A Letters Journal Exploring
the Frontiers of Physics

OFFPRINT

Burstiness and memory in complex systems

K.-I. Goh and A.-L. Barabási

EPL, **81** (2008) 48002

A Letters Journal Exploring
the Frontiers of Physics

February 2008

EPL, **81** (2008) 48002
doi: 10.1209/0295-5075/81/48002

www.epljournal.org

Burstiness and memory in complex systems

K.-I. Goh[1,2] and A.-L. Barabási[1,3]

[1] *Center for Complex Network Research and Department of Physics, University of Notre Dame*
Notre Dame, IN 46556, USA
[2] *Department of Physics, Korea University - Seoul 136-713, Korea*
[3] *Department of Physics, Biology, and Computer Science, Northeastern University - Boston, MA 02115, USA*

received 9 September 2007; accepted in final form 13 December 2007
published online 17 January 2008

PACS 89.75.-k – Complex systems
PACS 05.45.Tp – Time series analysis

Abstract – The dynamics of a wide range of real systems, from email patterns to earthquakes, display a bursty, intermittent nature, characterized by short timeframes of intense activity followed by long times of no or reduced activity. The understanding of the origin of such bursty patterns is hindered by the lack of tools to compare different systems using a common framework. Here we propose to characterize the bursty nature of real signals using orthogonal measures quantifying two distinct mechanisms leading to burstiness: the interevent time distribution and the memory. We find that while the burstiness of natural phenomena is rooted in both the interevent time distribution and memory, for human dynamics memory is weak, and the bursty character is due to the changes in the interevent time distribution. Finally, we show that current models lack in their ability to reproduce the activity pattern observed in real systems, opening up avenues for future work.

The dynamics of most complex systems is driven by the loosely coordinated activity of a large number of components, such as individuals in the society or molecules in the cell. While we witnessed much progress in the study of the networks behind these systems [1–4], advances towards understanding the system's dynamics have been slower. With increasing potential to monitor the time-resolved activity of most components of selected complex systems, such as time-resolved email [5–7], web browsing [8], and gene expression [9,10] patterns, we have the opportunity to ask an important question: is the dynamics of complex systems governed by generic organizing principles, or each system has its own distinct dynamical features? While it is difficult to offer a definite answer to this question, a common feature across many systems is increasingly documented: the burstiness of the system's activity patterns.

Bursts, vaguely corresponding to significantly enhanced activity levels over short periods of time followed by long periods of inactivity, have been observed in a wide range of systems, from email patterns [6] to earthquakes [11,12] and gene expression [9]. Yet, often a burstiness is more of a metaphor than a quantitative feature, and opinions about its origin diverge. In human dynamics, burstiness has been reduced to the fat-tailed nature of the response time distribution [6,7], in contrast with earthquakes and weather patterns, where memory effects appear to play a key role as well [13,14]. Once present, burstiness can affect the spreading of viruses [15] or resource allocation [16,17]. Also, deviations towards a regular, "anti-bursty" behavior in heartbeat indicate disease progression [18]. Given the diversity of systems in which it emerges, there is a need to place burstiness on a firmer quantitative basis. Our goal in this letter is to make a step in this direction, by developing a diagnosis tool that can help quantify the magnitude and potential origin of the bursty patterns seen in different real systems. Such a tool may also lend insights into the analysis of fractal and self-similar bursty signals [19].

Let us consider a system whose components have a measurable activity pattern that can be mapped into a discrete signal, recording the moments when some events take place, like an email being sent, or a protein being translated[1]. The activity pattern is random (Poisson process) if the probability of an event is time-independent. In this case the interevent time, τ, between two consecutive events follows an exponential distribution, $P_{\mathrm{P}}(\tau) \sim \exp(-\tau/\tau_0)$ (fig. 1a). An apparently bursty

[1] For systems with continuous signal, we may adopt a threshold method to transform it into a discrete one, and in many systems the statistical properties of the obtained signal are known to be threshold-independent [12,13].

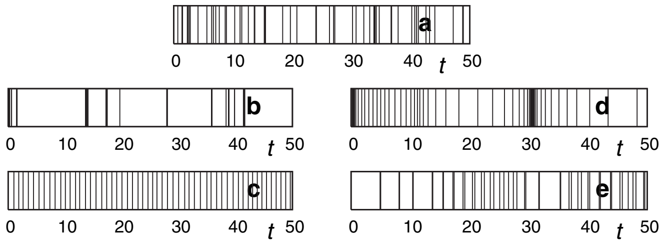

Fig. 1: (a) A signal generated by a Poisson process with a unit rate ($B=-0.05$, $M=0.02$; see text for the definitions of B and M parameters). (b,c) Bursty character through the interevent time distribution: A bursty signal ($B=0.44$, $M=-0.04$) generated by the power law interevent time distribution $P(\tau)\sim\tau^{-1}$ (b), and an anti-bursty signal ($B=-0.81$, $M=-0.02$) generated by the Gaussian interevent time distribution with $m=1$ and $\sigma=0.1$ (c). A bursty-looking signal can emerge through memory as well. For example, the bursty-looking signal shown in (d) ($M=0.90$) is obtained by shuffling the Poisson signal of (a) to increase the memory effect. A more regular looking signal, with negative memory, is obtained by the same shuffling procedure (e) ($M=-0.74$). Note that signals in (a), (d) and (e) have identical interevent time distribution, thus the same B-value.

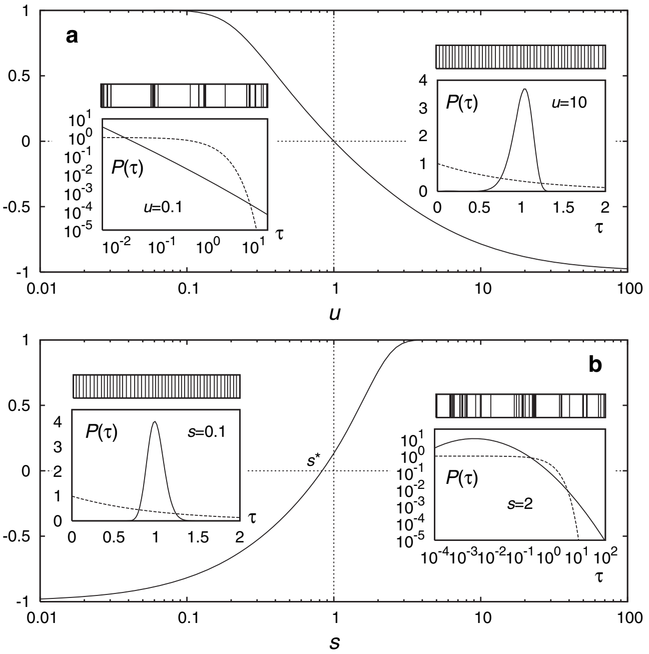

Fig. 2: The burstiness parameter B for (a) the stretched exponential and (b) log-normal interevent time distributions. Both distributions interpolate between a highly bursty ($B=1$), neutral ($B=0$), and a regular ($B=-1$) signal. Insets show the form of $P(\tau)$ in bursty and anti-bursty regime of each distribution along with a typical time signal generated with the corresponding $P(\tau)$. The dashed line in the insets refers the exponential distribution for the Poisson process.

(or anti-bursty) signal emerges if $P(\tau)$ is different from the exponential, such as the bursty pattern of fig. 1b, or the more regular pattern of fig. 1c. Yet, the change in the interevent time distribution is not the only way to generate a bursty signal. For example, the signals shown in fig. 1d,e have exactly the same $P(\tau)$ as in fig. 1a, yet they have a more bursty or a more regular character. This is achieved by introducing memory: in fig. 1d the short interevent times tend to follow short ones, resulting in a bursty look. In fig. 1e the relative regularity is due to the memory effect acting in the opposite direction: short (long) interevent times tend to be followed by long (short) ones. Therefore, the apparent burstiness of a signal can be rooted in two mechanistically different deviations from a Poisson process: changes in the interevent time distribution or memory. To distinguish these orthogonal effects, we consider two measures, the burstiness parameter B based on the interevent time distribution and the memory parameter M based on the interevent time correlations, that quantify the degree of each effect in real signals.

Distribution-based measure. – We may characterize the deviation from the Poisson signal in several ways. Perhaps the simplest measure in the literature would be the so-called coefficient of variation, defined as the ratio of the standard deviation to the mean, σ_τ/m_τ, where m_τ and σ_τ are the mean and the standard deviation of $P(\tau)$, respectively. It has a value 1 for a Poisson signal with the exponential $P(\tau)$, 0 for completely regular δ function-like $P(\tau)$, and ∞ for signals with a heavy-tailed $P(\tau)$ with infinite variance. Higher moments of the distribution such as skewness or kurtosis, or a more complicated measure based on the area between $P(\tau)$ and the exponential or that between its cumulative functions may also be used for this purpose. Here we use the coefficient of variation to define a burstiness parameter B as

$$B \equiv \frac{(\sigma_\tau/m_\tau - 1)}{(\sigma\tau/m_\tau + 1)} = \frac{(\sigma_\tau - m_\tau)}{(\sigma_\tau + m_\tau)}\ . \tag{1}$$

This definition is meaningful when both the mean and the standard deviation of $P(\tau)$ exist, which is always the case for real-world finite signals. When meaningful, B has a value in the bounded range $(-1,1)$, and its magnitude correlates with the signal's burstiness: $B=1$ is the most bursty signal, $B=0$ is neutral, and $B=-1$ corresponds to a completely regular (periodic) signal. For example, in fig. 2a we show B for the stretched exponential distribution,

$$P_{\mathrm{SE}}(\tau) = u(\tau/\tau_0)^{u-1}\exp[-(\tau/\tau_0)^u]/\tau_0\ , \tag{2}$$

often used to approximate the interevent time distributions of complex systems [20]. Here the smaller the parameter u is, the burstier is the signal, and for $u\to 0$, $P_{\mathrm{SE}}(\tau)$ follows a power law with the exponent -1, for which $B=1$. For $u=1$, $P_{\mathrm{SE}}(\tau)$ is simply the exponential distribution with $B=0$. Finally, for $u>1$, the larger u is, the more regular is the signal, and for $u\to\infty$, $P(\tau)$ converges to a

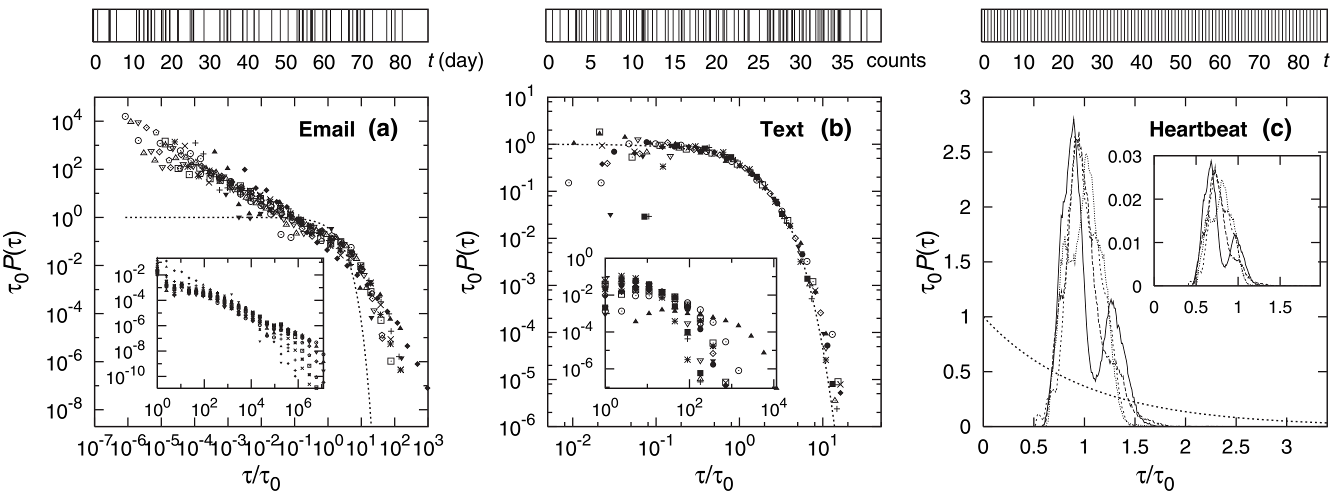

Fig. 3: Interevent time distributions $P(\tau)$ for some real signals. (a) $P(\tau)$ for the email activity of individuals from a University [4]. τ corresponds to the time interval between two emails sent by the same user. (b) $P(\tau)$ for the occurrence of a letter in the text of C. Dickens' *David Copperfield* [25]. Here τ corresponds to the number of letters between two consecutive occurrences of the same alphabet letter. (c) $P(\tau)$ of the cardiac rhythm of individuals [30]. Each event corresponds to the beat in the heartbeat signal and τ is the time interval between two consecutive heartbeats for an individual. In each panel, we also show for reference the exponential interevent time distribution (dotted). Unscaled interevent time distributions are shown in the inset for each dataset.

Dirac delta function with $B=-1$. We also show in fig. 2b the behavior of B for the log-normal distribution,

$$P_{\mathrm{LN}}(\tau)=\frac{1}{\tau s\sqrt{2\pi}}\exp\biggl(-\frac{[\ln(\tau)-\mu]^2}{2s^2}\biggr), \qquad (3)$$

also frequently used for the statistics of complex systems [21,22]. Here the larger the parameter s is, the larger is the variance of $P(\tau)$ hence the signal gets burstier ($B\to 1$). The smaller s is, the more regular is the signal, and B approaches to -1 as $s\to 0$. We note however that even though B becomes zero for a specific value s^* (fig. 2b), $P(\tau)$ does not become an exponential there, which is a caveat of the present measure.

Most complex systems display a remarkable heterogeneity: some components may be very active, and others much less so. For example, some users may send dozens of emails during a day, while others only one or two. To combine the activity levels of so different components, we can group the signals based on their average activity level, and measure $P(\tau)$ only for components with similar activity level. As the insets in fig. 3 show, the obtained curves are systematically shifted. If we plot, however, $\tau_0 P(\tau)$ as a function of τ/τ_0, where τ_0 is the average interevent time, the data collapse into a single curve $\mathcal{F}(x)$ (fig. 3), indicating that the interevent time distribution follows $P(\tau)=(1/\tau_0)\mathcal{F}(\tau/\tau_0)$, where $\mathcal{F}(x)$ is independent of the average activity level of the component, and represents a universal characteristic of the particular system [12,23,24]. This raises an important question: will B depend on τ_0? The burstiness parameter B is indeed invariant under the time rescaling as $\tilde{\tau}\equiv\tau/\tau_0$ and $\tilde{P}(\tilde{\tau})\equiv\tau_0 P(\tau)$ with a constant τ_0. Such an invariance enables us to assign to each system a characteristic burstiness parameter, despite the different activity level of its components. The scaling in fig. 3 could be a starting point of further theoretical work, aiming to answer how generic it is and what is the mechanism behind it. Currently, we have only partial answer to these questions for specific systems [23].

Correlation-based measure. – The way we can characterize the correlation properties of a signal is not unique either. The joint probability distribution parameterized by a time lag k, $P(\tau,\tau';k)$, defined as the probability density that we have two interevent times τ and τ' separated by k events, contains the most information about the two-point correlation properties. The autocorrelation function $C(k)=\langle(\tau_i-m_\tau)(\tau_{i+k}-m_\tau)\rangle/\sigma_\tau^2$, where $\langle\cdot\rangle$ means the average over the index i, is also widely used in many applications. A simple measure is offered by the correlation coefficient of consecutive interevent time values (τ_i,τ_{i+1}), defining the memory coefficient M as

$$M\equiv\frac{1}{n_\tau-1}\sum_{i=1}^{n_\tau-1}\frac{(\tau_i-m_1)(\tau_{i+1}-m_2)}{\sigma_1\sigma_2}\ , \qquad (4)$$

where n_τ is the number of interevent times measured from the signal and $m_1(m_2)$ and $\sigma_1(\sigma_2)$ are sample mean and sample standard deviation of τ_i's (τ_{i+1}'s), respectively ($i=1,\ldots,n_\tau-1$). Note that M is a biased estimator for $C(k=1)$, which is more appropriate for real-world finite signals, particularly if there are possible long-range correlations in the system. With this definition, the memory coefficient has a value in the range $(-1,1)$ and is positive when a short (long)

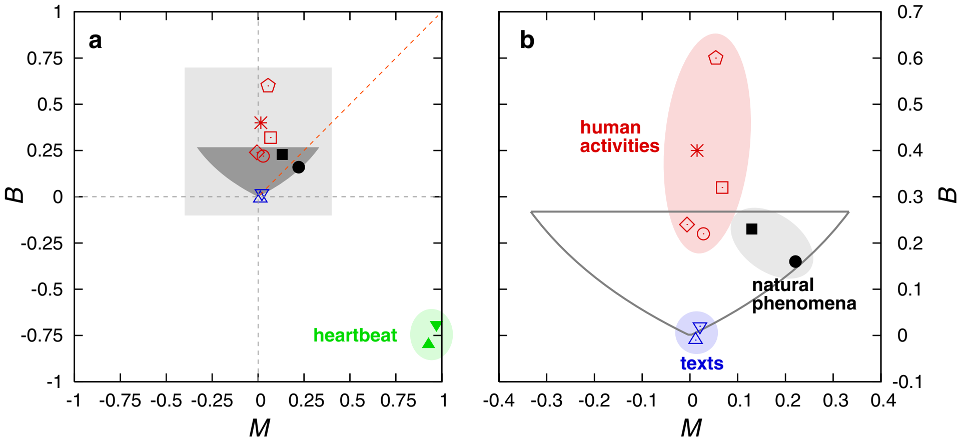

Fig. 4: (Color online) (a) The (M, B) phase diagram. Human activities (red) are captured by activity patterns pertaining to email ($\star$) [5], library loans ($\circ$) [7], and printing (⬠) [28] of individuals in Universities, call center record at an anonymous bank (□) [29], and phone initiation record from a mobile phone company ($\diamond$). Data for natural phenomena (black) are earthquake records in Japan ($\bullet$) [26] and daily precipitation records in New Mexico, USA (■) [27]. Data for written texts (blue) [25] are the English text of *David Copperfield* ($\triangle$) and the Hungarian text of *Isten Rabjai* by Gárdonyi Géza ($\triangledown$). Data for physiological behaviors (green) are the normal sinus rhythm (▼) and the cardiac rhythm with CHF (▲) of human subjects [30]. The dark-grey area is the region occupied by the 2-state model [34]. (b) Close-up of the most populated region (light-grey region in (a)). Data in each class are indicated by grouping with the respective dimmer color for the eye.

interevent time tends to be followed by a short (long) one, and it is negative when a short (long) interevent time is likely to be followed by a long (short) one. For example, the synthetic signals shown in figs. 1(a,d,e) with identical $P(\tau)$ have the memory coefficient $M = 0.02$ (neutral; a), $M = 0.90$ (positive memory; d) and $M = -0.74$ (negative memory; e), respectively.

Mapping complex systems on the (M, B)-space. – Given that the burstiness of a signal can have two qualitatively different origins, it is desirable to characterize real-world complex systems by quantifying both effects, using the corresponding B and M parameters to place them in a (M, B)-space (fig. 4). As a first example, we measured the spacing between the consecutive occurrences of the same letter in written texts of different kind, era, and language [25]. For these signals, we find $B \approx 0$, *i.e.*, the interevent time distribution follows closely an exponential (fig. 3b) and $M \approx 0.01$, indicating the lack of short-term memory. Thus, this signal is at the origin of the phase diagram (fig. 4). In contrast, natural phenomena, like earthquakes [26] and weather patterns [27] are in the vicinity of the diagonal, indicating that $P(\tau)$ and memory equally contribute to their bursty character. The situation is quite different, however, for human activities, ranging from email and phone communication to web browsing and library visitation patterns [5,7,8,28,29]. For these we find a high B and small or negligible M, indicating that while these systems display significant burstiness rooted in $P(\tau)$, memory plays a small role in their temporal inhomogeneity. This lack of memory is quite unexpected, as it suggests the lack of predictability in these systems in contrast with natural phenomena, where strong memory effects could lead to predictive tools. Finally, for cardiac rhythms describing the time interval between two consecutive heartbeats (fig. 3c) [30], we find $B_{\text{healthy}} = -0.69(6)$ for healthy individuals and $B_{\text{CHF}} = -0.8(1)$ for patients with congestive heart failure (CHF), both signals being highly regular. Thus the B parameter captures the fact that cardiac rhythm is more regular with CHF than in the healthy condition [18]. Furthermore, we find $M \approx 0.97$, indicating that memory also plays an important role in the signal's regularity.

The discriminative nature of the (M, B) phase diagram is illustrated by the clustering of the different systems in the plane: human-activity patterns locate themselves in the high-B, low-M region, natural phenomena near the diagonal, heartbeats in the high-M, negative-B region and texts near the origin, suggesting the existence of distinct classes of dynamical mechanisms driving the temporal activity in these systems. It will also be interesting to study how chaotic (real or model-generated) signals are placed in the (M, B)-plane, and whether there exist clear boundaries in the phase diagram separating systems into distinct classes.

Discussion. – Following the clustering of the empirical measurements in the phase diagram, a natural question emerges: to what degree can current models reproduce the observed quantitative features of bursty processes? Queueing models, proposed to capture human-activity patterns, are designed to capture the waiting times of the tasks, rather than interevent times [6,7,31–33]. Therefore, placing them on the phase diagram is not meaningful. A bursty signal can be generated by the 2-state model [34]. The 2-state model is a probabilistic automaton with two internal states q_0 and q_1, in each of which the system performs a Poisson process with rates λ_0 and λ_1, respectively. Each time the system switches its state (changes λ) with probability p, or remains in its current state with probability $(1-p)$. Thus the system alternates between two Poisson processes randomly, generating a bursty signal when $\lambda_0 \neq \lambda_1$. The B and M parameters for the 2-state model can be calculated analytically [35]. The region in the (M, B)-space occupied by the 2-state model with different λ rates and switching probability p is shown as the dark-grey area in fig. 4a, suggesting that the model could account for some of the observed behaviors by tuning its parameters. Yet, the agreement is misleading: for example, $P(\tau)$ of real bursty systems is often skewed and fat-tailed, which is not the case for the 2-state model for which we have the sum of two exponentials. This indicates that B and M parameters offer only a first-order discrimination for the origin of the burstiness. More sophisticated measures are needed to improve the comparison between models and real systems by, *e.g.*, using the full functional form of $P(\tau)$ and the autocorrelation function, or by developing measures to capture long-term correlations and non-linear effects present in real systems, such as those exhibiting self-organized criticality or chaotic behavior [36,37]. These topics deserve further investigation. This discrepancy also indicates the lack of satisfactory modeling tools to capture the detailed mechanisms responsible for the bursty activities seen in real complex systems, opening up possibilities for future work.

* * *

We would like to thank S. Havlin and A. Vázquez for helpful discussions, and an anonymous reviewer for comments on the manuscript. This work is supported by the S. McDonnell Foundation and the National Science Foundation under Grant No. CNS-0540348 and ITR DMR-0426737. K-IG is also supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2007-331-C00111).

REFERENCES

[1] Albert R. and Barabási A.-L., *Rev. Mod. Phys.*, **74** (2001) 47.

[2] Boccaletti S., Latora V., Moreno Y., Chavez M. and Hwang D.-U., *Phys. Rep.*, **424** (2006) 175.

[3] Newman M. E. J., Barabási A.-L. and Watts D. J. (Editors), *Structure and Dynamics of Complex Networks* (Princeton University Press, Princeton) 2006.

[4] Caldarelli G., *Scale-free networks* (Oxford University Press, Oxford) 2007.

[5] Eckmann J. P., Moses E. and Sergi D., *Proc. Natl. Acad. Sci. U.S.A.*, **101** (2004) 14333.

[6] Barabási A.-L., *Nature*, **207** (2005) 435.

[7] Vázquez A., Oliveira J. G., Dezső Z., Goh K.-I., Kondor I. and Barabási A.-L., *Phys. Rev. E*, **73** (2006) 036127.

[8] Dezső Z., Almaas E., Lukacs A., Racz B., Szakadat I. and Barabási A.-L., *Phys. Rev. E*, **73** (2006) 066132.

[9] Golding I., Paulsson J., Zawilski S. M. and Cox E. C., *Cell*, **123** (2005) 1025.

[10] Chubb J. R., Trcek T., Shenoy S. M. and Singer R. H., *Curr. Biol.*, **16** (2006) 1018.

[11] Bak P., Christensen K., Danon L. and Scanlon T., *Phys. Rev. Lett.*, **88** (2002) 178501.

[12] Corral A., *Phys. Rev. E*, **68** (2003) 035102(R).

[13] Bunde A., Eichner J. F., Kantelhardt J. W. and Havlin S., *Phys. Rev. Lett.*, **94** (2005) 048701.

[14] Livina V. N., Havlin S. and Bunde A., *Phys. Rev. Lett.*, **95** (2005) 208501.

[15] Vázquez A., Rácz B., Lukács A. and Barabási A.-L., *Phys. Rev. Lett.*, **98** (2007) 158702.

[16] Leland W. E., Taqqu M. S., Willinger W. and Wilson D. V., *IEEE/ACM Trans. Netw.*, **2** (1994) 1.

[17] Paxson V. and Floyd S., *IEEE/ACM Trans. Netw.*, **3** (1995) 226.

[18] Thurner S., Feurstein M. C. and Teich M. C., *Phys. Rev. Lett.*, **80** (1998) 1544.

[19] Lowen S. B. and Teich M. C., *Fractal-based point processes* (Wiley, Hoboken, NJ) 2005.

[20] Laherrère J. and Sornette D., *Eur. Phys. J. B*, **2** (1998) 525.

[21] Mitzenmacher M., *Internet Math.*, **1** (2004) 226.

[22] Stouffer D. B., Malmgren R. D. and Amaral L. A. N., e-print physics/0605027v1.

[23] Saichev A. and Sornette D., *Phys. Rev. Lett.*, **97** (2006) 078501.

[24] Candia J., González M. C., Wang P., Schoenharl T., Madey G. and Barabási A.-L., to be published in *J. Phys. A*, arXiv:0710.2939v2.

[25] Project Gutenberg, `http://gutenberg.org`.

[26] Japan University Network Earthquake Catalog, `http://wwweic.eri.u-okyo.ac.jp/CATALOG/junec/`.

[27] National Resources Conservation Service, `http://www.nm.nrcs.usda.gov/Snow/data/historic.htm`.

[28] Harder U. and Paczuski M., *Physica A*, **361** (2006) 329.

[29] Guedj I. and Mandelbaum A., `http://iew3.technion.ac.il/serveng/callcenterdata/`.

[30] PhysioBank, `http://www.physionet.org/physiobank/`.

[31] Oliveira J. G. and Barabási A.-L., *Nature*, **437** (2005) 1251.

[32] Vázquez A., *Phys. Rev. Lett.*, **95** (2005) 248701.

[33] Gabrielli A. and Caldarelli G., *Phys. Rev. Lett.*, **98** (2007) 208701.

[34] Kleinberg J., in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery Data Mining* (2002) pp. 91.

[35] Goh K.-I. *et al.*, unpublished.

[36] Bak P., Tang C. and Wiesenfeld K., *Phys. Rev. A*, **38** (1988) 364.

[37] Abarbanel H. D., Brown R., Sidorowich J. J. and Tsimring L. Sh., *Rev. Mod. Phys.*, **65** (1993) 1331.