The Self-Feeding Process: A Unifying Model for Communication Dynamics in the Web

Pedro O.S. Vaz de Melo
Universidade Federal de
Minas Gerais
Belo Horizonte, Brazil
olmo@dcc.ufmg.br

Christos Faloutsos
Carnegie Mellon University
Pittsburgh, USA
christos@cs.cmu.edu

Renato Assunção
Universidade Federal de
Minas Gerais
Belo Horizonte, Brazil
assuncao@dcc.ufmg.br

Antonio A.F. Loureiro
Universidade Federal de
Minas Gerais
Belo Horizonte, Brazil
loureiro@dcc.ufmg.br

ABSTRACT

How often do individuals perform a given communication activity in the Web, such as posting comments on blogs or news? Could we have a generative model to create communication events with realistic inter-event time distributions (IEDs)? Which properties should we strive to match? Current literature has seemingly contradictory results for IED: some studies claim good fits with power laws; others with non-homogeneous Poisson processes. Given these two approaches, we ask: which is the correct one? Can we reconcile them all? We show here that, surprisingly, both approaches are correct, being corner cases of the proposed Self-Feeding Process (SFP). We show that the SFP (a) exhibits a unifying power, which generates power law tails (including the so-called "top-concavity" that real data exhibits), as well as short-term Poisson behavior; (b) avoids the "i.i.d. fallacy", which none of the prevailing models have studied before; and (c) is extremely parsimonious, requiring usually only *one*, and in general, *at most two* parameters. Experiments conducted on eight large, diverse real datasets (e.g., Youtube and blog comments, e-mails, SMSs, etc) reveal that the SFP mimics their properties very well.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: [Markov processes, Stochastic processes, Time series analysis, Probabilistic algorithms]; H.4.3 [**Information systems applications**]: [Communication applications]

General Terms

Theory

Keywords

communication dynamics,inter-event times,generative model

1. INTRODUCTION

How long will it take until a next comment arrive on your Youtube[1] video given the past history of comments timestamps? Does the behavior of commenting on Youtube videos differ from the behavior of commenting on web blogs or online news websites? And how different these activities are from writing and receiving e-mails? The current availability of large datasets containing digitalized information about human communication dynamics has made it possible to propose a question that many thought was already answered: what is the timing of human communications[2, 20]? Thus, the focus of this work is to find patterns in inter-event times between real and modern communication activities of humans.

All the aforementioned communication activities are "point processes", and the simplest way to model them is by the Poisson Process (PP) [17]. Unfortunately, this simple and elegant model has proved unsuitable [34, 12, 15, 40], since analysis of real data have shown that humans have very long periods of inactivity and bursts of intense activity [2, 20], in contrast to the PP, where activities occur at a fairly constant rate. Although researches agree that the PP is not suitable, there is no consensus about the right model between two major schools of thought. The first viewpoint [2, 34] states that a power law [13] is an appropriate fit for the Probability Density Function (PDF) of the *inter-event time distribution* (IED), where bursts and heavy-tails in human activities are a consequence of a decision-based queuing process, when tasks are executed according to some perceived priority. The second viewpoint is that the IED is well explained by variations of the PP [25, 32, 31, 30, 23]. They are based on the fact that short-term communication events exhibits a Poissonian behavior [5, 21] and suggest a piecewise Poisson process: the first interval has a constant rate λ; for the next, change the rate, and continue.

Given these two approaches, we ask: which is the correct one? Can we reconcile them all? We show here that, surprisingly, both approaches are correct, being corner cases of the proposed Self-Feeding Process (SFP). The SFP generates a power-law-tail distribution for the inter-event time marginal, like [2], and it behaves as a PP in the short term, like [32]. Moreover, the SFP is also extremely *parsimonious*, requiring at most two parameters.

Additionally, unlike previous studies, we analyze the temporal correlations between inter-event times, illustrating the "i.i.d. fallacy" that has been routinely ignored until recently [22]. We show that, unlike the PP that generates independent and identically distributed (i.i.d.) inter-event times, individual sequences of communications tend to show a high dependence between consecutive inter-arrival times. This is the basis of the SFP model, which uses a Markovian approach to determine that the next inter-event time

[1] www.youtube.com

WWW 2013, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2035-1/13/05.

depends solely on the previous one. We validate the SFP model on eight diverse and large datasets from real and modern communication data, that can be divided into two groups. The first group contains five datasets extracted from Web applications in which several users comment on a given topic. The second group contains three datasets in which individuals perform and receive communication events. In summary, the main contributions of the SFP are as follows:

- **Unifying power**. It reconciles existing and contrasting theories in human communication dynamics[2, 32];
- **Temporal correlation**. It shows positive correlation between consecutive inter-event times [22];
- **Parsimony**. It requires usually one and at most two parameters.

Moreover, we would like to point out that our findings open a new perspective in understanding human communication dynamics both at the network (first group) and individual (second group) level. By knowing the typical human behavior, one can leverage a varied number of applications in different areas, such as popularity prediction of videos and news, identification of spammers and other anomalous behavior, resource allocation, among others.

The rest of the paper is organized as follows. In Section 2, we provide a brief survey of the related work that analyzed inter-event times between communications. In Section 3, we describe the eight datasets used in this work. In Section 4, we analyze the IED of individuals from these datasets and we show that the Odds Ratio function of their IEDs is well modeled by a power law. Later, in Section 5, we show that the typical behavior of inter-event sequences shows a positive correlation between consecutive inter-event times. In Section 6, we describe our proposed SFP model that provides an intuitive and simple explanation for the observed data. Then, in Section 7 we show that the SFP model also unifies existing theories on communication dynamics. Finally, we show the conclusions and future research directions in Section 8.

2. RELATED WORK

The accurate understanding of the human dynamics on the Web can benefit a large number of applications, such as query suggestions, crawling policies, advertising, result ranking, recommender systems, anomaly detection, among others. However, the dynamics of humans on the Web is very rich and varied, given the large number of activities one can perform online. For example, in [36] the authors developed methods for modeling the dynamics of the query and click behaviors seen in a large population of Web searchers. In [37], the authors analyzed the tendency a person has to comment on stories in the Web in order to connect users with stories they are likely to comment on. Moreover, [19] analyzed and modeled the temporal behavior of users in social rating networks, what may leverage the prediction of future links, ratings or community structures. Finally, in [27], the authors used stochastic models of user behavior on online news websites to predict the popularity of a given news based on early user reactions to new content. Our work tackle a more general human behavior, i.e., her/his communications activities, which may occur on news websites, online social networks, video channels, directly via e-mail and in many other ways.

The study of the time interval in which events occur in human activity is not new in the literature. The most primitive model is the classic Poisson process [17]. Although the most recent approaches have among themselves significant differences, they all agree that the timing of individuals systematically deviates from this classical approach. The Poisson process predicts that the time interval Δ_t between two consecutive events by the same individual follows an exponential distribution with expected value β and rate $\lambda = 1/\beta$, where

$$\Delta_t = -\beta \times \ln(U(0,1)), \qquad (1)$$

where $U(0,1)$ is a uniformly random distributed number between $[0,1]$. While in a Poisson process consecutive events follow each other at a relatively regular time, real data shows that humans have very long periods of inactivity and also bursts of intense activity [2].

Recently, Barabási et. al. [2, 34] proposed that a power law [13] is an appropriate fit for the Probability Density Function (PDF) of the *inter-event time distribution* (IED). They propose that bursts and heavy-tails in human activities are a consequence of a decision-based queuing process, when tasks are executed according to some perceived priority. In this way, most of the tasks would be executed rapidly while some of them may take a very long time. The queuing models generate power law distributions $p(X = x) \approx x^{-\alpha}$ with slopes $\alpha \approx 1$ or $\alpha \approx 1.5$.

The second modern approach claims that the IED is well explained by variations of the PP, such as the Interrupted Poisson [25] (IPP), Non-Homogeneous Poisson Process [32, 31, 30] and Kleinberg's burst model [23]. All these studies are based on the fact that short-term communication events exhibits a Poissonian behavior [5, 21] and suggest a piece-wise Poisson process: the first interval has a constant rate λ; for the next, change the rate (say, to zero, for the IPP, or to double-or-half for Kleinberg's model), and continue. Malmgreen et al. [32, 31, 30] proposes a non-homogeneous Poisson process, where the rate $\lambda(t)$ varies with time, in a periodic fashion (e.g., people answer emails in the morning; then go to lunch; then answer more e-mails, etc). This model explains the data at the cost of requiring several parameters and careful data analysis, being impractical for synthetic data generators, for instance. Later, the authors adapted this model to a more parsimonious version [30], but it still has 9 parameters.

3. DATA DESCRIPTION

In this work we analyze eight datasets that can be divided into two groups. The first group contains five datasets extracted from Web applications in which several users comment on a given topic. The datasets are extracted from five popular websites: Youtube, MetaFilter, MetaTalk, Ask MetaFilter and Digg. The second group contains three datasets in which individuals perform and receive communication events. In this group we have a Short Message Service (SMS), a mobile phone-call and a public e-mail dataset. For simplicity, we use the term "individual" to refer both to topics of the first group and users of the second group.

In the first group, we analyze a public online news dataset, containing a set of stories and comments over each story. More specifically, the data is from the popular social media site Digg and has 1,485 stories and over 7 million comments [11]. The Digg dataset is public for research interests and can be downloaded at `http://www.infochimps.com/datasets/diggcom-data-set`. We also analyze three publicly available datasets from the *Metafilter Infodump Project*[2], extracted from three discussion forums: MetaFilter[3] (Mefi), MetaTalk[4] (Meta) and Ask MetaFilter [5] (Askme). After

[2] downloaded on September 22nd from http://stuff.metafilter.com/infodump/

[3] http://www.metafilter.com/

[4] http://metatalk.metafilter.com/

[5] http://ask.metafilter.com/

disregarding topics which received less than 30 comments, the Mefi dataset has 8,384 topics and 1,471,153 comments, the Meta dataset has 2,484 topics and 503,644 comments and the Askme dataset has 498 topics and 65,950 comments.

Our final dataset from the first group was collected from the Youtube website using the Google's Youtube API[6]. We collected all the comments posted on the videos classified as *trending* by the API[7] from 22/Aug/2012 to 25/Sep/2012. We collected a total of 1,221,390 comments on 989 videos, but we use in our dataset only those videos with more than 30 comments and which the comments span for more than one week, a total of 610 videos and 1,008,511 comments. The full dataset can be downloaded at `www.dcc.ufmg.br/˜olmo/youtube.zip`.

In the second group, the mobile phone calls dataset contains more than 3.1 million customers of a large mobile operator of a large city, with more than 263.6 million phone call records registered during *one month*. From this same operator, we also have a SMS dataset of 300,000 users spanning six months of data, for a total of 8,784,101 records. These datasets from the mobile operator is under Non-Disclosure Agreement (NDA) and belong to the iLab Research at the Heinz College at CMU, but was already used in several papers [38, 39, 1]. We also analyze the public Enron e-mail dataset, consisting of 200,399 messages belonging to 158 users with an average of 757 messages per user [24]. The data is public and can be downloaded at `http://www.cs.cmu.edu/ enron/`.

4. MARGINAL DISTRIBUTION

In this work, we are first interested on the inter-event time distribution IED of the random variable Δ_k representing the time Δ_k between the $k - th$ and the $(k-1) - th$ communication events on a given topic (first group) or of an user (second group). For simplicity, we use the term "individual" to refer both to topics of the first group and users of the second group.

4.1 Odds Ratio Using the Cumulative Distribution Function

In Figure 1, we show the distribution of the time intervals Δ_k between communication events for a typical active user of the SMS dataset, with 44,785 SMS messages sent or received. The histogram is shown in Figure 1-a and, as we observe, this user had a significantly high number of events separated by small periods of time and also long periods of inactivity. Moreover, both the power law fitting, which in the best fit has an exponent of $\ -2$, and the exponential fitting, which is generated by a PP, deviates from the real data. The method we use to fit the power law is based on the Maximum likelihood estimation (MLE) described in [7].

In empirical data that spans for several orders of magnitude, which is the case of the IEDs, it is very difficult to identify statistical patterns in the histograms, since the distribution is considerably noisy at its tail [2, 32]. A possible option is to move away from the histogram and analyze the cumulative distributions, i.e., cumulative density function (CDF) and complementary cumulative density function (CCDF), which veil the data sparsity. However, by using the CDF, as we observe in Figure 1-b,we lose information in the tail of the distribution and, on the other hand, by using the CCDF, as we observe in Figure 1-c, we lose information in the head of the distribution.

In order to escape from these drawbacks, we propose the use of the Odds Ratio (OR) function combined with the CDF as it allows for a clean visualization of the distribution behavior both in the head and in the tail. This $OR(k)$ function is commonly used in the survival analysis [3, 29] and measures the ratio between the number of individuals who have not survived by time t and the ones that have survived. Its formula is given by:

$$Odds\ Ratio(t) = OR(t) \quad = \quad \frac{CDF(t)}{1 - CDF(t)}. \qquad (2)$$

In this paper, for a set of n inter-event times $\{\Delta_1, \Delta_2, ..., \Delta_n\}$, we calculate the odds ratio for each percentile $P_1, P_2, ..., P_{100}$ of the data. This avoids that minor deviations in the data harms the goodness of fit test we perform, which we explain in Section 4.2.

Thus, in Figure 1-d, we plot the OR for the selected user. From the OR plot, we can clearly see the cumulative behavior in the head and in the tail of the distribution. Also, observe again that both the exponential and the power law significantly deviate from the real data. Moreover, we can also observe that the OR of the inter-event times seems to entirely follow a linear behavior in logarithmic scales, having, then, a power law behavior with OR slope $\rho \approx 1$.

Again, in Figure 2, we plot the OR of a typical individual of each dataset. The OR plots clearly show the cumulative behavior in the head and in the tail of the distribution. Also, we can observe that the OR of the inter-event times seems to follow entirely the same linear behavior in logarithmic scales, having, then, an OR power law behavior. This implies that the marginal distribution of the IEDs is approximately equal to a log-logistic distribution [14], since this distribution shows a OR power law behavior.

4.2 Goodness of Fit

In this section, we check whether the OR of the IEDs of all individuals of our datasets can be explained by a power law. We perform a linear regression using least squares fitting on the OR of the IEDs of all individuals. Since we consider every percentile and the OR is a cumulative distribution, the linear regression is accurate. We performed a Kolmogorov-Smirnov goodness of fit test, but because of digitalization errors and other deviations, this test presented biased results. For instance, it rejects all fittings on distributions where the data is rounded up from seconds to minute values (e.g. 45 seconds to 60 seconds).

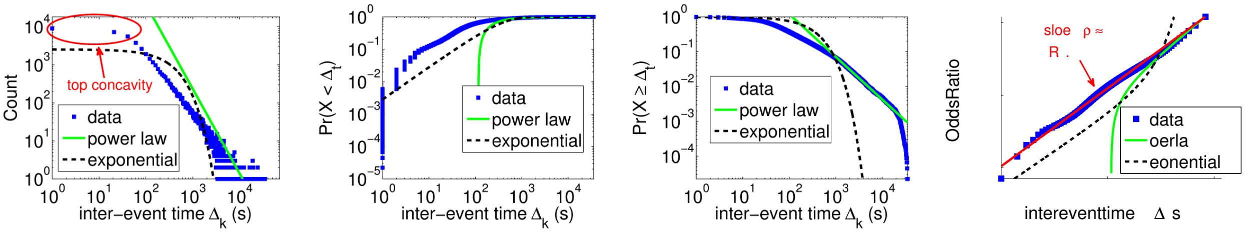
Figure 3 shows the histogram of the determination coefficient R^2 of the performed linear regressions. The determination coefficient R^2 is a statistical measure of how well the regression line approximates to the real data points. A $R^2 = 1.0$ indicates that the regression line perfectly fits the data. We observe that for the vast majority of individuals of our eight datasets, the R^2 is close to 1.0. More specifically, for the first group, the R^2 averages 0.97 for the Youtube, Askme and Digg datasets and 0.98 for the Mefi and Meta datasets. For the second group, the R^2 averages 0.99 for the phone dataset, 0.96 for the SMS dataset and 0.97 for the e-mail dataset. This allows us to state that for the vast majority of individuals, the OR of their IEDs is well fitted by a power law.

4.3 Typical Behavior

Since the IED of the majority of individuals is well modeled by an odds ratio power law, then we are able to characterize their behaviors by two values: the slope ρ and the median μ of the fitted OR power law. Observe in Figure 4 the PDF of the slopes ρ_i measured for every individual i of our eight datasets. Except the SMS dataset, the typical ρ_i for the majority of individuals is approximately 1. Moreover, observe in Figure 5 the PDF of the medians μ_i measured for every individual i of our eight datasets. Observe that, while the typical μ_i is around 1 hour for the second group, for the first group it varies from 3 to 8 minutes.

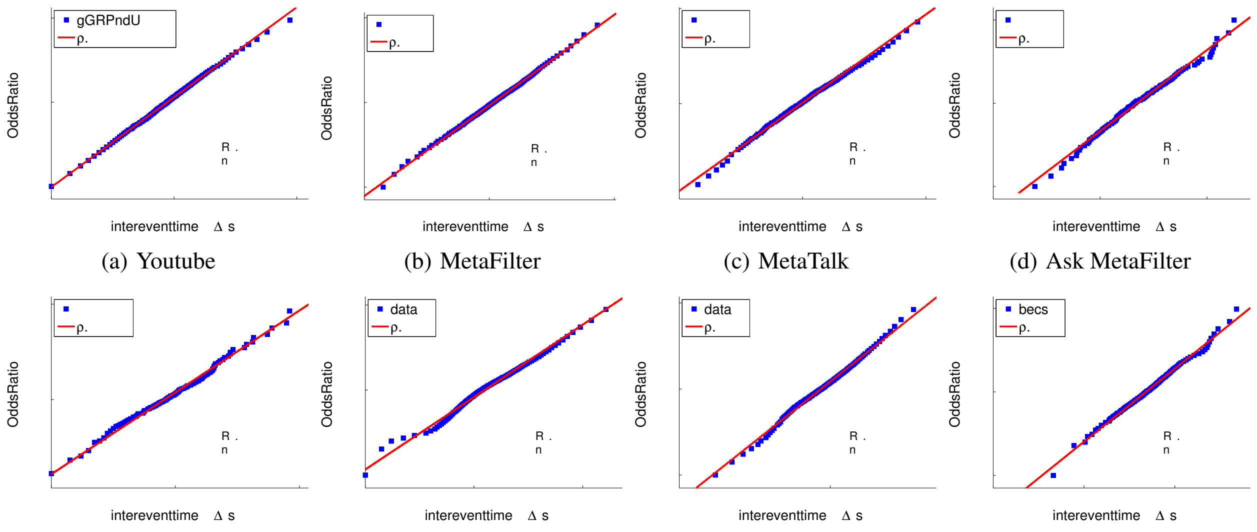
[6] https://developers.google.com/youtube/

[7] https://gdata.youtube.com/feeds/api/standardfeeds/on_the_web



(a) Histogram (b) CDF (c) CCDF (d) Odds Ratio

Figure 1: The inter-event times distribution of the most active individual of our eight datasets, with 44,785 SMS messages sent and received. We observe that both the power law fitting (PL fitting) with exponent 2 and the exponential fitting, generated by a PP, deviate from the real data. We also observe that the OR is very well fitted by a straight line with slope ≈ 1.



(a) Youtube (b) MetaFilter (c) MetaTalk (d) Ask MetaFilter

(e) Digg (f) SMS (g) Phone (h) E-mail

Figure 2: The Odds Ratio plot for one typical active individual of each dataset. Observe that an odds ratio power law, represented by a straight line with slope ρ in a log-log scale, is an appropriate fit for all individuals.

5. TEMPORAL CORRELATION

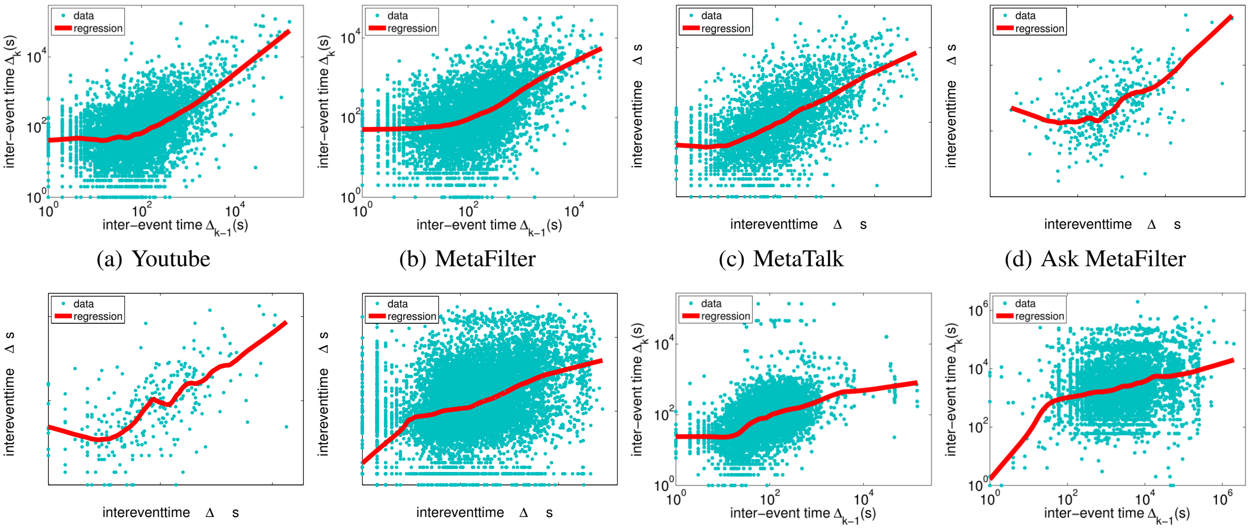
Although most previous analysis focus solely on the marginal IED, a subtle point is the *correlation* between successive inter-event times (Δ_{k-1} and Δ_k). What we illustrate here is that the independence between Δ_k and Δ_{k-1} does **not** hold for the eight datasets we analyzed in this work.

In Figure 6, we plot, for the same typical users of Figure 2, all the pairs of consecutive inter-event times (Δ_{k-1}, Δ_k). We also show the regression of the data points using the LOWESS smoother [8]. While the PP, as for any other i.i.d. process, the regression is a flat line with slope 0, for the eight typical users Δ_k tends to grow with Δ_{k-1}. This means that if I called you five years ago, my next phone call will be in about five years later. In short, there is a strong, positive dependency between the current inter-event time (Δ_k) and the previous one (Δ_{k-1}), clearly contradicting the independence assumption.

We formally investigate if two consecutive inter-event times are correlated analyzing the autocorrelation [4] of all the time series involving the inter-event times Δ_k of the individuals of our datasets. Autocorrelation refers to the correlation of a time series with its own past and future values. A positive autocorrelation, which is suggested by Figure 6, might be considered a specific form of "persistence", i.e., a tendency for a system to remain in the same state from one observation to the next.

We test if all the Δ_k time series of every individual of our datasets are random or autocorrelated. For this, we define the hypothesis test H_0 that a series $S = \{\Delta_0, \Delta_1, ..., \Delta_n\}$ of inter-event times is random. If S is random, then its autocorrelation coefficient $AC_l \approx 0$ for all lags $l > 0$, where a lag l is used to compare, in this case, values of Δ_k and Δ_{k-l}. More formally, if AC_l is between the 95% confidence interval for S to be random, then we accept H_0 that S is random. As we show in Figure 7, we reject the null hypothesis H_0 that the inter-event times of the individual of Figure 1 is random, since all $AC_l, 1 < l \leq 10$ are outside the confidence interval.

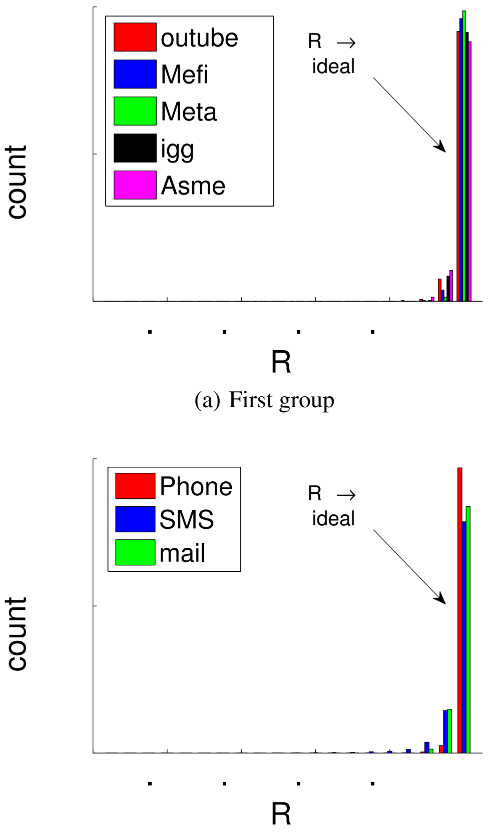
Since we are interested only in the case where the lag $l = 1$, we propose an alternative hypothesis test H_1 that the first-order autocorrelation coefficient AC_1 is greater than 0. If AC_1 is greater than the confidence interval for randomness, then we accept H_1 that the series is not random, i.e., there is a dependence between Δ_k and Δ_{k-1}. In Figure 8, we show the empirical probability $P(H_1)$ of accepting H_1 for individuals with a given number of events n of a given dataset. As we observe, as the number of communication events n grows and becomes significant, the probability of accepting H_1 increases rapidly. This strongly suggests that, on the con-



(a) Youtube (b) MetaFilter (c) MetaTalk (d) Ask MetaFilter

(e) Digg (f) SMS (g) Phone (h) E-mail

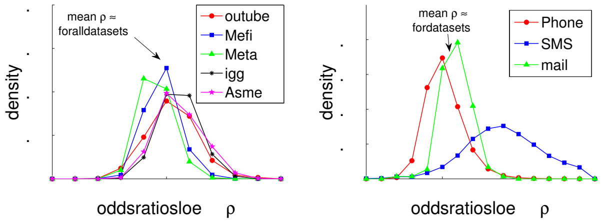
Figure 6: I.i.d. fallacy: dependence between Δ_k and Δ_{k-1}. Each point represents a pair of consecutive inter-event times (Δ_{k-1}, Δ_k) registered for a typical active individual of each dataset. The red line is a regression of the data points using the LOWESS smoother.



(a) First group

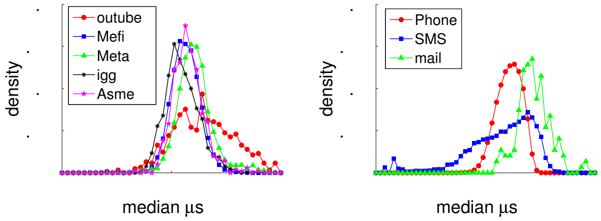
(b) Second group

Figure 3: The goodness of fit of our proposed model. We show the histograms of the R^2s measured for every user in the eight datasets. These histograms consider bins of size 0.05**. Thus, observe that the** R^2 **value for the great majority of individuals is located in the last bin, from** 0.95 **to** 1**.**



(a) First group (b) Second group

Figure 4: The PDF of the slopes ρ_i measured for every user u_i of our eight datasets. Except the SMS dataset, the typical ρ_i for the majority of individuals is approximately 1**.**



(a) First group (b) Second group

Figure 5: The PDF of the medians μ_i measured for every user of our eight datasets. Observe that the typical μ_i is around 3 and 8 minutes for the first group and around 1 **hour for the second group.**

trary of what happens with the i.i.d. inter-event times distribution generated by the Poisson Process or simply sampling from a log-logistic distribution, in real data there is a dependence between Δ_k and Δ_{k-1}. This also agrees with a recent work [35], which reports that daily series of calls made by a customer exhibits long memory.

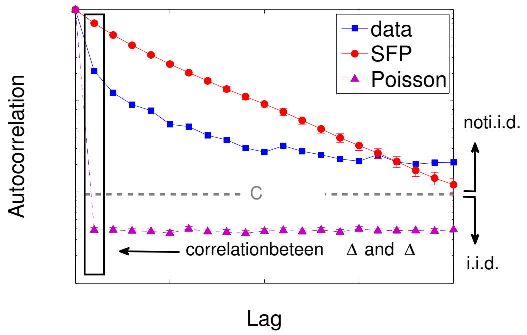
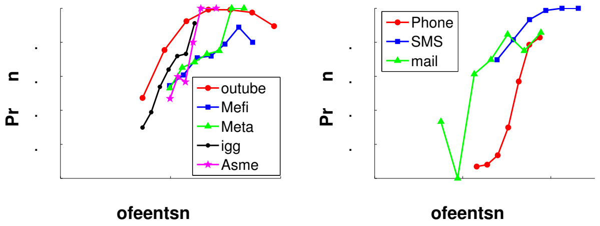


Figure 7: The sample autocorrelation for the same individual of Figure 1 and for synthetic data generated by the SFP and a PP with the same number of communication events and median.

Thus, in summary we can state that

$$E(\Delta_k|\Delta_{k-1}) = f(\Delta_{k-1}) \tag{3}$$

where f is a function that describes the dependency between Δ_k and Δ_{k-1}.



(a) First group (b) Second group

Figure 8: The empirical probability of an individual's inter-event times to be autocorrelated given his/her number of events. Note that as the number of events grows, the probability of having an autocorrelated series increases rapidly for the eight datasets.

6. THE SELF-FEEDING PROCESS

Given all the above evidence (OR power law; i.i.d. fallacy) and all the previous evidence (power law tails by Barabási; short-term regular behavior as the PP), the question is whether can we design a generator which will match all these properties? Our requirements for the ideal generator are the following:

R1: Realism – marginals The model should generate OR power law marginal IED;

R2: Realism – locally-Poisson: The model should behave as a Poisson Process within a short window of time;

R3: Avoid the i.i.d. fallacy Two consecutive inter-event times should be correlated;

R4: Parsimony It should need only few parameters, and ideally, just one or two.

At a high level, our proposal is that the next inter-arrival time will be an exponential random variable, with rate that *depends on the previous* inter-arrival time. It is subtle, but in this way our generator behaves like Poisson in the short term, gives power-law tails in the long term, generates OR power law marginals and is extremely parsimonious: just one parameter, the median μ of the IED. We call this model the *Self-Feeding Process* (SFP).

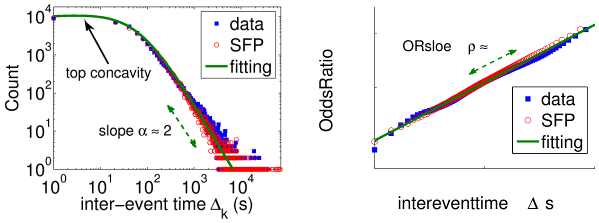
We propose the generator as follows

Model 1. *Self-Feeding Process SFP (μ).*
//μ is the desired median of the marginal PDF

$$\boxed{\begin{array}{lll} \Delta_1 & \leftarrow & \mu \\ \Delta_k & \leftarrow & \textit{Exponential}\ (\textit{mean}\ \beta = \Delta_{k-1} + \mu/e) \end{array}}$$

where μ is the only parameter of the model, being the desired median of the IED. The part μ/e has to be higher than 0 to avoid Δ_k to converge to 0 and has to be divided by the Euler's number e to make the median of the generated IED around the target median μ (more details in the Appendix A). This type of model is not new in the literature [41, 10] but they have not been extensively studied, perhaps due to the lack of empirical data fitting the implied distribution.

In Figures 9-a and 9-b we compare, respectively, the histogram and the OR of the inter-event times generated by the SFP model, all values rounded up, with the inter-event times of the individual of Figure 1. Notice that the distributions are very similar and both are well fitted by a log-logistic distribution, which looks like a hyperbola, thus addressing both the power-law tail, as well as the "top-concavity" that real data exhibits. For a generalized SFP model, that generates IEDs with different OR slopes, and more details about the log-logistic distribution, please see the Appendix B. Moreover, for an analysis over the temporal correlations generated by the SFP, see the Appendix A.4.



(a) Histogram (b) Odds ratio

Figure 9: Comparison of the marginal distribution of the inter-event times generated by the SFP model with the inter-event times of the user of Figure 1. Observe that both the histogram (a) and the OR (b) are almost identical.

The SFP model naturally generates an odds ratio power law with slope $\rho = 1$, which is the slope that characterizes the majority of the users of our datasets (see Figure 4). To the best of our knowledge, this is the first work that studies the IED of human communications using such a varied, modern and large collection of data. Despite the fact that the means of communications are intrinsically different, having their own idiosyncrasies, we have observed that the IED of individuals of these systems have the same characteristics, i.e., they follow an odds ratio power law behavior. Moreover, when the OR slope $\rho = 1$, the power law exponent of the PDF is $\alpha = -2$ (see the Appendix B.5 for details). This is the same IED slope α reported in [18, 40] as a result of fluctuations in the execution rate and in particular periodic changes. It has been argued that seasonality can only robustly give rise to heavy-tailed IEDs when the exponent $\alpha = 2$. However, we again point out that the proposed (Generalized) SFP model (see the Appendix A) can generate IEDs with power-law slopes α varying in the range $(-\infty, -1)$, agreeing with all the empirical studies we are aware of. Moreover, we point out that the typical values of the parameters μ and ρ can be easily extracted from the distributions shown in Figures 5 and 4.

7. THE UNIFYING POWER OF THE SFP

Finally, we would like to emphasize the unifying power of the SFP. Several works [21, 32, 31, 30, 25, 23] claim that in the short term, real data behave as regular as a PP. Our model also captures that, since successive inter-event times are exponentially distributed, with similar (but not identical) rates. Thus, one of the major contributions of this work is the unification of the two seemingly-conflicting viewpoints we mentioned earlier. The proposed SFP model unifies both theories by generating Poisson-like traffic in the short term, with smoothly varying rate, like the second viewpoint, and also generates a power-law tail distribution (see the Appendix B.5), even matching the top-concavity that power laws cannot match, like the first modern approach of Barabási [2].

In Figure 10, we explicitly show the SFP's unifying power. We compare synthetic data generated by the SFP model using the same odds ratio slope ρ, median μ and number of events of the user of Figure 1 with the real data from this user. Notice the bursts of activity and also the long periods of inactivity, in the first two columns of Figure 10. Also notice that both synthetic and real traffic significantly deviate from Poisson (sloping lines in Figures 10-b and 10-f) but are similar between themselves. However, in the short term, both real and synthetic data behave like Poisson, being practically on top of the black dashed lines of Figures 10-d and 10-h.

8. CONCLUSIONS

In this paper, we propose the SFP model, which reconciles previous approaches for human communication dynamics. The SFP is a parsimonious generator that requires at most two intuitive parameters, and yet it has several desirable properties:

- Realism: it matches very well the properties of the IEDs of eight large, diverse and real systems, such as online forums, Youtube comments, online news, e-mails, SMSs and phone calls;
- Unification Power: it reconciles seemingly-contradicting theories on human communication dynamics. Our model exhibits power law tail behavior and burstiness in the long term, as well as Poisson-like behavior in the short term;
- It avoids the "i.i.d. fallacy": inter-event times are not independent, i.e., the time needed for the next event to arrive depends on the time the previous event took to arrive. Our model is the first to capture this very subtle point.

Moreover, there are two additional contributions: (i) the proposal to use the so-called "Odds Ratio" function using the cumulative distribution function – most of our real data seems to obey a power-law in their Odds-Ratio function, even when their PDF deviates from a power-law; (ii) the proposal to use the log-logistic distribution, which has power-law tail, but also exhibits the so-called "top-concavity", that real data seem to have.

9. REFERENCES

[1] L. Akoglu, P. O. S. Vaz de Melo, and C. Faloutsos. Quantifying reciprocity in large weighted communication networks. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 2012, Kuala Lumpur*, 2012.

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

[3] S. Bennett. Log-logistic regression models for survival data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 32(2):165–171, 1983.

[4] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis, Forecasting, and Control*. Prentice-Hall, Englewood Cliffs, New Jersey, third edition, 1994.

[5] J. Cao, W. S. Cleveland, D. Lin, and D. X. Sun. Internet traffic tends to poisson and independent as the load increases. Technical report, Bell Labs Technical Report, 2001.

[6] F. Chierichetti, R. Kumar, P. Raghavan, and T. Sarlos. Are web users really markovian? In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 609–618, New York, NY, USA, 2012. ACM.

[7] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661+, Feb 2009.

[8] W. S. Cleveland. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

[9] D. Cox and V. Isham. *Point Processes*. Monographs on Applied Probability and Statistics. Taylor & Francis, 1980.

[10] D. R. Cox. Some Statistical Methods Connected with Series of Events. *Journal of the Royal Statistical Society. Series B (Methodological)*, 17(2):129–164, 1955.

[11] M. De Choudhury, H. Sundaram, A. John, and D. D. Seligmann. Social synchrony: Predicting mimicry of user actions in online social media. In *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04*, pages 151–158, Washington, DC, USA, 2009. IEEE Computer Society.

[12] J.-P. Eckmann, E. Moses, and D. Sergi. Entropy of dialogues creates coherent structures in e-mail traffic. *Proceedings of the National Academy of Sciences of the United States of America*, 101(40):14333–14337, October 2004.

[13] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM '99: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, pages 251–262, New York, NY, USA, 1999. ACM.

[14] P. R. Fisk. The graduation of income distributions. *Econometrica*, 29(2):171–185, 1961.

[15] S. Garriss, M. Kaminsky, M. J. Freedman, B. Karp, D. Mazières, and H. Yu. Re: Reliable email. In *Proceedings of the Third USENIX/ACM Symposium on Networked System Design and Implementation (NSDI'06)*, pages 297–310, 2006.

[16] S. S. Gokhale and K. S. Trivedi. Log-logistic software reliability growth model. In *HASE '98: The 3rd IEEE International Symposium on High-Assurance Systems Engineering*, pages 34–41, Washington, DC, USA, 1998. IEEE Computer Society.

[17] F. A. Haight. *Handbook of the Poisson distribution [by] Frank A. Haight*. Wiley New York,, 1967.

[18] C. A. Hidalgo. Scaling in the inter-event time of random and seasonal systems. *PHYSICA A*, 369:877, 2006.

[19] M. Jamali, G. Haffari, and M. Ester. Modeling the temporal dynamics of social rating networks using bidirectional effects of social relations and rating patterns. In *Proceedings of the*

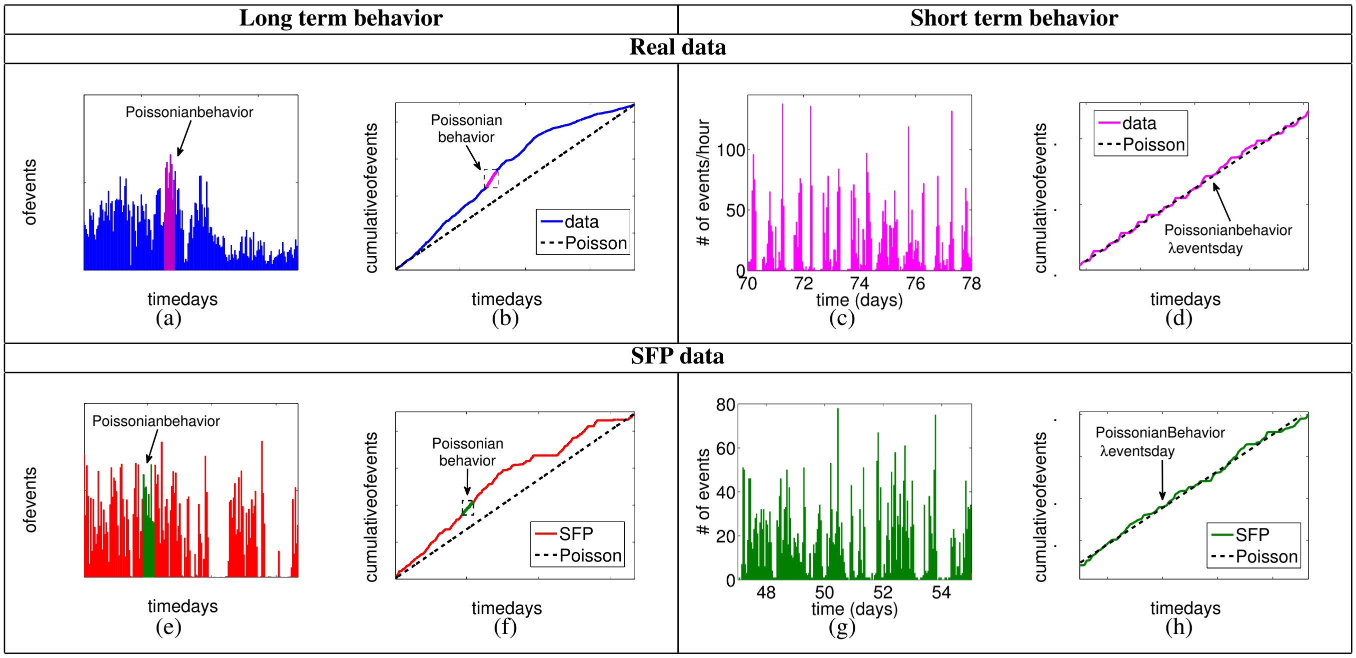


Figure 10: Unification Power of SFP: non-Poisson/bursty in the long term, but Poisson in the short term. Real data: Traffic of the user of Figure 2-1, showing event-count per unit time (a and c) and respective cumulative event-count (b and d). SFP data: synthetic traffic generated by the SFP model (with matching μ, ρ and event-count). Observe that (1) both time series are visually similar; (2) both are bursty in the long run (spikes; inactivity) (3) both are Poisson-like in the short term (last two columns)

20th international conference on World wide web, WWW '11, pages 527–536, New York, NY, USA, 2011. ACM.

[20] H. Jiang and C. Dovrolis. Why is the internet traffic bursty in short time scales? In *Proceedings of the 2005 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS'05)*, pages 241–252, 2005.

[21] T. Karagiannis, M. Molle, M. Faloutsos, and A. Broido. A nonstationary Poisson view of Internet traffic. In *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, volume 3, pages 1558–1569 vol.3, 2004.

[22] M. Karsai, K. Kaski, A.-L. Barabási, and J. Kertész. Universal features of correlated bursty behaviour. *Scientific Reports*, 2, May 2012.

[23] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of the eighth ACM SIGKDD*, KDD '02, pages 91–101, New York, NY, USA, 2002. ACM.

[24] B. Klimt and Y. Yang. Introducing the enron corpus. In *CEAS'04: The First Conference on Email and Anti-Spam*, 2004.

[25] A. Kuczura. The interrupted poisson process as an overflow process. *The Bell System Technical Journal*, 52:437–448, 1973.

[26] J. F. Lawless and J. F. Lawless. *Statistical Models and Methods for Lifetime Data (Wiley Series in Probability & Mathematical Statistics)*. John Wiley & Sons, January 1982.

[27] K. Lerman and T. Hogg. Using a model of social dynamics to predict popularity of news. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 621–630, New York, NY, USA, 2010. ACM.

[28] M. O. Lorenz. Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, 9:209–219, 1905.

[29] T. Mahmood. Survival of newly founded businesses: A log-logistic model approach. *JournalSmall Business Economics*, 14(3):223–237, 2000.

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

[31] R. D. Malmgren, D. B. Stouffer, A. S. L. O. Campanharo, and L. A. N. Amaral. On universality in human correspondence activity. *SCIENCE*, 325:1696, 2009.

[32] R. D. Malmgren, D. B. Stouffer, A. E. Motter, and L. A. N. Amaral. A poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences*, 105(47):18153–18158, November 2008.

[33] C. S. M.I. Ahmad and A. Werritty. Log-logistic flood frequency analysis. *Journal of Hydrology*, 98:205–224, 1988.

[34] J. G. Oliveira and A.-L. Barabasi. Human dynamics: Darwin and Einstein correspondence patterns. *Nature*, 437(7063):1251, Oct. 2005.

[35] M. Owczarczuk. Long memory in patterns of mobile phone usage. *Physica A: Statistical Mechanics and its Applications*, Oct. 2011.

[36] K. Radinsky, K. Svore, S. Dumais, J. Teevan, A. Bocharov, and E. Horvitz. Modeling and predicting behavioral dynamics on the web. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 599–608, New York, NY, USA, 2012. ACM.

[37] E. Shmueli, A. Kagian, Y. Koren, and R. Lempel. Care to comment?: recommendations for commenting on news stories. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 429–438, New York, NY, USA, 2012. ACM.

[38] P. O. S. Vaz de Melo, L. Akoglu, C. Faloutsos, and A. A. F. Loureiro. Surprising patterns for the call duration distribution of mobile phone users. In *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*, pages 354–369, 2010.

[39] P. O. S. Vaz de Melo, C. Faloutsos, and A. A. Loureiro. Human dynamics in large communication networks. In *SIAM Conference on Data Mining (SDM)*, pages 968–879. SIAM / Omnipress, 2011.

[40] A. Vazquez, J. G. Oliveira, Z. Dezso, K.-I. Goh, I. Kondor, and A.-L. Barabasi. Modeling bursts and heavy tails in human dynamics. *Phys Rev E Stat Nonlin Soft Matter Phys*, 73:036127, 2006.

[41] H. Wold and U. universitet. Statistiska institutionen. *On Stationary Point Processes and Markov Chains*. Selected publications - University of Uppsala, Department of Statistics. Swedish and Danish Actuarial Societies, 1948.

APPENDIX

A. THE GENERALIZED SFP MODEL

A.1 Model Definition

In Figure 4 we showed the slopes ρ of the OR fitting for the IEDs of all individuals of our datasets. It is fascinating that the typical ρ_i for the individuals of seven of our eight datasets is approximately 1, the same slope generated by the SFP model. Several individuals though, mainly from the SMS dataset, have a much higher value of ρ, close to $\rho \approx 2$. To accommodate that and all the variance seen in the data, we introduce our Generalized SFP model, which needs just one parameter more, ρ. Thus, we have:

Model 2. *Generalized Self-Feeding Process $SFP(\mu, \rho)$.*

$$\boxed{\begin{array}{rcl} \delta_1 & \leftarrow & \mu \\ \delta_t & \leftarrow & \textit{Exponential (mean: } \beta = \delta_{t-1} + \mu^{\rho}/e) \\ \Delta_k & \leftarrow & \delta_t^{1/\rho}. \end{array}}$$

Note the auxiliary variable δ_t, which stores the inter-event times without the influence of ρ.

A.2 Parameters

Before reaching the full SFP model described in the paper, we had a simpler version of it, relying on a different parametrization scheme:

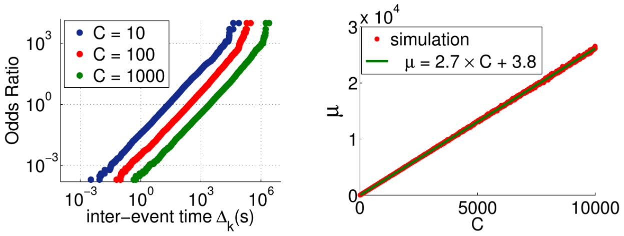
Model 3. *Self-Feeding Process SFP (C,a).*

$$\boxed{\begin{array}{rcl} \delta_1 & \leftarrow & C \\ \delta_t & \leftarrow & \textit{Poisson Process}(\beta = \delta_{t-1} + C) \\ \Delta_k & \leftarrow & \delta_t^{a}, \end{array}} \tag{4}$$

where C is the location parameter and a is the shape parameter that defines the odds ratio slope ρ. An easy and direct way to define the relationships between this model's parameters and the distribution properties μ and ρ is through simulations.

Thus, the first point we consider is the median μ of the inter-event times generated by the SFP model when $a = 0$. When $OR(x) = 1$, x is the median μ of the distribution. Thus, in Figure 11-a, we plot the OR for different values of C. We observe that changing the value of C changes μ and, consequently, the location of the distribution, but maintains its slope. We also see that μ is close but different than the value of C.

In order to investigate the relationship between C and μ, we run simulations of the model for all integer values of C between [1,10000]. As we observe in Figure 11-b, the median μ of the inter-event times distribution (IED) varies linearly with C according to a slope of ≈ 2.72, that can be approximated by Euler's number e, in a way that $\mu \propto e \times C$. This allows us to generate inter-event times with a determined μ when the slope $\rho = 1$. We ignore the constant factor 3.8 because its 95% confidence interval is $(-8.596, 16.3)$, which contains zero.



(a) The OR of the IED for different values of C

(b) μ as a function of C

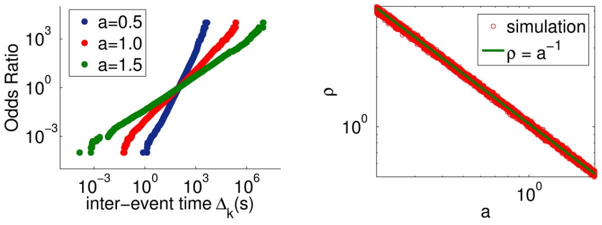
Figure 11: Changing the value of C changes the location of the distribution. The median of the distribution μ varies linearly with C, $\mu = a \times C + b$, with $a = 2.719$ and $b = 3.8$. The 95% confidence interval for a is $(2.715, 2.723)$ and for b is $(-8.60, 16.3)$. Since the confidence interval for b contains 0, b is not significant.

Now we know how to generate inter-event times with different medians μ using the parameter $C = \mu/e$ of SFP. The next step is to verify how the SFP model can generate IEDs with a desired slope $\rho \neq 1$. Considering that up to this point the SFP model generates a set of inter-event times I_1 with a slope 1, the idea is to use an exponent a to transform I_1 into I_ρ, which is an IED with a different slope ρ. When we elevate each $\Delta_k \in I_1$ to the power of $a \neq 1$, the resulting slope ρ becomes different than 1, as we see in Figure 12-a. In the same way we did for C, we run simulations of the model for 1000 different values of $a \in [0.1, 2]$. As we observe in Figure 12-b, there is an inverse relationship between a and ρ, i.e., $\rho = a^{-1}$. Moreover, since the median of the distribution is also elevated to the power of a, we have to elevate the parameter μ to the power of $\rho = a^{-1}$ to preserve the median.

A.3 The need for the constant μ/e in SFP

Lemma 1. *The constant $C = \mu/e > 0$ of Model 3 is needed to assure that the inter-event times generated by the SFP model will not converge to zero.*

Proof. If we remove the constant C from Model 2, $\Delta_k = (\Delta_{k-1}) \times (-\ln(U(0,1)))$, or Δ_k will be equal to Δ_{k-1} multiplied by a random number X extracted from the exponential distribution with parameter $\beta = \lambda = 1$. If $(X = \frac{1}{k} \mid k > 1)$, then Δ_k will be equal to Δ_{k-1} divided by k. The probability of X to be $\frac{1}{k}$ is $P(X = \frac{1}{k}) = e^{-\frac{1}{k}} = \frac{1}{\sqrt[k]{e}}$. On the other hand, the probability of multiplying Δ_k by k and, therefore, return Δ_{k+1} to Δ_{k-1} value is $P(X = k) = e^{-k} = \frac{1}{e^k}$. Given



(a) The OR of the IED for different values of a

(b) ρ as a function of a

Figure 12: Changing the value of a changes the slope ρ of the distribution in a way that $\rho = a^{-1}$.

these probabilities, observe that $P(X = \frac{1}{k}) = \frac{1}{\sqrt[k]{e}} > P(X = k) = \frac{1}{e^k}, \forall k > 1$. From this, we conclude that the expected value of Δ_k when $t \to \infty$ is 0. With C in the equation, even when $\Delta_{k-1} = 0$, $\Delta_k = -C \times \ln(U(0, 1)$, that is a classic Poisson process with $\beta = C$, and, obviously, does not converge to 0. □

A.4 Lower Temporal Correlation

The SFP model is build upon a direct dependence between consecutive inter-event times. Because of that, the correlation between consecutive inter-event times is significantly higher than real data. While the average Pearson's correlation coefficient for real data is approximately 0.4, for synthetic data generated by the SFP model it is approximately 0.7. In order to generate more realistic data, we suggest a slight modification of the SFP process. Instead of generating the next inter-event time (Δ_k) based on the immediate previous one (Δ_{k-1}), we propose that it should be generated from a ϵ-th previous one ($\Delta_{k-\epsilon}$). This can be done by extracting ϵ from an exponential distribution with mean $\beta = 1$ and making its ceiling, so the lower bound for ϵ is 1. In summary, the SFP model is changed as follows:

Model 4. *Self-Feeding Process* SFP*(μ).*
//μ is the desired median of the marginal PDF

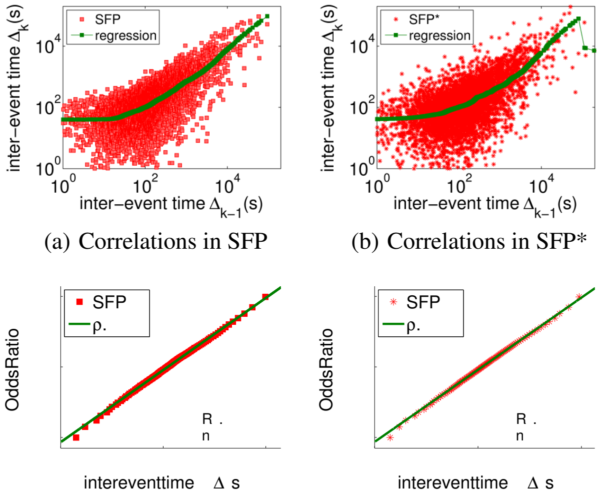
$$\boxed{\begin{array}{rcl} \Delta_1 & \leftarrow & \mu \\ \epsilon & \leftarrow & \lceil \mathit{Exponential}\ (\mathit{mean}\ \beta = 1) \rceil \\ \Delta_k & \leftarrow & \mathit{Exponential}\ (\mathit{mean}\ \beta = \Delta_{k-\epsilon} + \mu/e) \end{array}}$$

Observe in Figure 13 that the synthetic data generate by the SFP* has a lower correlation (0.43) between consecutive inter-event times than the original one (0.70). Despite of that, the odds ratio generated by the SFP* is still a power law with slope $\rho \approx 1$.

B. THE SFP STATIONARY DISTRIBUTION

B.1 Analytical Result

The SFP model is within the general class of Wold processes defined as processes with Markov-dependent interevents intervals [41, 10]. These processes have not been extensively studied in the literature, perhaps due to the mathematical difficulties in deriving their probabilistic properties. Consider the existence of a stationary distribution for the generalized SFP model. A stationary PDF $f(x)$ of the Markov chain δ_t must satisfy



(a) Correlations in SFP

(b) Correlations in SFP*

(c) Odds Ratio for SFP

(d) Odds Ratio for SFP*

Figure 13: Comparison between the synthetic data generated by the SFP and the SFP*. Observe that the synthetic data generate by the SFP* has a lower correlation (0.43) between consecutive inter-event times than the SFP (0.70). Despite of that, the odds ratio generated by the SFP* is still a power law with slope $\rho \approx 1$.

$$\begin{aligned} f(x) &= \int_0^\infty f(y \to x) f(y) dy \\ &= \int_0^\infty \frac{1}{y + \mu^\rho/e} \exp(-x/(y + \mu^\rho/e)) f(y) dy \end{aligned}$$

This integral equation has no obvious analytical solution but in the next sections we show via simulations of the point process that $f(x)$ is very well approximated by a log-logistic density. This mathematical difficulty is common in the previous attempts to model data with Wold processes. Even if a consistent density $f(x)$ and a transition kernel $f(y \to x)$ are given, properties are, in general, difficult to obtain [9].

B.2 Fitting Synthetic Data

In Figure 14-a, we plot the histogram of 100,000 time intervals Δ_k generated by the SFP model with $\mu = e$. Moreover, in Figure 14-b, we plot the OR for the same time intervals. While a classic PP generates an exponential distribution, we observe that the generated data by the SFP perfectly fits a distribution with an Odds Ratio function that is a power law with slope $\rho = 1$. Thus, we propose the following conjecture:

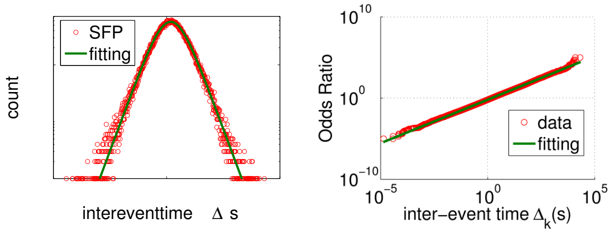
Conjecture 1. *The SFP model generates a log-logistic distribution with $\rho = \sigma = 1$,*

where σ is the shape parameter of the log-logistic distribution.

We have several and significant evidences that the SFP generates a log-logistic distribution, but at this moment we do not have a formal analytical proof that this is true.

B.3 The SFP Markov Chain

The SFP can be naturally considered as a Markov Chain (MC), since it is a sequence of random variables $\Delta_1, \Delta_2, \Delta_3, ...$ with the



(a) Histogram (marginal). (b) OR (marginal).

Figure 14: Inter-event times Δ_k generated by the SFP. The generated Δ_ks are perfectly fitted by a log-logistic distribution with the slope $\rho = \sigma = 1$.

Markov property, namely that, given the present inter-event time, or state, the future and past inter-event times, or states, are independent. Thus, here we model the SFP as a time-homogeneous Markov chain with a finite state space to give another evidence that the SFP has a stationary distribution and that it is very likely that this distribution is the log-logistic.

Originally, the SFP can be considered as a continuous-time MC, but for simplicity, we build a discrete-time Markov chain in a way that each state $\Delta_i = \{\Delta_1, \Delta_2, \Delta_3, ...\}$ represents the inter-event times with values within the interval $(i-1, i]$. For instance, considering the granularity in seconds, if the current inter-event time is 3.8 seconds, then the MC is in the state Δ_4. Also for simplicity, we build a finite-state MC with a maximum number of states n, i.e., the states go from Δ_1 to Δ_n. The MC will be in state Δ_n every time the current inter-event time is within the interval (n, ∞).

Thus, considering a n-state MC build from the SFP model, the transitions probabilities $p_{i,j}$ of going from state Δ_i to Δ_j are given in the following way:

$$p_{i,j} = \begin{cases} CDF_{exp}(\text{x=j},\beta\text{=i+C}) - CDF_{exp}(\text{x=j-1},\beta\text{=i+C}) & \text{if j<n} \\ 1 - CDF_{exp}(\text{x=j},\beta\text{=i+C}) & \text{if j=n,} \end{cases}$$

where $CDF_{exp}(x,\beta)$ is the cumulative distribution function of the exponential distribution on x with mean β and $C = \mu/e$, given by the SFP (Equation 1). Observe in Figure 15 that the probability density function of the log-logistic is virtually identical to the one of the stationary distribution of the SFP Markov Chain. This is another strong indication that the SFP generates log-logistically distributed data.

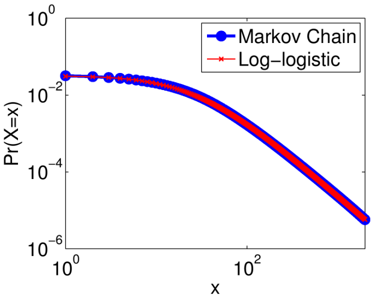


Figure 15: The probability density function of the log-logistic distribution and the stationary distribution of the SFP MC.

It is important to point out that in [6] the authors showed that the behavior of Web users is not Markovian, i.e., a user's next action does not depends only on her/his current state. Our assumption differs from this one because we assume that users have Markovian behavior in communications, while [6] studied whether users have Markovian behavior while navigating on the Web.

B.4 Log-logistic Distribution

The log-logistic distribution was first proposed by Fisk [14] to model income distribution, after observing that the OR plot of real data in log-log scales follows a power law $OR(x) = cx^{\rho}$. In summary, a random variable is log-logistically distributed if the logarithm of the random variable is logistically distributed. The logistic distribution is very similar to the normal distribution, but it has heavier tails. In the literature, there are examples of the use of the log-logistic distribution in survival analysis [3, 29], distribution of wealth [14], flood frequency analysis [33], software reliability [16] and phone calls duration [38]. A commonly used log-logistic parametrization is [26]:

$$\begin{aligned} PDF_{LLG}(x) &= \tfrac{e^z}{\sigma x(1+e^z)^2}, \\ CDF_{LLG}(x) &= \tfrac{1}{1+e^{-z}}, \\ z &= (\ln(x) - \ln(\mu))/\sigma, \end{aligned} \qquad (5)$$

where $\sigma = 1/\rho$, the slope of our SFP model, and μ is the same. Moreover, when $\sigma = 1$, it is the same distribution as the Generalized Pareto distribution [28] with shape parameter $\kappa = 1$, scale parameter μ and threshold parameter $\theta = 0$.

B.5 SFP has Power Law Tail

The universality class model proposed by Barabási [2] states that the IED has a power law tail. The proposed SFP model agrees with this model in a way that:

Lemma 2. *If Conjecture 1 is correct, then the SFP model generates an IED that converges to a power law when $x \to \infty$, i.e., $\lim_{x\to\infty} \frac{PDF_{LLG}(x)}{x^{-\alpha}} = k$, where k is a constant greater than* 0.

Proof. Considering the Probability Density Function of the log-logistic distribution showed in Equation 5, if we set the location parameter $\mu = 1$ for simplicity, $e^z = x^{1/\sigma}$. Then, $PDF_{LLG}(x)$ can be simplified to

$$PDF_{LLG}(x) = \frac{x^{\frac{1}{\sigma}-1}}{\sigma(1 + x^{\frac{1}{\sigma}})^2}.$$

When $x \to \infty$, the addition of 1 in the denominator can be disregarded, resulting in the following simplification:

$$PDF_{LLG}(x) = \frac{x^{-(1+1/\sigma)}}{\sigma}, x \to \infty.$$

Thus, when $x \to \infty$, the IED generated by the SFP model is a power law with slope

$$\alpha = -(1 + 1/\sigma) = -(1 + \rho). \qquad (6)$$

Observe again Figure 14-a and note the power law tail. □

C. SFP CODE

Below we show the *Python* code for the SFP generator.

```
def SFP(n, mu, rho=1):
    #first inter-event time
    deltat = mu
    #list of inter-event times
    Deltat = []
    for i in range(1, n):
        #Poisson Process which Beta=deltat+mu/e
        deltat = -(deltat+(mu**rho)/math.e)
        deltat = deltat * math.log(random.random())
        Deltat.append(deltat**(1/rho))
    return Deltat
```