Sustainable growth in complex networks

This article has been downloaded from IOPscience. Please scroll down to see the full text article.
2011 EPL 96 58005
(http://iopscience.iop.org/0295-5075/96/5/58005)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 82.130.119.48
The article was downloaded on 10/02/2012 at 13:25

Please note that terms and conditions apply.
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 \( \alpha, \beta, \gamma \), 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.
For our analysis, we have used snapshots of intervals of version control systems which record all changes made. 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 $\alpha$; i.e.

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

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
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, \( k \rightarrow k + 1 \), we assume

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

(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. \( \sigma \) and \( r \) are constants described below. The transition rate corresponding to the deletion of links, \( k \rightarrow k - 1 \), is also assumed to be proportional to the degree of the node,

\[
\omega[k \rightarrow k - 1] = \left( \sigma - \frac{r}{2} \right) k.
\]  

(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:

\[
\frac{\partial n(k, t)}{\partial t} = -\frac{\partial}{\partial k} \left[ \sigma k^2 n(k, t) \right] + \frac{\partial^2}{\partial k^2} \left[ \sigma k^2 n(k, t) \right]
\]  

(5)

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

\[
\dot{k}_i(t) = -r k_i(t) + \sigma k_i(t) \xi_i(t).
\]  

(6)

This describes the known law of proportional growth [19,20], where \( r \) is the mean growth (drift) and \( \sigma \) 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 \( \gamma \) 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 \( \sigma \) 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:

\[
\begin{align*}
\dot{n}(1, t) &= \delta_{1,i(0)} - \frac{k_0(t)}{K(t)} n(1, t), \\
\dot{n}(k, t) &= \delta_{k,i(0)} + \frac{k_0(t)}{K(t)} \{ (k-1)n(k-1, t) - kn(k, t) \}
\end{align*}
\]

(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 \( K(t) = k_0(t) \) holds. The ensemble average \( \langle K(t) \rangle \) over many realisations of the network growth process is then given by

\[
\langle K(t) \rangle = \langle k_0 \rangle [\langle k_0 < t \rangle + t \cdot \text{Prob}[k_0(t) > t]].
\]

(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

\[
\langle K(t) \rangle = \frac{\alpha - 1}{2 - \alpha} + \frac{t^{2-\alpha}}{2-\alpha}.
\]

(9)

Asymptotically, we find that the total number of links grows in time or with network size \( t = N \), respectively, as a power law, \( K(t) \propto t^2 \), 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 \( \langle \delta_{k,i(0)} \rangle = \text{Prob}[k = k_0(t)] = (\alpha - 1)/k^\alpha \) and similar arguments as in eq. (9), we find that

\[
\langle k_0(t) \rangle = \frac{t^{2-\alpha}}{2-\alpha} + \frac{\alpha - 1}{\alpha - 2}.
\]

(10)

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

\[
\begin{align*}
\dot{n}(1, t) &= (\alpha - 1) - \frac{n(1, t)}{\zeta(\alpha) t}, \\
\dot{n}(k, t) &= \frac{\alpha - 1}{k^\alpha} + \frac{(k-1)n(k-1, t) - kn(k, t)}{\zeta(\alpha) t}.
\end{align*}
\]

(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 \( \alpha \) is small, the first case dominates and the expected degree distribution is simply given by

\[
\langle n(k, t) \rangle = \frac{(\alpha - 1)}{k^\alpha} t.
\]

(12)

On the other hand, if \( \alpha \) 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 generalised central limit theorem, is expected to occur for \( \alpha > 3 \) 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 \( \alpha \): \( \gamma = \alpha \) if \( \alpha < 3 \); \( \gamma = 3 \), otherwise.

To conclude, our analytical approach has provided a firm relation between the three different exponents \( \alpha \), \( \beta \), \( \gamma \), 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 (\( \beta \rightarrow 1 \)). 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.
on the addition of links and nodes at a constant (or in those areas where network dynamics is not based on a new strand of research for network growth processes in different areas. We argue that this may open and could also be used to recast known network dynamics beyond the specific application for software growth 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 CR121.1/25298. CJT acknowledges partial support from SBF (Swiss Confederation) through research project C09.0055.

REFERENCES

[18] Gamma E. et al., Design Patterns: Elements of Reusable Object-Oriented Software (Addison-Wesley) 1995.