Eur. Phys. J. B **75**, 389–394 (2010)
DOI: 10.1140/epjb/e2010-00139-9

THE EUROPEAN PHYSICAL JOURNAL B

Regular Article

# Poissonian bursts in e-mail correspondence

C. Anteneodo[1,a], R.D. Malmgren[2,3], and D.R. Chialvo[4]

[1] Departamento de Física, PUC-Rio and National Institute of Science and Technology for Complex Systems, Rua Marquês de São Vicente 225, CEP 22453-900 RJ, Rio de Janeiro, Brazil
[2] Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL 60208, USA
[3] Datascope Analytics, Evanston, IL 60201, USA
[4] Department of Physiology, Feinberg Medical School, Northwestern University, 303 East Chicago Ave. Chicago, IL 60611, USA

Received 8 July 2009 / Received in final form 22 November 2009
Published online 6 May 2010 – 

**Abstract.** Recent work has shown that the distribution of inter-event times for e-mail communication exhibits a heavy tail which is statistically consistent with a cascading Poisson process. In this work we extend this analysis to higher-order statistics, using the Fano and Allan factors to quantify the extent to which the empirical data are more correlated – bursty – than a Poisson process. Our analysis demonstrates that the correlations in the empirical data are indistinguishable from those of randomly reordered time series, illustrating that any correlations in the data are not due to the precise ordering of events. We further find that correlations in synthetic time series generated from a cascading Poisson process agree quite well with the correlations observed in the empirical data. Finally, we rescale the empirical time series to confirm that e-mail correspondence is no more correlated than expected from a suitably chosen Poisson process.

The assessment of human activity patterns is crucial for many applications, such as optimization of information traffic, service scheduling and human resource planning. The temporal dynamics of e-mail correspondence has sparked recent interest [1–11] because of its importance as a communication medium and the availability of large databases that precisely record human activity. Recent research [2–5,7,12] has shown that the probability distribution of the time elapsed between consecutively sent e-mails by a single user – the inter-event time distribution – exhibits heavy tails, suggesting that human activity might be more bursty than expected from an uncorrelated Poisson process [2].

One explanation for the observed heavy-tailed inter-event time distribution is a cascading Poisson process [7,13,14]. In this model, there is a primary non-homogeneous Poisson process, which explicitly incorporates daily and weekly modulations, each of whose events triggers a secondary process which is also Poissonian but with a much larger characteristic rate. According to this model, "bursts" of e-mail activity occur in non-overlapping homogeneous Poisson cascades (as opposed to the overlapping cascades of Refs. [15–17], for instance) separated by long periods of inactivity defined by the primary process. The resulting inter-event time distribution predicted by the model is heavy-tailed due to the mixture of several different scales of rates of activity.

Although the cascading Poisson process has been shown to be statistically consistent with the empirical inter-event time distributions of several individuals [7], it is unclear whether higher-order statistical patterns (e.g. correlations) are present in the data and whether the cascading Poisson process adequately captures these correlations. Here, we investigate the higher-order statistical structure of e-mail correspondence and we compare it with the higher-order statistics that is expected from a cascading Poisson process. We show that the bursty patterns in e-mail communication are no more correlated than expected from a suitably chosen Poisson process.

We study here a database considered previously [1–5,7,13,14] comprised of 3188 e-mail accounts over an 83-day period at a European University. This database consists of several accounts that sent over 1000 messages and several accounts that do not send any e-mails at all. To avoid having our analysis distorted by accounts that are listservs or spammers or by accounts that send too few messages to capture a reasonable snapshot of their communication activity, we restrict our analysis to the correspondence from 394 users that sent at least 40 messages over 83 days and are not likely to be listservs (for details on the preprocessing procedure, see the supporting information of Ref. [7]). These users are "typical" in the sense that they use e-mail frequently enough to constitute a reasonable portion of e-mail communication patterns. Throughout this manuscript we refer to these 394 users as the "empirical" time series.

We also analyze here "synthetic" time series' that are generated from the cascading Poisson process [7]. In the cascading Poisson process, cascades of activity are initiated by a non-homogeneous Poisson process with rate

[a] e-mail: celia@fis.puc-rio.br

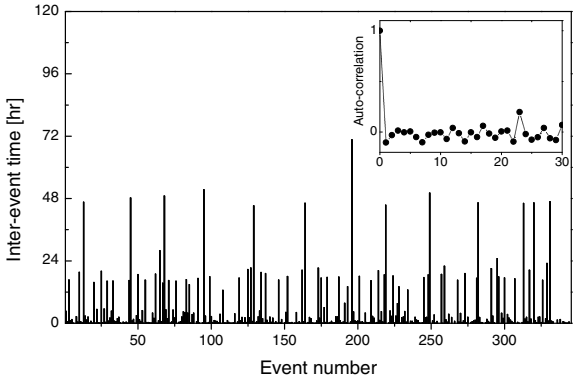


**Fig. 1.** Time series of inter-event times for a representative user (User 467), and its autocorrelation function (inset).

$\rho_p(t)$ that depends on time in a periodic manner. The rate of the non-homogeneous Poisson process is proportional to the daily and weekly distributions of cascade initiation, $p_d(t)$ and $p_w(t)$:

$$\rho_p(t) = N_w p_d(t) p_w(t), \tag{1}$$

where $\rho_p(t)$ has a weekly periodicity and the proportionality constant $N_w$ is the average number of cascades of activity per week. Once the cascade of activity is initiated, $N_a$ additional events occur during a homogeneous Poisson process with rate $\rho_a \gg \max_t \rho(t)$ where $N_a$ is drawn from some distribution $p(N_a)$. Once the $N_a$ events have occurred in the active interval, the activity of the individual is again governed by the non-homogeneous Poisson process defined by equation (1). The overall rate $\rho(t)$ of the cascading Poisson process is therefore a mixture of the periodic rate $\rho_p(t)$ and the rate during cascades of activity $\rho_a$. In this manuscript, we use the parameters from reference [7] to generate 30 "synthetic" time series' for each of the 394 users.

Before delving into a detailed analysis of all 394 users, we would first like to illustrate our analysis on a single user, User 467, which exhibits a behavior close to the average behavior and is prominently displayed throughout reference [7]. User 467 exhibits some features that are typical of human correspondence (Fig. 1). First, User 467 has long pauses between e-mails on the order of 16 and 48 h, which are related to the fact that User 467 usually sends e-mails while at work during the work week. These long pauses are captured by $\rho_p(t)$ in the cascading Poisson process. Second, User 467 has several short time intervals between e-mails that are uncorrelated. These short time intervals are captured by the cascades of activity in the cascading Poisson process. Here, we wish to quantify the extent to which User 467's activity patterns are correlated; that is, how bursty is User 467's activity?

One might be tempted to characterize the burstiness of User 467's behavior through multivariate distributions of inter-event times. Quantifying burstiness from multivariate distributions, however, is difficult to assess when the time series' have few events, as in the present case. Instead, we can analyze the counting statistics [18] of User 467's e-mails. Specifically, we use the Fano and Allan factors, two suitable metrics for point processes that provide reliable results for time series with few events [15–20], to gain insight into the higher-order statistical structure of e-mail correspondence.

The Fano and Allan factors are calculated by dividing the whole observation time interval into $W$ non-overlapping adjacent time windows of equal length $T$ and counting the number of events $N_k$ in the $k$th time window. The Fano factor ($FF$) is the ratio of the variance to the mean of the number of events in each time window, $FF = (\langle N_k^2\rangle - \langle N_k\rangle^2)/\langle N_k\rangle$, and it represents a measure of the dispersion of the resulting time series relative to a homogeneous Poisson process with the same rate. The Allan factor ($AF$) quantifies the difference in variance of counts of adjacent time windows, $AF = (\langle (N_{k+1} - N_k)^2\rangle)/(2\langle N_k\rangle)$, and it is a measure of the correlation of counts between successive time windows relative to the expectation from a homogeneous Poisson process with the same rate. Both of these measures quantify different aspects of what might constitute bursty activity: larger than Poissonian dispersion and larger than Poissonian correlations.

If the time series were generated by a homogeneous Poisson process, then the number of counts $N_k$ in each time window would be independent and identically distributed random variables drawn from a Poisson distribution. In such a case, $FF(T) = AF(T) = 1$, regardless of the time window length $T$ (Fig. 2, gray triangles). Deviations from unity therefore quantify departures from uncorrelated Poissonian statistics. For example, $FF(T) > 1$ would indicate that the time series is more bursty than expected from a homogeneous Poisson process at a particular time-scale $T$.

We begin by analyzing the Fano and Allan factors as a function of the length of the time window for User 467, to be complemented later in the manuscript with an analysis of all 394 users. The results for User 467 are plotted in Figure 2 (black circles). Notice that for time windows shorter than a few minutes the point process of e-mails is essentially Poissonian, denoted by the fact that both indices remain close to unity. For longer times the $FF$ and $AF$ curves noticeably depart from unity, suggesting that there might be some non-Poissonian effects in e-mail communication.

To further evaluate the empirical time series' departure from Poissonian statistics, we analyzed the surrogate time series obtained by randomly reordering the sequence of inter-event times to remove any correlations between consecutive inter-event times. If the empirical time series exhibits the same behavior in $FF$ and $AF$ as the shuffled time series, then the observed departure from Poissonian statistics is only due to the distribution of inter-event times and not due to their particular ordering; that is, the inter-event times are no more bursty than expected by chance. Indeed, we find that the empirical $FF$ and $AF$ curves are not discernably different from their shuffled counterparts (Fig. 2, white squares), indicating that the observed "departure" from Poissonian statistics is merely an artifact of the heavy-tailed inter-event time distribution, and not due to some intrinsic burstiness.

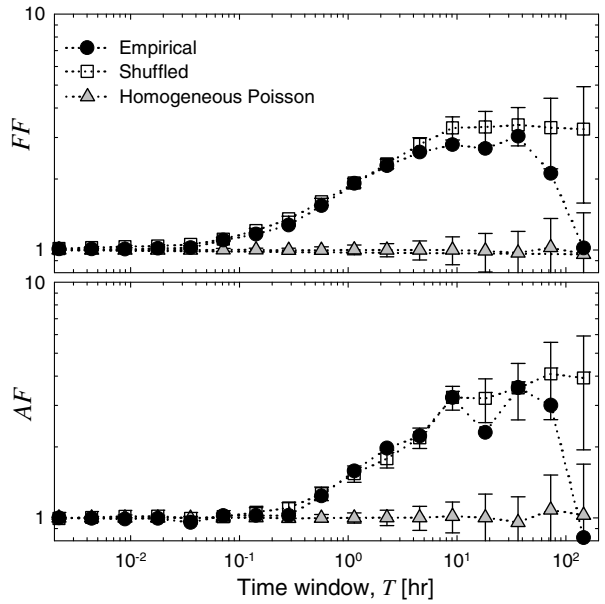


**Fig. 2.** Fano and Allan factors as a function of the time window $T$ for User 467. For comparison, we plot the mean and standard deviation of the Fano and Allan factors for 30 realizations of shuffled time series as well as the mean and standard deviation of the Fano and Allan factors for 30 realizations of the homogeneous Poisson process with the same rate. Large deviations observed at long time-scales are due to poor statistics (e.g. $W = 8$ in the last time window).

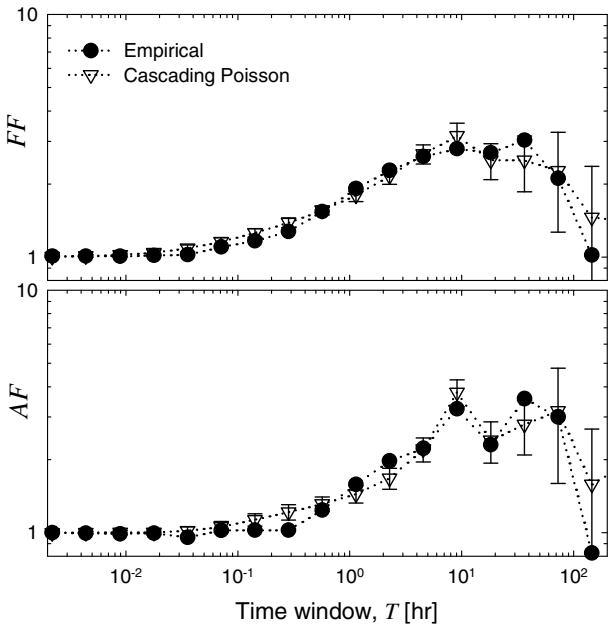


**Fig. 3.** Fano and Allan factors as a function of the length of the counting time window $T$, for the empirical time series for User 467 together with the mean value and standard deviation of Fano and Allan factors computed for 30 realizations of synthetic data generated from a cascading Poisson process over 83 days with best-estimate parameters obtained in references [4,5]. Large deviations observed at long time-scales are due to poor statistics (e.g. $W = 8$ in the last time window).

We next proceed to compare the Fano and Allan factors for User 467 with the Fano and Allan factors of synthetic time series generated from the cascading Poisson process [7]. Here, we generate synthetic time series over 83 days using the best-estimate parameters for this user from reference [7]. As Figure 3 shows, the empirical and synthetic $FF$ and $AF$ curves agree quite well, indicating that the cascading Poisson process is capturing not only the inter-event time distribution (already discussed at length in Ref. [7]) but also the higher-order statistical features of e-mail communication.

The analysis thusfar suggests that the correlations in the empirical data are well-approximated by a cascading Poisson process. If this is, in fact, the case, we should be able to rescale time such that the resulting point process appears to have originated from a homogeneous Poisson process with unit rate. Specifically, any non-homogeneous Poisson process with occurrence rate $\rho(t)$, can be mapped onto a homogeneous Poisson process through a simple transformation of the timescale [18], namely

$$\tilde{t} = \int_0^t \rho(s)ds. \qquad (2)$$

In this new time scale, the Poisson process has unit rate, $\tilde{\rho}(\tilde{t}) = 1$. In the particular case of a cascading Poisson process with known best-estimate parameters $\rho_p(t)$ and $\rho_a$ and where we know which events are associated with which cascade of activity [7], we can rescale the times between consecutive events by their respective rates. The results presented in Figure 4 confirm our hypothesis: the

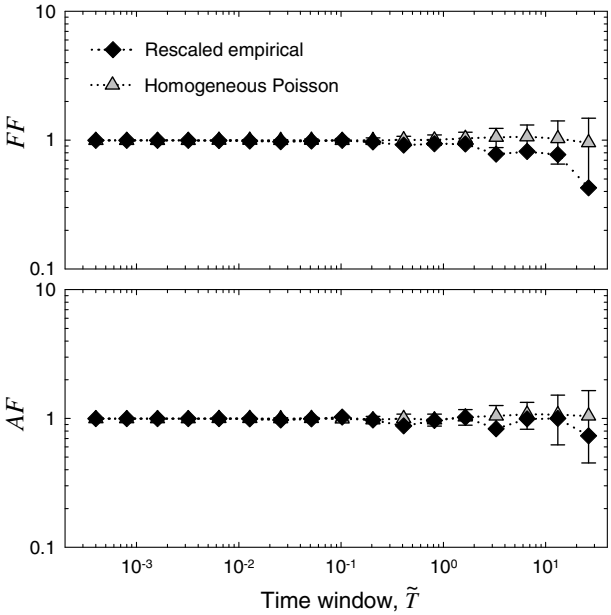


**Fig. 4.** Fano and Allan factors as a function of the (dimensionless) time window $\tilde{T}$ for the rescaled time series of User 467. For comparison, we also plot the mean and standard deviation of the Fano and Allan factors for 30 realizations of a homogeneous Poisson process with unit rate over an equivalent duration of time. The large deviations observed at long time-scales are due to poor statistics (e.g. $W = 8$ in the longest time window).

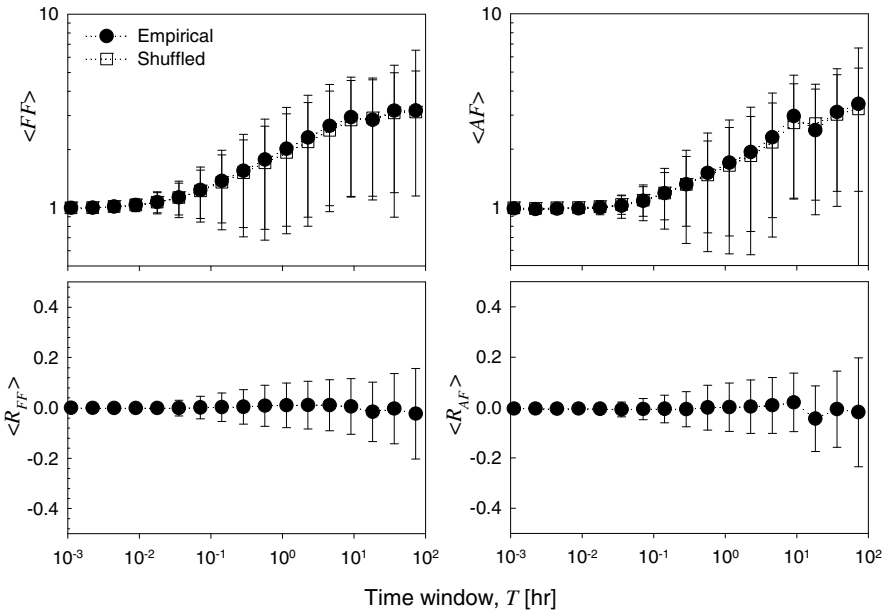


**Fig. 5.** Summary of the agreement between empirical and shuffled time series. The top panels show the mean Fano and Allan factors, averaged over all 394 users, and their standard deviations (whiskers), as a function of the time window length $T$. The bottom panels show the population average of the logarithmic residuals between the empirical and shuffled $FF$ and $AF$ curves.

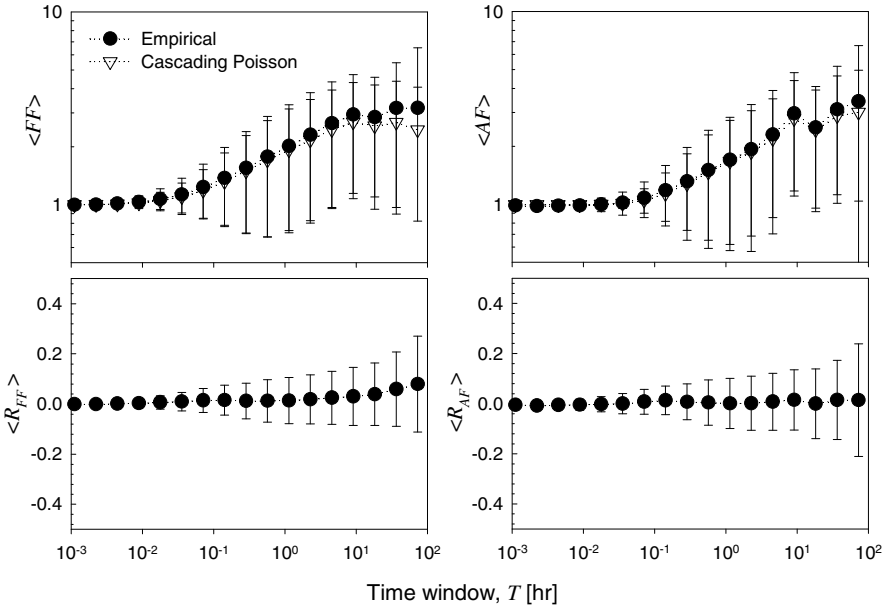


**Fig. 6.** Summary of the agreement between empirical and synthetic time series. The top panels show the mean Fano and Allan factors, averaged over all 394 users, and their standard deviations (whiskers), as a function of the time window length $T$. The bottom panels show the population average of the logarithmic residuals between the empirical and synthetic $FF$ and $AF$ curves.

rescaled inter-event time sequence exhibits $FF$ and $AF$ values close to unity for time window lengths up to about ten units in this homogenized time scale, in visual agreement with the results for a homogeneous Poisson process with unit rate.

We reiterate here that the rescaled results in Figure 4 need not be unitary. If there were non-Poissonian bursts in the empirical time series, we would observe systematic deviations from unity, suggesting that a cascading Poisson process is insufficient for capturing the correlations in User 467's e-mail correspondence. The fact that the $FF$ and $AF$ are unitary for the rescaled time series provides strong evidence that the "bursts" in activity are Poissonian in nature.

We now move beyond the analysis of User 467 to present the results of our analysis for all 394 users. In lieu of displaying these results for each of the 394 users, we summmarize these results by averaging the $FF$ and $AF$ curves for all 394 users under consideration (Figs. 5 and 6). The population-level correlations exhibit the same

basic patterns as User 467: for short time windows (up to a few minutes), the $FF$ and $AF$ are unitary, a fingerprint of a homogeneous Poisson process; and, for longer time windows, both exhibit the same deviation from unity as User 467. The visual agreement between the empirical and shuffled $FF$ and $AF$ curves (Fig. 5, top panels) indicates that these correlations are trivially related to the distribution of inter-event times, not to their particular ordering. Note that the large error bars are due to the large variation in human activity patterns, not a disagreement between the empirical and shuffled curves.

To account for the wide variation in human activity patterns, we compute the logarithmic residual $R_{X,i}(T) = \log_{10}[X_{e,i}(T)/\langle X_{s,i}(T)\rangle]$ between the empirical $X = FF, AF$ curve for user $i$ ($X_{e,i}(T)$) and the corresponding average shuffled $X = FF, AF$ curve for user $i$ ($\langle X_{s,i}(T)\rangle$). By examining the average logarithmic residuals across the population, we can detect whether there is any systematic deviation between the empirical and synthetic $FF$ and $AF$ curves at a particular time scale. The average logarithmic residual curves do not significantly deviate from zero, indicating that the empirical and shuffled $FF$ and $AF$ curves agree quite well (Fig. 5, bottom panels) and that the apparent correlations in the $FF$ and $AF$ curves are not due to the precise ordering of events.

The comparison between the empirical and synthetic $FF$ and $AF$ curves yields identical conclusions (Fig. 6). First, the cascading Poisson process visually captures the correlations quantified by the $FF$ and $AF$ curves. Second, we do not find any significant deviations between the empirical $FF$ and $AF$ curves and their synthetic counterparts at any time scale $T$. These results suggest that a cascading Poisson process is adequately capturing the higher-order statistical structure of human activity.

Finally, we rescale the empirical time series for each of the 394 users using equation (2) and we examine the mean $FF$ and $AF$ curves for these rescaled time series (Fig. 7). These rescaled $FF$ and $AF$ curves remain close to unity at all dimensionless time windows $\tilde{T}$, indicating that the rescaled $FF$ and $AF$ curves exhibit no more dispersity and correlations than expected from a homogeneous Poisson process.

Insofar as the Fano and Allan factors can detect burstiness in point process time series', our analysis conclusively demonstrates that e-mail communication patterns are no more bursty than expected from a suitably chosen Poisson process. We have demonstrated this in three complementary ways. First, e-mail communication patterns are no more bursty than expected by chance (Figs. 2 and 5), a necessary condition for a Poisson process [18]. Second, the burstiness in e-mail communication patterns is captured by a cascading Poisson process (Figs. 3 and 6), suggesting that this model adequately captures the dynamics. Third, the rescaled $FF$ and $AF$ curves do not exhibit any signs of burstiness, exactly as one would expect for a Poisson process.

Note that it did not have to be this way. Our analysis could have revealed non-Poissonian bursts in e-mail correspondence, suggesting that a cascading Poisson process is not sufficient to reproduce complex human activity patterns. The fact that we do not observe a bursty signature in e-mail correspondence offers strong evidence that human activity patterns are, in fact, Poissonian. Moreover, our analysis suggests that no additional mechanisms are necessary to describe e-mail activity patterns. E-mail activity is parsimoniously described by three simple mechanisms: people typically work during the same days of the week, people typically sleep during the same times of the day, and people typically continue to send e-mails once they have started sending e-mails.

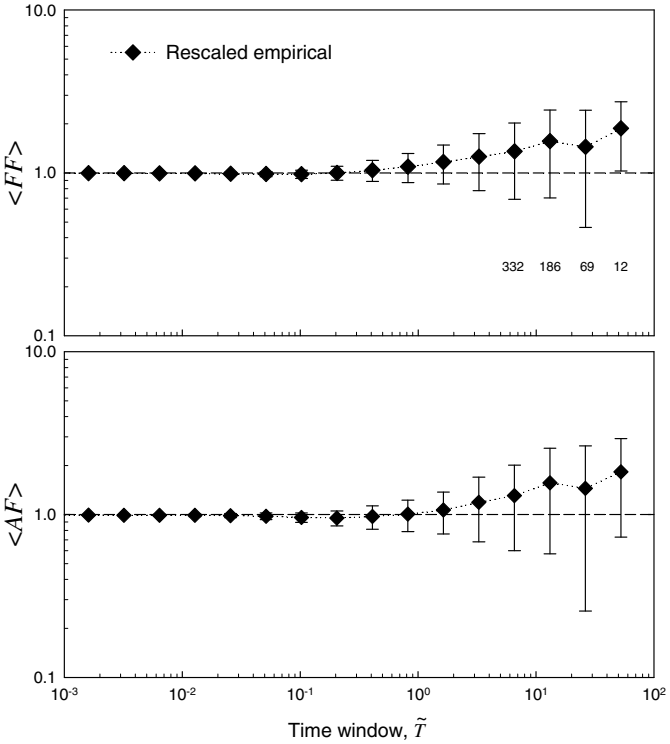


**Fig. 7.** Mean and standard deviation (whiskers) of Fano and Allan factors as a function of the (dimensionless) time window length $\tilde{T}$ for the rescaled time series, averaged over all 394 users. Dashed lines where drawn as a reference. Because the number of events (hence the average rate) is different for each user, not all contribute to the largest time windows. The number of timeseries that do contribute is indicated by the small numbers in the upper panel, when smaller than 394.

We are grateful to J.-P. Eckmann for providing the data. C.A. acknowledges Northwestern University for the kind hospitality and Brazilian agencies CNPq and Faperj for partial financial support, D.R.C. acknowledges support by NIH NINDS of USA (Grants NS58661).

## References

1. J.P. Eckmann, E. Moses, D. Sergi, Proc. Natl. Acad. Sci. USA **101**, 14333 (2004)
2. A.-L. Barabási, Nature **435**, 207 (2005)

3. A. Vazquez, J.G. Oliveira, Z. Dezso, K.I. Goh, I. Kondor, A.-L. Barabási, Phys. Rev. E **73**, 036127 (2006)
4. D.B. Stouffer, R.D. Malmgren, L.A.N. Amaral, preprint [arXiv:physics/0510216]
5. D.B. Stouffer, R.D. Malmgren, L.A.N. Amaral, preprint [arXiv:physics/0605027]
6. A.-L. Barabási, K.-I. Goh, A. Vazquez, preprint [arXiv:physics/0511186]
7. R.D. Malmgren, D.B. Stouffer, A.E. Motter, L.A.N. Amaral, Proc. Natl. Acad. Sci **105**, 18153 (2008)
8. H. Ebel, L.-I. Mielsch, S. Bornholdt, Phys. Rev. E **66**, 035103 (2002)
9. R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, A. Arenas, Phys. Rev. E **68**, 065103 (2003)
10. G. Kossinets, D.J. Watts, Science **311**, 88 (2006)
11. A. Vazquez, Physica A **373**, 747 (2007)
12. T. Zhou, X.P. Han, B.H. Wang, *Science Matters: Humanities as Complex Systems*, edited by M. Burguete, L. Lam (World Scientific Publishing, Singapore 2008) preprint [arXiv:physics/0801.1389v1]
13. R.D. Malmgren, J.M. Hofman, L.A.N. Amaral, D.J. Watts, *Proc. ACM SIGKDD*, 607 (2009)
14. R.D. Malmgren, D.B. Stouffer, A.S.L.O. Campanharo, L.A.N. Amaral, Science **325**, 1696 (2009)
15. F. Grüneis, M. Nakao, M. Yamamoto, Biol. Cybernetics **62**, 407 (1990)
16. F. Grüneis, Physica A **123**, 149 (1984)
17. F. Grüneis, H.J. Baiter, Physica A **136**, 432 (1986)
18. D.R. Cox, V. Isham, *Point Processes* (Chapman and Hall, London, 1980)
19. S. Thurner, S.B. Lowen, M.C. Feurstein, C. Heneghan, H.G. Feichtinger, M.C. Teich, Fractals **5**, 565 (1997)
20. C. Anteneodo, D.R. Chialvo, Chaos **19**, 033123 (2009)