Sustainable growth in complex networks

2011 EPL 96 58005

(http://iopscience.iop.org/0295-5075/96/5/58005)

A Letters Journal Exploring the Frontiers of Physics

December 2011

EPL, **96** (2011) 58005
doi: 10.1209/0295-5075/96/58005

www.epljournal.org

Sustainable growth in complex networks

C. J. Tessone[(a)], M. M. Geipel and F. Schweitzer

Chair of Systems Design, D-MTEC, ETH Zurich - CH-8032 Zurich, Switzerland

received 27 June 2011; accepted in final form 19 October 2011
published online 25 November 2011

PACS 89.75.Hc – Networks and genealogical trees
PACS 05.40.-a – Fluctuation phenomena, random processes, noise, and Brownian motion
PACS 89.20.Ff – Computer science and technology

Abstract – Based on the analysis of the dependency network in 18 Java projects, we develop a novel model of network growth which considers both preferential attachment and the addition of new nodes with a heterogeneous distribution of their initial degree, k_0. Empirically we find that the cumulative distributions of initial and final degrees in the network follow power law behaviours: $1-P(k_0) \propto k_0^{1-\alpha}$, and $1-P(k) \propto k^{1-\gamma}$, respectively. For the total number of links as a function of the network size, we find empirically $K(N) \propto N^{\beta}$, where $\beta \in [1.25, 2]$ (for small N), while converging to $\beta \sim 1$ for large N. This indicates a transition from a growth regime with increasing network density towards a sustainable regime, which prevents a collapse due to ever increasing dependencies. Our theoretical framework allows us to predict relations between the exponents α, β, γ, which also link issues of software engineering and developer activity. These relations are verified by means of computer simulations and empirical investigations. They indicate that the growth of real Open Source Software networks occurs on the edge between two regimes, which are dominated either by the initial degree distribution of added nodes, or by the preferential attachment mechanism. Hence, the heterogeneous degree distribution of newly added nodes, found empirically, is essential to describe the laws of sustainable growth in networks.

How do *real* networks grow? Tracing the complete history of empirical networks is difficult and rare. A noticeable exception are dependency networks in Open Source Software (OSS) projects. In this letter we provide empirical evidence that their evolution is governed by the addition of nodes with a very skewed initial degree distribution. This differs from most common network growth models and leads us to develop a novel model of network growth which extends existing analytical approaches to encompass the role of inhomogeneities.

So far, many modelling approaches, most notably the preferential attachment, simply assume that i) at any time step τ a constant number of nodes is added to the network, that ii) each new node is linked to the network with a constant number of links, and that iii) neither nodes nor links are deleted [1–3]. If such assumptions hold, this results in a growth $N(\tau) \propto \tau^{\eta}$ of the total number of nodes in the network, and $K(\tau) \propto \tau^{\lambda}$ of the total number of links, where both exponents $\eta \simeq 1$ and $\lambda \simeq 1$. Such a network growth could be called *sustainable*, in contrast to the two limiting cases of a) accelerated growth [4–6], if $\lambda/\eta > 1$, or b) saturated growth, if $\lambda/\eta < 1$. Both of these growth processes are not sustainable in the long run as they either lead to collapse or to stagnation [7]. There is, at least for observable intermediate time scales, also empirical evidence of networks growing with increasing link density, for example the World Wide Web [8,9]. Here we focus on OSS as another example.

The evolution of software was recently investigated by means of statistical physics, ranging from the self-organised software dynamics [10,11], to the analysis of the network topology of the dependency networks [12–14], and to network motifs in software networks [15]. However, results obtained for $N(\tau)$ or $K(\tau)$ refer to macroscopic properties, which are compatible with a large variety of "microscopic" assumptions about node and link additions (or deletions). More importantly, the kinetic exponents may change over time and may reach 1 only asymptotically, which would point to changes in the growth mechanism on intermediate time scales. Eventually, in addition to the total number of nodes and links, there are other characteristics of the network structure and dynamics which need to be predicted and to be verified empirically. In this letter, we address these problems both theoretically and empirically by i) developing a detailed model of network growth which includes the heterogeneous degree

[(a)]E-mail: tessonec@ethz.ch

Table 1: Empirical results obtained for 18 Open Source Java projects. N gives the maximum number of nodes (classes) at the date of the last snapshot taken; α, γ are the exponents for the initial and final degree distribution. β is the value of the exponent describing the growth of the total number of links as a function of network size.

Project	N	α	β	γ
eclipse	28898	2.7(1)	1.06(4)	2.6(1)
springframework	7707	3.5(1)	1.02(4)	2.9(1)
fudaa	7610	2.7(1)	1.1(1)	2.7(1)
jpox	7259	2.49(8)	1.08(2)	2.44(8)
architecturware	7110	2.7(1)	1.00(3)	2.8(1)
jena	6619	3.5(1)	0.99(3)	2.9(1)
hibernate	5938	2.5(2)	1.03(3)	2.5(1)
sapia	4129	3.44(8)	1.00(2)	3.0(1)
rodin-b-sharp	4077	2.8(1)	1.03(2)	2.6(1)
azureus	4051	2.9(2)	1.14 (5)	2.6(2)
jedit	3997	2.9(1)	1.01(1)	2.93(8)
jaffa	3854	3.0(3)	1.1(1)	2.7(3)
jmlspecs	3590	2.4(2)	0.97(6)	2.6(2)
openxava	3000	3.2(2)	1.04(4)	2.9(2)
phpeclipse	2881	2.8(1)	1.02(2)	2.73(8)
personalaccess	2687	3.1(1)	0.95(6)	2.9(1)
xmsf	2576	2.2(1)	1.08(3)	2.3(1)
aspectj	1856	2.5(1)	1.03(4)	2.5(1)

distribution of newly added nodes, and ii) by verifying the predictions of our general model against a novel data set of growing networks.

We start by describing the empirical findings, to motivate the assumptions of the network growth model, later. We have used a dataset of 18 OSS projects (see table 1), which are programmed in Java. The network consists of *nodes*, which are Java classes (each file corresponds to one class), and *links* representing dependencies between these classes. For example, one class can call a function defined in another class, use another class or inherit another one. During software evolution, new classes are added to the project and are linked to existing classes based on principles defined in *software engineering*. So, if we are able to reveal a universal dynamics underlying such growth processes, this is a remarkable result on its own. For the time-dependent evolution of the OSS projects, we rely on version control systems which record all changes made. For our analysis, we have used snapshots of intervals of 30 days, for a project life span between 2.7 and 8.2 years —which goes much beyond the few snapshots available for previous investigations of OSS growth [13,14,16,17]. Nevertheless, we may use these studies as a point of reference, as they also study some topological properties, such as the cumulative degree distribution $P(k)$.

In order to derive analytical results about the latter, let us define $n(k,\tau)$ as the degree distribution, *i.e.* the number of nodes with *total* degree k at time τ. Obviously $2\,K(\tau) = \sum_{k=1}^{N(\tau)} k\, n(k,\tau)$. The complementary cumulative degree distribution at time τ is then given by $P(k,\tau) = 1 - \sum_{l<k} n(l,\tau)/N(\tau)$. We can replace the real time τ by using the scaling $\tau \propto N^{1/\eta}$, which means $K(N) \propto N^{\beta}$, where $\beta = \lambda/\eta$. This procedure implies that the number of nodes is increasing, *i.e.* the deletion of nodes is not considered. Figure 1 illustrates the empirical results for these quantities by showing examples of four OSS projects of very different size, while table 1 contains the detailed information for all projects investigated.

Looking at the final complementary cumulative degree distribution $1 - P(k)$ obtained for the maximum size of the project, we clearly identify a power law $1 - P(k) \propto k^{1-\gamma}$ (fig. 1, left panel), which characterises the *structure* of the *final product*. Dependent on the size of the project, values between 2 and 3 are found, with a tendency towards values closer to 3. For the growth of the OSS projects (fig. 1, middle panel) we obtain slightly bend curves for the four projects which indicate that the exponents β change over time, as can be also seen in fig. 2. To calculate β, for every project the total degree as a function of system size was split into different windows (of size 500) for each of which β was estimated. Starting at values of $\beta \in [1.25, 2]$, they converge to smaller values around $\beta \cong 1$, which implies that fewer dependencies are added as the network ages. Thus, we observe a transition from accelerated to sustainable growth. The right panel of fig. 1 further presents the most interesting empirical finding that, different from the above mentioned assumptions about preferential attachment and most modelling approaches, newly added nodes have a very heterogeneous initial degree k_0. In fact we observe a power law for the complementary cumulative *initial degree* distribution $P(k_0) \propto k_0^{1-\alpha}$, where α is related to the initial conditions of the software growth, *i.e.* to *software design*. It remains to reveal the inherent relations between the three exponents α, β, γ which is done by the following analytical approach.

We assume that nodes are added to the project at a constant rate, *i.e.* $\eta = 1$. Moreover, time t is given by the total number of nodes, $t = N$. For the dynamics of the degree distribution we postulate the following rate equation:

$$\begin{aligned}\dot{n}(k,t) \;=\; & \delta_{k,k_0(t)} + n(k-1,t)\,\omega[k-1\to k] \\ & + n(k+1,t)\,\omega[k+1\to k] \\ & - n(k,t)\left\{\omega[k\to k-1] + \omega[k\to k+1]\right\}. \quad (1)\end{aligned}$$

This is a first-order approximation of the dynamics. The term $\delta_{k,k_0(t)}$ in eq. (1) describes the addition of a new node with an initial degree exactly equal to k_0. In accordance with our empirical findings, this degree is randomly drawn from a truncated power law distribution $g(k_0)$ with exponent α; *i.e.*

$$\mathrm{Prob}[k_0(t) = k] = \min((\alpha-1)/k^{\alpha}, t-1). \quad (2)$$

This broad initial degree distribution indicates the role of *modularity* and *anticipation to change* in software engineering [18], where different problems have to be isolated and solved separately. From this distribution, it

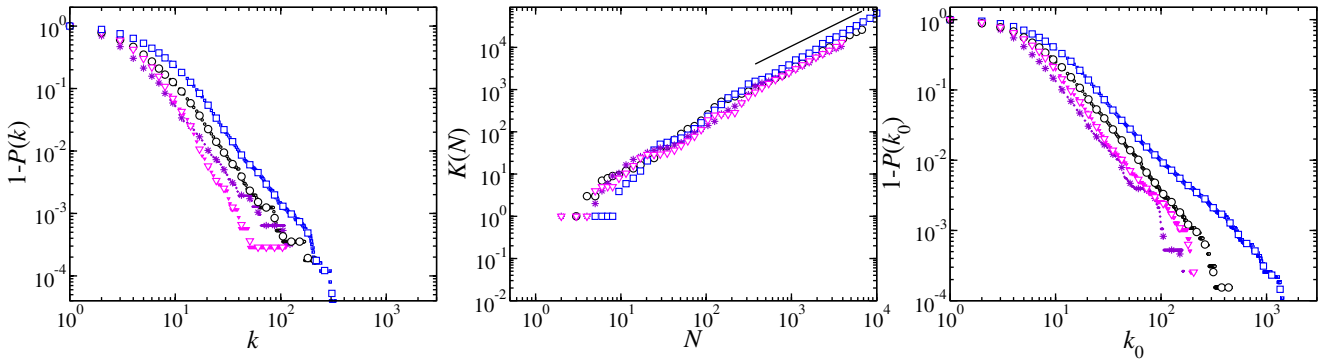

Fig. 1: (Colour on-line). Left: final complementary accumulated degree distribution $1-P(k) \propto k^{1-\gamma}$. Right: initial complementary cumulated degree distribution $1-P(k_0) \propto k^{1-\alpha}$. Middle: total number of links $K(N) \propto N^{\beta}$ as a function of the network size N. Colours indicate four different OSS projects: architecturware (black circles), eclipse (blue squares), jEdit (violet stars), sapia (magenta triangles). See table 1 for more details. The small symbols, represent the complete empirical datasets, while the large ones the binned data.

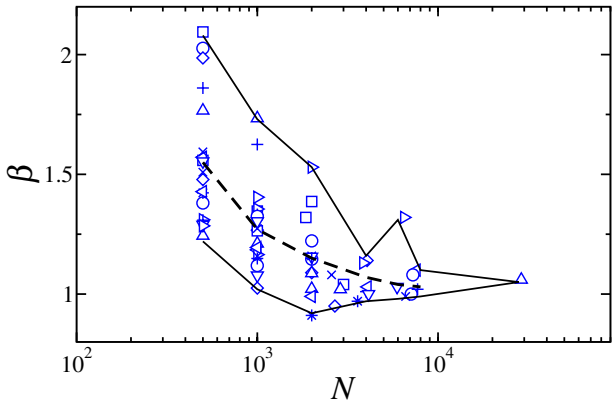

Fig. 2: (Colour on-line) Evolution of exponent β. Different symbols represent the 18 OSS projects given in table 1, while the dashed line gives the median of β obtained from all projects.

is apparent that modules spanning multiple scales have to be written to implement the functionality expected from software projects.

For the transition rate of growth processes, $k \to k+1$, we assume

$$\omega[k \to k+1] = \left\{ \frac{k_0(t)}{K(t)} + \left(\sigma + \frac{r}{2}\right) \right\} k. \qquad (3)$$

This rate is proportional to k, *i.e.* it is based on preferential attachment. Without that assumption, the process would result in a single-scale network which is not in accordance with the empirical studies above. Two different processes are included in this transition rate: in the first term a newly added node links to $k_0(t)$ existing ones, which are selected with a probability proportional to their relative degree k/K, while in the second term links between existing nodes are created. σ and r are constants described below. The transition rate corresponding to the deletion of links, $k \to k-1$, is also assumed to be proportional to the degree of the node,

$$\omega[k \to k-1] = \left\{ \sigma - \frac{r}{2} \right\} k. \qquad (4)$$

This formalism provides a complete model for the dependency dynamics in software. If the network dynamics due to the *addition of nodes* with heterogeneous degree k_0, does not play any role, *i.e.* $k_0(t) = 0$, the dynamics is only governed by the *addition/deletion of links* distributed between existing nodes. Then, the rate equation (1), in the continuous limit and for large N, can be transformed into the following Fokker-Planck equation:

$$\partial_t n(k,t) = -\partial_k \left[r\, k\, n(k,t) \right] + \partial^2_{kk} [\sigma^2\, k^2\, n(k,t)] \qquad (5)$$

which is equivalent to the following Langevin dynamics for the degree k_i of a single node i,

$$\dot{k}_i(t) = -r\, k_i(t) + \sigma\, k_i(t)\, \xi_i(t). \qquad (6)$$

This describes the known *law of proportional growth* [19,20], where r is the mean growth (drift) and σ is the variance of the normalised random force $\xi_i(t)$. It is well known [21] that this dynamics, if coupled with birth and death processes, yields a power law distribution $1-P(k) \propto k^{1-\gamma}$, with γ equal to 2, *i.e.* known as Zipf's law. In fact, this law was empirically confirmed for the *in-degree* distribution of *Linux packages* [16] as well as for *Java projects* [17] whereas the *out-degree* distribution, at least for the latter dataset, clearly followed a *log-normal* distribution.

In this letter, we are interested in another limiting case of the dynamics given by eq. (1) where the addition/deletion of links among existing nodes expressed by σ and r are negligible. *I.e.*, we emphasise network growth based on the addition of nodes with a broad initial degree distribution, $g(k_0)$. This assumption is fully justified for the broad distribution of initial degrees found empirically. This regime is described by the governing principle of software engineering, *incrementality*, where new functionality is added on top of the existing one. Moreover, there is a tendency to *open/closed* design [18]: once implemented, classes have a fixed interface with only internal changes not affecting their connectivity with others. In this case,

the dynamics is fully described by the following set of equations:

$$\begin{aligned}\dot{n}(1,t) &= \delta_{1,k_0(t)} - \frac{k_0(t)}{K(t)} n(1,t),\\ \dot{n}(k,t) &= \delta_{k,k_0(t)} + \frac{k_0(t)}{K(t)}\{(k-1)n(k-1,t) - k\, n(k,t)\}\end{aligned} \tag{7}$$

and the initial condition $n(k,0) = n_0\, \delta_{k,n_0-1}$. *I.e.* initially a small number of nodes (*e.g.* $n_0 = 2$) with a degree of $n_0 - 1$ is assumed, which describes a small, fully connected network to start with. From this set of equations, we first derive the dynamics for the total number of links, $K(t)$. By definition, for a single network realisation $\dot{K}(t) = k_0(t)$ holds. The ensemble average $\langle K(t)\rangle$ over many realisations of the network growth process is then given by

$$\langle \dot{K}(t)\rangle = \langle k_0 | k_0 < t\rangle + t \cdot \mathrm{Prob}[k_0(t) > t]. \tag{8}$$

The first term represents the expected value of $k_0(t)$ restricted to $k_0(t) < t$, which applies if the number drawn from the distribution $g(k_0)$ is lower than the current network size ($t = N$) and, thus, the newly added node is able to establish as many links as drawn from the distribution. If this is not the case, *i.e.* $k_0(t) > t$ the node can only create at most $t-1$ links, which is described by the second term. By recasting the power law distribution for $g(k_0)$, we get

$$\langle \dot{K}(t)\rangle = \frac{\alpha - 1}{2-\alpha} + \frac{t^{2-\alpha}}{2-\alpha}. \tag{9}$$

Asymptotically, we find that the total number of links grows in time or with network size $t = N$, respectively, as a power law, $K(t) \propto t^{\beta}$, with the exponent $\beta = 3 - \alpha$ if $\alpha < 2$; $\beta = 1$, otherwise.

By applying the ensemble average to eqs. (7), we are further able to find a mean-field approximation for the dynamics of the degree distribution $n(k,t)$. Using $\langle \delta_{k,k_0(t)}\rangle = \mathrm{Prob}[k = k_0(t)] = (\alpha - 1)/k^{\alpha}$ and similar arguments as in eq. (9), we find that

$$\langle k_0(t)\rangle = t^{2-\alpha}\frac{1}{2-\alpha} + \frac{\alpha - 1}{\alpha - 2}. \tag{10}$$

By analysing the solution of eqs. (9), (10) we find two different regimes for the ratio $\langle k_0(t)\rangle / \langle K(t)\rangle$: i) if $\alpha > 2$, then $\langle k_0(t)\rangle \propto (\alpha-1)/(\alpha-2)$ and $\langle K(t)\rangle \propto (\alpha-1)/(\alpha-2)$; ii) if $\alpha < 2$, $\langle k_0(t)\rangle \propto t^{2-\alpha}/(\alpha-2)$ and $\langle K(t)\rangle \propto t^{3-\alpha}$. Both regimes, however, yield identical result, *i.e.* $\langle k_0(t)\rangle / \langle K(t)\rangle = \zeta(\alpha) t$, with $\zeta(\alpha)$ a normalisation constant. Thus, we can rewrite eqs. (7) as

$$\begin{aligned}\langle \dot{n}(1,t)\rangle &= (\alpha - 1) - \frac{\langle n(1,t)\rangle}{\zeta(\alpha)\, t},\\ \langle \dot{n}(k,t)\rangle &= \frac{(\alpha-1)}{k^{\alpha}} + \frac{(k-1)\langle n(k-1,t)\rangle - k\langle n(k,t)\rangle}{\zeta(\alpha)\, t}.\end{aligned} \tag{11}$$

These equations reveal a competition between two different processes: the growth of the network caused by the addition of links with a broad initial degree distribution (first term) and the growth of a node's degree caused by a mechanism akin to preferential attachment (second term). If α is small, the first case dominates and the expected degree distribution is simply given by

$$\langle n(k,t)\rangle = \frac{(\alpha - 1)}{k^{\alpha}} t. \tag{12}$$

On the other hand, if α is large and the addition of new nodes with a heterogeneous initial degree distribution can be neglected, we recover the usual Barabási-Albert model with $n(k,t) \propto k^{-3}$. That means if the initial degree follows a Gaussian distribution which, according to the generalized central limit theorem, is expected to occur for $\alpha \geqslant 3$ our model recovers the standard scale-free behaviour with exponent $\gamma = 3$. Thus, we have found two different regimes for the final degree distribution, which depend of the exponent of the initial degrees distribution α: $\gamma = \alpha$ if $\alpha < 3$; $\gamma = 3$, otherwise.

To conclude, our analytical approach has provided a firm relation between the three different exponents α, β, γ, which can be tested in two different ways: i) by computer simulations of the full dynamics for various network sizes N and distribution of initial degrees; ii) by comparison with the empirical findings from the 18 OSS projects. The results are shown in fig. 3. They confirm that the analytical approximations are indeed valid and in good agreement both with the computer simulations and the empirical results. Most interestingly, they reveal that the growth dynamics of real OSS networks is on the edge between two regimes: for $\alpha < 3$, the initial degree distribution and hence the addition of new nodes would dominate the whole growth process, whereas for $\alpha > 3$ the preferential attachment of links between existing nodes would dominate. Moreover, all the projects in our dataset show $\alpha > 2$. As the empirical findings verify, none of these regimes fully cover real software growth. In particular, the heterogeneous degree distribution of newly added nodes cannot be neglected.

Eventually, we wish to point to the self-organising dynamics observed in OSS which turns an initially accelerated network growth ($\beta > 1$) into a sustainable one ($\beta \to 1$). For mature projects, this transition prevents software growth from collapsing caused by the non-linearly increase of dependencies between classes. This raises the question whether this transition indicates a shift from developing the core functionality (during the first steps of the project) to actually using it (after the project has grown). Such an explanation is in line with the observed transition from an increasingly connected network to a sparser one, as the network grows. We emphasise, however, that even in such a scenario the initial degree distribution has proven to be the key ingredient in the network evolution.

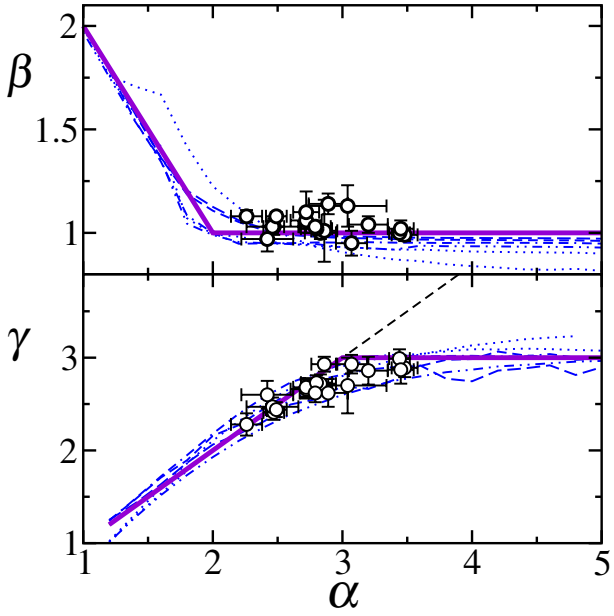

Fig. 3: (Colour on-line) Upper panel: exponents β for the growth of the total number of links. Lower panel: exponents γ of the power law degree distribution of the final network, both as a function of the exponent α of the initial degree distribution. The different thin lines correspond to simulations of the model described for various network sizes: (dotted lines) $N = 2 \times 10^3$, (dashed lines) $N = 10^5$. The thick lines indicate the analytical results. Marks with error bars correspond to the empirical results for the 18 projects.

The exponent β also describes the effort, or social activity, of developers adding new classes to the software. The larger this exponent, the more collaboration is needed to ensure further project growth, either explicitly through interaction between devolopers, or implicitly through documentation. It is a remarkable finding that this exponent is closely related to the other two exponents α, γ, describing a very different "dimension" of the software evolution, namely software engineering. This may shed new light on the underlying principles of software design and software project management.

Finally, we point out that the growth mechanism based on a heterogeneous initial degree distribution was largely overlooked in previous studies of complex network evolution. To the best of our knowledge, only in [22] network growth with power law initial degree distribution was succinctly studied as a limiting case of a local search mechanism; however, its emerging properties were not computed.

We emphasise that the results presented in this letter go beyond the specific application for software growth and could also be used to recast known network dynamics processes in different areas. We argue that this may open a new strand of research for network growth processes in those areas where network dynamics is not based on the addition of links and nodes at a constant (or single-scaled) rate. For examples, network evolution governed by scale-free processes can be related to self-organised criticality, or similar processes. With our approach we have demonstrated how such processes can create network dynamics at all possible scales, generating growth by means of avalanches that completely reshape the outcome network.

* * *

We acknowledge financial support from the Swiss National Science Foundation through grant CR12I1_125298. CJT acknowledges partial support from SBF (Swiss Confederation) through research project C09.0055.

REFERENCES

[1] Barabási A.-L. and Albert R., *Science*, **286** (1999) 509.
[2] Dorogovtsev S., Mendes J. and Samukhin A., *Phys. Rev. Lett.*, **85** (2000) 4633.
[3] Krapivsky P., Redner S. and Leyvraz F., *Phys. Rev. Lett.*, **85** (2000) 4629.
[4] Dorogovtsev S. and Mendes J., *Phys. Rev. E*, **63** (2001) 025101.
[5] Gagen M. and Mattick J., *Phys. Rev. E*, **72** (2005) 016123.
[6] Smith D., Onnela J. and Jones N., *Phys. Rev. E*, **79** (2009) 056101.
[7] Mattick J. and Gagen M., *Science* , **307** (2005) 856.
[8] Faloutsos M., Faloutsos P. and Faloutsos C., *On power-law relationships of the Internet topology*, in *SIGCOMM '05* (ACM New York) 1999, pp. 251–262.
[9] Leskovec J., Kleinberg J. and Faloutsos C., *ACM Trans. Knowl. Discov. Data*, **1**, issue No. 1 (2007) 2.
[10] Challet D. and Lombardoni A., *Phys. Rev. E*, **70** (2004) 046109.
[11] Gorshenev A. A. and Pis'mak Y. M., *Phys. Rev. E*, **70** (2004) 067103.
[12] Valverde S., Ferrer Cancho R. and Sole R. V., *Europhys. Lett.*, **60** (2002) 512.
[13] Myers C. R., *Phys. Rev. E*, **68** (2003) 046116.
[14] Valverde S. and Solé R. V., *Europhys. Lett.*, **72** (2005) 858.
[15] Valverde S. and Solé R. V., *Phys. Rev. E*, **72** (2005) 026107.
[16] Maillart T., Sornette D., Spaeth S. and von Krogh G., *Phys. Rev. Lett.*, **101** (2008) 218701.
[17] Kohring G., *Adv. Complex Syst.*, **12** (2009) 565.
[18] Gamma E. *et al.*, *Design Patterns: Elements of Reusable Object-Oriented Software* (Addison-Wesley) 1995.
[19] Simon H. and Bonini C., *Am. Econ. Rev.*, **46** (1958) 607.
[20] Saichev A., Malevergne Y. and Sornette D., *Theory of Zipf's Law and Beyond* (Springer, Berlin) 2009.
[21] Clauset A., Shalizi C. R. and Newman M. E. J., *SIAM Rev.*, **51** (2009) 661.
[22] Rozenfeld H. D. and ben Avraham D., *Phys. Rev. E*, **70** (2004) 056107.