A Letters Journal Exploring the Frontiers of Physics

November 2012

EPL, **100** (2012) 48004
doi: 10.1209/0295-5075/100/48004

www.epljournal.org

Scaling behavior of online human activity

Zhi-Dan Zhao[1], Shi-Min Cai[1(a)], Junming Huang[2], Yan Fu[1] and Tao Zhou[1]

[1] *Web Sciences Center, School of Computer Science and Engineering, University of Electronic Science and Technology of China - Chengdu 610054, PRC*
[2] *Institute of Computing Technology, Chinese Academy of Sciences - Beijing, 100190, PRC*

received 14 July 2012; accepted in final form 9 November 2012
published online 3 December 2012

PACS 89.75.Da – Systems obeying scaling laws
PACS 05.45.Tp – Time series analysis
PACS 89.65.-s – Social and economic systems

Abstract – The rapid development of the Internet technology enables humans to explore the web and record the traces of online activities. From the analysis of these large-scale data sets (*i.e.*, traces), we can get insights about the dynamic behavior of human activity. In this letter, the scaling behavior and complexity of human activity in the e-commerce, such as music, books, and movies rating, are comprehensively investigated by using the detrended fluctuation analysis technique and the multiscale entropy method. Firstly, the interevent time series of rating behaviors of these three types of media show similar scaling properties with exponents ranging from 0.53 to 0.58, which implies that the collective behaviors of rating media follow a process embodying self-similarity and long-range correlation. Meanwhile, by dividing the users into three groups based on their activities (*i.e.*, rating per unit time), we find that the scaling exponents of the interevent time series in the three groups are different. Hence, these results suggest that a stronger long-range correlations exist in these collective behaviors. Furthermore, their information complexities vary in the three groups. To explain the differences of the collective behaviors restricted to the three groups, we study the dynamic behavior of human activity at the individual level, and find that the dynamic behaviors of a few users have extremely small scaling exponents associated with long-range anticorrelations. By comparing the interevent time distributions of four representative users, we can find that the bimodal distributions may bring forth the extraordinary scaling behaviors. These results of the analysis of the online human activity in the e-commerce may not only provide insight into its dynamic behaviors but may also be applied to acquire potential economic interest.

Copyright © EPLA, 2012

Introduction. – The behavior involving the human daily activities is one of the highest complexity and most complicated things because it is driven by countless unknown facts. Mining the human dynamics from these recorded large-scale data sets has become much more important for understanding their behavior patterns, and modeling the human dynamic behaviors helps us to explain many socioeconomic phenomena and find significant applications ranging from resource allocation, transportation control, epidemic prediction to personalized recommendation [1,2]. Thanks to the development of the information technology, massive Internet data and resources make us easily realize the empirical analysis and modeling of human activity. One of the most attractive observation is the heavy-tailed nature of the interevent time distribution, which implies that the bursts of rapidly occurring events are separated by long periods of inactivity. Examples of empirical studies include the email communication [3], the surface mail communication [4], the cell-phone communication [5], the online activities [6–10], and so on. To understand the heavy-tailed phenomena of the human dynamic behavior, the experts have put forward many mechanisms, such as the highest-priority-first queue model [3], the varying interest [11], the memory effects [12] and the human interactions [13–15] to mimic the temporal bursts.

On the other hand, the techniques of time series analysis are applied to investigate the evolutional data in the real world. One of the most popular techniques is the detrended fluctuation analysis (DFA) proposed by Peng *et al.* [16,17], which can effectively quantify the long-range power-law correlations embedded in the nonstationary time series (or self-similarity process). It provides a simple scaling exponent α to represent the correlation

(a)E-mail: shimin.cai81@gmail.com

characters of the time series, and thus is applied to various research fields including heart rate dynamics [17–21], financial time series [22–26], particle condensation [27], Internet traffic [28], musical rhythm spectra [29], etc. In recent times Rybski *et al.* [30–32] have applied the DFA to study the long-term correlations of the communication patterns (*i.e.*, interactive activities) in online social networks, which are associated with the source of general Gibrat's law in economics. Meanwhile, Costa *et al.* [33] recently proposed a technique, namely multiscale entropy (MSE), to quantify the information complexity of physiologic time series over multiple scale. They further used the MSE to analyze the human heartbeat [34], which suggested that the time asymmetry is a fundamental property of healthy. In the following works, the MSE was widely applied to analyze the EEG signals, which indicated that the human brain variability increases with maturation [35] and the Alzheimer disease patients usually had lower sample entropy on the small and medium time scales [36]. In the environmental field, Li and Zhang [37] analyzed the long-term daily flow rates of the Mississippi River, and found that the sample entropy for flow rates generally monotonously increases with scale factor and the complexity was beginning to decrease since 1940. Moreover, based on the interevent time distributions and memory, Goh *et al.* [38] used the orthogonal measures to quantify the burstiness in many real interevent time series, and found that the origin of burstiness in human activity was much more correlated with the changes in the interevent time distributions.

The e-commerce is composed of the online business trades among humans and rating information based on the Internet technology, in which the dynamic behaviors of human activity (*i.e.*, trading or rating records) are involved with a large amount of useful knowledge for acquiring potential economic interest. In this letter, we first focus on the interevent intervals (*i.e.*, time series) of the human activity in e-commerce, and empirically investigate its scaling properties and information complexities both in the collective and individual levels based on the measures including the DFA, MSE and interevent time distribution. The rich results include the following: 1) The interevent time series of rating behaviors restricted to the types of media show similar scaling properties which implies that the collective behaviors of the rating media follow a process embodying self-similarity and long-range correlation. 2) Different scaling exponents of the interevent time series can be observed from the collective behaviors restricted to the users' activities (*i.e.*, rating per unit time), yet they suggest stronger long-range correlations existing in these collective behaviors. 3) The information complexities of collective behaviors are obviously distinguishable from the users' activities. 4) The extremely small scaling exponents (indicating long-range anticorrelation) of a few representative individual users are mainly brought by the bimodal interevent time distributions.

Materials. – The experimental data set is randomly sampled from Douban, which is a companion of e-commerce. It is similar to the Social Networking Services (SNS) that allows registered users to record information and create content related to movies, books, and music, yet it also can make a personalized recommendation for the registered users. We focus on users who perform more than 1000 rating actions on all the three types of media, which results in a set of 65 individuals. In the data set, we can find series of important history knowledge of registered users, such as user ID, item ID, rate, time stamp, item type, etc. Note that the sample time resolution is second, and we here focus on the interevent time series defined as the intervals between two consecutive rating actions.

Methods. – Here we apply the DFA and MSE methods to quantitatively understand the scaling behavior and complexity of human activity in e-commerce. In order to keep our description as self-contained as possible, we should introduce the DFA and MSE methods briefly.

Detrended fluctuation analysis. – We describe the process of the DFA which involves the following steps [16,39,40].

i) Starting with a time series $t(i)$, where $i \in [1, N]$ and N is the length of the series, we first integrate the series $t(i)$ and obtain $y(k) \equiv \sum_{i=1}^{k} [t(i) - \langle t \rangle]$, where $\langle t \rangle$ is the mean. Meanwhile, $y(k)$ is divided into N/l nonoverlapping boxes, each containing l interevent intervals.

ii) In the k-th box, we use a polynomial function $y_l(k)$ of order n to represents the local trend. In the experiments, the order is selected as $n = 2$, and the algorithm is denoted as DFA-2.

iii) We calculate the variance of the residual time series after the detrending procedure,

$$F(l) \equiv \sqrt{\frac{1}{N} \sum_{k=1}^{N} [y(k) - y_l(k)]^2}. \tag{1}$$

iv) Altering the box size l and repeating the detrending procedure, we can obtain the variances $F(l)$ as a function of box size l. A power-law relations between $F(l)$ and l is $F(l) \sim l^{\alpha}$. α is a real value in the bounded range from 0 to 1, where $\alpha > 0.5$ means that the time series is correlated, $\alpha = 0.5$ suggests that the time series is the same as the white noise (*i.e.*, no correlation), and $\alpha < 0.5$ indicates that the time series is anticorrelated.

Multiscale entropy. – We look back over the MSE method. The MSE is based on the simple observation that the complex signals generally exhibit the dynamics deviating far from perfect regularity and their multiscale complexity. The procedure of MSE is described as follows [33]:

i) For a given time series, $x_1, x_2, \ldots, x_N$, where N is the total number of time series. we divide it into nonoverlapping boxes with length l.

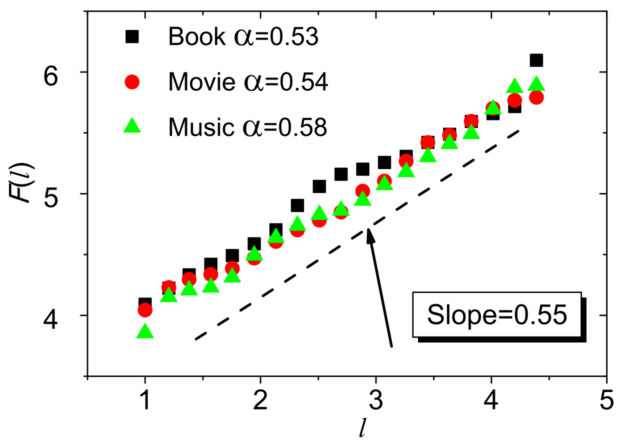


Fig. 1: (Color online) The results of the interevent time series measured by DFA. The symbols represent the types of media, books (black squares), movies (red circles), and music (green triangles). The dashed line represents a guide for the eyes. Similar scaling behaviors suggest that the human activity in systems follows a long-correlated self-similar process.

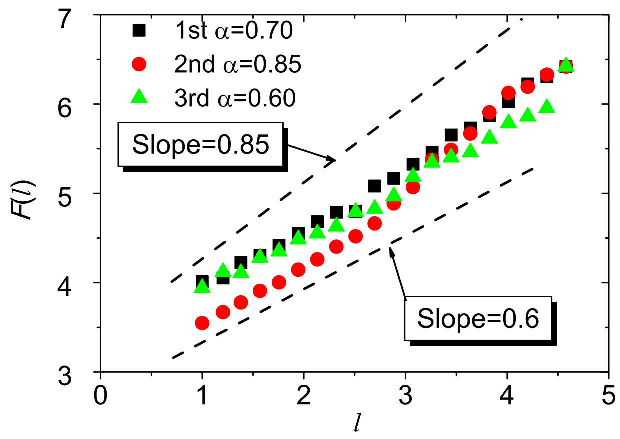


Fig. 2: (Color online) The results of the interevent time series measured by DFA. The interevent time series are constructed from the intervals between consecutive actions in the three groups, 1st (black squares), 2nd (red circles), and 3rd (green triangles), respectively. The scaling exponents are 0.7, 0.85, and 0.65 corresponding to 1st, 2nd, and 3rd groups, respectively.

ii) The averages of the time series inside each box are deemed as the elements of a coarse-grained time series,

$$s_j^l = \frac{1}{l}\sum_{i=(j-1)l+1}^{jl} x_i, \quad 1 \leqslant j \leqslant N/l. \tag{2}$$

By altering the box length l, we can obtain many coarse-grained time series, which characterize the original time series at multiple scales.

iii) The sample entropy [41] is used to measure each coarse-grained time series. Thus, we can find the relation between the entropy measure and the scale factor (*i.e.*, the box length l).

Collective level. – The time stamp of the data set has 1 second precision. Our focus is the interevent interval τ between consecutive actions, *i.e.*, rating a media by a certain user in Douban. The interevent time series are composed of these intervals in many ways. From the view of the whole system, we first investigate the interevent time series restricted to the types of medias. Figure 1 shows the results of the interevent time series measured by DFA, in which we observe that the scaling exponent α fluctuates in the small interval $[0.53, 0.58]$. The values of α reveal similar scaling behaviors for all types of medias and suggest that the interevent time series of rating media in the system evolves a process embodying self-similarity and long-range correlation. Furthermore, scaling exponents slightly greater than 0.5 also imply the weak memory of the signals, which is consistent with the previous results found in other human activities [38].

Although the human activity in the whole system approximately obeys a common scaling law, we should pay attention to the effect of the individual user activity on their scaling behaviors because the user activity is strongly associated with a good understanding of human dynamics [7,10]. The activity of an arbitrary user i, is defined as $A_i = n_i/T_i$, where n_i is the number of actions and T_i is the time difference between the first and last action [42]. We sort these users in increasing order according to their activities, and then divide them into three groups which are indicated by low (1st), mid (2nd) and high (3rd) activity. The number of users in each group is 21 (1st), 22 (2nd), and 22 (3rd). The interevent time series are constructed from the intervals between consecutive actions done by users in the three groups, respectively. Their results measured by DFA are shown in fig. 2, which suggests that the scaling behaviors are different in the three groups. For the 2nd group, the scaling exponent $\alpha = 0.85$ indicates that the interevent time series opposes much stronger long-range correlations than those of the other two groups. However, all the scaling exponents are much larger than 0.5, which suggests that long-range correlations generally exist in the interevent time series regardless of their activities.

Additionally, to verify memory effects of online human activities does not arise from the power-law distributions of the interevent time series, we reshuffle the original interevent time series to disturb their long-range correlations, and measure these shuffled ones by the DFA method. In table 1, we show the remarkable differences between the scaling exponents of the original and shuffled time series, which suggest that memory effects generally exist in online human activities.

To further determine the differences among the three groups, we use the MSE method to quantify the information complexities of the interevent time series. There are two guidelines, the higher information complexity is in correspondence with the larger sample entropy and the monotonic increase of the sample entropy indicates much more information of the interevent time series at

Table 1: The scaling exponent α of original and shuffled time series measured by DFA. Specifically, the first column corresponds to the scaling exponents α of original time series, and the second column indicates those of the shuffled time series, respectively. Their difference is remarkable, which validates that memory effects generally exist in the online human activities.

	Orginal α	Shuffled α
Book	**0.53**	0.51
Movie	**0.54**	0.50
Music	**0.58**	0.50
1st	**0.70**	0.50
2nd	**0.85**	0.52
3rd	**0.60**	0.51

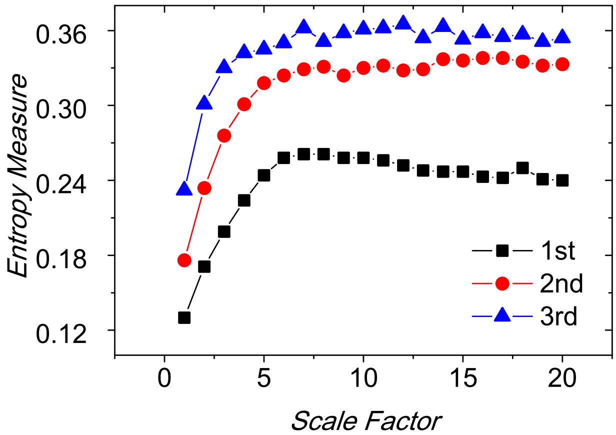


Fig. 3: (Color online) The results of the interevent time series measured by MSE. The sample entropy are obviously different from the activity of user behaviors, and much more information is contained at large scale factors. The symbols indicate 1st (black squares), 2nd (red circles) and 3rd (blue triangles) groups, respectively.

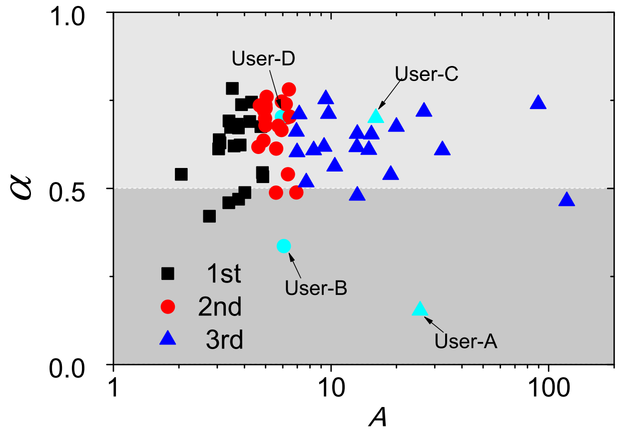


Fig. 4: (Color online) A scatter plot shows the detrended fluctuation analysis of the time series of different users' activities. Each point corresponds to a different user, indicating that there are significant differences between the scaling exponents and users' activity. "Black squares", "red circles" and "blue triangles" denote the individuals in 1st, 2nd and 3rd activity group, respectively.

Table 2: The basic statistical features of the four selected typical users. The first column corresponds to the scaling exponents α, the second column indicates the activities, and the third column represents the frequency, respectively.

	α	Activity (day)	Frequency
User A	0.1534	25.68	3354
User B	0.3369	6.07	3369
User C	0.7034	5.94	3557
User D	0.7001	16.08	9725

large scale factors, to compare them. Figure 3 shows that the sample entropy increases with the growth of the user activity at all scale factors, and for all interevent time series they first increase at small scale factors and then approximately stabilize at constant values, which suggests that the interevent time series constructed from the more active users become more complex and contain the much more information at large scale factors. These results demonstrate that more actions done by the users lead to relative homogeneous interevent time series (*i.e.*, less extreme intervals). We should notice that although the information complexities are different in the three groups, yet they do not directly associate with the degrees of long-range correlation.

Individual level. – Although obvious differences exist in the users' behaviors at the collective level, we still need to understand them and explore the underlying mechanism of the individual user behavior. Therefore, the interevent time series constructed from the total 65 individual user behaviors are further investigated by DFA. The users belonging to different groups are denoted by different symbols. In fig. 4, the scaling exponents are as a function of the activity of users, which shows that there is no direct correlation between them. Furthermore, most of the scaling exponents greater than 0.5 demonstrate the existence of long-range correlations in the interevent time series of human activity in e-commerce. However, there are also abnormal user behaviors (*e.g.*, users A and B seen in fig. 4) suggested by scaling exponents much less than 0.5.

This phenomenon urges us to observe and study these interevent time series constructed from the abnormal user behavior. Table 2 shows the basic statistical features of the abnormal users A and B as well as those of two normal users C and D. From table 2, we can observe that similar activities may show completely different scaling behaviors (*e.g.*, users B and C) and similar scaling behaviors do not mean that they have the same the activities (*e.g.*, users C and D). We note that the frequency denotes the event number of user behavior in table 1.

To uncover the origin of the observed differences among scaling behaviors of individual users, we first present

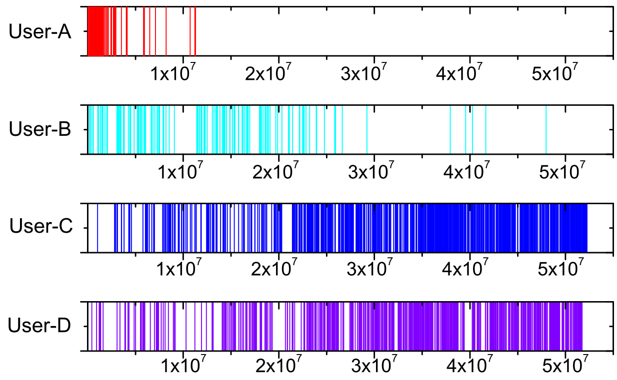
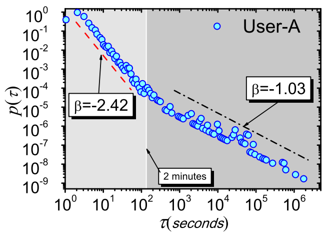
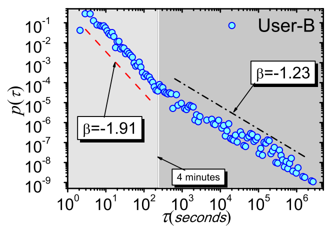


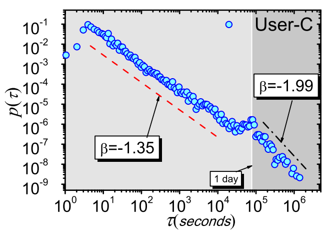
Fig. 5: (Color online) The four online active patterns of users corresponding to table 2. The horizontal axis denotes time, and each vertical line indicates an individual event.



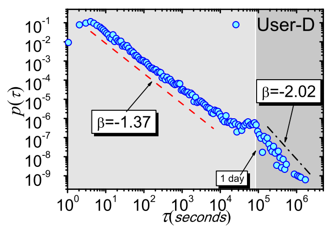
(a) User-A



(b) User-B



(c) User-C



(d) User-D

Fig. 6: (Color online) Interevent time distributions.

the rating activity evolving with time in fig. 5. We can straightly find that the active patterns of users A and B are very different from those of users C and D. Concretely, many more actions occurred in users A and B at the initial stage, which results in shorter time intervals, and the actions become much less or even absent (*e.g.*, user A) when the time evolves.

We propose the query as to what the specific active pattern for four users is, and therefore statistically illustrate the interevent time distributions of the four users in fig. 6, where the interevent time series are defined as the time intervals between consecutive actions by a certain user. The results confirm that the interevent time distributions of these online human activities follow a power-law form, $p(\tau) \sim \tau^{-\beta}$. However, we should note that there is a cutoff at the minute scale for the the interevent time distribution of users A and B, respectively, while for those of users C and D, the blurry cutoffs are on the day scale. The scales of minute and day are a typical decay length of online interests, for example, user actions usually appear within a day in the microblogging systems [10]. Based on the phenomena of bimodal interevent time distributions, it can be found that the power-law exponents β of users A and B change from big to small, yet these cases are quite reversed for users C and D. These changing trends of the bimodal interevent time distributions vividly describe the differences among active patterns of individual user behavior, which suggests that the short and long interevent intervals for users A and B alternately occur and short (or long) interevent intervals for users C and D continuously emerge. Thus, we think that the scaling behaviors strongly associate with these changing trends of the interevent time distributions for human activity in e-commerce although the potential dynamic mechanisms of online individual activity are similar.

Conclusions. – We conclude that our empirical analysis, including the scaling behaviors and information complexities of human activity (*i.e.*, rating the media including music, books, and movies) comprehensively investigated by using DFA and MSE methods, provides a good understanding of the behavior patterns of human activity in e-commerce. We also find that, for all rating behaviors corresponding to the types of media, they display similar scaling properties with exponents ranging from 0.53 to 0.58, which implies that the collective behavior pattern of rating media follows a process embodying self-similarity and long-range correlation. Furthermore, by dividing the users into three groups based on their activity, we observe that the scaling exponents in the three groups are different, yet they both suggest that stronger long-range correlations exist in the collective behaviors. Meanwhile, the information complexities of human activity quantified by MSE confirm the differences of scaling behaviors in these three groups. Moreover, we study the behavior patterns of human activity at the individual level, and find that the behaviors of a few users have extremely small scaling exponents associated with long-range anticorrelations. By comparing the distributions of the interevent time of four representative users, we think that the different scaling behaviors are brought forth by the bimodal forms of the interevent time distributions.

* * *

We thank Ming Tang for the valuable discussion. This work is jointly supported by the NNSFC (Grant Nos. 90924011, 60933005, 61004102, 11105025), China Postdoctoral Science Foundation (Grant No. 20110491705), the Specialized Research Fund for the Doctoral Program of Higher Education (Grant No. 20110185120021). ZDZ appreciates the financial support of the Fundamental Research Funds for the Central Universities (Grant Nos. ZYGX2011YB024, ZYGX2012J075).

REFERENCES

[1] Barabási A.-L., *IEEE Control Syst. Mag.*, **27** (2007) 33.
[2] Zhou T., Han X.-P. and Wang B.-H., *Towards the Understanding of Human Dynamics* (World Scientific Publishing, Singapore) 2008.
[3] Barabási A.-L., *Nature*, **435** (2005) 207.
[4] Oliveira J. G. and Barabási A.-L., *Nature*, **437** (2005) 1.
[5] Candia J., González M. C., Wang P., Schoenharl T., Madey G. and Barabási A.-L., *J. Phys. A: Math. Theor.*, **41** (2008) 224015.
[6] Dezsö Z., Almaas E., Lukács A., Rácz B., Szakadát I. and Barabási A.-L., *Phys. Rev. E*, **73** (2006) 066132.
[7] Zhou T., Kiet H. A. T., Kim B. J., Wang B.-H. and Holme P., *EPL*, **82** (2008) 28002.
[8] Gonçalves B. and Ramasco J., *Phys. Rev. E*, **78** (2008) 026123.
[9] Radicchi F., *Phys. Rev. E*, **80** (2009) 026118.
[10] Zhao Z.-D. and Zhou T., *Physica A*, **391** (2012) 3308.
[11] Han X.-P., Zhou T. and Wang B.-H., *New J. Phys.*, **10** (2008) 073010.
[12] Vázquez A., *Physica A*, **373** (2007) 747.
[13] Oliveira J. and Vázquez A., *Physica A*, **388** (2009) 187.
[14] Min B., Goh K.-I. and Kim I.-M., *Phys. Rev. E*, **79** (2009) 056110.
[15] Wu Y., Zhou C., Xiao J., Kurths J. and Schellnhuber H., *Proc. Natl. Acad. Sci. U.S.A.*, **107** (2010) 18803.
[16] Peng C.-K., Buldyrev S., Havlin S., Simons M., Stanley H. E. and Goldberger A. L., *Phys. Rev. E*, **49** (1994) 1685.
[17] Peng C.-K., Mietus J., Hausdorff J. M., Havlin S., Stanley H. E. and Goldberger A. L., *Phys. Rev. Lett.*, **70** (1993) 1343.
[18] Peng C.-K., Havlin S., Stanley H. E. and Goldberger A. L., *Chaos*, **5** (1995) 82.
[19] Ivanov P. C., Bunde A., Amaral L. A. N., Havlin S., Fritsch-Yelle J., Baevsky R. M., Stanley H. E. and Goldberger A. L., *Europhys. Lett.*, **48** (1999) 594.
[20] Ivanov P. C., Amaral L. A. N., Goldberger A. L., Havlin S., Rosenblum M. G., Stanley H. E. and Struzik Z. R., *Chaos*, **11** (2001) 641.
[21] Ashkenazy Y., Ivanov P. C., Havlin S., Peng C.-K., Goldberger A. L. and Stanley H. E., *Phys. Rev. Lett.*, **86** (2001) 1900.
[22] Liu Y., Gopikrishnan P., Cizeau P., Meyer M., Peng C.-K. and Stanley H. E., *Phys. Rev. E*, **60** (1999) 1390.
[23] Gopikrishnan P., Plerou V., Gabaix X and Stanley H. E., *Phys. Rev. E*, **62** (2000) R4493.
[24] Yamasaki K., Muchnik L., Havlin S., Bunde A. and Stanley H. E., *Proc. Natl. Acad. Sci. U.S.A.*, **102** (2005) 9424.
[25] Ivanov P. C., Yuen A., Podobnik B. and Lee Y.-K., *Phys. Rev. E*, **69** (2004) 056107.
[26] Jiang Z.-Q., Chen W and Zhou W.-X., *Physica A*, **388** (2009) 433.
[27] Tang M. and Liu Z. H., *Physica A*, **387** (2008) 1361.
[28] Cai S.-M., Fu Z.-Q., Zhou T., Gu J. and Zhou P.-L., *EPL*, **87** (2009) 68001.
[29] Levitin D., Chordia P. and Menon V., *Proc. Natl. Acad. Sci. U.S.A.*, **109** (2012) 3716.
[30] Rybski D., Buldyrev S. V., Havlin S., Liljeros F. and Makse H. A., *Proc. Natl. Acad. Sci. U.S.A.*, **106** (2009) 12640.
[31] Rybski D., Buldyrev S. V., Havlin S., Liljeros F. and Makse H. A., *Eur. Phys. J. B*, **84** (2011) 147.
[32] Rybski D., Buldyrev S. V., Havlin S., Liljeros F. and Makse H. A., *Sci. Rep.*, **2** (2012) 560.
[33] Costa M., Goldberger A. and Peng C.-K., *Phys. Rev. Lett.*, **89** (2002) 068102.
[34] Costa M., Goldberger A. L. and Peng C.-K., *Phys. Rev. Lett.*, **95** (2005) 198102.
[35] McIntosh A. R., Kovacevic N. and Itier R. J., *PLoS Comput. Biol.*, **4** (2008) e1000106.
[36] Escudero J, Abásolo D., Hornero R., Espino P. and López M., *Physiol. Meas.*, **27** (2006) 1091.
[37] Li Z and Zhang Y. K., *Stoch. Environ. Res. Risk Assess.*, **22** (2008) 507.
[38] Goh K.-I. and Barabási A.-L., *EPL*, **81** (2008) 48002.
[39] Hu K., Ivanov P., Chen Z., Carpena P. and Stanley H. E., *Phys. Rev. E*, **64** (2001) 011114.
[40] Chen Z., Ivanov P., Hu K. and Stanley H. E., *Phys. Rev. E*, **65** (2002) 041107.
[41] Richman J. S. and Moorman J. R., *Am. J. Physiol.*, **278** (2000) H2039.
[42] Ghoshal G. and Holme P., *Physica A*, **364** (2006) 603.