Dynamics of Citation Networks

Gábor Csárdi[1,2]

[1] Department of Biophysics, KFKI Research Institute for Particle and Nuclear Physics of the Hungarian Academy of Sciences, Budapest, Hungary
[2] Center for Complex Systems Studies, Kalamazoo College, Kalamazoo, MI 49006, USA
csardi@rmki.kfki.hu

Abstract. The aim of this paper is to give theoretical and experimental tools for measuring the *driving force* in evolving complex networks. First a discrete-time stochastic model framework is introduced to state the question of how the dynamics of these networks depend on the properties of the parts of the system. Then a method is presented to determine this dependence in the possession of the required data about the system. This measurement method is applied to the citation network of high energy physics papers to extract the in-degree and age dependence of the dynamics. It is shown that the method yields close to "optimal" results.

1 Introduction

The network concept is an abstract representation. A simple network (or graph, the two are the same for our purposes) is simply a homogeneous relation over a set. The relation can be symmetric (undirected networks) or asymmetric (directed networks). While this is an adequate definition of a network, usually we imagine a network as an interconnected set of vertices (also called nodes), while the connections are called edges or arcs. In a neural network the vertices represent neurons and the edges the synapses between them; this network is clearly asymmetric, the synapse 'leads' from the presynaptic cell to the postsynaptic one. In a citation network, the vertices represent ('are') scientific papers published in journals and the edges are citations from one paper to another, forming again a directed network. In a collaboration network two vertices representing researchers are connected by an edge if they have published at least (say) one joint paper in a journal, this network is an undirected one.

There has been an upsurge in the field of complex networks recently; networked representations of various complex systems have shed light to a number of structural and dynamical phenomena. The main advantage of the network schema is that its simplicity makes it universal: every large enough system consists of many – structurally, dynamically and/or functionally interconnected parts. For recent reviews written by researchers in different fields see [1,2,3,4].

In this work we address evolving networks, and study the dynamical process of adding and removing vertices and edges to/from the the network. Particularly we

S. Kollias et al. (Eds.): ICANN 2006, Part I, LNCS 4131, pp. 698–709, 2006.

are interested in the question of how the structural and non-structural properties on the vertices determine the place of the next edge addition.

There have been a number of network evolution models in the literature recently, the most successful being the preferential attachment model proposed by Barabási and Albert [5]. They suggest a simple mechanism in which the rate for attaching new edges to a node is proportional to its number of adjacent edges at each time step. This model is thought to be valid for very different kinds of networks (showing the ease of the universal network representation) based on indirect evidence: the scale-free degree distribution. It is observed that in many networks the distribution of the vertex degree (which is simply the number of adjacent edges for a vertex) is a power-law distribution; and the BA-model is known to generate power-law degree distributions [6], so it is likely (or at least possible) that this simple mechanism is at work in many networks. It is shown however that the scale-free degree distribution can be obtained without preferential attachment, by assuming vertex intrinsic fitness, see [7]. It is also true that there may be several *underlying* causes producing preferential attachment [8,9]. Only a few studies addressed the direct observation, ie. somehow measuring the actual attachment probabilities in the evolving network as a function of the vertex degree or vertex fitness, see [10,11,12] for examples.

This neglect is partially caused by the lack of data. For calculating the actual degree distribution of a network we only need to know the *current* structure of it, ie. the binary relation defining which vertices a given vertex connects to. For studying the process of vertex and edge addition and deletion however we need to know the structure of the network at any time in the past. (At least this would be the ideal case.) It comes not as a surprise that we usually don't have this data, except in a few cases. This indicates that the rare dynamical data is very important and can be used to validate various network evolution models. Our work discussed here serves as an example for such a study.

This paper is organized as follows. In Sect. 2 we introduce a model framework and a measuring method for extracting the dependence of the network dynamics on the hypothetical dynamical parameters. In Sect. 3 we show two applications for the model and method: measuring the dynamics of scientific citation networks, and predicting the number of future citations for scientific citation networks. Finally in Sect. 4 we discuss our results and other possible applications.

2 Methods

The networked representation of a dynamic complex system is an evolving graph: vertices join to the system, they form new connections, some old connections break and perhaps some vertices are removed from the network. In each time step the network has a configuration in which the vertices and edges exhibit various structural properties. Further on, the vertices and edges may also exhibit some intrinsic properties we don't intend to ignore: in a neural network some neurons

are pyramidal cells, others are interneurons and this distinction is important for most purposes.

A very natural question is the following: what structural and/or intrinsic properties determine the evolution of a given network? Another question coming hand in hand with this: how is it possible to describe the form of the dependence? (If it is possible at all.) In the rest of this section we will give a model framework and method for answering these questions in some special cases.

Let us focus on the simplest kind of evolving networks first: citation networks. We do this for two reasons. First, citation networks are simple in the sense that all outgoing edges of a vertex are added to the network right after adding the vertex itself, in the same time step. Second, there is data available for citation networks of scientific papers.

A number of important structural properties may play significant roles in the evolution of a particular citation network: the in-degree of the vertices, their transitivity (ie. if every vertex citing vertex v also cites vertex w so far, then it is likely that this will happen in the future as well). Some intrinsic properties of the vertices are also thought to be important: the topic of a paper, since it is likely that two topically close papers will cite each other; or the age of the papers since it is a reasonable assumption that out-of-date (or common knowledge) papers are not or only rarely cited.

2.1 Preferential Attachment

Let us now define the framework in which our questions can be stated formally.

The first structural property we will address is the in-degree of the vertices. Let us assume that the probability that at time step t an outgoing edge (e) of a newly added v vertex will cite a given w vertex depends on the in-degree of w, and the in-degree of other vertices in the network:

$$P[e \text{ cites } w](t) = \frac{A(d_w(t))}{\sum_{i \in V(t)} A(d_i(t))} \ . \tag{1}$$

Here $d_w(t)$ is the in-degree of vertex w in time step t and $V(t)$ is the set of all vertices in time step t. The $A(\cdot)$ attachment kernel function defines the dependence of the network dynamics on the in-degree of the vertices. In this simple framework this function stochastically governs the network evolution. The preferential attachment model suggests that for many networks this function is simply $A(k) = k + 1$. There are also other models which fit into this framework, see [13,14,15].

Similarly, the probability that in time step t an e edge of a newly added v vertex cites *any* other vertex with in-degree k is given by:

$$P[e \text{ cites a } k \text{ in-degree vertex}](t) = P_e(k) = \frac{N_k(t)A(k)}{\sum_{i \in V(t)} A(d_i(t))} \ . \tag{2}$$

$N_k(t)$ is the number of k-degree nodes in the network at time step t.

From this formula we can extract $A(k)$:

$$A(k) = \frac{P_e(k)S(t)}{N_k(t)} \tag{3}$$

by using the notation $S(t) = \sum_{i \in V(t)} A(d_i(t))$. From the data we can estimate $P_e(k)$, so if we manage to determine $S(t)$ then $A(k)$ can be estimated as well. For $S(t)$ we can use the following simple iterative approach: first we assume that $S(t)$ is constant and estimate $A(k)$ for each k. Then by using this estimation we calculate the next approximation of $S(t)$ which in turn allows us to better estimate $A(k)$. While the convergence of this iteration is hard to prove, in practice it converges fast.

In Sect.3 we show applications for the in-degree dependence of the network dynamics in scientific citation networks.

2.2 Preferential Attachment and Aging

Let us now assume that an additional intrinsic vertex property, the *age* of the vertex, also contributes to the network evolution. For simplicity from now on we measure "time" by the addition of the new vertices, ie. in each time step a single vertex is added to the network; we denote vertices by the time step of their addition, ie. vertex 1 is added in the first time step, vertex 2 in the second, etc. This implies that in time step t the age of vertex i is simply $t - i$.

Similarly to the previous section the probability that edge e of vertex v added in time step t cites vertex w is given by

$$P[e \text{ cites } w] = \frac{A(d_w(t), l_w(t))}{\sum_{i \in V(t)} A(d_i(t), l_i(t))} \ . \tag{4}$$

The probability that edge e of vertex v added in time step t cites some vertex with in-degree k and age l is

$$P[e \text{ cites a } k \text{ in-degree, } l \text{ age vertex}] = \frac{A(k,l)N_{k,l}(t)}{S(t)} \ . \tag{5}$$

Using the data for estimating $P_e(k,l)$ and the iteration technique introduced in the previous section we can extract $A(k,l)$, the function governing the evolution of the network.

2.3 Validating the Method

For validating this measurement method and software, we've applied it to various toy networks generated by different attachment rules, ie. different built-in $A(k)$ and $A(k,l)$ functions.

To validate the in-degree based method we've generated networks by the Barabási-Albert model and compared the measured $A(k)$ function to the expected linear dependence. These test networks had 300,000 nodes each having

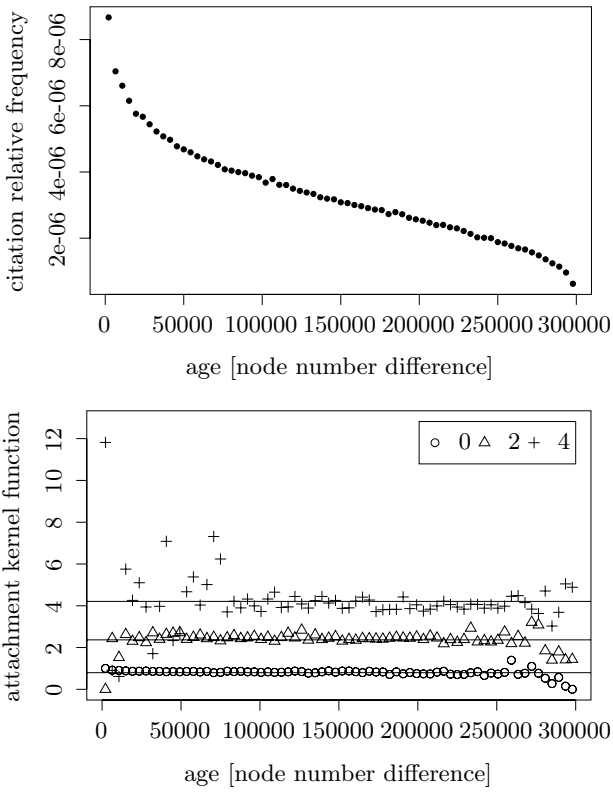

Fig. 1. The naive (upper) and non-naive (lower) methods for measuring the age dependence of the $A(k, l)$ function. The network was generated according to the Barabási-model, has 300,000 nodes and the out-degree of each node is 2. The age axes are binned into 70 units. The lower plot shows the measured $A(k, l)$ functions for various k values. The horizontal lines were added by least square fitting the data points, they can be considered as the "correct" values of the $A(k, l)$ function.

out-degree two. The measurement yielded the expected 1.0 exponent with minimal error (± 0.05).

Next we've checked the in-degree and age based method and software by similar toy networks. The measurement method very well reproduced the expected attachment rules. These experiments however have also shown that the method cannot predict the "rare" events in the evolution. As there are almost never any young nodes with high in-degree in the network, the $A(k, l)$ function for large k and small l values cannot be estimated well.

Although one might argue that for the age dependence of the $A(k, l)$ function a simpler approach could be used, we show here that this is not the case. A naive approach would simply consider the distribution of the age differences (citation lags) between the citing and the cited node as the age dependent component of $A(k, l)$, however this is clearly biased: small citation lags are overrepresented in

the network because of two reasons. The first is that young nodes are more likely to be cited when the network is still small because there is less competition in the network. If older nodes are also present then the competition in higher as the network is also bigger. Second, young nodes have simply more chance to get cited, as they are present in small and big networks as well.

Figure 1 shows the two types of measurement of the age dependence of $A(k, l)$ for a simple Barabási network. While it is clear that there is no age dependence in this model, the histogram of the citation lags does not show a horizontal line. Our proposed measurement method correctly finds that $A(k, l)$ is independent of the age of the nodes.

3 Applications

3.1 The Pace of Science

In this section we apply the method described in the previous one to a scientific citation network, consisting of 28632 high energy physics papers with 367790 directed edges among them. This data is available online from the homepage of the 2003 KDD Cup (`http://www.cs.cornell.edu/projects/kddcup/datasets.html`).

First we've cleared up the dataset by removing forward citations. A forward citation means that a paper cites a more recent one. This is possible either because of errors in the database or because some papers were updated (with new citations) after their first submission without changing their original submission date.

Then the dynamics of the network (ie. the $A(\cdot)$ function) was measured in terms of the in-degree and the age of the nodes. The age of the papers was simply defined by assigning numbers to them in the order their first submission date and binning these numbers into 70 units.

After the extraction of the $A(k, l)$ function the measured data has shown that the effects of k and l can be separated, and $A(k, l)$ can be written in the form

$$A(k, l) = A_k(k) \cdot A_l(l) \ . \tag{6}$$

This separation supports the assumptions made by various network models with aging, see works by [16] and [17]. The measured $A_k(k)$ and $A_l(l)$ functions can be seen in Figs. 2 and 3. They can be well fitted by $A_l(l) = l^{-\beta}$ and $A_k(k) = k^{\alpha} + 1$, with $\alpha = 1.11$ and $\beta = 1.13$. This α value is close to the celebrated linear preferential attachment phenomenon, thought to be universal, although rarely measured directly.

The fact that the β exponent is close to one shows that ceteris paribus the "importance" of a paper is inversely proportional to its age. This defines the "pace" of science.

3.2 Citation Prediction

The ACM Special Interest Group on Knowledge Discovery and Data Mining organizes a conference each year and together with the conference they also

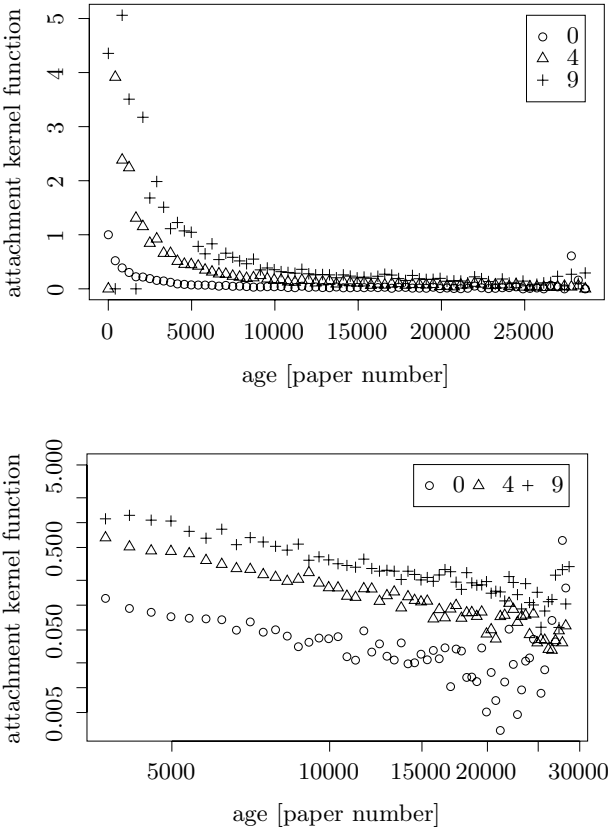

Fig. 2. The age dependence of the attachment kernel function of the high energy physics network for various in-degrees. The upper plot has linear, the lower one logarithmic axes. The lower plot clearly shows that the aging is well described by a power-law decrease independently of the degree.

host a data mining competition called KDD Cup. In 2003 the first task of the KDD Cup was to predict the citations to the papers in the high energy physics database. This database contains high energy papers submitted to the arXiv e-print archive between 1992 and July 31, 2003. The deadline for the KDD Cup submission was before April 30, 2003 and the citations made by papers in the next three months had to be predicted.

The evaluation of the prediction algorithms was done by considering only papers receiving at least six citations during the period February 1, 2003 – April 30, 2003. For these papers first the target vector, the difference between the citations received between May and August and between February and May were calculated. The specific task was the prediction of this vector. The error of the prediction was simply defined by the sum of the absolute value of the difference of the prediction and the target vector.

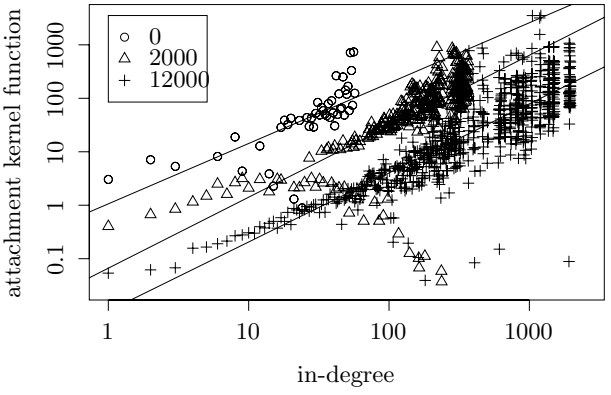

Fig. 3. The degree dependence of the attachment kernel function for various node ages. The lines are simple least square fits for the data points. The axes are logarithmic. The plot shows that the in-degree dependent part of the $A(k, l)$ function can be reasonable well estimated by an increasing power-law function, independently of the age of the nodes.

While the method in the previous section is not developed for citation prediction, is can be used for that in the following way. We can measure the dynamics (ie. the $A(\cdot)$ function) of the network up to *now* and assuming that this function will be the same in the future we can simulate the growth of the network according to the measured dynamics and see a possible realization of how the network will look like (say) three months later. By generating many realizations and taking the average number of citations a node received in these realizations we can predict the "average" expected evolution of the network.

Another important reason to do the prediction task with our proposed method is that we can compare the error of the measured $A(\cdot)$ function to other $A(\cdot)$ functions to evaluate it. If a given $A_1(\cdot)$ function proves to be a better predictor than another $A_2(\cdot)$ attachment kernel that would mean that the former one is based on more relevant properties than the latter.

At the 2003 KDD Cup, the error of the winner algorithm was 1329. The totally random network evolution, when each new node connects to a number of randomly selected nodes yields on the average an error of 3463. This value was obtained by averaging hundred totally random realizations. These error values can be used as baselines to place the error of the predictions of our method.

First we measured the $A(\cdot)$ function based on the in-degree of the nodes solely and found that the

$$A(k) = k^{\alpha} + 1 \tag{7}$$

form gives a reasonable good fit with the measured data. We fitted this form by a simple weighted least square method and got $\alpha \approx 0.85$. The prediction with this $A(k)$ function yielded an error about $2473.51(\pm 4.39)$. These values were obtained by generating 100 realizations five times, the error is simply the standard deviation of the five predictions.

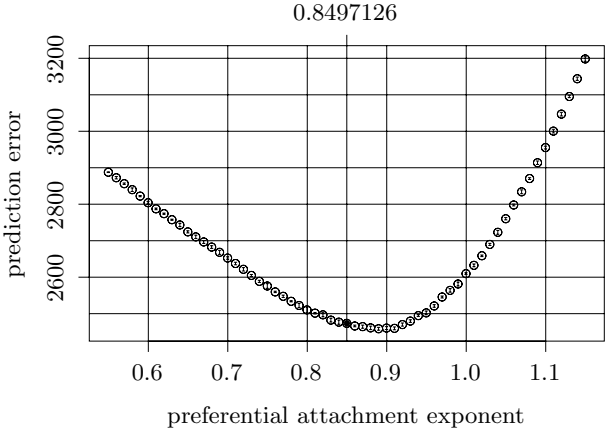

Fig. 4. Prediction error for different α values in (7). The plot was obtained by running five times 100 realizations for each α value, the error bars show the standard deviation of the five predictions. The measured 0.85 exponent is close to the optimal 0.89 value.

To evaluate our dynamics measurement method we've calculated predictions with other α exponents as well, and found that the $\alpha = 0.85$ value is very close to the "optimal" exponent, optimal in terms of the error of this prediction.

Instead if using solely the in-degree as the predictor, now we will also add the age of the nodes, and by applying the measurement method described in the previous section we measure the $A(k, l)$ function (as before k being the in-degree and l being the age of a node) governing the dynamics of the network. The measured $A(k, l)$ function can be reasonably well fitted by the following form:

$$A(k, l) = (k^{\alpha} + 1)\, l^{-\beta} \ . \tag{8}$$

This form assumes that the effect of in-degree and age can be separated, our data supports this assumption. By fitting this form using weighted least square fits we arrive to the exponents: $\alpha \approx 1.14$ and $\beta \approx 1.14$. By using these values in generating possible realizations of the HEP network for the prediction we get a prediction error 1732.76 ± 6.19. The fact that this prediction is much better than the "in-degree-only" one, indicates that the age of the nodes makes an important contribution to the edge-dynamics of the evolving network.

Note that the exponent of the preferential attachment is lower if we don't use the age of the papers as a property, $\alpha_k \approx 0.85$ versus $\alpha_{k,l} \approx 1.14$. This is clearly because in the former the effect of the aging is "built in" into the preferential exponent and since aging works *against* preferential attachment it makes the exponent smaller. Some works suggest that the preferential attachment mechanism can be present even in network not showing the scale-free degree distribution because there is another, opposite effect working in the system, such as limits for the number of edges a node can acquire or because the nodes lose their "attractiveness" by getting older, ie. aging, see [18]. To our knowledge the work

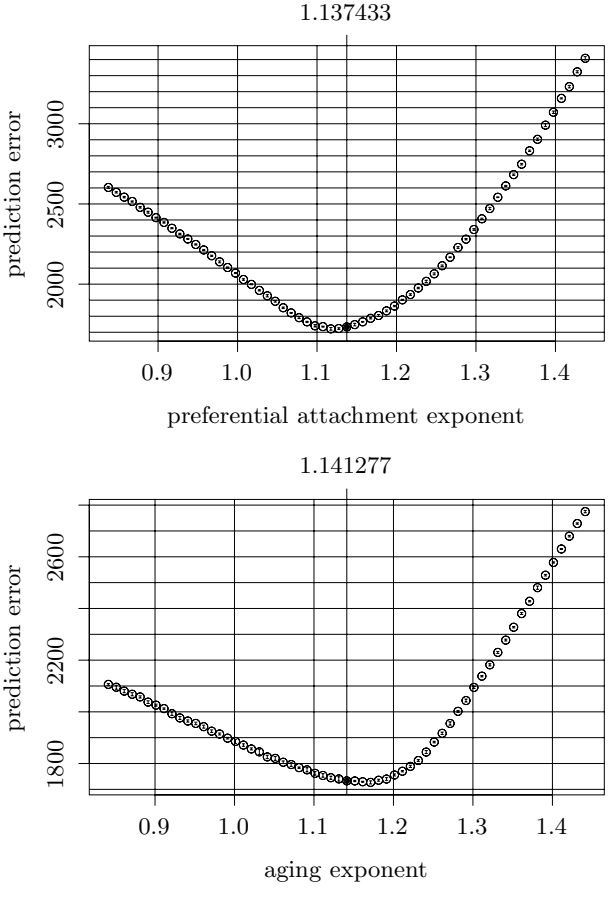

Fig. 5. Prediction error for different preferential attachment exponents (α, upper plot) and aging exponents (β, lower plot). For both exponents the dynamics measurement method gives solutions close to the optimal ones.

presented in this paper is the first one giving experimental evidence for this assumption.

4 Discussion

We have presented a model framework and a measurement method for defining and determining the dynamics of citation networks based on the properties of their nodes.

We've applied this method to a network of high energy physics papers and extracted the $A(k, l)$ function which stochastically governs the evolution of the network in terms of the in-degree and age of the nodes. Without assuming any favored form for this function we found that it can be estimated as the product

of the in-degree dependent $A_k(k)$ and the age-dependent $A_l(l)$ function. The in-degree dependent part shows slightly superlinear preferential attachment while the age-dependent part shows power-law decrease.

We've evaluated the results given by the measurement method by predicting the citations received by important papers in the last three months of the high energy physics papers database and found that the measured preferential attachment and aging exponents are close to the "optimal".

We believe that the framework and method presented in this paper is a useful tool for researchers of any field interested in the evolution of complex systems. Also, it can be generalized for general evolving networks with node and edge additions and deletions, our experiments show promising results in this direction.

The citation prediction study presented here can be a general way for evaluating the description of a system based on various properties, just like we've shown that adding the age of the nodes to the considered properties resulted a much better citation prediction.

Acknowledgement

The authors would like to thank Jan Tobochnik, Katherine J. Strandburg, Péter Érdi, László Zalányi and Tamás Kiss for their cooperation. This work was funded in part by the EU FP6 Programme under grant numbers IST-4-027173-STP and IST-4-027819-IP and by the Henry R. Luce Foundation.

References

1. Newman, M.E.J.: The structure and function of complex networks. SIAM Review **45** (2003) 167–256
2. Watts, D.J.: The "new" science of networks. Annual Review of Sociology **30** (2004) 243–270
3. Barabási, A.L., Oltvai, Z.N.: Network biology: Understanding the cells's functional organization. Nature Reviews Genetics **5** (2004) 101–113
4. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.U.: Complex networks: Structure and dynamics. Physics Reports **424** (2006) 175–308
5. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**(5439) (1999) 509–512
6. Mitzenmacher, M.: A brief history of generative models for power law and lognormal distributions. Internet Mathematics **1** (2004) 226–251
7. Caldarelli, G., Capocci, A., Rios, P., Muñoz, M.: Scale-free networks from varying vertex intrinsic fitness. Physical Review Letters **89** (2002) 258702
8. Kleinberg, J.M., R., K.S., Raghavan, P., Rajagopalan, S., Tomkins, A.: The web as a graph: Measurements, models and methods. In: Proceedings of the International Conference on Combinatorics and Computing, no. 1627 in Lecture Notes in Computer Science, Springer (1999)
9. Berger, N., Borgs, C., Chayes, J.T., D'Souza, R.M., Kleinberg, R.D.: Competition-induced preferential attachment. In: Proceedings of the 31st International Colloquium on Automata, Languages and Programming. (2004) 208–221

10. Jeong, H., Néda, Z., Barabási, A.L.: Measuring preferential attachment for evolving networks. Europhys. Lett. **61** (2003) 567–572
11. Redner, S.: Citation statistics from 110 years of physical review. Physics Today **58** (2005) 49
12. Roth, C.: Measuring generalized preferential attachment in dynamic social networks. arxiv:nlin.AO/0507021 (2005)
13. Krapivsky, P.L., Redner, S.: Organization of growing random networks. Phyisical Review E **63** (2001) 066123
14. Ergun, G., Rodgers, G.J.: Growing random networks with fitness. Physica A **303** (2002) 261–272
15. G., B., Barabási, A.L.: Competition and multiscaling in evolving networks. Europhysics Letters **54** (2001) 436–442
16. Dorogovtsev, S.N., Mendes, J.F.F.: Evolution of networks with aging of sites. Phys. Rev. E **62**(2) (2000) 1842–1845
17. Zhu, H., Wang, X., Zhu, J.Y.: Effect of aging on network structure. Phys. Rev. E **68** (2003) 056121
18. Amaral, L.A.N., Scala, A., Barhélémy, M., Stanley, H.E.: Classes of small-world networks. Proc. Natl. Acad. Sci. USA **97**(21) (2000) 11149–11152