From Time Series to Co-Evolving Functional Networks: Dynamics of the Complex System 'Wikipedia'

Mirko Kämpf[1], Jan. W. Kantelhardt[1], Lev Muchink[2]

1 Institut für Physik, FG Theoretische Physik, Martin-Luther-Universitat Halle-Wittenberg, 06099 Halle (Saale), Germany
2 Leonard N. Stern School of Business, New York University, USA

Introduction

Internet-based social networks (as novel information and communication platforms) often reflect the dynamics of changing interests and activities in society by characteristic usage patterns. Here we study the dynamics of user *access-rate time series* and *edit-interval time series* for all articles in the online encyclopedia "Wikipedia" with access rates exceeding 255 per hour at least once. While other research on social networks mainly focuses on the development of their structure, we also study the usage of the elements (Wikipedia articles) for information spread. In particular, we characterize the fluctuation behaviour of the both, access-rate and edit-interval time series. For describing the reoccurrence of bursts exceeding certain thresholds we investigate the *statistics of return intervals* between these bursts. We find stretched exponential distributions of return intervals with identical parameters for all thresholds in access-rate time series, while edit time series show a simple exponential distribution of return intervals. To characterise the fluctuation behaviour of the access-rate time series we apply – after removing the daily and weekly periodicities – the *detrended fluctuation analysis* (DFA) method. We find that most access-rate time series are characterized by long-term correlations with fluctuation exponents $\alpha \approx 0.9$.

To understand the complex processes underlying these different dynamics of access rates and edit intervals, we characterize and compare three organizational and dynamical networks associated with 'Wikipedia' in the second part of the work: (i) the *network of direct links* between Wikipedia articles, (ii) the *usage network* as determined from cross-correlations between access-rate time series of many pairs of articles, and (iii) the *edit network* as determined from co-incident edit events. The major goal is to find correlations between components of these three networks that characterize the dynamics of information spread in the complex system. The *network reconstruction* is done by two different approaches. For access-rate time series, we use the cross-correlation coefficient at time delay zero between both time series of a selected group of nodes, linked to a central node, in combination with statistical significance tests. The link strengths for the corresponding edit time series are determined by the event synchronisation between all pairs of articles. We find that – even though the dynamics of article access-rate and edit-interval time series are characteristically different – there are indications of a co-evolution of the corresponding dynamic functional networks. Obvious differences between both reconstructed networks are also shown. The results help in understanding the complex process of collecting, processing, validating, and distributing information in self-organised social networks.

Dataset

We study Wikipedia access-rate and edit-interval time series recorded during a period of 10 months in 2009, focussing on all articles with access rates exceeding 255 per hour at least once. Our dataset is a collection of log records of all edit activities (in an SQL database, 1 second time resolution) and the hourly counted number of accesses (downloads) of each page (in a binary database, time resolution of access rates is 1 hour). We begin with characterizing the properties of each single page (article) [1]. We observe daily and weakly activity patterns in the hourly access rates for most Wikipedia pages in addition to apparently random fluctuations and bursting activity, see Figure 1. These weekly trends are removed from the raw data of access times.

To characterize the properties of extreme events in the access-rate data we extract the width of pronounced bursts that exceed the average access-rate be more than a factor of two. The average durations of each burst before (t_{before}) and after (t_{after}) the maximum are compared to each other, see Figure 2. We find two regimes of different scaling behaviour for short events with length of less than

10 h and long events with a length of more than 10 hours. This property is independent from the selected threshold. The bursts are characterized by power-law or exponential increases and decreases of activity, and they can be classified as 'endogenous' (with significant precursory activity) or 'exogenous' (extrinsically caused) events [1,2,3].

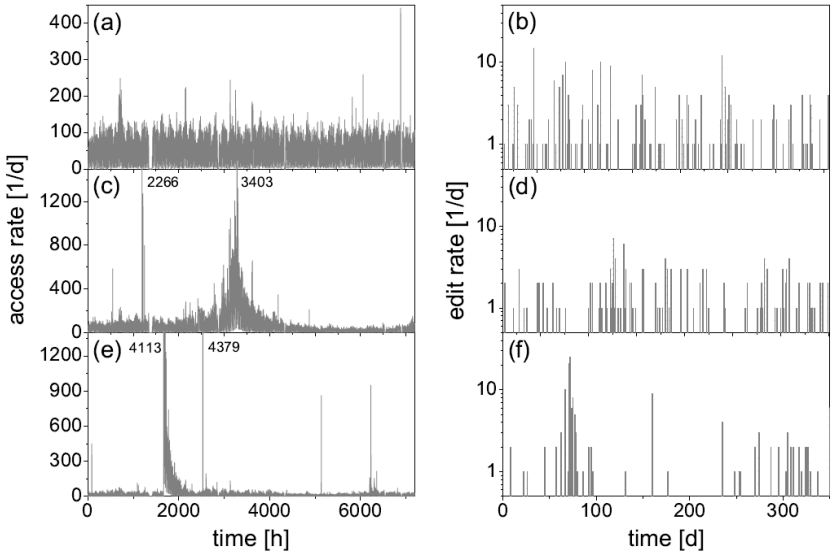

Figure 1: Examples of Wikipedia access statistics for three selected articles with (a,b) rather stationary access rates (topic 'Illuminati (book)'), (c,d) an apparently endogenous burst of activity (peak on Sunday, May 7, 2009, topic 'Heidelberg'), and (e,f) an exogenous burst of activity (topic 'Amoklauf Erfurt' (shooting rampage)) has a peak on Wednesday, March 11, 2009, as another shooting rampage occurred in Winneden). The left parts show the complete hourly access rate time series (from January 1, 2009, till October 21, 2009; e. g. for 42 weeks=294 days=7056 hours) with numbers in the plot giving the height of peaks truncated to show baseline fluctuations. The gap around t=1200h, is a systematical disruption and was found in all records. The right parts show edit-event data for the three representative articles. The plots (b,d,f) show the number of edits per day; the articles were edited 270, 163, and 157 times in total during the recording period. (Figure taken from [1].)

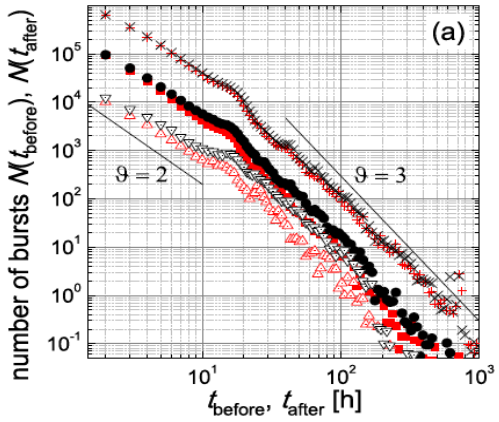

Figure 2: Number of bursts versus their duration before (t_{before}, red symbols) and after (t_{after}, black symbols) the maximum of the burst. Data for maxima at least 2 times (crosses for t_{before} and plus signs for t_{after}), 5 times (filled squares and circles), and 15 times (open triangles up and down) above the average access time of that hour were considered. The straight lines representing power-laws with slopes ϑ = 2 and 3 are shown for comparison. (Figure taken from [1].)

Characterization of single-article (node) properties

To characterize the fluctuation behaviour of the access time series we applied – after removing the daily and weekly periodicities – the detrended fluctuation analysis (DFA) method [4,5]. We also compare results for different detrending orders to see if there are relevant effects of trends or non-stationarities (such as bursts). We find that most article access time series are characterized by long-term (power-law) correlations, see Figure 3(a) to (c). The histograms in Figs. 3(d) and (e) show that the power-law scaling behaviour, $F(s) \sim s^{\alpha}$, is quite universal with fluctuation scaling exponents rather narrowly distributed around $\alpha \approx 0.9$ for articles with quite stationary access rates and rather non-stationary, bursty fluctuations. Note that α is related to the correlation exponent γ characterising the

power-law decay of the auto-correlation function: $C(s) \sim s^{-\gamma}$ by $\gamma = 2 - 2\alpha$. There is only a weak dependence of α on the total number of accesses (i.e., the "importance" of the articles), see Fig. 3(f).

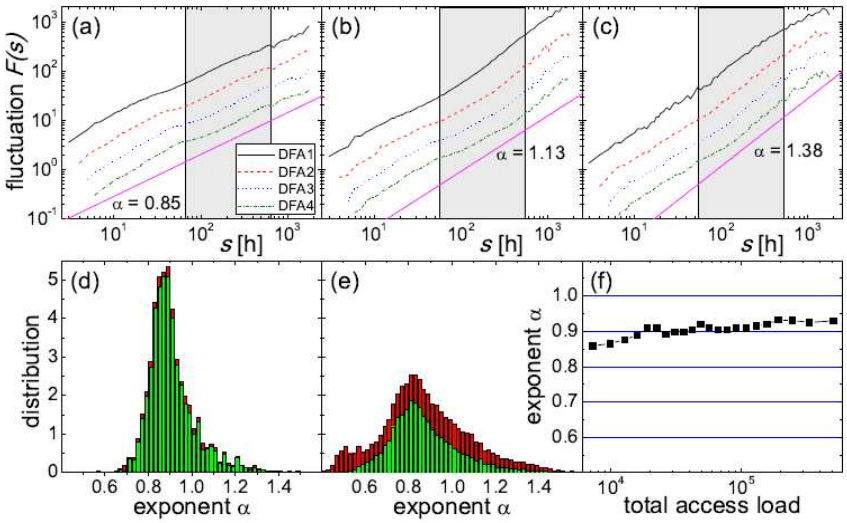

Figure 3: (a,b,c) DFA of access-count time series of representative Wikipedia articles (the same as for Fig. 1). The results for different detrending orders show very similar behaviour: DFA1 (black solid line), DFA2 (red dashed line), DFA3 (blue dotted line), and DFA4 (green dash-dotted line); data for DFA2 to DFA4 shifted by multiple factors of 2 for clarity. The straight lines below the data have the indicated slopes and are shown for comparison. On small time scales ($s < 50$h) the effective scaling exponents α are sometimes a bit smaller due to incompletely removed daily rhythms (periodic trends with 24h period, see Fig. 3(b) in particular). Parts (d,e) show distributions of the scaling exponents for the regime 55h < s< 550h, separately for (d) stationary and (e) non-stationary access-rate time series using DFA1. The green fractions of the bars indicate the ratio of time series for which very good power-law fits with correlation coefficients $r > 0.98$ were obtained. (f) Dependences of the average DFA1 scaling exponents on the total access volume that occurred within the considered 42 weeks. (Figure taken from [1].)

For describing the reoccurrence of bursts exceeding certain thresholds we investigate the statistics of the return intervals between these bursts [6,7], see Figure 4.

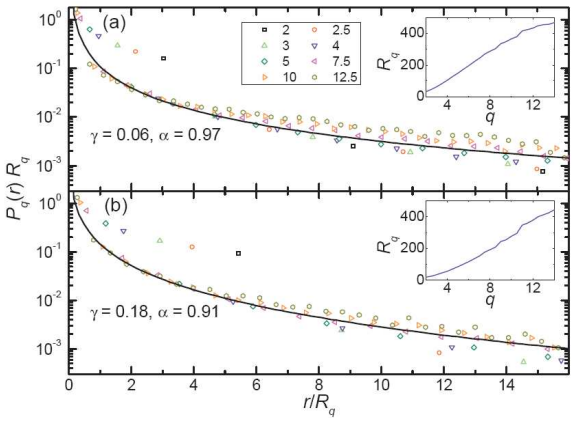

Figure 4: Normalized distributions $P_q(r)R_q$ of return intervals r between bursts exceeding the different thresholds q given in the legend for Wikipedia access-rate data with hourly resolution for (a) English data and (b) German data. (Figure taken from [1].)

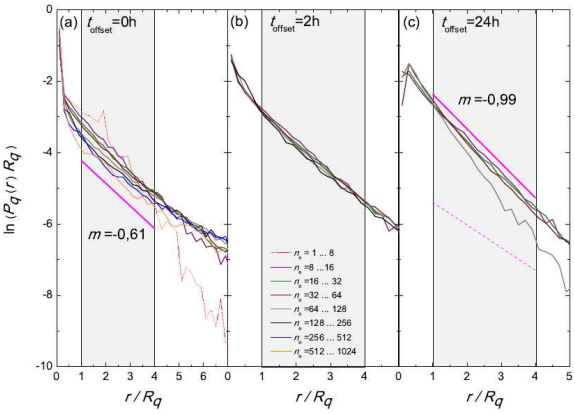

Figure 5: Normalized distributions $P_q(r)R_q$ of return intervals r between article edit events for a selection of 13.416 English Wikipedia articles with different total numbers of edits n_e (see legend, there z is the number of articles in each group. The data follow a simple exponential distribution except for drastically increased probabilities of very short periods between edits (first data points in each curve). The straight line is a simple exponential, but with a smaller prefactor and slightly different slope due to normalization changed by the large first value of each curve. (Figure taken from [1].)

We find stretched exponential distributions of return intervals with identical parameters for all thresholds in access-rate time series. Edit time series, on the other hand show a simple exponential distribution of return intervals, see Figure 5. The results are also compared regarding different languages in Wikipedia. The results for article access-rates are in full agreement with DFA results shown in Fig. 3. For more details we refer to [1].

After the detailed characterisation of the properties of single pages, we want to understand the causes of the different dynamics of the edit and access processes. Therefore we use a network representation of the interrelations of the articles to study these different processes.

Construction of Functional Networks

As human interaction and information spread via on-line networks is becoming increasingly important for our contemporary technological society we should not regard the Wikipedia system as a collection of independent web pages. We therefore reconstruct and compare three organizational and dynamical network structures associated with Wikipedia in the following second part of our work.

The analysis of the static link network is just one aspect of the whole system. By looking at a dynamic link structures, we can obtain a second aspect or a second subsystem. The whole Wikipedia community uses the system, while it is edited and while it changes its underlying properties. Because of this, we want to isolate the different views of the interconnected processes of growing, changing and using the network.

In particular, we study (i) the network of the direct links between Wikipedia articles of various languages, (ii) the usage network as determined from cross-correlations between click-count time series of many pairs of articles, and (iii) the edit network as determined from co-incident edit events, see Figure 6. The major goal is to find correlations between components of these three subsystems which can be seen as networks which characterize the dynamics of information spread in the complex system.

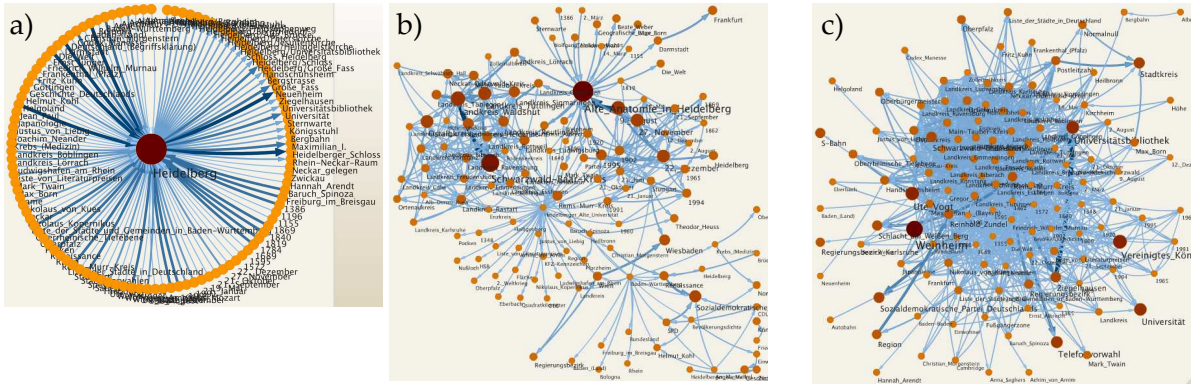

Figure 6: Comparison of (a) the static link network, (b) the correlation network based on access activity for the whole recording period and (c) the correlation network of edit activity for a subnet of about 120 Wikipedia pages linked to the page with the name "Heidelberg". The figures were generated using the map.equation tool [8].

The process of reconstruction is done by two different approaches. For access-rate time series, we use the cross-correlation coefficient at time delay zero [9] between both time series of a selected group of nodes, linked to a central node, in combination with statistical significance tests. The link strengths for the corresponding edit time series are determined by the event synchronisation between all pairs of articles [10,11]. Obvious differences between both reconstructed networks are apparent in the exemplary functional networks, see Figure 6.

In addition, we can observe dynamic changes in the reconstructed networks, which reflect changes of interest focus in society. These presentations (as shown in Figure 7) help in understanding the complex process of collecting, processing, validating, and distributing information in self-organised social networks.

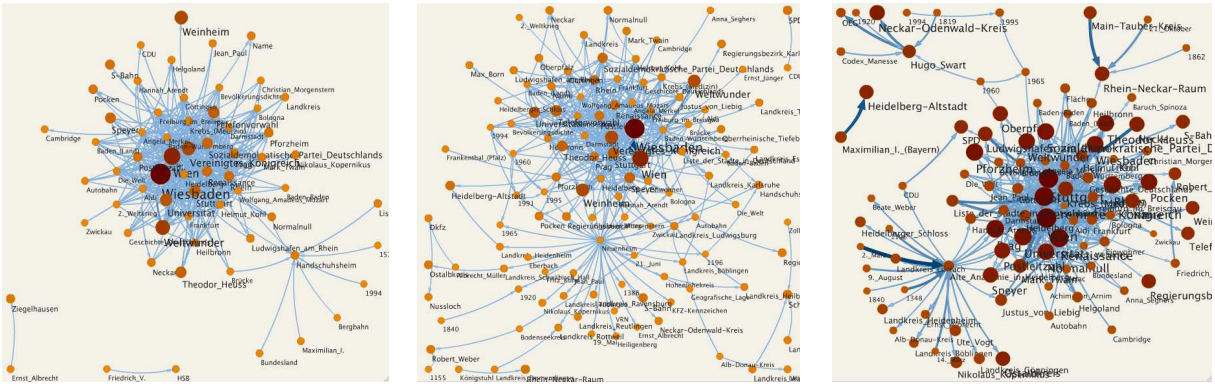

Figure 7: For one selected central node (Wikipedia page for the city of Heidelberg) the time series for all linked nodes are extracted and access-rate cross-correlation link strengths are calculated for three different time frames. One can see clearly, that the correlation between single nodes in the context of a central node changes in time. The figures were generated using the map.equation tool [8].

Outlook

A deeper analysis of the complex processes underlying the Wikipedia system will be possible as soon as we have a generic method for generating such networks based on measureable data sets. The properties of the data recordings also have an influence on the used algorithms, e.g. the number of edit events is much smaller than the access events. Such differences lead to several variations of the definition of the link or correlation strength. Dependent on the properties of each single subsystem different algorithms have to be used and adapted.

We have to select or define useful measures, for example the clustering coefficient, the degree distributions or the average path length of the calculated networks. Based on these properties we may see, what external influences could lead to a phase transition in the underlying system. A study of dependencies between properties of correlation networks, calculated from measured time series, will allow new approaches in the research field of socio-technical-complex-systems. A study of the relations or interactions between connected subsystems also leads to the emerging field of "network of networks" [12,13].

References

[1] M. Kämpf, S. Tismer, J. W. Kantelhardt, and L. Muchnik; **Burst event and return interval statistics in Wikipedia access and edit data**, submitted to Physica A (2011).

[2] L. Mitchell, M.E. Cates; **Hawkes Process as a model of social interactions : a view on video dynamics**; J. Phys. A: Math. Theor. **43** (2010) 045101.

[3] R. Crane, D. Sornette; **Robust dynamic classes revealed by measuring the response function of a social system**; PNAS **105** (2008) 15649-15653.

[4] C. K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger; **Mosaic organization of DNA nucleotides**; Phys. Rev. E **49** (1994) 1685.

[5] J.W. Kantelhardt, S.A. Zschiegner, E. Koscielny-Bunde, S. Havlin, A. Bunde, and H.E. Stanley**; Multifractal detrended fluctuation analysis of nonstationary time series**; Physica A **316** (2002) 87.

[6] J.F. Eichner, J.W. Kantelhardt, A. Bunde, and S. Havlin; **Statistics of return intervals in long-term correlated records**; Phys. Rev. E **75** (2007) 011128.

[7] A. Bunde, J.F. Eichner, J.W. Kantelhardt, and S. Havlin; **Long-Term Memory: A Natural** Mechanism for the Custering of Extreme Events; Phys. Rev. Lett. **94** (2005) 048701.

[8] M. Rosevall, D. Axelsson, C.T. Bergstrom; **The map equation**, Eur. Phys J. Special Topics **178** (2009) 13-23.

[9] J.F. Donges, Y. Zuo, N. Marvan and J. Kurths; **Complex networks in climate dynamics : Comparing linear and non-linear network construction methods**, Eur. Phys J. Special Topics **174** (2009) 157-179.

[10] R.Q. Quiroga, T. Kreuz, and P. Grassberger;**Event synchronization: A simple and fast method to measure synchronicity and time delay patterns**, Phys. Rev. E **66** (2002) 041904.

[11] N. Malik, B. Bookhagen, N. Marwan, and J. Kurths; **Analysis of spatial and temporal extreme monsoonal rainfall over South Asia using complex networks**, Clim Dyn (in press 2011), DOI 10.1007/s00382-011-1156-4

[12] S.V. Buldyrev, R. Parshani, G. Paul, H.E. Stanley, and S. Havlin; **Catastrophic cascade of failures in interdependent networks**; Nature **464** (2010) 1025-1028.

[13] S. Havlin, D.Y. Kenett, E. Ben-Jacob, A. Bunde, R. Cohen, H. Hermann, J. W. Kantelhardt, J. Kertesz, S. Kirkpatrick, J. Kurths, Y. Portugali, and S. Solomon; **Challenges of network science: Applications to infrastructures, climate, social systems and economics**, Eur. Phys. J. ST (in print, 2012).