Sustainable growth in complex networks

Claudio J. Tessone, Markus M. Geipel, F. Schweitzer

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

(Dated: July 9, 2010)

Based on the empirical analysis of the dependency network in 18 Java projects, we develop a novel model of network growth which considers both: an attachment mechanism and the addition of new nodes with a heterogeneous distribution of their initial degree, k_0. Empirically we find that the cumulative degree distributions of initial degrees and of the final network, follow power-law behaviors: $P(k_0) \propto k_0^{1-\alpha}$, and $P(k) \propto k^{1-\gamma}$, respectively. For the total number of links as a function of the network size, we find empirically $K(N) \propto N^\beta$, where β is (at the beginning of the network evolution) between 1.25 and 2, while converging to ~ 1 for large N. This indicates a transition from a growth regime with increasing network density towards a sustainable regime, which prevents a collapse because of ever increasing dependencies. Our theoretical framework is able 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 either dominated 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.

PACS numbers: 89.75.Hc, 05.40.-a, 89.20.Ff

How do *real* networks grow? Nodes and links are not always added at a constant rate. Instead, their numbers could be drawn from a broad distribution, spanning almost the size of the network. This inhomogeneity considerably impacts the network growth, but it was not covered in existing analytical approaches. Hence, this problem is addressed in this Letter. We provide a novel model of network growth which is solved analytically and verified empirically by studying the evolution of several Open Source Software projects.

During its evolution, a network can have a non-linear growth of its set of nodes, or edges. However, many modeling approaches, most notably the preferential attachment, simply assume that (i) at any time step 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 or links are deleted [1–3]. If such assumptions hold, this would result 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 $\eta \simeq 1$ 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 of (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]. But there is, at least for the intermediate observable time scales, also empirical evidence of networks growing with increasing link density, K/N for example the World Wide Web [8].

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 distribution of newly added nodes (instead of adding nodes with the same degree), 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 new assumptions of our network growth model, later. We have used a dataset of 18 Open Source Software (OSS) projects (see Table 1), which are programmed in Java. The complex network consists of *nodes*, which are Java classes (each file corresponds to one class), and *directed links* representing dependencies between these classes. For example, one class can call a function defined in another class, or extend a functionality of another class. 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 universal dynamics underlying such growth processes, this is a remarkable result on its own. For the time dependent evolution of the OSS projects, we can 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 [9, 10].
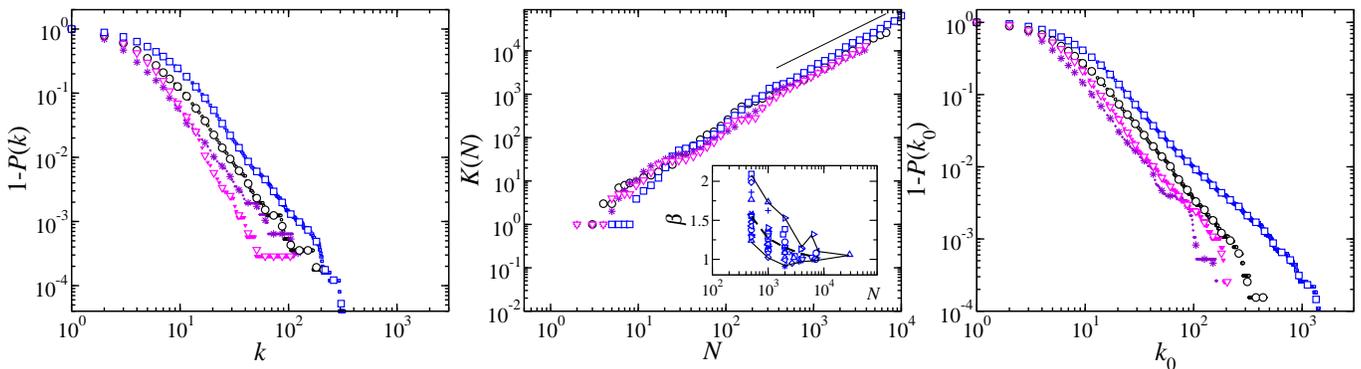
FIG. 1. (Color on-line). (left) Final complementary accumulated degree distribution $P(k) \propto k^{1-\gamma}$, (right) initial complementary cumulated degree distribution $P(k_0) \propto k^{1-\alpha}$, (center) total number of links $K(N) \propto N^{\beta}$ as a function of network size N. Colors indicate four different OSS projects: Architecturware (black circles), Eclipse (blue squares), JEdit (violet stars), Sapia (magenta triangles). See Table I for more details. The small symbols, represent the complete empirical datasets, while the large ones the binned data. The inset in the central panel shows the evolution of β for the 18 projects, during the growth process. In this inset, the different symbols represent the evolution of β for the different projects, while the dashed line represents the median of β for the complete set.

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 $K(\tau) = \sum_{k=1}^{N(\tau)} k\, n(k, \tau)$. 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 remove 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 increasing, i.e. the deletion of nodes is not considered in the empirical analysis. Figure 1 illustrates the empirical results for these quantities by showing four OSS projects of very different size.

Looking at the final complementary cumulative degree distribution $P(k)$, obtained for the maximum N of the project, we clearly identify a power-law $P(k) \propto k^{1-\gamma}$ (left panel of Fig. 1), which is equivalent to a degree distribution $n(k) \propto k^{-\gamma}$ The exponents γ which characterize the *structure* of the *final product* are given in Table I. Dependent on the size of the project, we find values between 2 and 3, with a clear tendency towards values closer to 3. For the growth of the OSS projects (center panel of Fig. 1), we obtain slightly bend curves for the six projects, which indicate that the exponents β changed over time (shown in the inset of the center panel). For every project, the total degree as a function of system size was split into different windows (of size 500) and an estimation of the exponent β was performed for each window. Starting at values between 1.25 and 2, they converge to smaller values of about 1, i.e. we observe a transition from accelerated to sustainable growth. The final values of β are given in Table I. Note that β is a measure of the *activities of the developers*, i.e. it characterizes a social

process. The right panel of Fig. 1 eventually presents the most interesting empirical finding that, different from the above mentioned assumptions about preferential attachment and most modeling 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*. The values found are presented in Tab. I. 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. time t is given by the total number of nodes, $t = N$, or conversely, $\eta = 1$. For the dynamics of the degree distribution we postulate the following rate equation:

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

This is a first order approximation of the dynamics based on the addition/deletion of one node at a time. The term $\delta_{k, k_0(t)}$ in Eq. (2) describes the addition of a new node with an initial degree exactly equal to k. 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. $\text{Prob}[k_0(t) = k] = \min\left((\alpha-1)/k^\alpha, t-1\right)$. For the transition rate of growth processes, $k \to k+1$, we assume

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

This rate is proportional to k, i.e. it is based on preferential attachment. Without that assumption, the process

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)

TABLE I. 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 asymptotic growth of the total number of links as a function of network size.

would result in a single-scale network which is not in accordance with the empirical studies above. The preferential attachment can occur by means of two different processes: The first one occurs if a newly added node with $k_0(t)$ new links to existing nodes, which are selected with a probability proportional to their relative degree k/K. The second process describes the addition of links between existing nodes, where 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,

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

Let us now elucidate the emerging dynamics by splitting it into two different limiting cases. The first one occurs when the growth of the network based on the *addition of nodes* with heterogeneous degree k_0, does not play any role. I.e. k_0 can be set to zero, for every time step. In this case the dynamics is only governed by the *addition/deletion of links* distributed between existing nodes, which follows the preferential attachment/deletion rule. Then, the rate equation (2), in the continuous limit and for large N, can be easily translated into the following Fokker-Planck equation:

$$\partial_t n(k,t) = r\, k\, n(k,t) + \partial_{kk}^2 \left(\sigma^2\, k^2\right) n(k,t) \qquad (4)$$

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

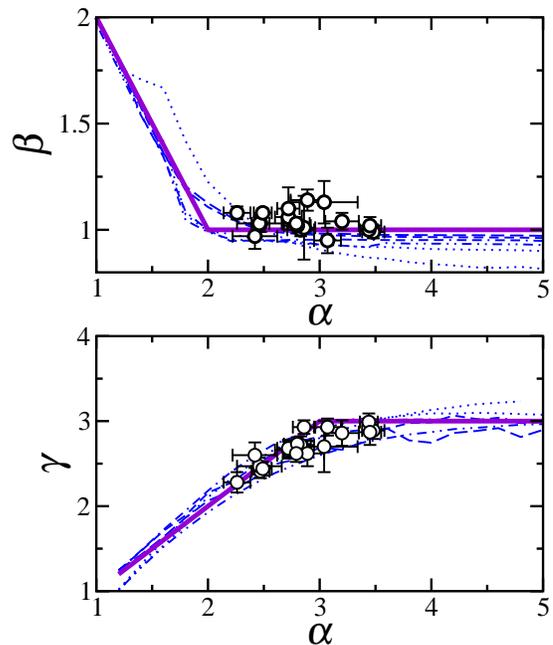$$\dot{k}_i(t) = -r\, k_i(t) + \sigma\, k_i(t)\, \xi_i(t), \qquad (5)$$

FIG. 2. (upper panel) Exponents γ of the power-law degree distribution of the final network, (lower panel) exponents β for the growth of the total number of links 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 ($N = 2 \times 10^3$ –dotted line– to $N = 10^5$ –dashed line–). The thick lines indicate the analytical results of Eqs. (9) and (12). Marks with error bars correspond to the empirical results for the 18 projects, given in Table I.

This describes the known *law of proportional growth* [11, 12], where r is the mean growth (drift) and σ is the variance of the normalized random force $\xi_i(t)$. It is well known [13] that such processes in the long run lead to a power-law distribution $n(k) \propto k^{-\gamma}$, i.e. to Zipf's law for the cumulative degree distribution $P(k) \propto k^{1-\gamma}$, with γ equal to 2. In fact, Zipf's law was empirically confirmed for the *in-degree* distribution of *Linux packages* [9] as well as for *Java projects* [10]. However, the *out-degree* distribution, at least for the latter dataset, clearly follows a *log-normal* distribution. After all, because this limit case only considers the growth of the number of links, but not of the number of nodes, it only provides a limited understanding of real software dynamics (and random addition/deletion processes are only one of many different ways to obtain Zipf's law).

Therefore, in this Letter, we are more interested in the second limiting case which ignores the addition/deletion of links among existing nodes –i.e. σ, r are negligible–, while emphasizing the network growth based on the addition of nodes with a broad initial degree distribution, $g(k_0)$. This assumption is fully justified in the case of broad distributions of initial degrees, as found empirically. This dynamics is fully described by the following

set of equations:

$$\dot{n}(1,t) = \delta_{1,k_0(t)} - \frac{k_0(t)}{K(t)} n(1,t) \tag{6}$$

$$\dot{n}(k,t) = \delta_{k,k_0(t)} + \frac{k_0(t)}{K(t)} \big\{ (k-1)n(k-1,t) - k\, n(k,t) \big\}$$

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 realization $\dot{K}(t) = k_0(t)$ holds. The ensemble average $\langle K(t) \rangle$ over many realizations of the network growth process is then given by:

$$\left\langle \dot{K}(t) \right\rangle = \langle k_0 | k_0 < t \rangle + t \cdot \mathrm{Prob}[k_0(t) > t]. \tag{7}$$

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 after some calculation:

$$\left\langle \dot{K}(t) \right\rangle = \int_1^t dk\, g(k)\, k + t \int_t^\infty dk\, g(k) = \frac{\alpha - 1}{2 - \alpha} + \frac{t^{2-\alpha}}{2 - \alpha}. \tag{8}$$

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

$$\beta = \begin{cases} 3 - \alpha & \text{if } \alpha < 2 \\ 1 & \text{if } \alpha \geq 2 \end{cases}. \tag{9}$$

By applying the ensemble average to Eqs. (6), 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 Eqs. (8-9), we find that

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

By analyzing the solution of Eqs. (8)-(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 $\zeta(\alpha)$ being a normalization constant. Thus, we can rewrite Eqs. (6)

as,

$$\langle \dot{n}(1,t) \rangle = (\alpha - 1) - \frac{\langle n(1,t) \rangle}{\zeta(\alpha)\, t} \tag{11}$$

$$\langle \dot{n}(k,t) \rangle = \frac{(\alpha - 1)}{k^\alpha} + \frac{(k-1)\langle n(k-1,t) \rangle - k\langle n(k,t) \rangle}{\zeta(\alpha)\, t}$$

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 $\langle \dot{n}(k,t) \rangle = t(\alpha - 1)/k^\alpha$. 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 $n(k,t) \propto k^{-3}$. Thus, we have found two different regimes for the final degree distribution, which depend of the exponent of the initial degrees distribution α:

$$\gamma = \begin{cases} \alpha & \text{if } \alpha < 3 \\ 3 & \text{if } \alpha \geq 3 \end{cases}. \tag{12}$$

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: (a) by computer simulations of the full dynamics for various network sizes N and initial conditions (α), (b) by comparison with the empirical findings from the 18 OSS projects. The results are shown in Figure 2. 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. 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-organizing dynamics observed in OSS, which turns an initially accelerated network growth ($\beta > 1$) into a sustainable one ($\beta \to 1$) found for mature projects. This prevents a collapse of the software growth due to non-linearly increasing dependencies between classes. Interestingly, β, which describes the effort (social activity) of developers adding new classes to the software, is found to be 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.

[1] A.-L. Barabási and R. Albert, Science, **286**, 509 (1999).

[2] S. Dorogovtsev, J. Mendes, and A. Samukhin, Phys. Rev. Lett., **85**, 4633 (2000).

[3] P. L. Krapivsky, S. Redner, and F. Leyvraz, Phys. Rev. Lett., **85**, 4629 (2000).

[4] S. N. Dorogovtsev and J. F. F. Mendes, Phys. Rev. E, **63**, 025101 (2001).

[5] M. J. Gagen and J. S. Mattick, Physical Review E, **72**, 016123+ (2005).

[6] D. M. D. Smith, J. P. Onnela, and N. S. Jones, Physical Review E, **79**, 056101+ (2009).

[7] J. S. Mattick and M. J. Gagen, Science, **307**, 856 (2005), ISSN 1095-9203.

[8] M. Faloutsos, P. Faloutsos, and C. Faloutsos, in *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication* (ACM New York, NY, USA, 1999) pp. 251–262.

[9] T. Maillart, D. Sornette, S. Spaeth, and G. von Krogh, Phys. Rev. Lett., **101**, 218701 (2008).

[10] G. A. Kohring, Adv. in Complex Systems, **12**, 565 (2009).

[11] Simon, H. and Bonini, C.P., American Economic Review, **46**, 607 (1958).

[12] A. Saichev, Y. Malevergne, and D. Sornette, *Theory of Zipf's law and beyond* (Springer, Berlin, 2009).

[13] A. Clauset, C.R. Shalizi, and M.E.J. Newman, SIAM Review, **51**, 661 (2009).