

Chapter 1

Estimating the dynamics of kernel-based evolving networks

Gábor Csárdi

Center for Complex Systems Studies, Kalamazoo, MI, USA and
Department of Biophysics, KFKI Research Institute for Particle
and Nuclear Physics of the Hungarian Academy of Sciences,
Budapest, Hungary
csardi@kzoo.edu

Katherine Strandburg

DePaul University – College of Law, Chicago, IL, USA

László Zolányi

Department of Biophysics, KFKI Research Institute for Particle
and Nuclear Physics of the Hungarian Academy of Sciences,
Budapest, Hungary

Jan Tobochnik

Department of Physics and Center for Complex Systems Studies,
Kalamazoo College, Kalamazoo, MI, USA

Péter Érdi

Center for Complex Systems Studies, Kalamazoo College,
Kalamazoo, MI, USA

In this paper we present the application of a novel methodology to scientific citation and collaboration networks. This methodology is designed for understanding the governing dynamics of evolving networks and relies on an *attachment kernel*, a scalar

function of node properties, which stochastically drives the addition and deletion of vertices and edges. We illustrate how the kernel function of a given network can be extracted from the history of the network and discuss other possible applications.

1.1 Introduction

The network representation of complex systems has been very successful. The key in this success is universality in at least two senses. First, the simplicity of representing complex systems as networks makes it possible to apply network theory to very different systems, ranging from the social structure of a group to the interactions of proteins in a cell. Second, these very different networks show universal structural traits such as the small-world property and the scale-free degree-distribution.

Usually it is assumed that the life of most complex systems is defined by some – often hidden and unknown – underlying governing dynamics. These dynamics are the answers to the question ‘How does it work?’ and a fair share of scientific effort is taken to uncover this dynamics.

In the network representation the life of a (complex or not) system is modeled as an evolving graph: sometimes new vertices are introduced to the system while others are removed, new edges are formed, others break and all these events are governed by the underlying dynamics.

This paper is organized as follows. In Section §1.2 we define a framework for studying the dynamics of two types of evolving networks and show how this dynamics can be measured from the data. In Section §1.3 we present two applications and finally in Section §1.4 we discuss our results and other possible applications.

1.2 Modeling evolving networks by attachment kernels

In this section we introduce a framework in which the underlying dynamics of evolving networks can be estimated from knowledge of the time dependence of the evolving network.

This framework is a discrete time model, where time is measured by the different events happening in the network. An event is a structural change: vertex and/or edge additions and/or deletions. The interpretation of an event depends on the system we’re studying; see Section §1.3 of this paper for two examples.

The basic assumption of the model is that edge additions depend on some properties of the vertices of the network. This property can be a structural one such as the degree of a vertex or its clustering coefficient but also an intrinsic one such as the age of a person in a social network or her yearly net income. The model is independent of the *meaning* of these properties.

The vertex properties drive the evolution of the network stochastically through an attachment kernel, a function giving the probabilities for any new edges which might be added to the network.

In this paper we specify the model framework for two special kinds of networks: citation and non-decaying networks, more general results will be published in forthcoming publications.

1.2.1 Citation networks

Citation networks are special evolving networks. In a citation network in each time step (event) a single new node is added to the network together with its edges (citations). Edges between “old” nodes are never introduced and there are no edge or vertex deletions either.

For simplicity let us assume that the $A(\cdot)$ attachment kernel depends on a only one property of the potentially cited vertices, their degree. (The formalism can be generalized easily to include other properties as well.) We assume that the probability that at time step t an edge e of a new node will attach to an old node i with degree d_i is given by

$$P[e \text{ cites } i] = \frac{A(d_i(t))}{\sum_{k=1}^t A(d_k(t))} \quad (1.1)$$

The denominator is simply the sum of the attachment kernel functions evaluated for every node of the network in the current time step.

With this simple equation the model framework for citation networks is defined: we assume that in each time step a single new node is attached to the network and that it cites other, older nodes with the probability given by (1.1).

For a given citation network we can use this model to estimate the form of the kernel function based on data about the history of a network. In this paper we only give an overview of this estimation process, please see [] for the details.

Based on (1.1) the probability that an e edge of a new node at time t cites an old node with degree d is given by

$$P[e \text{ cites a } d\text{-degree node}] = P_e(d) = \frac{A(d)N_d(t)}{S(t)}, \quad S(t) = \sum_{k=1}^t A(d_k(t)) \quad (1.2)$$

$N_d(t)$ is the number of d -degree nodes in the network in time step t . From here we can extract the $A(d)$ kernel function:

$$A(d) = \frac{P_e(d)S(t)}{N_d(t)} \quad (1.3)$$

If we know $S(t)$ and $N_d(t)$, then by estimating $P_e(d)$ based on the network data we have an estimate for $A(d)$ via (1.3), and by doing this for each edge and d degree, in practice we can have a reasonable approximation of the $A(d)$ function for most d values. (Of course we cannot estimate $A(d)$ for those degrees which were never present in the network.)

It is easy to calculate $N_d(t)$ so the only piece missing for the estimation is that we need $S(t)$ as well, however this is defined in terms of the measured $A(d)$ function. We can do an iterative approach to make better and better approximations for $A(d)$ and $S(t)$. First we assume that $S_0(t) = 1$ for each t and measure $A_0(d)$ which can be used to calculate the next approximation of $S(t)$, $S_1(t)$ yielding $A_1(t)$ via the measurement, etc. In practice this procedure converges fast, after five iterations the difference between successive $A_n(d)$ and $A_{n+1}(d)$ estimations is very small.

1.2.2 Non-decaying networks

Non-decaying networks are more general than citation networks because connections can be formed between older nodes as well. It is still true however that neither edges nor nodes are ever removed from the network.

Similarly to the previous section, we assume that the attachment kernel depends on the degree of the vertices, but this time on the degree of both vertices involved in their potential connection. The probability of forming an edge between nodes i and j in time step t is given by

$$P[i \text{ and } j \text{ will be connected}] = \frac{A(d_i(t), d_j(t))}{\sum_k^{N(t)} \sum_{l \neq k}^{N(t)} (1 - a_{kl}(t)) A(d_k(t), d_l(t))} \quad (1.4)$$

The denominator is the sum of the attachment kernel function applied to all possible (not yet realized) edges in the network. $a_{kl}(t)$ is 1 if there is an edge between nodes k and l in time step t and 0 otherwise.

By using a similar argument as in the previous section the measurement can be done by estimating $A(d^*, d^{**})$ via

$$P[e \text{ connects } d^* \text{ and } d^{**} \text{ degree nodes}] = P_e(d^*, d^{**}) = \frac{A(d^*, d^{**}), N_{d^*, d^{**}}(t)}{S(t)}, \quad (1.5)$$

$N_{d^*, d^{**}}(t)$ is the number of not yet realized edges between d^* and d^{**} degree nodes in time step t , and

$$S(t) = \sum_k^{N(t)} \sum_{l \neq k}^{N(t)} (1 - a_{kl}(t)) A(d_k(t), d_l(t)) \quad (1.6)$$

$$A(d^*, d^{**}) = \frac{P_e(d^*, d^{**}) S(t)}{N_{d^*, d^{**}}(t)} \quad (1.7)$$

For approximating $S(t)$ a similar iterative approach can be used as the one introduced in the previous section.

1.3 Applications

In this section we briefly present results for two applications for the model framework and measurement method. For other applications and details see [].

1.3.1 Preferential attachment in citation networks

The preferential attachment model [1] gives a mechanism to generate the scale-free degree-distribution often found in various networks. In our framework for citation networks it simply means that the kernel function linearly depends on the degree:

$$A(d) = d + a, \quad (1.8)$$

where a is a constant.

By using our measurement method, it is possible to measure the kernel function based on node degree for various citation networks and check whether they evolve based on this simple principle.

Let us first consider the network of high-energy physics papers from the arXiv e-print archive. We used data for papers submitted between January, 1992 and July, 2003, which included 28632 papers and 367790 citations among them. The data is available online at <http://www.cs.cornell.edu/projects/kddcup/datasets.html>.

First we've applied the measurement method based on the node degree to this network and found that indeed, the attachment kernel of the network is close to the one predicted by the preferential attachment model, that is

$$A_{\text{HEP}}^*(d) = d^{0.85} + 1 \quad (1.9)$$

gives a reasonably good fit to the data. See the measured form of the kernel in Fig. 1.1.

The small exponent for d is in good agreement with the fact that the degree distribution of this network decays faster than a power-law.

Next, we've applied the measurement method by using two properties of the potentially cited nodes: their degree and age, the latter is simply defined as the difference of the current time step and the time step when the node was added. We found that the two variable $A(d, a)$ attachment kernel has the following form:

$$A_{\text{HEP}}^{**}(d, a) = (d^{1.14} + 1) a^{-1.14}. \quad (1.10)$$

This two-variable attachment kernel gives a better understanding of the dynamics of this network: the citation probability increases about linearly with the degree of the nodes and decreases as a power-law with their age. Note that these two effects were both present in the degree-only dependent A^* attachment kernel, this is why the preferential attachment exponent was smaller there ($0.85 < 1.14$).

Similar results were obtained for the citation network of the US patents granted between 1975 and 1999 containing 2,151,314 vertices and 10,565,431 edges:

$$A_{\text{patent}}^*(d) = d^{0.79}, \quad A_{\text{patent}}^{**}(d, a) = (d^{1.2} + 1) a^{-1.6}. \quad (1.11)$$

These two studies show that the preferential attachment phenomenon can be present in a network even if it does not have power-law degree-distribution because there is another process – *aging* in our case – which prevents nodes from gaining very many edges.

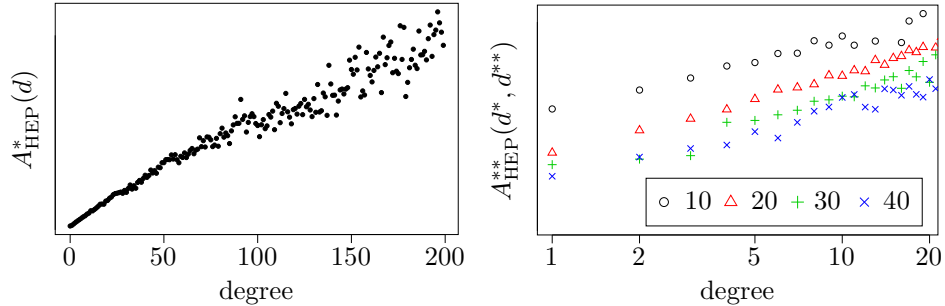


Figure 1.1: The two measured kernel functions for the HEP citation network. The left plot shows the degree dependent kernel, the right the degree and age dependent kernel. On the right plot four sections along the degree axis are shown for different vertex ages.

1.3.2 The dynamics of scientific collaboration networks

In this section we briefly present the results of applying our methods to a non-decaying network: the cond-mat collaboration network. In this network a node is a researcher who published at least one paper in the arXiv cond-mat archive between 1970 and 1997 (this is the date when the paper was submitted to cond-mat, not the actual publication date, but most of the time these two are almost the same). There is an edge between two researchers/nodes if they’ve published at least one paper together. The data set contains 23708 papers, 17636 authors and 59894 edges.

We measured the attachment kernel for this network based on the degrees of the two potential neighbors. See Fig. 1.2 for the $A_{\text{cond-mat}}(d^*, d^{**})$ function.

We’ve tried to fit various functional forms to the two-dimensional attachment kernel function to check which is a better description of the dynamics. See Fig. 1.3 for the shape of the fitted functions and Table 1.1 for the functional forms and the results.

The best fit was obtained by

$$A'_{\text{cond-mat}}(d^*, d^{**}) = c_1 \cdot (d^* d^{**})^{c_2} + c_3 \quad (1.12)$$

where the c_i are constants.

1.4 Discussion

We’ve briefly presented a methodology for understanding the evolution of networks through kernel functions and showed how the kernel functions can be extracted from network data.

We’ve discussed two applications for this methodology: first the “fitting” of the preferential attachment model to a network of scientific citations and then

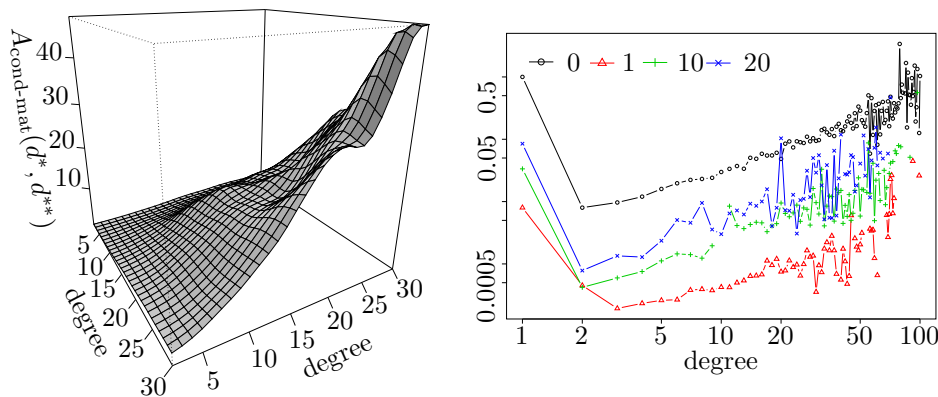


Figure 1.2: The attachment kernel for the cond-mat collaboration network, the surface plot was smoothed by applying a double exponential smoothing kernel to it. The right plot has logarithmic axes. The right plot shows that the kernel function has high values for zero-degree nodes, this might be because a new researcher will usually write a paper with collaborators and thus will have a high probability of adding links to the network.

determining how the evolution of a scientific collaboration network depends on the degree of the vertices.

The methodology outlined here is general and can be successfully applied to any kind of evolving network where time dependent data is available. By defining the kernel function in terms of the potentially important vertex properties one can check whether these properties really significantly influence network evolution: if a kernel function is not sensitive to one of its arguments that suggests that this argument does not have an important contribution. Another possible application would be to identify changes in the dynamics of a system by doing the measurements in sliding time windows, see [] for an example.

1.5 Acknowledgement

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. The authors also thank Mark Newman for providing the cond-mat data set.

Bibliography

- [1] BARABÁSI, Albert-László, and Réka ALBERT, “Emergence of scaling in random networks”, *Science* **286**, 5439 (1999), 509–512.

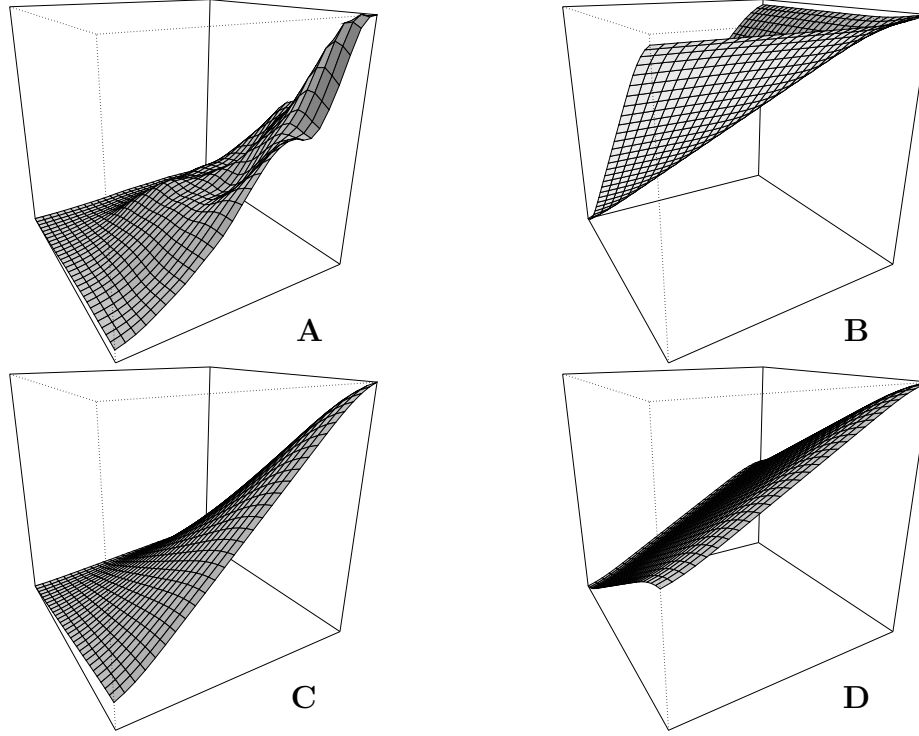


Figure 1.3: **A** shows the smoothed measured kernel function for the collaboration network, **B**, **C** and **D** are fitted functional forms shown in the first three lines of Table 1.1. The best fit is clearly obtained by the multiplicative fit.

Fitted form	Fitted parameters	Fit Error	Fitting method
$c_1 \max(d^*, d^{**}) + c_2$	$c_1 = 1.26, c_2 = -10.56$	107357.6	Nelder-Mead
$c_1 d^* d^{**} + c_2$	$c_1 = 0.0697, c_2 = -2.11$	4300.2	Nelder-Mead
$c_1(d^* + d^{**}) + c_2$	$c_1 = 1.08, c_2 = -18.98$	31348.9	Nelder-Mead
$c_1 d^* d^{**} + c_2(d^* + d^{**}) + c_3 \max(d^*, d^{**}) + c_4$	$c_1 = 0.0783, c_2 = -0.12,$ $c_3 = -0.093, c_4 = 1.50$	3532.9	BFGS
$c_1(d^* d^{**})^{c_2} + c_3$	$c_1 = 0.016, c_2 = 1.22,$ $c_3 = 0.58$	3210.4	SANN

Table 1.1: Four optimization methods were run for each functional form to minimize the least square difference: BFGS, Nelder-Mead, CG and SANN, the results of the best fits are included in the table. See [] for the details of these methods.