Description and Prediction of Slashdot Activity

Andreas Kaltenbrunner[*,‡]
andreas.kaltenbrunner@upf.edu

Vicenç Gómez[*,‡]
vgomez@iua.upf.edu

Vicente López[*,‡]
vicente.lopez@barcelonamedia.org

[*]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

We perform a statistical analysis of user's reaction time to a new discussion thread in online debates on the popular news site Slashdot. First, we show with Kolmogorov-Smirnov tests that a mixture of two log-normal distributions combined with the circadian rhythm of the community is able to explain with surprising accuracy the reaction time of comments within a discussion thread. Second, this characterization allows to predict intermediate and long-term user behavior with acceptable precision. The prediction method is based on activity-prototypes, which consist of a mixture of two log-normal distributions, and represent the average activity in a particular region of the circadian cycle.

1. Introduction

Human communication behavior has experienced important changes during the last decades. The daily use of email, chats, discussion forums, blogs, etc. has changed the way we interact with each other. It has never been easier for an ordinary person to reach a large and growing audience. Nevertheless, there are some underlying principles in our communication behavior that remain invariant and can be observed without considering semantic issues. E.g. the reaction and inter-event times of human communication seem to be governed by heavy-tailed distributions [26]. This fact has been reported for modern communication forms such as email [14], online chats [6] and forum discussions [15] as well as traditional communication in form of letters [22]. For more examples see [28] and references therein. However, which type of heavy-tailed distribution provides the best explanation of the data is still an open problem. Some favor power-law distributions [21] while others incline more to log-normal (LN) distributions [18]. See for example the discussion over an email dataset between Barabási [2] and Stouffer et al. [27] and the remarks of Mitzenmacher [20] about similar controversies in other areas of science.

In this work we analyze the reaction time of many-to-many communication in form of online debates at the popular technology-news website Slashdot[1]. The site, created in 1997, publishes frequently short news **posts** and allows its readers to **comment** on them, which provokes online-discussions that may trail for days. A moderation system upholds the quality of discussions by discouraging spam and offensive comments [16]. In [15] it was shown that the distribution of the time differences between a post and its comments, i.e. the post-comment-interval (PCI), fits well a LN distribution, but the quality of the fit strongly depends on the circadian rhythm of the site.

Here we use the same dataset as in [15], which represents one year of activity on Slashdot and consists of about 10^4 news posts which received more than $2 \cdot 10^6$ comments (see [15] for more details on the dataset) and extend this previous work in two directions. First, we improve the approximation quality of the PCIs by using distributions which diminish the dependency on the circadian cycle. This is either achieved by the use of double log-normal (DLN) distributions, as for example used in [27] to explain the waiting time in email conversation, or by multiplying a LN distribution with a periodic function. The best results are obtained if both methods are combined.

Although Slashdot holds much closer ties to web message boards and newsgroups, we can find some related studies about the comments to posts on weblogs [19, 8]. The amounts of comments per post and per blog follow heavy-tailed distributions, but only 30% of the blogs (15% percent of the posts) received comments [19]. According to Duarte et al. [8], 55% of the discussions appearing in these blogs can be classified as many-to-many communication. Among other temporal patterns of the comments, their study also analyzes the aggregate of all PCI-distributions, which is fit by a Weibull distribution.

In the second part of this work we propose a method to predict the activity generated on Slashdot. Our goal is to approximate how many comments a given post will receive

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

using only the first few minutes/hours of the activity it generates as evidence. This is related to the problem of predictive inference of future responses, for which in the case of LN models several analytical studies [9, 11] found estimators of the single future response density (i.e. the probability of the next comment) using a subset of the data.

There exists extensive literature concerning the prediction of Internet traffic or, in particular, web traffic demand (see [10, 23, 1, 25, 3] for just a few examples). The most elaborated approaches apply multi-resolution analysis using wavelet transform decompositions to characterize the data at different temporal resolutions [23, 1]. The resulting components are then used to model a time dependent function, from which short lookahead predictions (from one up to five minutes), or long term trends can be obtained. Several methods have been proposed to model such functions. Some of them rely on time series analysis techniques [4]. Especially the auto-regressive integrated moving average model or variants are widely used [10, 23], although other function approximation techniques, spanning from linear fits [3] to recurrent neural networks [1], have been applied as well to obtain predictions. Direct applications of those methods cover dynamic resource allocation [24], congestion control [12], or security issues [13].

Our work differs from those approaches in many ways. For instance, the temporal data considered in those studies consists of aggregated traffic measurements from web, Ethernet, or IP backbone traces, with different data protocols or applications merged in one common traffic stream. In contrast, the post induced activity existing in Slashdot occurs at a higher logical level. Moreover, the stationarity and linearity assumptions usually made by time series methods are clearly not valid in our case. The LN temporal profile is highly non-stationary and has a transient temporal nature. Nevertheless, since we know the trigger of enhanced activity (i.e. the publishing of a post) and the underlying generative model, we can extrapolate the future activity. But, instead of a simple parameter estimation of a truncated LN distribution, which is very sensitive to small fluctuations in the data and thus not the best technique for prediction purposes, we use different prototypes of posts, each of them representing the average behavior in different regions of the circadian cycle, rescale these prototypes to adjust best the initial activity of a given new post, and take the remaining part of the prototype as prediction.

2. Approximations of the User Activity

In this section we compare the quality of different types of approximations of the PCI-distribution of a post. First we explain in detail the four different probability distributions used to approximate the data and the statistical test applied to compare their performance.

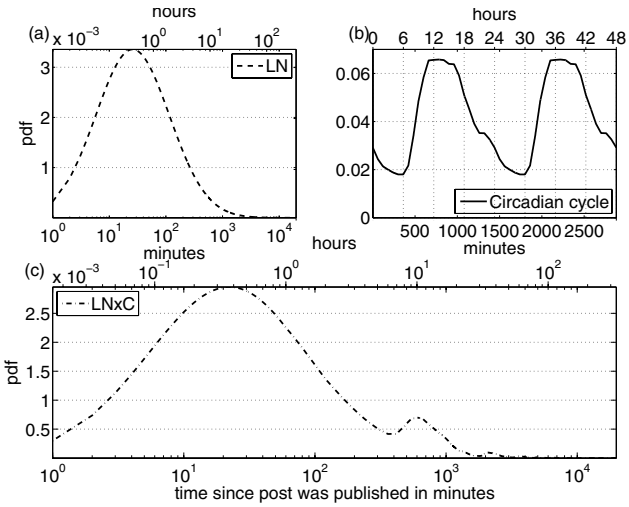

Figure 1. Transformation of LN into LNxC.

2.1. Statistical Preliminaries

Since we model discrete data with minute precision we use discretized versions of the following probability density functions (pdf). The simplest distribution used is the log-normal (LN) distribution whose pdf is:

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

We also use a double log-normal (DLN) distribution which is a mixture of two independent log-normals. Its pdf is thus:

$$\begin{aligned} f_{DLN}(t;\theta) &= k f_{LN}(t;\mu_1,\sigma_1) + (1-k) f_{LN}(t;\mu_2,\sigma_2) \\ \text{where } \theta &= (\mu_1,\sigma_1,k,\mu_2,\sigma_2). \end{aligned} \quad (2)$$

The third and forth distributions used are generated from the previous ones by point-wise multiplication of their pdfs with a periodic continuation of an "ad hoc" circadian cycle[2] as shown in Figure 1b. The cycle is approximated by the normalized mean number of comments per hour of the day, which is then linearly interpolated to achieve minute resolution. Alternatively, higher dimensional interpolation could be used, but the differences are negligible for our purposes.

The starting point of the periodic function coincides with the moment the post is published. After the multiplication we have to normalize to obtain the final pdf. We denominate the two resulting probability distributions LNxC and DLNxC. This procedure is visualized in Figure 1. Figure 1a shows an example of a LN-pdf. After multiplying it with the periodic continuation of the circadian activity cycle (Figure 1b) and renormalizing we obtain the LNxC-pdf (Figure 1c).

[2] In the calculation of the cycle we account for the daylight saving time, an effect which has been neglected in [15].

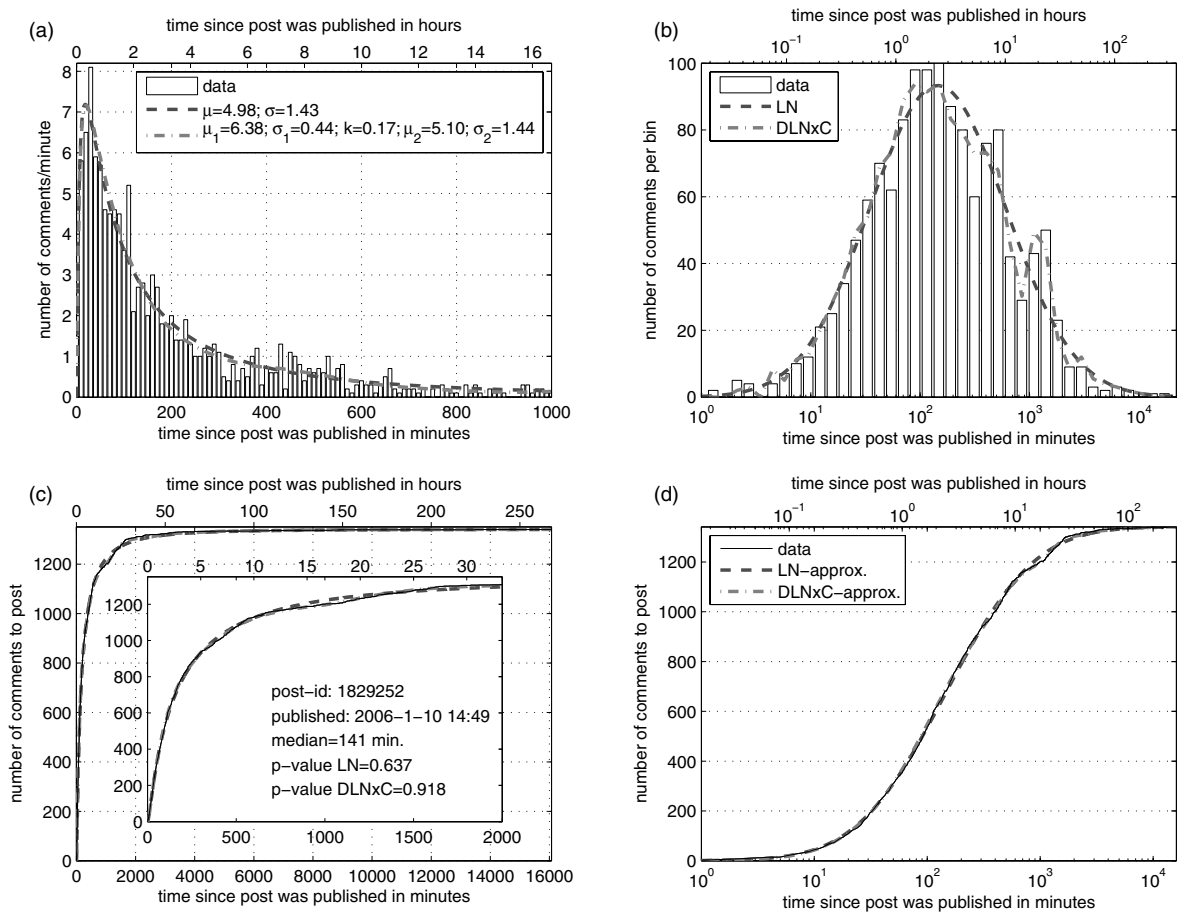

Figure 2. Approximation with LN (dashed line) and DLNxC (dashed doted) of the PCI-distribution (solid lines and bars). (a) Comments per minute (bin-with=2) for the first 1000 minutes after the post's publishing. (b) Same as (a) but in logarithmic scale. (c) The cumulative distribution of the data shown in (a). Inset shows a zoom on the first 2000 minutes. (d) Same as (c) but in logarithmic scale.

To find the optimal parameters of these distributions for a given post we use maximum likelihood estimation [5], which is performed by minimizing the negative logarithm of the likelihood function with `fminsearch` in MATLAB.

To test whether for a given post its PCIs are distributed according to one of the above described distributions, we use the Kolmogorov-Smirnov (KS) test with the following hypotheses:

H_0**:** The PCI is a sample of distribution F.

H_1**:** The hypothesis H_0 is not true.

F stands for the tested probability distribution (either LN, LNxC, DLN or DLNxC). The test is based on finding the maximal difference between the cumulative distribution functions (cdf) of data and approximation. With this maximum and the number of samples (i.e. the number of comments in our case) we can calculate the p-value of the KS-test. It gives us the probability of obtaining a result as different from F as the data. In other words: the greater the p-value, the closer is the fit with the test distribution. The hypothesis H_0 is accepted if the p-value is greater than the chosen level of significance α_0 (usually set to 0.05 or 0.01). For more details see for example [5].

2.2. Two example Posts

Before we compare the results of all four distributions explained in the previous section we examine the fit of two example posts with either a LN or a DLNxC-distribution in detail. The simpler one, the LN, was already used in [15] and gave good results for posts published between 6am and 16pm. Figure 2 shows such a post. We observe that the PCI-distribution is fit well by both distributions. In Fig-

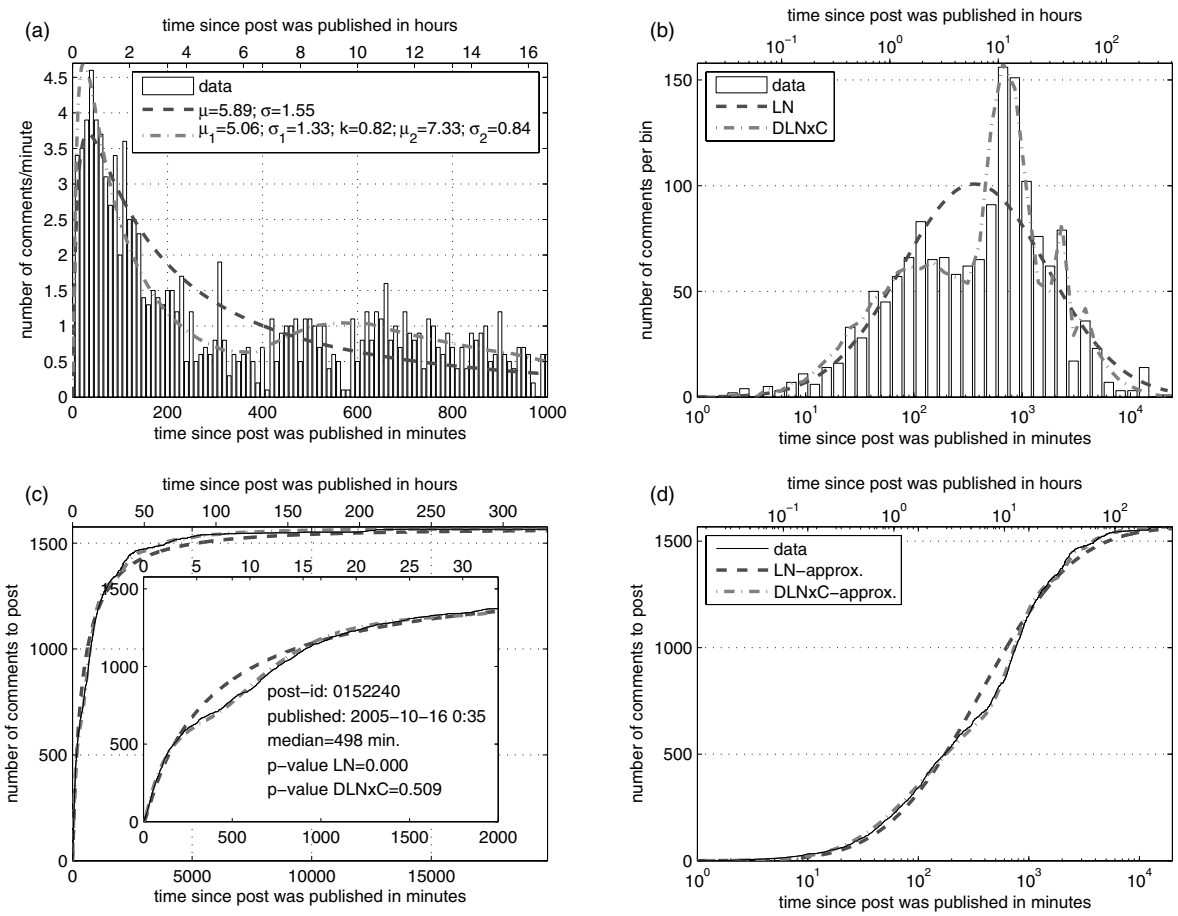

Figure 3. DLNxC improves the fit for a post published late at night. Description as in Figure 2.

ure 2a, which shows only the first 1000 minutes of activity provoked by this posts, it is hard to decide which of the two approximations is better. Nevertheless, we notice that in logarithmic scale[3] (Figure 2b) the oscillations of the PCI-distribution are better approximated by a DLNxC (gray dashed-doted line). The same effect can be observed in the PCI-cdf (Figure 2c and Figure 2d) where a small bump after about 1000 minutes is adjusted well by a DLNxC. This is reflected by the KS-test, which accepts the LN-fit with a p-value of 0.637 and the DLNxC-fit with $p = 0.918$. Although both p-values are far greater than the usual threshold of $\alpha_0 = 0.05$ for acceptance of the null hypothesis, they indicate that the DLNxC-fit is much closer to the data.

Moreover, the DLNxC leads to excellent results even for those cases where the KS-test rejects the LN-hypothesis. An example of such a post, which was published late at night and suffers thus distortions due to the circadian cycle is shown in Figure 3. Here the LN-hypothesis is rejected with a very low p-value ($< 10^{-10}$). However, a DLNxC-fit would be accepted with a p-value of 0.509. Already in linear scale (Figure 3a) it becomes clearly visible that the DLNxC is much closer to the data. This impression is enhanced by the PCI-distribution in logarithmic scale (Figure 3b) and the corresponding PCI-cdf (Figures 3c and 3d). The DLNxC-fit adapts well to the oscillations of activity.

We will show in section 2.4 that the hour a post is published determines whether it can be approximated well by only a single LN or needs either a DLN or a DLNxC.

2.3. Approximation of all posts

After these two examples, we perform a KS-test for all posts and all types of distributions presented in section 2.1 to analyze their approximation quality. The cdf of the p-values of these tests are shown in Figure 4. The previously used LN-approach (dashed line) gives the worst results. A significant improvement is achieved if we use a LN plus circadian cycle (LNxC)-distribution (black continuous line). But the best results are obtained for the two types of double log-normal distributions. Both DLN and DLNxC-fits have

[3]Note that the bin-width in log-scale increases with time, which causes different locations of the peak of activity in Figures 2a, 2b and 3a, 3b.

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

Table 1. Percentage of rejected 0-Hypotheses

p-values much higher than their log-normal counterparts. Now the improvement achieved by using the circadian cycle is much smaller, the curves of DLNxC (dashed-dotted line) and DLN-curve (gray continuous line with circles) nearly coincide. If we fix the level α_0 with either 0.01 or 0.05

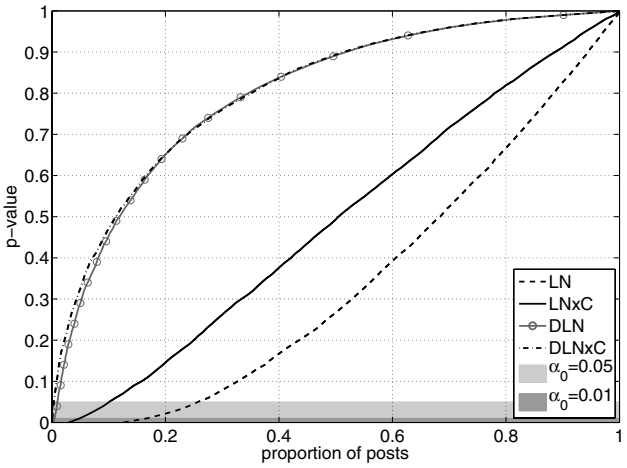

Figure 4. Results of KS-tests on all posts

(shown as gray areas in Figure 4), we can quantify the percentage of posts for which the KS-test rejects the null hypothesis (see Table 1). While a single LN-distribution only explains in 83% of the cases, both double log-normal variants are a valid model of the data for more than 99% of all posts. The best results are obtained for DLNxC which is only rejected for 11 of 10016 posts ($\alpha_0 = 0.01$).

The small difference between the outcome of the KS-tests for DLN- and DLNxC-distributions suggests that the DLN-fit might already account for the main part of the variations caused by the circadian rhythm. This is confirmed by Figure 5 which shows the dependence of the p-values on the publishing hour of the posts. The left panel compares LN (dark gray) and DLN-approximations (light gray) and the right panel their oscillating extensions. Clearly, the quality of LN and LNxC-approximation depends on the hour of the day a post is published, although this dependence diminishes for the case of LNxC. On the contrary, both types of double log-normal distributions show only minor variations due to the publishing hour of the post, but again DLNxC is slightly more constant than DLN.

2.4. Two waves of activity

The fact that a combination of two LN distributions (LN_1 and LN_2) allows a good approximation of the PCI suggests that the activity provoked by a post consists of two major waves. The first one starts directly after the post is published and the second one after the next increase of the circadian cycle. To verify this claim we combine all posts of our dataset which have been published during the same hour of the day into an aggregate post. For example, to obtain the first aggregate post we sum the PCI-distributions of all posts published between 1am and 2am. In this way we obtain 24 aggregate posts, which we approximate with DLN-distributions. The normalized PCI-cdfs of the 24 aggregates (black solid lines) and their DLN-approximations (gray dashed lines) are shown in Figure 6d.

The parameters of these 24 DLN-approximations can be observed in the top three subplots of Figure 6. We notice that μ_1 and σ_1 (continuous lines in Figure 6a and 6c) of the first LN-distribution (LN_1) experience only minor variations due to the posting hour. LN_1 corresponds to the first wave of activity. The mixing parameter k, μ_2 and σ_2 (dashed lines), on the other hand, vary significantly. Figure 6b shows that k experiences a cyclic behavior, similar to the circadian activity cycle (Figure 1b). The location in time of the maximum and minimum of both cycles approximately coincide. The value of k reaches its maximum around 3pm, which indicate that for posts published at this time of the day most of the activity can be modelled by LN_1. At the same time μ_2 reaches its maximum and σ_2 its minimum. The difference between the medians $\exp(\mu_1)$ and $\exp(\mu_2)$ of LN_1 and LN_2 is of about 16 hours, which tells us that LN_2 models the activity of those users which comment the post during the following day, i.e. during the next high-phase of the circadian cycle. For publishing times later than 15pm the value of k decreases successively, while σ_2 increases, which implies that the proportion of the total number of comments received during this second wave of activity increases as well. Parallel to this rise, μ_2 decays as the time-difference between the publishing of the post and the next rise of activity decreases. This trend is stopped around 5am in the morning when the proportion of comments provided by the first wave of activity increases again. Between 9am and 14pm, during the high-phase of activity, the values of μ_1 and μ_2 are very similar making a separation of the two waves very difficult. The activity can be approximated well during this time window with only a single LN distribution. A DLN leads here only to minor improvements, which makes its parameters hard to interpret during this interval.

The gray areas in Figure 6d, representing the activity within $\exp(\mu_{1,2} \pm \sigma_{1,2})$ centered around the medians μ_1 and μ_2 of the two LN-distributions, visualize the influ-

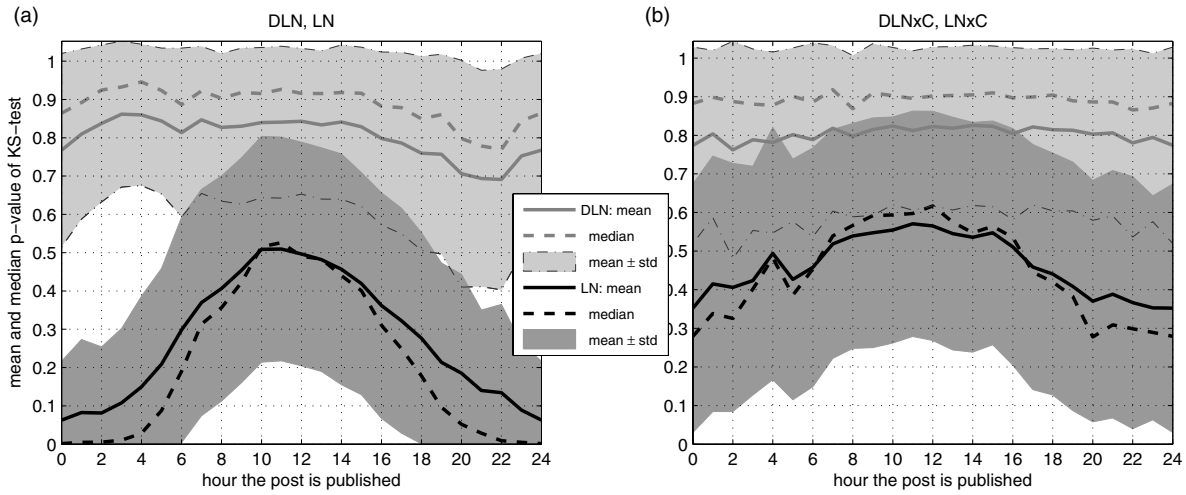

Figure 5. Results of KS-tests per publishing hour of post of (a) DLN and LN, (b) DLNxC and LNxC.

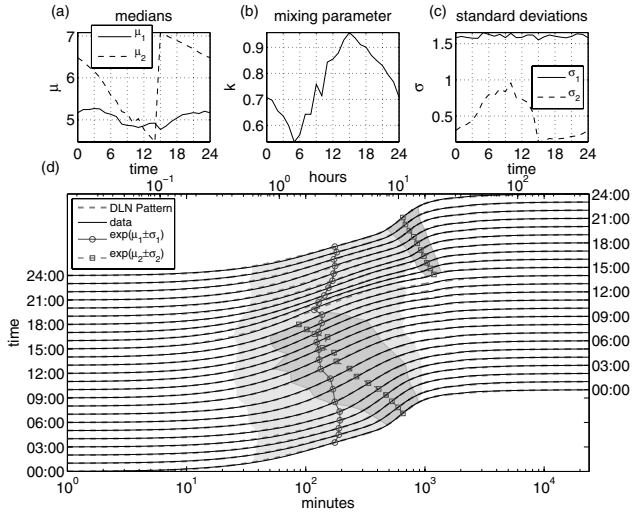

Figure 6. PCI of aggregate posts and parameters of DLN approx. by publication hour.

ence of above described two waves of activity in the DLN-approximation.

This analysis gives us further insight why the post of Figure 2 published around 2pm can be approximated well by a single LN distribution, while the post of Figure 2, published at around midnight needs a DLN or (a DLNxC) to account for the two waves of activity.

3. Prediction of user activity

After having successfully approximated the user activity in the previous section, we apply these findings to predict the activity a post provokes.

3.1. Task description and error measure

We want to solve the following problem: at time t we want to predict the comment activity in the following s minutes of a post that has been published x minutes ago and has received until now N comments. This means we take as evidence for our prediction a **data window** $[t-x, t]$ and predict the number of comments M in the **prediction window** $(t, t+s]$. If we are interested in the total number of comments, we just set the upper bound of the prediction window equal to the length of the time window a post is open to receive comments (about 14 days). Although the average duration of activity of a post is of about 5.57 (stdv=3.86) days, this overestimation increases the error only by a small amount since 97% of all comments are received within the first 2 days after a post has been published. Compare with Figure 6d.

To measure the quality of the prediction we use a standard relative error measure ϵ, which is defined in the following way:

$$\epsilon = \left| \left(M_{\text{predicted}} - M_{\text{real}} \right) / M_{\text{real}} \right| \qquad (3)$$

3.2. Prediction algorithm

Since the PCI-distribution can be well approximated, one would expect that the prediction task just reduces to a problem of parameter estimation using only a truncated version of the PCI-distribution. However, since we deal with heavy-tailed distributions, a great part of the probability mass, decisive to determine the parameters in noisy data, lies outside of the truncated region. This implies that parameter estimation is extremely prone to small fluctuations in truncated data, especially in the case of DLN distributions. Even if we use only a simple LN distribution, the

results are not very promising due to this reason (data not shown). We therefore use instead the following technique which overcomes this problem.

In analogy to the 24 aggregate posts of section 2.4, we create 24 activity prototypes, one for every hour of the day. To calculate them, we choose the oldest 10% of all the posts (published between 26-08-05 and 05-10-05), which represent the "training" set of the prediction.[4] This allows us to simulate the prediction of a post using only data of older posts. The first prototype is represented by the DLN approximation[5] of the aggregate of all training-posts published between 1am and 2am. The remaining 23 prototypes are obtained in an analogous way. The prediction of a post simply consists in rescaling the prototype corresponding to the post's publishing hour in such a way that the cdf of the comments in the data window is best approximated.

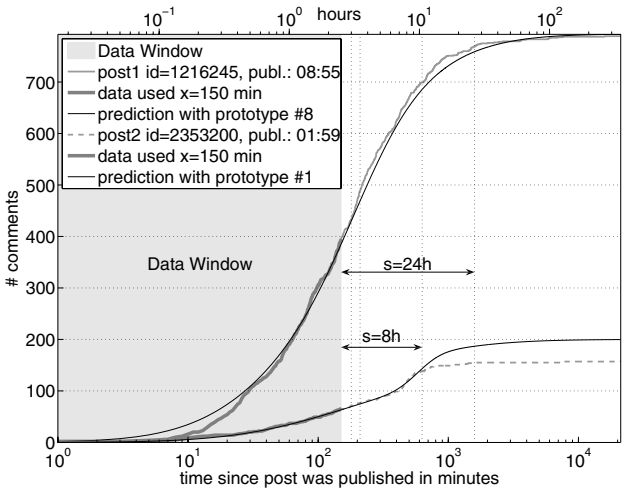

Figure 7. Two examples of prediction with predefined prototypes of activity.

3.3. Illustrating Examples

Two prediction examples are shown in Figure 7. We use here the first 150 minutes of the activity (dark gray-lines in gray area) as evidence to estimate the comments the post will receive afterwards. The two prototypes used for the prediction (black lines) are chosen according to the publishing time of the post and rescaled to adjust best the input data. The PCI (light gray continuous line) of post1 is predicted quite well by the prototype, whereas the prototype of post2 (dashed continuous line) predicts well only the following first 8 hours, afterwards the activity is overestimated. Table 2 shows the exact values of the error measure ϵ for different lengths of the prediction window. It is interesting to note that the error for a short prediction window length of 30 minutes is higher than for an intermediate length of 8 hours. This shows the sensitivity of the error measure to small fluctuations, which are more noticeable when the number of predicted messages is small, i.e. the shorter the prediction window is.

s	30min	1h	8h	24h	total
post1	18.42%	10.64%	0.99%	0%	0.92%
post2	14.29%	0%	9.59%	36.67%	44.91%

Table 2. Error ϵ of the posts of Figure 7.

3.4. Performance of the prediction

In this section we analyze the quality of the prediction. We compare the mean and the 90% quantile of the error ϵ of all posts for different lengths of data and prediction time-windows in Figure 8. The best results are obtained for a 24 hour prediction (dash-dotted lines with squares), for which the average error is around 36% and the 90%-quantile is situated at about 70%. This values look quite high at first sight, but are much lower than those of a simple approximation which assumes that every post causes more or less the same amount of activity, i.e. using the mean activity caused by the posts as representative for all posts. Under these circumstances the mean error would be situated between 150% and 200% (gray dashed line with ∇-markers), which in absolute numbers corresponds to 100 and 50 comments respectively[6], and the 90% quantile lies between 370% and 440% (data lies outside of the scale of Figure 8b). Globally, the performance of our method is reasonable. Only in the case of a short length of the prediction window $s = 1$h combined with a long data window ($x \geq 120$min) the performance decays, which becomes visible in the 90% quantile and is caused by a large number of posts with a very low number of comments in the prediction window.

The mean proportions over all posts of the number of comments received during different lengths of data (columns) and prediction windows (lines) are shown in Table 3. We observe that, for example, if we use a data window length of 30min which corresponds in the mean to only 15% of the activity we are able to predict with reasonable accuracy the remaining parts of activity. It is even

[4] Other sizes of the training set lead to very similar results as long as they contain at least a minimal amount of data for all 24 classes. In this case they contain between 11 and 77 posts, distributed similar as the circadian activity cycle.

[5] As an alternative we could use DLNxC, but as in the case of the PCI-approximation, the results are not significantly better and the circadian cycle is harder to estimate with less posts.

[6] Note that a greater relative error corresponds to a lower absolute error, since the number of comments in the prediction window is lower for longer data windows. Compare with Table 3.

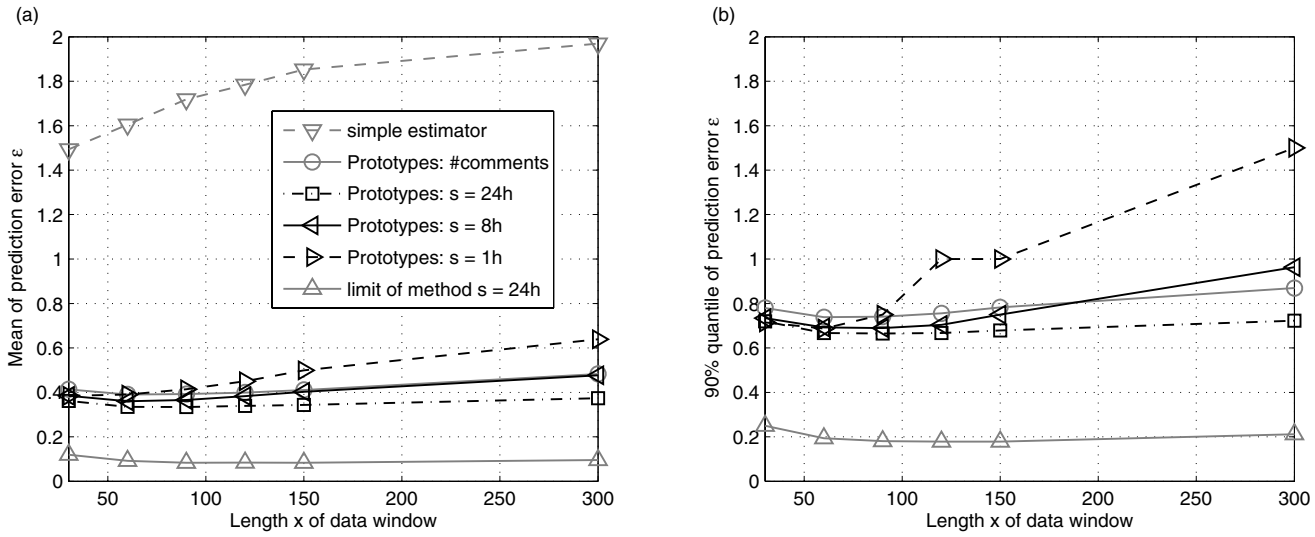

Figure 8. Mean and 90% quantiles of the prediction error of different approaches.

x	30min	60min	90min	120min	150min	300min
$[0, x]$	15.0% (± 07.5)	26.9% (± 10.9)	35.8% (± 12.7)	42.6% (± 13.7)	48.0% (± 14.3)	64.1% (± 14.8)
$[x, x + 1\text{h}]$	20.8% (± 07.5)	15.7% (± 05.5)	12.2% (± 04.5)	9.7% (± 03.9)	8.0% (± 03.5)	3.9% (± 02.5)
$[x, x + 8\text{h}]$	60.1% (± 11.0)	49.2% (± 10.2)	41.4% (± 09.9)	35.6% (± 09.7)	31.1% (± 09.6)	19.1% (± 09.2)
$[x, x + 24\text{h}]$	78.5% (± 07.8)	66.8% (± 10.1)	58.1% (± 11.4)	51.5% (± 12.2)	46.2% (± 12.6)	30.9% (± 12.7)
% of lifetime	0.73% (± 0.97)	1.46% (± 1.95)	2.18% (± 2.92)	2.91% (± 3.89)	3.64% (± 4.87)	7.28% (± 9.74)

Table 3. Mean ($\pm$ stdv) proportions of activity for different data (columns) and prediction windows (rows). Last row: mean proportion of data window length compared to the total lifetime of activity.

more surprising that the time-span used for prediction in this case corresponds to less than 1% of the duration of activity. These values are shown in last line of Table 3.

To calculate a lower bound for our prediction method, we use the fact that the prediction error is caused by a combination of two rather different circumstances: *(i)* prototypes that do not fit well the shape of a particular PCI-distribution and *(ii)* wrong rescaling factors. To quantify the latter influence we assume that we know the parameters of the DLN-distribution which approximates best the entire PCI-distribution at forehand and only have to adjust it with the data in the prediction window. The result of this minimum possible error caused only by rescaling is shown in Figure 8 as gray continuous line with $\triangle$-markers. It is situated around 10% (Figure 8a) and is lower than 20% for 90% of the posts (Figure 8b).

We also notice that the prediction quality increases with the number of comments of the post, since fluctuation errors are more important with a small number of comments, as Figure 9 illustrates. We plot the mean error for all posts which receive more than a certain number of comments (the x-axis) in the prediction window. The prediction error first decreases successively and stabilizes then at a minimum error of approximately 30%.[7] The same effect can be observed in the 90% quantiles which decrease to less than 60% or for the number of comments in the data window. Reasonable accuracy can be obtained if it contains more than 5 comments (data not shown). We thus conclude that the length of the windows is less important for prediction accuracy than the number of comments they contain.

How strong is the dependence of these results on the data used to generate the prototypes? To answer this we compare the DLN-prototypes (calculated with 10% of the posts) and the DLN fit of the aggregates of all posts (shown in Figure 6d). First, we measure the accuracy of the two fits to model the aggregate posts by rescaling the DLN-fits according to the total number of comments in the aggregate and calculate then the "prediction" error ϵ of the number of comments in a $[x, x + 24\text{h}]$ window[8]. The mean of these

[7] The increase of the curve for a data window length of 300 (gray continuous with circles) is an artifact caused by the low number of samples with more than 70 comments in the prediction window in this case.

[8] Although this is formally no prediction-error it gives a good estimate of the quality of the approximation.

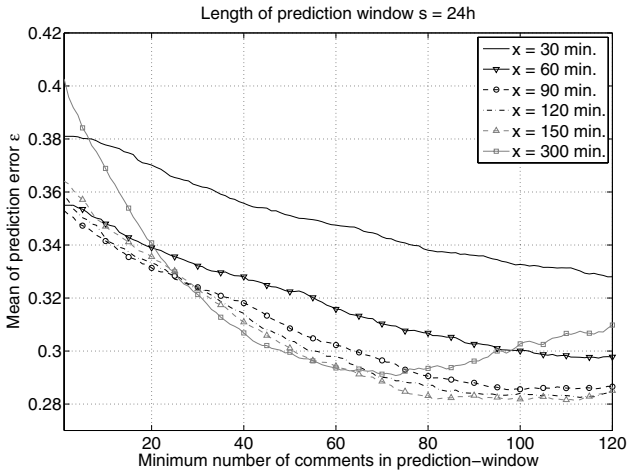

Figure 9. Dependence of the mean error on the number of future comments.

errors over all 24 classes are compared in the lower two rows of Table 4. As expected the errors of the DLNs of the prototypes are slightly bigger than those of the DLNs of the aggregates. However, if we calculate the same measure for the individual posts instead of the aggregates (upper two rows of Table 4), we observe nearly no difference in using the entire data or only a sufficient large subset to calculate the prototypes, whose performance can thus be considered as quite stable to variations.

x	30min	60min	120min	150min
$s = 24$h	Mean of accuracy on all posts			
DLN-Protot.	6.93%	10.30%	16.04%	18.29%
DLN-Aggreg.	6.99%	10.11%	15.34%	17.41%
$s = 24$h	Accuracy on aggregate of all posts			
DLN-Protot.	2.36%	3.85%	4.95%	5.30%
DLN-Aggreg.	0.52%	0.84%	1.12%	1.15%

Table 4. Accuracy of DLN-prototypes.

4. Conclusions

In the first part of this study we compare the quality of four different approximations of the PCI-distribution. We observe that DLN distributions provide an excellent explanation for the discussion activity provoked by a post on Slashdot. The quality of the fit is independent of the publishing hour of the post, contrary to what is is observed if only a single LN distribution is used [15]. We can conclude that a post provokes two major waves of activity, which correspond to two LN distributions. The first wave starts directly after the post is published and the second one after the next increase of the circadian cycle. Since more such oscillations with smaller amplitudes occur during the lifetime of a discussion, a slight improvement of the fit can be achieved with a combination of DLN distributions and the circadian cycle.

DLN distributions were used as well in [27] to explain waiting times in email conversation. It seems thus that the observed phenomenon is quite general and we would expect to find it as well in other aspects of human communication. For instance, in the access-distributions of news-posts [7], a damped periodic pattern similar to the one analyzed here has been reported. As in many other studies, a power-law model is assumed, without being contrasted with the LN hypothesis. A recent study [2] proposed a model to explain this waiting or reaction times under the premise that they fit power-law distributions. However, to achieve reasonable accuracy of those fits, the heads and tails of the distributions were not considered, while a DLN fit of the same data allowed a characterization of the entire dataset without the need of cutoffs [27]. We believe that a theoretical understanding of the presented phenomena is thus still an open question and further research towards a (double) LN model for human communication behavior is needed.

A second question we investigated here is whether the approximation of the PCI-distributions can be used to predict the reaction a post will provoke in the community. The proposed method stores several prototypes of activity, each of them covering the entire life-time of a post, and consists in rescaling a prototype, which is determined by the publishing hour of a post. This technique is fast and flexible in the sense that one can predict at an arbitrary moment in the lifetime of a post the expected number of comments it will receive afterwards during a time window of likewise arbitrary length.

The transient profile of in the PCI-cdfs (e.g. the sharp initial raise) makes accurate prediction nearly impossible using standard time-series methods. Nevertheless, although its average error is relatively high, our approach predicts the magnitude of the expected reaction to a post already after a short time-period, when only a small fraction of its total number of comments has been received. The method could allow, for instance, dynamic pricing or placing of online advertisements according to the expected reaction to a post, or early adaptation of online marketing campaigns. using the viral marketing concept [17].

It should be easy to build a real-time system which predicts the total writing activity of the site. Such a system would consist of as many predictors as active posts which are updated every Δt minutes. At every updating event all the predictors would first incorporate their evidences (the number of comments received within the last Δt minutes) and recalculate their predicted activity by rescaling their corresponding prototype properly. Eventually, some predic-

tors could be removed if their posts had been "closed" in the meantime and then included in the training set to generate improved prototypes. Other predictors could be incorporated if new posts had appeared within the last Δt minutes. Such a system might provide estimates of the total activity by just summing up the predictions of all the existing posts.

It seems natural to apply this approach as well on page request data to predict server loads, where it should lead to better results since its error decreases for larger datasets. We expect the number of readers per minute to follow approximately the same distribution as the number of comments but in a larger scale. This is supported by a study of visits of news-pages on an Hungarian website [7], which revealed patterns quite similar to the PCI-distribution on Slashdot. Unfortunately we do not have access to the server-logs of Slashdot to verify this claim, but we are optimistic to be able to apply our technique soon on similar data.

5. Acknowledgments

This work was partially funded by Càtedra Telefónica de Producció Multimèdia de la Universitat Pompeu Fabra.

References

[1] A. Aussem and F. Murtagh. Web traffic demand forecasting using wavelet-based multiscale decomposition. *International Journal Of Intelligent Systems*, 16(2):215–236, 2001.

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

[3] Y. Baryshnikov, E. G. Coffman, G. Pierre, D. Rubenstein, M. Squillante, and T. Yimwadsana. Predictability of web-server traffic congestion. In *Proceedings of the 10th IEEE International Workshop on Web Content Caching and Distribution (WCW'05)*, pages 97–103, Washington, DC, USA, 2005. IEEE Computer Society.

[4] G. E. P. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1994.

[5] M. H. DeGroot and M. J. Schervish. *Probability and Statistics*. Addison-Wesley, New York, 3rd edition, 2002.

[6] C. Dewes, A. Wichmann, and A. Feldmann. An analysis of internet chat systems. In *IMC '03: Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement*, pages 51–64, New York, NY, USA, 2003. ACM Press.

[7] Z. Dezso, E. Almaas, A. Lukacs, B. Racz, I. Szakadat, and A. L. Barabási. Dynamics of information access on the web. *Physical Review E*, 73:066132, 2006.

[8] F. Duarte, B. Mattos, A. Bestavros, V. Almeida, and J. Almeida. Traffic Characteristics and Communication Patterns in Blogosphere. In *Proceedings of the 1st International Conference on Weblogs and Social Media (ICWSM'06)*, Boulder, Colorado, USA, March 2007.

[9] A. Fernandez. Bayesian inference from type II doubly censored Rayleigh data. *Statistics and Probability Letters*, 48(4):393–399(7), 2000.

[10] N. Groschwitz and G. Polyzos. A time series model of long-term NSFNET backbone traffic. In *Proceedings of the IEEE International Conference on Communications (ICC'94)*, volume 3, pages 1400–4., May 1994.

[11] S. B. P. Hafiz M. R. Khan, M. Safiul Haq. Bayesian prediction for the log-normal model under Type II. censoring. Technical Report 0405-17, CAMS, NJIT, 2005.

[12] V. Jacobson. Congestion avoidance and control. In *ACM SIGCOMM '88*, pages 314–329, Stanford, CA, Aug. 1988.

[13] J. Jiang and S. Papavassiliou. Detecting Network Attacks in the Internet via Statistical Network Traffic Normality Prediction. *Journal of Network and Systems Management*, 12:51–72, March 2004.

[14] A. Johansen. Probing human response times. *PHYSICA A*, 338:286, 2004.

[15] A. Kaltenbrunner, V. Gómez, A. Moghnieh, R. Meza, J. Blat, and V. López. Homogeneous temporal activity patterns in a large online communication space. In *Proceedings of the BIS 2007 Workshop on Social Aspects of the Web (SAW 2007)*. Poznan, Poland, 2007.

[16] C. Lampe and P. Resnick. 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*, New York, NY, USA, 2004. ACM Press.

[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 Press.

[18] E. Limpert, W. A. Stahel, and M. Abbt. Log-normal distributions across the sciences: Keys and clues. *Bioscience*, 51:341–352, 2001.

[19] G. Mishne and N. Glance. Leave a reply: An analysis of weblog comments. In *WWW2006, 3rd Annual Workshop on the Weblogging Ecosystem*, Edinburgh, UK, 2006.

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

[21] M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323–351, 2005.

[22] J. G. Oliveira and A. L. Barábasi. Human dynamics: Darwin and Einstein correspondence patterns. *Nature*, 437:1251–1251, 2005.

[23] K. Papagiannaki, N. Taft, Z. L. Zhang, and C. Diot. Long-term forecasting of Internet backbone traffic. *IEEE Transactions On Neural Networks*, 16:1110–1124, 2005.

[24] J. Platt. A resource-allocating network for function interpolation. *Neural Comput.*, 3(2):213–225, 1991.

[25] A. Sang and S. qi Li. A predictability analysis of network traffic. *Computer Networks*, 39(4):329–345, 2002.

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

[27] D. B. Stouffer, R. D. Malmgren, and L. A. N. Amaral. Log-normal statistics in e-mail communication patterns. e-print physics/0605027, 2006.

[28] A. Vázquez, J. G. Oliveira, Z. Dezso, K. I. Goh, I. Kondor, and A. L. Barabási. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73:036127, 2006.