutes (bin-with= 2 for better visualization) for the first 200 minutes after the post has been published. **(b)** Same as (a) in logarithmic scale. **(c)** The cumulative distribution of the data shown in (a). Inset shows a zoom on the first 2000 minutes. **(d)** Same as (c) in logarithmic scale.

3.2.2 Comparison of posts

With the log-normal shape of the PCI-distribution identified, we focus on the quality of this approximation in general. We therefore calculate the error measure ε of the fit for all posts which received comments. The resulting distribution of ε can be seen in Figure 5a. For 87% of the posts the approximation error ε is lower than 0.05, and for 29% lower than 0.02.

If we take a closer look at the data, we notice a dependence of ε on the publishing-hour of a post (Figure 5b). The best fit is reached when the post is published between 6am and 11am. Then the mean error increases successively until 11pm to stay high during the night and recover again in the early morning.

This behavior can be understood looking at the daily activity cycle (Figure 1b). The less time the community has to comment on a post during the time-window of high activity, the greater is the need to comment on it the next time the high activity phase is reached, and hence the expected log-normal behavior is altered. Figure 4 (bottom) gives examples of such a late post (published at 10:35pm).

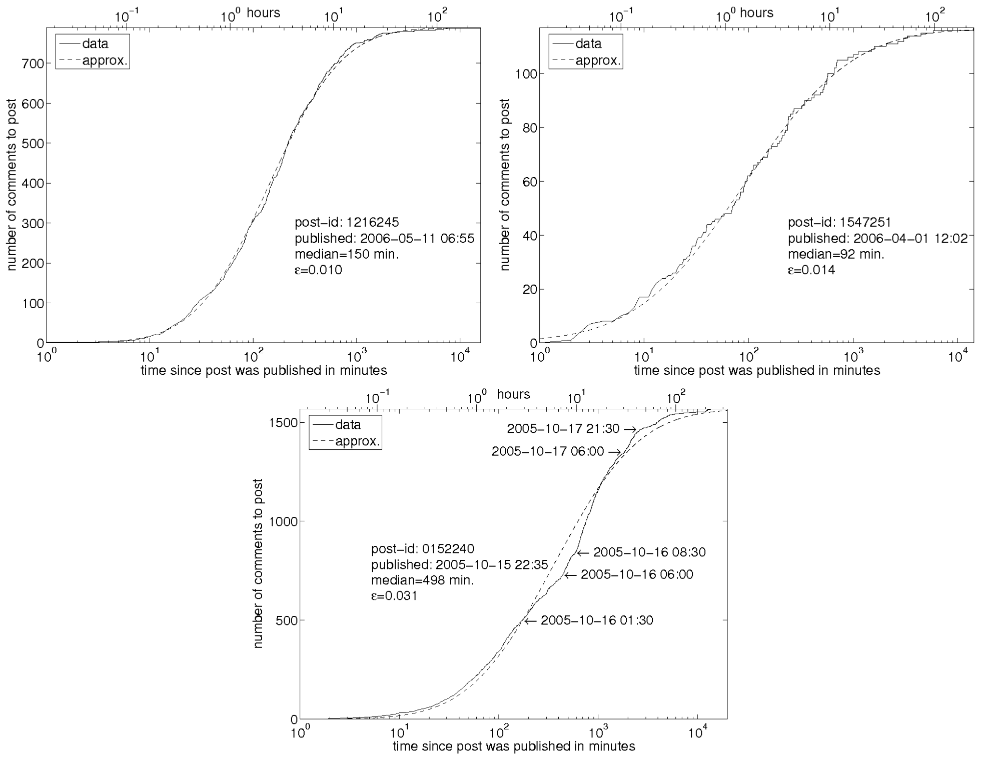

Figure 4. Log-normal approximation of the PCI-distribution of 3 different posts.

The good quality of the approximation allows us to describe the activity triggered by a post with only two parameters, the median[3] and the geometric standard deviation σ_g of the PCI-pdf, commonly used to compare log-normally distributed quantities (Limpert et al. 2001). Figure 6 shows the distribution of these quantities. The inset shows σ_g, which is centered around 1.036 and very similar for all posts. The median of the post-induced activity on the other hand shows more variations, but is rather short (for 50% of the posts it is below 2.5 hours, for 90% below 6.5 hours) compared to the maximum PCI (approx. 12 days). We can thus conclude that although the total activity a post generates covers a large time interval the major part of the activity happens within the first few hours after the post's publication.

3.3 User dynamics

In this section we analyze the activity on Slashdot taking the authorship of the comments into account. We first study the distribution of activity among all the users participating in the debates and then focus on the temporal activity patterns of single users.

[3] Note that the median coincides with the geometric mean for a log-normally distributed random variable.

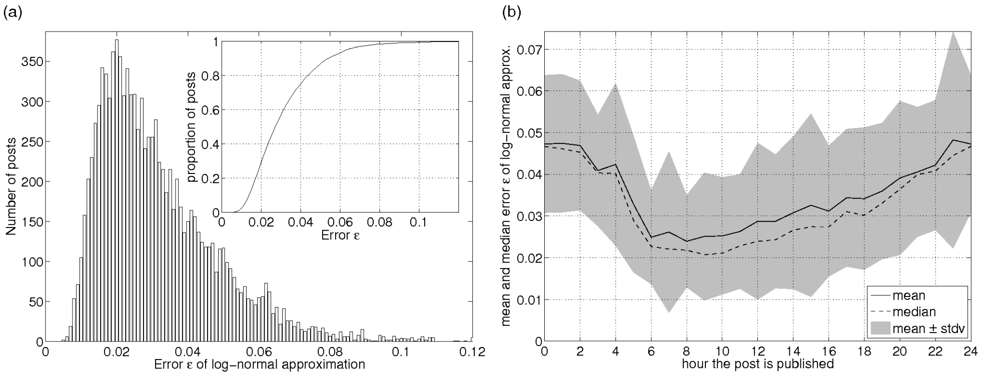

Figure 5. **(a)** Errors ε of the log-normal approximation of the PCI-cdf (bin-width $= 10^{-3}$). Inset shows the corresponding cdf. **(b)** Dependence of mean and median of the approximation error ε on the hour the post is published.

3.3.1 Global user activity

The activity of all users is best illustrated by the distribution of the number of comments per user. It is shown in double-logarithmic scale in Figure 7a. The obtained distribution follows quite closely a straight line, suggesting a power-law probability distribution governing this relation. We note that 53% of the users write 3 or less comments whereas only 93 users (0.1%) write more than 1000 comments. Indeed, after applying linear regression as in other studies (Faloutsos et al. 1999, Albert et al. 1999) we obtain a quite large correlation coefficient $R^2 = -0.97$ for an exponent of $\gamma = -1.79$.

However, if we apply rigorous statistical analysis as proposed in Goldstein et al. (2004) the picture changes. First, we estimate the power-law exponent computing the less biased maximum likelihood estimator (MLE). The resulting exponent $\gamma = -1.5$ differs significantly from the previous one and is illustrated in Figure 7 (dashed-line). Although Figure 7a tempts one to accept the power-law hypothesis, the cdf shown in Figure 7b discards it. It is thus not surprising that the Kolmogorov-Smirnov test forces us to reject the power-law hypothesis with statistical

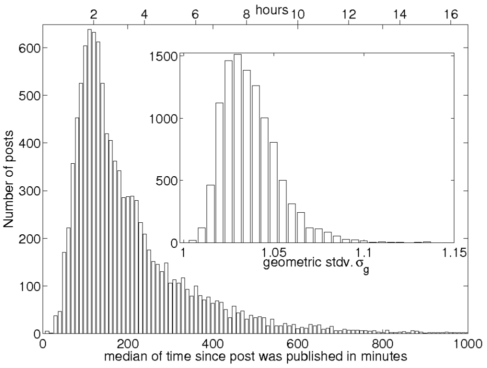

Figure 6. Histograms of medians (bin-width = 10) and geometric standard deviations (inset, bin-width = 0.005) of the PCI-distributions.

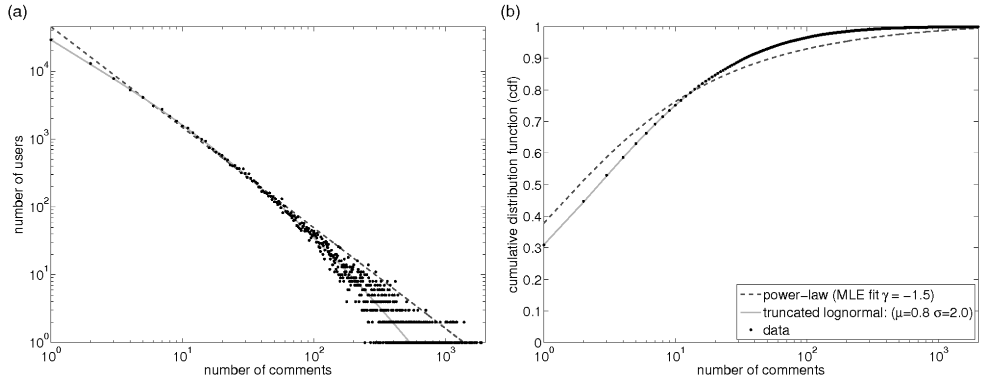

Figure 7. **(a)** Histogram of the number of comments per user and **(b)** and its corresponding cdf.

significance at the 0.1% level.

As an alternative hypothesis to describe the data we propose a (truncated) log-normal probability distribution, shown in Figure 7 as grey-solid-line. Its parameters are found using the MLE. Clearly, the fit is better using this hypothesis. We remark that in many studies some data points (considered outliers) are discarded to improve the power-law fit. Here, in contrast, the truncated log-normal approximation can characterize the entire data-set.

3.3.2 Single user dynamics

After characterizing the user activity at a general level, we investigate the temporal behavior patterns of single users . The analysis concentrates on the two most active users (to protect their privacy we call them user1 and user2). Table 2 shows the number of commented posts and the total number of comments these two users published during the time-span covered by our data.

Table 2. Contributions of the two most active users.

	user1	user2
commented posts	1189	1306
comments	3642	3350

We focus on the distribution of the PCIs of all of their comments as well as on their inter-comment-interval (ICI) distribution, i.e. the time-difference between two comments of the same user.

The PCI-cdf (see Figure 8a) of the two users can also be approximated by a log-normal distribution, although the fit is worse than in the case of the post-induced comment activity. Again we notice a clear dependence of the quality of the fit on the activity cycle (shown in the insets of Figure 8a). The approximation is much better for user1, whose daily and especially weekly activity cycles are much more balanced than those of user2. The activity of the latter user concentrates almost exclusively on the working hours from Monday to Friday. Hence his PCI-distribution shows a clear decrease after 8 but increases again after 16 hours. This increase is less pronounced if only the first comment to a post is considered (data not shown), indicating that the user frequently rechecks the posts he commented the day before to participate again in an ongoing discussion.

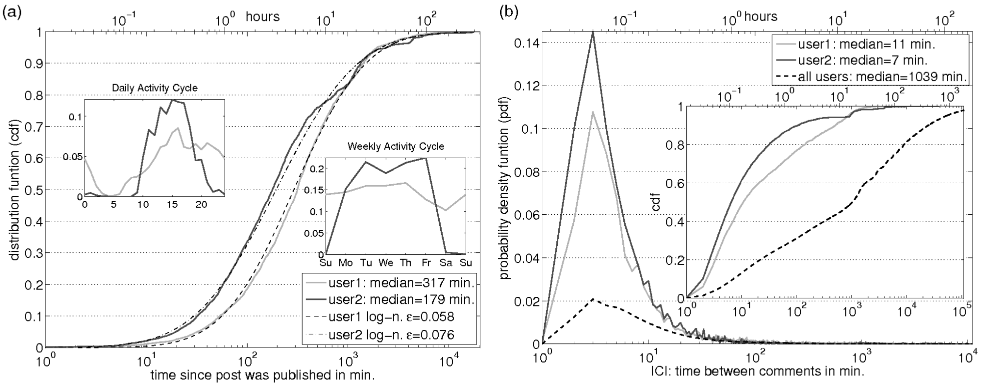

Figure 8. Activity patterns of the two most active users: **(a)** PCI-distributions, insets shows daily and weekly activity cycles. **(b)** Distribution of the inter-comment intervals (ICI) compared with the whole population (dashed line).

The same effect can be observed in their ICIs, which are illustrated in Figure 8b. There the cdf (inset of Figure 8b) of user1 shows an even more pronounced increase around an ICI of 16 hours. We further observe that the ICI-pdf peaks for both users as well as for the whole population at 3 minutes. This is probably caused by an anti-troll filter (Malda 2002), which should prevent a user from commenting more than once within 120 seconds. The medians of the ICI-distributions of user1 and user2 are rather short (11 and 7 minutes respectively) compared to the median of the whole population (about 17 hours), indicating that the two users engage in discussions frequently during their activity phase.

4. DISCUSSION

The special architecture of the technology-related news website Slashdot allowed us to analyze the temporal communication patterns of an online society without considering semantic aspects. The site activity is driven by news-posts which provoke communication activity in the form of comments.

Despite the great amount of users participating in the discussions, close to 10^5 in the data we have studied, and the diversity of themes (games, politics, science, books,etc.) some simple patterns can be identified, which repeat themselves over and over again. One of these patterns appears in the shape of the distribution of time differences between a post and its comments (the PCIs). It can be well approximated by a log-normal distribution (Figures 3 and 4) for most of the posts. The only remarkable deviations from these approximations are caused by oscillatory daily and weekly activity patterns (Figure 1), which become less noticeable if a post is published early in the morning (Figure 5a).

In single user behavior an akin pattern appears in the PCI-distribution of all of the comments a user writes to several posts (Figure 8a). Again deviations are caused by the circadian cycle. Another interesting pattern can be observed analyzing the ICI of single-users, i.e. the time-span between two consecutive comments of a certain user. In the case of the two most active users (Figure 8b) the ICI-distributions are very similar, which further supports our hypothesis of the existence of homogeneous temporal patterns on Slashdot.

We would expect that the time-spans between publishing and reading of a post also follow a log-normal pattern. This could be easily verified checking the server logs of Slashdot or access-times of an external homepage linked by a Slashdot post. Such a study has been performed to show the Slashdot effect (Adler 1999), but the scale of the data presented does not allow to draw significant conclusions. Further investigation is needed to verify this claim.

Log-normal temporal patterns similar to those described above were found in person-to-person communication by Stouffer et al. (2006), who investigated the waiting and inter-event times of an e-mail activity dataset. A second coincidence between their study and our findings is that the number of comments (or e-mails in their case) can be well approximated by the same distribution (a truncated log-normal in this case). The temporal patterns of the e-mail data were previously claimed to show power-law behavior, which would be explained by a queuing model (Barabási 2005). Although this model might allow insight into other types of human activity (Vázquez et al. 2006) it is not able to account for the observed log-normal behavior patterns. We hope therefore to encourage further research towards a theoretical understanding of the underlying phenomena responsible for this apparently quite general human behavior pattern.

Our results indicate that communication activity on Slashdot can be described using only two parameters, i.e. the median and the geometric standard deviation (Figure 6). The medians are very low compared to the overall duration of the activity provoked by a post. Although the posts might be available for commenting during more than 10 days, the first few hours decide whether they will become highly debated or just receive some sporadic comments. We would therefore expect that the simplicity of the approximation together with the high initial activity should make an accurate prediction of the expected user behavior feasible at an early phase after a post has been put online. The accuracy of such forecasting is subject of current research and will be published elsewhere.

An early characterization of the activity triggered by a post could be applied, for instance, on dynamic pricing or placing of online advertisements or on the improvement of online marketing. The success of a campaign might be predicted already after a short time-period, thus allowing an early adaptation of the strategy of information diffusion. In this context the viral marketing concept (Leskovec et al. 2006) which relies on personal communication might be the most promising field.

In our opinion, the regular communication activity patterns described in this work may be relevant in two aspects. The first, simpler one, is related to applications where a better understanding of information trade in the web translates easily into a better description, and even quantification, of Internet audience. But a second, more complex, aspect is related to the human "communicative" behavior uncovered at present time: Internet based communication capabilities. We face a new, large scale, all-to-all public space in which a novel kind of social behavior arises, a scenario that we do not yet fully understand. However, we should not forget that the new activity is being largely recorded and the data can be available for research. The work presented in this contribution is a good example of how those data can be collected and analyzed to give, at least, a quantitative description of the behavior. This is a first step towards a more ambitious target: to develop "ab initio" models for the population dynamics of message interchange, which is also the goal of our current research.

ACKNOWLEDGMENTS

This work has been partially funded by Càtedra Telefónica de Producció Multimédia de la Universitat Pompeu Fabra.

REFERENCES

Adler, S., 1999. The Slashdot Effect, an analysis of three Internet publications. Published online.

Albert, R. et al, 1999. The diameter of the world wide web. *Nature* **401**:130.

Baoill, A. Ó., 2000. Slashdot and the Public Sphere. *First Monday* **5**(9).

Barabási, A. L., 2005. The origin of bursts and heavy tails in human dynamics. *Nature* **435**:207–211.

Dewes, C. et al, 2003. An analysis of Internet chat systems. In *IMC '03: Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement*, pp. 51–64, New York, NY, USA. ACM Press.

Dezso, Z. et al, 2006. Dynamics of information access on the web. *Physical Review E* **73**:066132.

Faloutsos, M. et al, 1999. On Power-law Relationships of the Internet Topology. In *SIGCOMM*, pp. 251–262.

Goldstein, M. L. et al, 2004. Problems with fitting to the power-law distribution. *The European Physical Journal B* **41**(2):255–258.

Habermas, J., 1962/1989. *The Structural Transformation of the Public Sphere : Inquiry into a Category of Bourgeois Society*. Cambridge, MA: MIT Press.

Harder, U. and Paczuski, M., 2006. Correlated dynamics in human printing behavior. *Physica A* **361**:329–336.

Henderson, T. and Bhatti, S., 2001. Modelling user behaviour in networked games. In *MULTIMEDIA '01: Proceedings of the 9th ACM International Conference on Multimedia*, pp. 212–220, New York, NY, USA. ACM Press.

Johansen, A., 2004. Probing Human Response Times. *PHYSICA A* **338**:286.

Kleban, S. D. and Clearwater, S. H., 2003. Hierarchical Dynamics, Interarrival Times, and Performance. In *SC '03: Proceedings of the 2003 ACM/IEEE conference on Supercomputing*, p. 28, Washington, DC, USA. IEEE Computer Society.

Lampe, C. and Resnick, P., 2004. Slash(dot) and burn: Distributed Moderation in a Large Online Conversation Space. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 543–550, New York, NY, USA. ACM Press.

Leskovec, J. et al, 2006. The dynamics of viral marketing. In *EC '06: Proceedings of the 7th ACM conference on Electronic commerce*, pp. 228–237, New York, NY, USA. ACM Press.

Limpert, E. et al, 2001. Log-normal distributions across the sciences: Keys and clues. *Bioscience* **51**:341–352.

Mainardi, F. et al, 2000. Fractional calculus and continuous-time finance II: the waiting-time distribution. *Physica A* **287**:468–481.

Malda, R., 2002. Slashdot FAQ: Comments and Moderation. http://slashdot.org/faq/com-mod.shtml#cm2000.

Masoliver, J. et al, 2003. Continuous-time random-walk model for financial distributions. *Physical Review E* **67**:021112.

Mitzenmacher, M., 2004. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics* **1**(2):226–251.

Newman, M. E. J., 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* **46**:323–351.

Paxson, V. and Floyd, S., 1995. Wide area traffic: The failure of Poisson modeling. *IEEE-ACM Transactions On Networking* **3**:226–244.

Poor, N., 2005. Mechanisms of online public sphere: The web site Slashdot. *Journal of Computer-Mediated Communication* **10**(2).

Sigman, K., 1999. Appendix: A primer on heavy-tailed distributions. *Queueing Systems* **33**:261–275.

Stouffer, D. B. et al, 2006. Log-normal statistics in e-mail communication patterns. e-print physics/0605027.

Vázquez, A. et al, 2006. Modeling bursts and heavy tails in human dynamics. *Physical Review E* **73**:036127.

APPENDIX A ERROR MEASURE ε

We use the following distance measure to calculate the error of log-normal approximation of the data. The distance between approximation and data is only calculated for the time-bins (i.e. minutes) where a post actually receives a comment to avoid a distortion of the error measure by the periods with low comment activity.

Definition 1. *Let $\mathbb{T}$ be the set of time-bins where a post receives at least one comment and T its cardinality. We define then the approximation error ε of a function $f(t)$ approximating $g(t)$ (both defined for all $t \in \mathbb{T}$) as the normalized ℓ^1-norm of $f(t) - g(t)$:*

$$\varepsilon = \sum_{t \in \mathbb{T}} \frac{|f(t) - g(t)|}{T} \tag{1}$$

If $f(t)$ and $g(t)$ are cumulative probability density functions (i.e. $0 \leq f(t) \leq 1$ and $0 \leq g(t) \leq 1$), it follows that $0 \leq \varepsilon \leq 1$.