Toward a Next Generation of Network Models for the Web

Hans Akkermans
The Network Institute,
VU University Amsterdam VUA
and AKMC, the Netherlands
Hans.Akkermans@akmc.nl

Rena Bakhshi
The Network Institute,
VU University Amsterdam VUA
Amsterdam, the Netherlands
rbakhshi@few.vu.nl

ABSTRACT

It is generally thought that the World Wide Web belongs to the class of complex networks that is scale-free: the distribution of the number of links that nodes have follows a power law ('rich-get-richer' effect). This phenomenon is explained by a combination of theoretical-computational and empirical analysis based on stochastic network models. However, current network models embody a number of assumptions and idealizations that are not valid for the Web. Better and richer network models are needed, in association with a much more refined and in-depth empirical data gathering and analysis. In particular, the understanding of the dynamics leaves much to desire. In this paper we present a dynamic network model that avoids a number of unrealistic idealizations commonly introduced. We show how properties such as average degree and power laws are the outcome of dynamic network parameters. Exemplified by a Wikipedia case study, we show how these dynamic parameters might be empirically measured directly. We falsify several widely held ideas about the emergence of power laws: (i) that they are related to growing networks; (ii) that they are related to (linear) preferential attachment; (iii) that they may hold strictly. Power laws do not have the status of a first principle in networks: if they hold, they are just conditional and approximate empirical regularities.

Categories and Subject Descriptors

J.4 [**Social and Behavioural Sciences**]: [Web Science]

Keywords

power law, dynamic network models, nonlinear preferential attachment, degree distributions, Wikipedia hyperlink network

1. INTRODUCTION

The World Wide Web is a complex network thought to be scale-free: the number of links that nodes have obeys a power law (more popularly known as the rich-get-richer effect). This phenomenon has been extensively studied by a combination of theoretical-computational and empirical analysis based on statistical network models. These studies have led to the common view that the emergence of power laws is linked to two factors: (i) growth of networks, in conjunction with (ii) (linear) preferential attachment [2].

However, as critically reviewed in Sec. 2, current network models embody a number of assumptions and idealizations that are not valid for the Web. This has implications for the Web Science research agenda. Results from general network science might not be as rock solid as they seem when applied to the Web. Moreover, although the Web is often touted as highly dynamic, our understanding of dynamic network aspects leaves much to desire. Thus, we argue that as part of the research program investigating the Web-as-a-network, better and richer network models are needed, in association with a much more refined and in-depth empirical data gathering and analysis.

We contribute to such a research program, by presenting and analysing a dynamic network model that avoids a number of unrealistic idealizations common in network science (Sec. 3). Measurable properties such as average degree and degree distributions are shown to be the outcome of dynamic parameters of the network.

At the empirical side of this research program, we discuss in Sec. 4 how dynamic network parameters might be measured directly (rather than just their outcomes). This is a type of empirical network study that has not been done yet, but is necessary to do as it delivers basic but essential dynamic information (such as typical time scales for different event types in the network). Exemplified by a Wikipedia case study, we show that in principle it is indeed possible to extract more refined empirical data about dynamic network behaviour — information that is very much needed to validate or falsify theoretical network model hypotheses.

2. THIN ICE

Despite significant network research efforts in the past 15 years, our understanding of the Web-as-a-network must be assessed as still very limited. Many aspects that are central in social network research are still insufficiently understood and explained in the case of the Web, both for the Web graph itself and for the many social or sociotechnical networks it hosts. Examples of underresearched or underexplained social network effects and mechanisms are the tendency to reciprocate links, the tendency to form triangles (via mutual acquaintances, also variously referred to as transitivity or triadic closure) and the clustering coefficient (the number of triangles in the network relative to the number of pure pairs).

Formal and empirical methods for analysis of social network mechanisms as mentioned above already have a long standing in the social and behavioural sciences [21]. These methods typically focus on the structural, static aspects of networks. More recently, tools that can also analyze the dynamics of networks have been developed, especially the Groningen-Oxford model called SIENA [19]. A limitation of the social science methods in the context of Web is that they have originally been designed for relatively small social networks (order hundreds of nodes). Unfortunately, it turns out to be non-trivial to scale them up to the very large-scale networks (where one is easily talking order millions or more) that inhabit the Internet and the Web. Only very recently attempts to do so are emerging. They infuse the more traditional methods from the sociological tradition with methods borrowed from natural science (such as the mean-field approach) [6].

The network aspect that undoubtedly has been studied most extensively is that of the degree distribution, i.e., the number of links that nodes have, and the associated phenomenon of power laws. The thrust of this research originates from statistical natural science (so has a very large-scale perspective to begin with), but interaction of this type of work with that stemming from the sociological tradition(s) is still hardly discernible in the scientific literature, both in statistical physics and in the social sciences. Web Science might offer an important contribution if it succeeds in bridging these two hitherto separate traditions.

As we will now show, even in the (relatively simple) case of degree distributions and power laws, the state of Web knowledge is unsatisfactory. The existence of power laws is generally considered to be empirically well established also for the Web. Clauset et al. [8] however discuss many caveats regarding empirical studies proclaiming power laws. One of them is that there are several distributions that come close to power laws in certain regions, and that it is empirically difficult to distinguish between them. The present article will below offer additional theoretical grounds that support this: if power laws occur at all, they will only be approximate.

The emergence of power laws has theoretically been made plausible by the Price-Barabási family of generative network models (for overviews see [2] and [17], Ch. 14). These theoretical network models however make strong assumptions and idealizations that are questionable for the study of the Web:

1. The network is indefinitely growing: in every time step a new node is added, and it is these new nodes that create the links.

2. The probability to create links increases linearly with the degree (preferential attachment).

3. The possibilities of link removal and of network churn (nodes not only entering but also leaving the network) are ignored.

4. Studies of power-law emergence limit themselves to the regime where a steady-state solution holds, i.e. after the passage of a very long (formally, infinite) time.

These assumptions and idealizations seem pretty appropriate for citation networks (the area for which generative network models were indeed originally developed), but they are much less so for studies of the Web and the Internet. Assumptions 1 and 3 are simply not valid for the Web, assumption 4 focuses investigations on the non-dynamic regime only, and assumption 2 considers only one candidate hypothesis. All in all, theoretical network studies have succeeded in showing that emergence of power laws is a *possibility* as a result of growth of networks combined with preferential attachment; but they fall short of an actual *demonstration* of the necessity and/or universality of power laws as a first principle (for which they are often taken, though). The Web conditions have not been properly taken into account, and therefore these studies can not be considered conclusive.

It is already qualitatively easy to see why things might be different, see Figure 1. The left pane shows the situation considered by the vast majority of generative network model studies. The link creation process considered on its own will drive the system to ever higher values of the degree k. The (only) counteracting mechanism that drives the degree down is the steady introduction of new nodes with zero or low degree. If there is a balance between these two forces that holds for every value of k (in an appropriately differentiated way for each k), the emergence of a power law is possible. Link attachment linearly proportional to the degree is one such possibility according to generative network model studies.

The right panel shows the general case. Nothing changes for link creation, but more counteracting mechanisms come into play that drive the degree down. To dynamically obtain a power law, the link creation process must be balanced by the joint process of low-degree node addition, link removal, plus node removal. Conceptually it is clear that this is possible in principle, but it is not clear that power laws will emerge under the same conditions (linear preferential attachment). In this paper we consider the general situation of the right pane, and we will show that conclusions for the general case differ from the established ones specific for growing networks.

3. ENRICHED NETWORK MODELLING AND ANALYSIS

Berners-Lee *et al.* [4, 5] point out in their call for an interdisciplinary Web Science agenda that there is a clear need for better mathematical modelling of the Web. Below we present and analyze one such model (based on the exciton model of [1]) that is dynamic, includes the possibility of link loss and network churn (so does not make the growing network assumption), and is able to handle nonlinear forms of attachment. Furthermore, it is formulated in terms of continuous time which makes it more amenable to empirical measurement and testing (as measurements are carried out in continuous rather than discrete-event time, a point also made in [19] for empirical social network analysis). Importantly, these are all features that increase the realism of network models in the context of the Web but that almost all generative network models lack.

If we include all these features, but otherwise retain the stochastic nature of the existing network formation models, one can write down a so-called master equation for the dynamics of the degree distribution:

$$\begin{aligned}\tfrac{d}{dt}q_k(t) = &+\lambda^+_{k-1}\cdot q_{k-1}(t)\\ &-[\lambda^+_k+\mu^-_k+w_0+f_0]\cdot q_k(t)\\ &+\mu^-_{k+1}\cdot q_{k+1}(t)+f_0\cdot\delta_{k,0}\end{aligned} \tag{1}$$

Here, $q_k(t)$ is the probability that a node has degree k at (real) time t, λ^+_k and μ^-_k denote the rates (transition probabilities per unit of time) of link creation and loss, respectively, and network churn is modelled here by a process of node loss with rate w_0 and a process of (initially linkless, $k = 0$) node creation with a rate f_0. Note that these processes are all stochastic, so the (discrete) events in the network take place at irregular times (similar to incoming calls to a helpdesk). For example, network churn is technically modelled here as a process of irregular bursts known as shot noise. The rates represent the average waiting time between events (and so are measurable quantities, at least in principle, see Sec. 4). The link loss

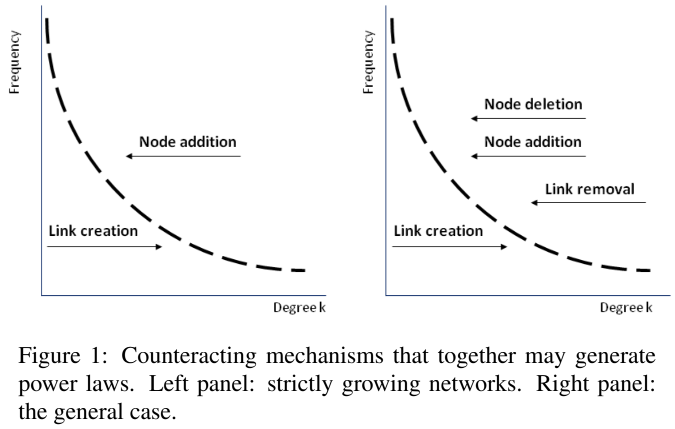

Figure 1: Counteracting mechanisms that together may generate power laws. Left panel: strictly growing networks. Right panel: the general case.

rate μ_k^- is a bit special as it is the sum of direct link deletion and link loss due to churn (node removal):

$$\mu_k^- = \lambda_k^- + 2w_0 \cdot k, \tag{2}$$

where λ_k^- is the intrinsic link deletion rate and the second term represents the contribution as a result of node removal (the factor of 2 models the fact that not only the links of the lost node itself are removed, but also of the nodes it was connected to).

The master equation says that the rate of change of the occupation probability of a degree k by nodes in the network (the left-hand side) is the net total of all transitions per unit of time leading into that state k minus the sum total of all transitions per unit of time moving out of that state. Link creation, link loss and network churn all contribute to these changes. One would generally expect that all these processes influence whether or not a power law emerges, and if so, what their impact is on the shape, as determined by the power-law exponent.

The above exciton model is much more general than the Price-Barabási type of generative network models, and in fact subsumes them: the usual generative network models are recovered by setting the rates for link deletion (λ^-) and node removal (w_0) equal to zero for all k, and by setting the node creation rate f_0 equal to one. Although our model is much more general, it still admits of several important results that can be obtained in an explicit analytical form. Here we state a few of them.

Modelling Nonlinearities. The key dynamic parameters in Eq. (1) are the various transition rates: together they fully determine the dynamic evolution of the network including important variables such as the average degree and the degree distribution. Moreover, they have a clear conceptual interpretation in terms of the mean waiting time between network events of a given type (e.g. link creation, removal) and, much more so than the transition probabilities that are put central in the generative network model studies, they are accessible to direct empirical measurement. The network science tradition is to assume that transitions increase *linearly* with the degree k, known as preferential attachment: 'popular' or 'rich' (i.e. high-degree) nodes have a higher probability to obtain new links than 'poor' low-degree nodes.

However, there is no *a priori* reason that transitions must be linear in the degree, even though this is the case that hitherto almost exclusively has been studied. To study the possible effects of nonlinearities, we therefore assume a parametrized polynomial plus broken-power shape for the transition rates, as follows:

$$\lambda_k^{\pm} = k^{\gamma} \cdot [a_0^{\pm} + a_1^{\pm} \cdot k + a_2^{\pm} \cdot k^2 + \mathcal{O}(k^3) \ \cdots], \tag{3}$$

with $0 \leq \gamma < 1$. This parametrization subsumes and extends the investigations done so far. It subsumes the constant (random graph) and linear (preferential attachment) cases; the factor with γ enables to study broken powers (the case studied by Krapivsky *et al.* [14, 13]), and it enables to study for example quadratic and other superlinear or sublinear shapes.

Here we will focus on contrasting the results for the traditional linear case, what happens if we include link removal and network churn, and furthermore what happens if nonlinearities of a quadratic form are introduced. The broken-power case will be separately studied in the Appendix.

Linear Preferential Attachment & Strictly Growing Networks. As a first model example we consider the situation that link creation is linear in the degree k (so $\gamma = 0$, $a_{k\geq 2}^{\pm} = 0$ in Eq. (3)), and that both links and nodes are never removed ($\lambda_k^- = w_0 = 0$ in Eq. (1)). Essentially, we consider the situation of the left pane of Figure 1 with a linear preferential attachment rate ($\lambda_k^+ = a_0^+ + a_1^+ \cdot k$), so that in effect one is assuming that the Web behaves in a way similar to citation networks. This is the situation assumed in the vast majority of generative network model studies.

In [1] we introduced and explained a novel method computing litmus tests for the emergence of power laws, based on the principle of detailed balance. It enables to analyze whether or not a power law ultimately emerges in the network and in addition what the power-law exponent is, *without* having to know the stationary solution itself (which is much more complicated but is what all generative network model studies do).

If we apply this method, we find that the stationary solution is approximately [1] a power-law distribution $\left(\frac{1}{k}\right)^{\alpha}$ with a power-law exponent α given by:

$$\alpha_{\,\mathrm{linearpref,growing}} = 1 + \frac{f_0}{a_1^+}. \tag{4}$$

One observes that the long-term outcome for the degree distribution is determined by the dynamic parameters of the network (in this case f_0 and a_1^+).

Specifically, the type of power law that emerges is *determined by the relative speeds* of the counteracting processes that are at play (here, link creation and linkless new node introduction). This confirms quantitatively what one would expect qualitatively. If node addition is fast (relatively many events per unit of time) compared to link creation, the power law exponent will be higher and the degree distribution steeper than when it is relatively slow.

The above result can be easily specialized so as to recover the major results of the first wave of generative network studies [3, 12, 10, 9]. In these studies it is typically assumed that one node is added per unit of time, and so we can set $f_0 = 1$ in Eq. (4). Furthermore, it is assumed that the linear link creation function for attachment relates to the probabilities (rather than the rates) for attachment and therefore must normalized to unity (in our notation, this implies the assumption $a_1^+ + a_0^+ = 1$).

If we introduce these assumptions, Eq. (4) simplifies to:

$$\alpha_{\,\mathrm{generative,linearpref,growing}} = 1 + \frac{1}{a_1^+} = 2 + \frac{a_0^+}{a_1^+}, \tag{5}$$

which is indeed the result according to the link-copying model of Kleinberg *et al.* [12], the early citation network studies by Derek de Solla Price [10, 9] and, if we adopt an even split in linking probability $a_1^+ = a_0^+ = \frac{1}{2}$, we obtain $\alpha = 3$, the original result by Barabási *et al.* [3].

We note that these works derive the power-law exponent in ways very different from our method. An important conceptual difference is that the generative network model studies take the attachment functions as referring to the transition probabilities, whereas in our exciton model they refer to the transition rates. Interestingly, for linear preferential attachment this difference does not matter and the quantitative results for the power law are the same. However, this difference is no longer inconsequential when one considers nonlinear attachment functions (see further below and the Appendix).

[1] To be more precise: the litmus test of [1] shows that the stationary solution is a (non-power law) distribution that coincides with a power law in the two leading orders when a series expansion of the distribution for the adjacent degrees is made with $\frac{1}{k}$ as the small parameter. Simply put, a power law emerges, but it is only approximate; and the approximation holds well only for the tail containing the nodes with high degree.

Linear Preferential Attachment & General Networks. That the Web really behaves like a citation network is unlikely. A novel and more realistic case is therefore to consider linear preferential attachment, but also allow network churn and link removal (as in the right pane of Figure 1). Remarkably, this more general and very relevant case has hardly received attention in the network science literature, with just a few and partial exceptions: [16] studied the inclusion of node deletion and [17] (Sec. 14.4.2) that of linear link removal.

Here we include all these processes simultaneously, following Eq. (1), and assume a linear link creation rate as above and linear link removal according to $\lambda_k^- = a_1^- \cdot k$). Applying the litmus test approach of [1] employing a power-series expansion in $1/k$ we see that link and node removal both significantly change the picture.

Following this method we get instead of Eq. (4):

$$\alpha_{\text{linearpref,general}} = 1 + \frac{f_0 + w_0}{a_1^+ - (a_1^- + 2w_0)}. \qquad (6)$$

So, as soon as link and/or node removal become significant compared to link creation, one obtains unrealistic values of the power-law exponent or the power-law behavior becomes even totally lost.

Quadratic Attachment & General Networks. As pointed out previously, there is no *a priori* reason that preferential attachment should be linear. It is just a convenient hypothesis, although warranted to make because it is basically the simplest one, next to constant attachment that does not work because it predicts a Poissonian 'bell-shape' degree distribution rather than a right-skewed one such as an approximate power law that is also observed empirically. But above we have demonstrated that the linear preferential attachment hypothesis also runs into trouble or is at least questionable in view of well-established empirical observations of the Web if we integrate realistic Web phenomena such as link removal and node churn into the theoretical network model.

This constitutes a trigger to investigate yet other hypotheses, in particular the possibility of nonlinear attachment functions. Beyond linear preferential attachment, most likely the next-simplest hypothesis is a quadratic shape for attachment. Accordingly, we now consider the case of quadratic link creation and removal rates: $\lambda_k^+ = a_0^+ + a_1^+ \cdot k + a_2^+ \cdot k^2$ and $\lambda_k^- = a_1^- \cdot k + a_2^- \cdot k^2$. Mathematically, we proceed in the same way as indicated above (although the mathematical bookkeeping in comparing terms in the power-series expansion again becomes a bit more tedious). In summary, this analysis leads to the following results for the quadratic attachment case.

In general, the stationary solution here is a non-power law distribution but one that is well approximated for not too small k by the equation:

$$q_k^{stat} = \frac{\nu}{k^\alpha} \cdot \left(\frac{a_2^+}{a_2^-}\right), \qquad (7)$$

where ν is a normalization constant, and

$$\alpha_{\text{quad,general}} = 2 + \frac{a_1^-}{a_2^-} - \frac{a_1^+}{a_2^+} + \frac{2w_0}{a_2^-}. \qquad (8)$$

In other words, the general long-term solution in the quadratic attachment case has a power-law like factor but with an exponential cut-off in the high-degree tail.

It follows that under certain conditions, namely if $a_2^+ = a_2^-$ implying a regime of non-growth whereby link removal and link creation are approximately balancing each other, we have a stationary solution that is pretty close to a power law. This is confirmed by considering the second moment $M_2 = \sum_0^\infty k^2 q_k^{stat}$ and requiring that it goes to infinity as should be the case in truly scale-free networks. If we apply this to Eq. (1) we obtain an additional constraint to the dynamic parameters, as a result of which the power-law exponent in the scale-free network limit becomes:

$$\alpha_{\text{quad,general,scale-free}} = 3 - \frac{(f_0 + w_0)}{(a_2^+ + a_2^-)}. \qquad (9)$$

Hence, in the quadratic attachment case, the power-law exponent will be close to 3 if node addition and deletion are small compared to the link creation and deletion rates, and it will be close to 2 when all process rates have the same order of magnitude. This fits the empirical observations regarding the Web very well, and so the quadratic attachment or other nonlinear hypotheses are certainly ones that are to be taken seriously. The above equations (8) and (9) nicely show how *all* processes of link creation, link deletion, node addition and node deletion influence the emergence of power laws as well as the numerical value of the power-law exponent in ways that have as yet not been uncovered by the traditional linear preferential attachment network studies. It moreover follows that approximate power-law behaviour is also possible in non-growing networks, in contrast to what is often thought [2, 16].

Average Degree as an Outcome of Dynamics. In the generative network model studies that have been carried out so far, the average degree of a network is always a fixed and constant parameter, input rather than output. This is unlikely to be realistic, as the average degree is expected to be not predetermined but to be the outcome of the dynamic processes and parameters in the network. This is in fact the case in our network model. If we take the first moment of Eq. (1) and take the long-time limit $t \longrightarrow \infty$, we obtain an expression for $M_1 = \sum_0^\infty k \cdot q_k^{stat}$. The average degree c predicted for the network follows from the key dynamic parameters as follows:

$$c = \frac{a_0^+}{a_1^- - a_1^+ + 3w_0 + f_0}. \qquad (10)$$

This equation subsumes both the linear and quadratic forms of attachment as discussed above (for the latter it is assumed that $a_2^+ = a_2^-$, i.e. rough balance between link addition and removal). Without going into detail, it is clear that (i) the average degree is not predetermined (as the generative network model studies have it) but is the outcome of dynamic parameters of the network; (ii) all processes of link creation, link removal, node addition and node removal are co-determining the average degree of a network.

In summary, in order to be able to select between the large variety of possible theoretical models, we need richer in-depth empirical data. It is not sufficient to have data on power-law behaviour as such, because this does not shed any light on the underlying mechanisms. Even, where one has attempted to measure preferential attachment directly [11], these are aggregated studies that are not able to distinguish between link/node creation and removal, but just measure the net result over a time window. But as the transition rates for the linkand node update processes fully determine the dynamics, a question is whether one cannot obtain direct empirical information on these transition rates themselves. This has to our knowledge never been studied, and is what we turn to now.

4. TRANSITION RATES: EMPIRICAL STUDY

Churn is an inherent part of the Web, and all node and link related events (addition and deletion) dynamically contribute to the changes in the complex network structure in a different way. But

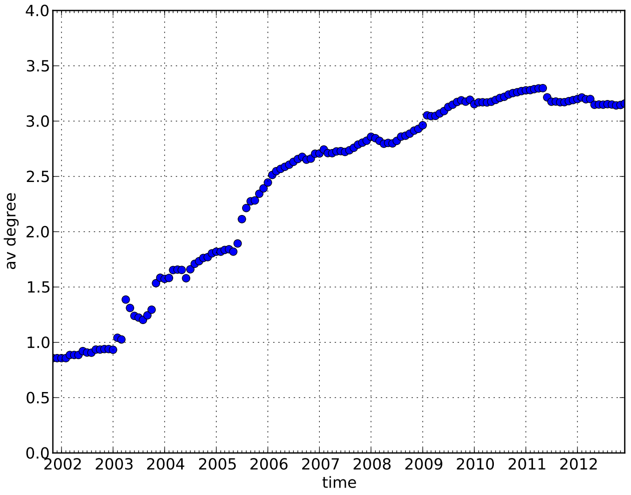

Figure 2: Evolution of the average number of links over time (on a monthly basis).

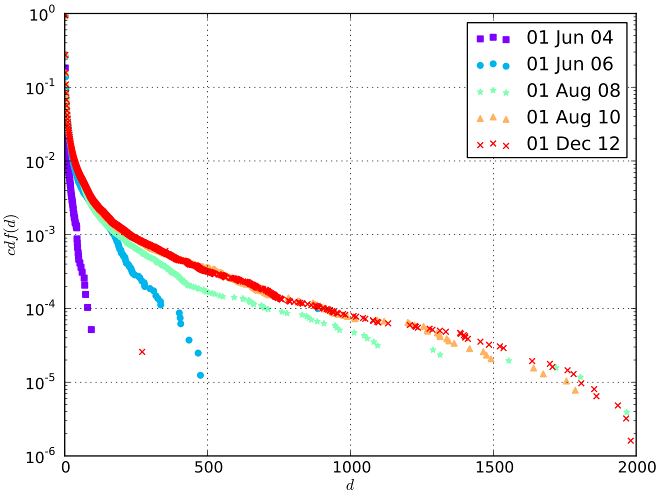

Figure 3: Evolution of the cumulative degree distribution of the simple Wikipedia. The y axis shows the fraction of nodes in the network having at least degree d.

as already stated, all prior studies only analyze the *net* result of these processes by aggregating them so that one cannot distinguish the differentiated impact of the separate processes (cf. Fig. 1). Empirically, we believe one can (and therefore should) go a step further. Therefore, below we carry out a simple empirical case study demonstrating that one can in fact obtain much more detailed insights into the different counteracting processes also directly from empirical data themselves. Using Wikipedia [22] as an example of a large-scale evolving network, we show how key dynamic network parameters might be measured or estimated directly from empirical data, and how for example churn influences the degree distribution.

Wikipedia is one of the large-scale dynamic evolving networks, due to the continuous collaborative editing and maintenance of its content by users. We conduct this case study on the so-called *Wikipedia hyperlink network*, where its articles are nodes in the network linked to each other by means of internal links (i.e., within Wikipedia). Its entire edit history is available on a monthly basis for download, and we use the edit history dump of the simple English language Wikipedia [18] as of January 1, 2013 for our measurements. Currently, the number of articles of this Wikipedia stands at 91,451. The edit history dump contains every revision (timestamp, full article text and user information about the editor of the revision) for every article of Wikipedia. Deleted pages, however, are not included into any public Wikipedia data dump due to some legal constraints.

The hyperlink network of Wikipedia can be represented as a simple directed graph $G = (V, E)$ evolving over time, with new pages appearing, pages deleted, and links added and removed on demand. In fact, links may change due to page edits or to vandalism. In the latter case malicious users sometimes overwrite entire pages, thereby temporarily removing or adding the links from vandalized pages. We can clearly see this flux of links in Figure 2: initially the network grows to become more connected, followed by periods of Wikipedia cleanup and expansion, when users collaboratively expand the content of existing pages or fix pages that do not adhere to Wikipedia standards on content and style. The average degree is not very high (in the range of 2–3), but is a dynamic variable as it clearly changes over time (where generative network models have it constant and mostly as a predetermined input variable).

As one would expect (although theoretical network models currently do not account for this), the degree distribution of the network also evolves over time. Figure 3 visualizes the changes in the cumulative degree distribution of the network over time. Each curve in the Figure represents the degree distribution of nodes in the indicated year. (Again, deletion of pages could not be included in these results.).

Subsequently, we have extracted a graph of the links between Wikipedia pages for every month, taking into account all revisions of every Wikipedia page, and from this we have attempted to directly measure the link creation, link removal and node addition rates (λ_k^+, λ_k^- and f_0 in the exciton model of Eq. (1)).

4.1 Measurement and Estimation Procedures

Let k be a degree of a Wikipedia page, i.e., the number of incoming (internal) links to the page. The pages with degree k of the Wikipedia hyperlink network form a subgraph $G_k(t) = (V_k(t), E_k(t))$ at time t. As described above, a page has its revision history, each of which has a timestamp and a set of hyperlinks to the other Wikipedia pages. We take Δt as an average time window between two consecutive revisions of any page through the entire Wikipedia history.

In order to capture link addition and link removal events of a page separately, we compare every pair of consecutive revisions of the page, computing the complement of the set of its links. Thus, we obtain the number of added links and the number of removed links. For all pages $v_k \in V_k(t)$ with (in)degree k, and for all their respective revisions $r(v_k) \in R_k$, the total number of links addition events $\varepsilon_k^+(t)$ at time t can be expressed as:

$$\varepsilon_k^+(t) = \sum_{\substack{v_k \in V_k,\\ r(v_k) \in R_k}} |r(v_k, t) \setminus r(v_k, t-1)| \tag{11}$$

where $r(v_k, t)$ is the latest revision of the page v_k, prior to time t. Similar to $\varepsilon_k^+(t)$, the total number of link removal events $\varepsilon_k^-(t)$ at time t is

$$\varepsilon_k^-(t) = \sum_{\substack{v_k \in V_k,\\ r(v_k) \in R_k}} |r(v_k, t-1) \setminus r(v_k, t)| \tag{12}$$

The rate of link addition λ_k^+ can then be computed as the number of links added to the Wikipedia pages averaged over the number of pages in the network and over time:

$$\lambda_k^+ = \frac{1}{|T|} \cdot \sum_{t \in T} \frac{e_k^+(t)}{|V_k(t)|} \tag{13}$$

Likewise, the rate of link removal λ_k^- is empirically computed, averaging the link removal events over all pages, their revisions, and over the entire period of the Wikipedia history:

$$\lambda_k^- = \frac{1}{|T|} \cdot \sum_{t \in T} \frac{e_k^-(t)}{|V_k(t)|} \tag{14}$$

For our empirical case study, we estimated that a suitable time step Δt for the simple Wikipedia from the entire period (01.01.2001 – 01.12.2012) is one month; in Eqs. (13) and (14), $|T| = 132$.

Finally, the page addition rate per time month is computed as the size of the complement of two sets: (i) the set of pages $V(t)$ existing at time t, and (ii) the set of pages $V(t-1)$ existing at time $t-1$, averaged over the number of months of the empirical data:

$$f_0 = \frac{1}{|T|} \cdot \sum_{t \in T} \frac{|V(t) \setminus V(t-1)|}{|V(t)|} \tag{15}$$

where $|T| = 132$.

Since the Wikipedia dumps with the complete edit history (or any other public Wikipedia dump) do not contain deleted pages, it is currently not possible to get a reliable measure for the rate of page deletion w_k.

4.2 Observations and Caveats

We now summarize some observations made in the course of our empirical study on the growing Wikipedia hyperlink network.

Results. We plot the resulting rates λ_k^+, λ_k^-, measured as described above. The scatterplots in Figure 4 show the rates of link addition λ_k^+ and link deletion λ_k^- per month for a degree k. The results suggest that overall the link addition rate is higher than the link removal rate. However, the link removal rate is approximately 20% of the link addition, so it cannot be neglected as most generative network studies do to date. Thus, link removal is a significant process that should be taken into account into theoretical network model studies. Moreover, as the higher magnification insets in Figure 4 suggest, the link addition rate has a local minimum around degree 50, while the same degree corresponds to a local maximum of the link removal rate; for this degree the link removal is only half of the link addition rate.

Furthermore, the plots (Figures 4, 5 and 6) suggest the combination of a 'bell-shape' curve, combining an increasing curve for the smaller degrees of the nodes, an intermediate regime resembling a plateau (up to 200, note that this is nearly two orders of magnitude larger than the average degree) and an exponentially decreasing curve for the long tail of the very high node degrees (≥ 200). The left inset figure in Figure 4 also suggests that the first peak in the link addition rate roughly coincides with the average degree (see Figure 2). Note that the regimes with degree < 200 will contain the vast majority of the nodes.

Furthermore, the node addition rate for Wikipedia is higher than the node removal rate (since the overall number of pages grows despite the page removals). We obtained from the empirical data that the page creation rate is $f_0 = 0.076301841$. The number is actually quite high, and suggests that the counteracting process such as page removal should be high too. As mentioned before, it is now not possible to measure w_k directly from the Wikipedia dump as the deleted pages are removed from the public data. As an estimate, we can only base ourselves for now on findings in [20] [Table 2.3], where page addition and removal rates have been measured for the English Wikipedia version for a duration of 4 months: 10.94% of pages added compared to 5.03% of pages removed, suggesting that the average node deletion rate w_0 is about half of the node addition rate f_0.

Choosing The Time Window. To observe the dynamics in networks directly via the empirical data, the time interval needs to be chosen very carefully. On the one hand, good statistics is required, so the time window should be large enough in order to have enough individual events for statistical estimates and be able to average out fluctuations or noise. On the other hand, the time window should be small enough to observe the significant changes in and emergent behaviour of the network. In other words, the issue here is to separate what is signal and what is noise in the empirical data. If the time window is too small, one is keeping in too much noise; if it is too big one is also averaging out meaningful dynamic

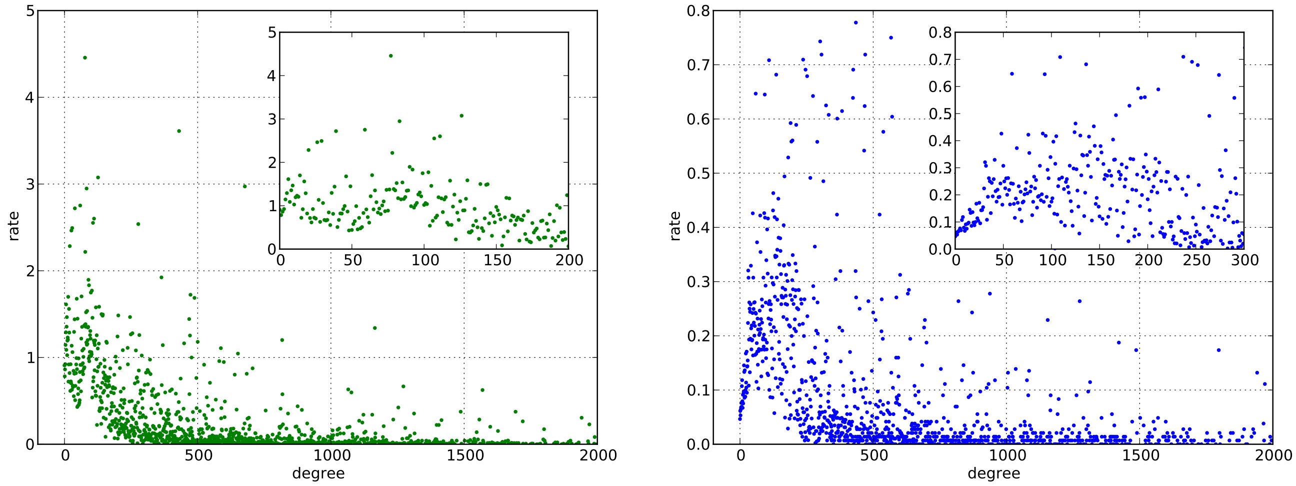

Figure 4: Empirically estimated average rates λ_k^+ (left) and λ_k^- (right) for different degrees k. Inset figures show peaks of the rates.

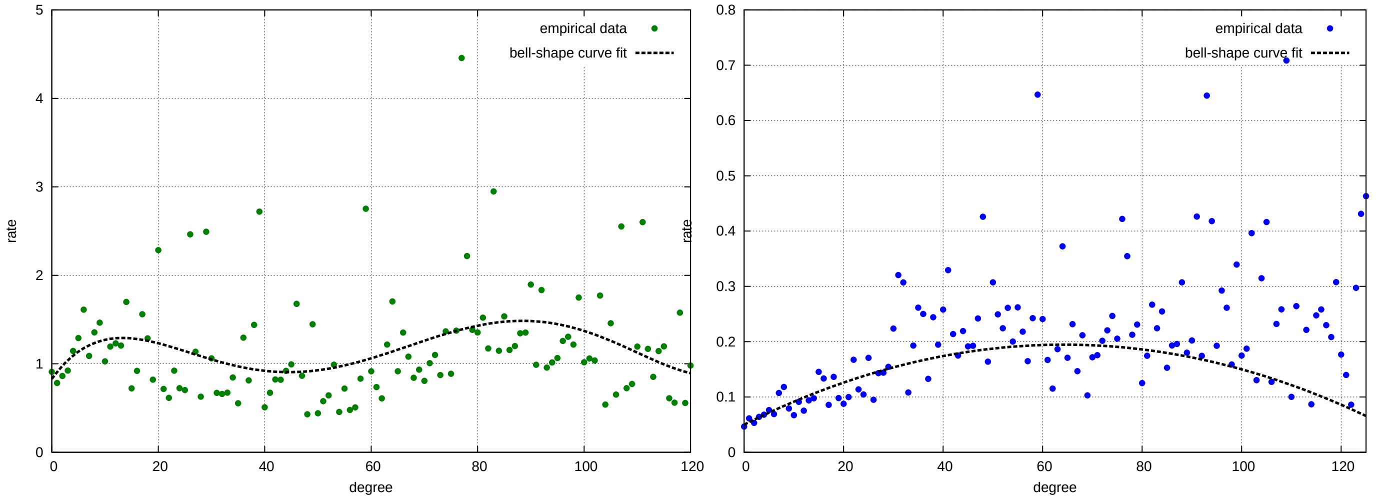

Figure 5: Scatter plots with fitted curves for average rates λ_k^+ (left) and λ_k^- (right) for degrees $k < 120$.

features of the network. This choice is therefore critical. It will depend on the specifics of the network under study (as the typical time scales of change in networks are not universal features but vary very much depending on the nature of the network), but the proper choice of the time window constitutes a general and very non-trivial issue in the empirical study of network dynamics [19].

In the present case study, for example, although Wikipedia pages are periodically edited causing changes in the (hyperlink) network structure, there are pages with much fewer revisions than the rest. We thus chose the time interval to be an average time window between two consecutive revisions of any page through the entire Wikipedia history, which is approximately one month.

Network Flux. As described above, we separately measure link addition and link removal events for different node degrees. We measure these fluctuations between time intervals of one month each. However, the average rates λ_k^+, λ_k^- that we derive in this way are approximate since we may miss out on intermediate short-term degree fluctuations within the time interval.

Most studies also disregard the churn in a network, notably, by assuming that nodes are added one-by-one in a sequence, but are never removed. Moreover, new nodes are assumed to have a fixed degree (zero, one, or equal to the average degree are the common cases in the theoretical model studies). Actually, in the case of Wikipedia, some pages may already have links to it before their creation and some do not. This suggests for theoretical studies of network dynamics that both the initial and boundary conditions are more complex than accounted for now by the network model studies.

Since any public data dump of Wikipedia does not contain deleted pages, it is not possible to directly measure the rate of pages deletion w_k. The only way to compute the rate of page deletion is by collecting Wikipedia dumps at different time points and compare the set of pages in them, thereby tracking the deleted pages.

Finally, the number of possible degrees of pages is very high, which results in a relatively sparse sample set for the rates estimation. For example, there are very few Wikipedia pages that have an extremely high number of the links ($k \geq 500$). The average rates

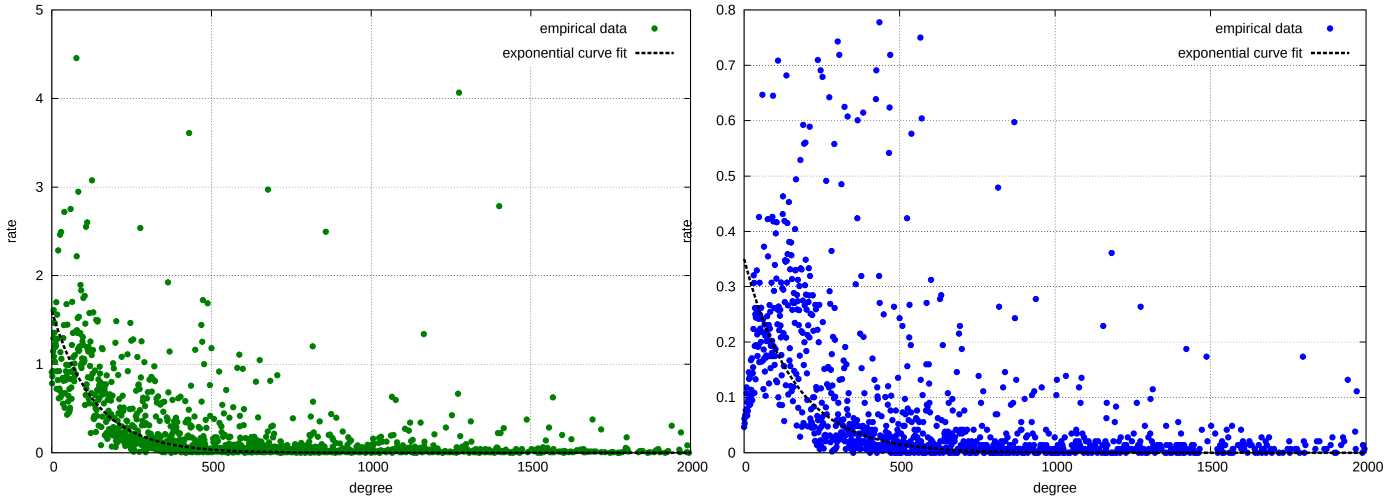

Figure 6: Scatter plots and fitted curves for average rates λ_k^+ (left) and λ_k^- (right) for different degrees k.

for these degrees are not representative. Nevertheless, they create a very long tail in the graph (see Figures 4 and 6). In order to get more accurate average rates it might help to consider ranges (bins) of the intervals for node degrees, where we adapt the size of an interval in such a way that every interval covers a statistically suitable minimum number of samples (i.e. nodes with the respective degrees).

Link Asymmetry. Large networks are typically asymmetric. In fact, the Wikipedia hyperlink network is a directed graph: if page A refers to page B in the Wikipedia, it does not imply that page B also refers to page A. Thus, a page may have a different number of incoming links (in-degrees) and outgoing links (out-degrees). The popularity of a Wikipedia page is best estimated by its in-degree k, which we used as the basis to compute the rates λ_k^+ and λ_k^-. The asymmetry of in-degrees and out-degrees has been previously discussed and measured in [12, 7]. It is however relatively straightforward to include link asymmetry into our model of Eq. (1), by generalizing it to a two-dimensional random walk; the state space of such a model will be spanned by the (out-degree, in-degree) pair of the nodes.

In summary, our empirical investigations lead to the following overall picture:

- It is indeed possible to extract direct and differentiated empirical information from network data regarding the rates of link creation, link removal, and node addition and deletion.

- The appropriate measurement and interpretation procedures need careful scrutiny and come with a number of caveats; a key reason (but not the only one) is that the empirical data show a lot of scatter.

- Nevertheless, a pretty strong conclusion is that link removal and network churn do play a non-negligible empirical role in the dynamics of networks, in ways that theoretical network model studies to date have not accounted for.

- Moreover, the data suggest that (even strongly) nonlinear and complex forms of link attachment as well as link removal might occur in practice.

Although more case studies on other dynamic networks are called for, we believe to have shown that this line of work is necessary and fruitful to set new tasks for and requirements on theoretical and computational network models. Furthermore, we deem it pretty likely that theoretical network models must become much more sophisticated before they can properly handle empirically observed network dynamics, even for relatively simple network variables such as average degree and the degree distribution.

5. CONCLUSION

To better understand the dynamic mechanisms at work in the Web, we must develop a more advanced set of theoretical and empirical instruments for dissecting the Web-as-a-network — instruments that are able to cut much deeper.

Theoretically, we have shown that nonlinear attachment is a serious possibility to explain observed power-law behaviour, that power-law behaviour is possible in non-growing networks, but also that there can be many regimes in networks where power laws are only approximate or even absent.

In order to be able to select between the many theoretical possibilities, we need measurement programs leading to more refined and in-depth data.

Empirically, we have shown that it is actually possible to directly measure dynamic network parameters such as link addition, link removal and node churn.

Although these studies have a tentative nature, they also point to the possibility of nonlinear forms of attachment.

Given the complexity of the Web and the variety of the social/sociotechnical networks it hosts, it is in our view necessary to have a research program of more in-depth and dynamics-oriented theoretical-computational studies as well as stronger and richer empirical data gathering and testing, in a way that mutually informs both.

Acknowledgements. We thank Jerome Kunegis, Steffen Staab and other colleagues from the Web Science and Technology Institute of the University of Koblenz-Landau, D., for various useful discussions and exchanges.

6. REFERENCES

[1] H. Akkermans. Web dynamics as a random walk: how and why power laws occur. In *Proc. of ACM Web Science Conf.*, WebSci '12, pages 1–10. ACM, 2012.

[2] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, Jan 2002.

[3] A.-L. Barabási and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286:509–512, 1999.

[4] T. Berners-Lee, W. Hall, J. Hendler, N. Shadbolt, and D. J. Weitzner. Creating a Science of the Web. *Science*, 313(5788):769–771, 2006.

[5] T. Berners-Lee, W. Hall, J. A. Hendler, K. O'Hara, N. Shadbolt, and D. J. Weitzner. A Framework for Web Science. *Foundations and Trends in Web Science*, 1(1):1–130, 2006.

[6] J. M. Birkholz, R. Bakhshi, R. Harige, P. Groenewegen, and M. van Steen. Scalable Analysis for Large Social Networks: the data-aware mean-field approach. In *Proc. of Conf. on Social Informatics (SocInfo)*, volume 7710 of *LNCS*, pages 406–420. Springer, 2012.

[7] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the Web. *Comput. Netw.*, 33(1-6):309–320, 2000.

[8] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-Law Distributions in Empirical Data. *SIAM Rev.*, 51(4):661–703, 2009.

[9] D. De Solla Price. A General Theory of Bibliometric and Other Cumulative Advantage Processes. *J. Amer. Soc. Inform. Sci.*, 27:510–515, 1965.

[10] D. De Solla Price. Networks of Scientific Papers. *Science*, 149:292–306, 1976.

[11] H. Jeong, Z. Neda, and A.-L. Barabási. Measuring preferential attachment in evolving networks. *Europhysics Letters*, 61:567–572, 2003.

[12] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The Web as a Graph: Measurements, Models, and Methods. In *Proc. Conf. on Combinatorics and Computing (COCOON)*, volume 1627 of *LNCS*, pages 1–17. Springer, 1999.

[13] P. L. Krapivsky and S. Redner. A statistical physics perspective on Web growth. *Computer Networks*, 39(3):261–276, 2002.

[14] P. L. Krapivsky, G. J. Rodgers, and S. Redner. Degree Distributions of Growing Networks. *Phys. Rev. Lett.*, 86:5401–5404, Jun 2001.

[15] J. Kunegis and M. Blattner. Preferential attachment in online networks: Measurement and explanations. Submitted.

[16] C. Moore, G. Ghoshal, and M. E. J. Newman. Exact solutions for models of evolving networks with addition and deletion of nodes. *Phys. Rev. E*, 74:036121, Sep 2006.

[17] M. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[18] Wikipedia (simple english). `simple.wikipedia.org`.

[19] T. Snijders, G. van de Bunt, and C. Steglich. Introduction to actor-based models for network dynamics. *Social Networks*, 32:44–60, 2010.

[20] G. Urdaneta. *Collaborative Wikipedia Hosting*. PhD thesis, VU University Amsterdam, 2012.

[21] S. Wasserman and K. Faust. *Social network analysis: Methods and applications*. Cambridge University Press, 1994.

[22] Wikipedia, The Free Online Encyclopedia. `www.wikipedia.org`.

APPENDIX

Why Nonlinearity Cannot Be Ruled Out

In this Appendix we take a critical look at the studies that have been done so far on nonlinear attachment and the conclusions they draw. As pointed out earlier in this paper, there is no convincing *a priori* reason that attachment *should be* linear. Both the present paper as well as other recent Web Science investigations [15] offer theoretical and empirical reasons to believe that (minimally) the hypothesis of nonlinear attachment is worthwhile to investigate carefully.

Since almost all theoretical network model studies employ the assumption of linear attachment, there are not that many on nonlinear attachment. To the best of our knowledge, only a set of studies by Krapivsky *et al.* (e.g., [14, 13]) have studied some of these matters in depth.

Their studies investigate the possibility that link creation is modelled by a broken-power function k^γ, called the attachment kernel A_k , where γ is any real number inbetween zero and two. If $\gamma = 1$ we have the usual linear attachment, if $\gamma < 1$ we have sublinear attachment, and if $\gamma > 1$ we have superlinear attachment.

In very brief summary, these theoretical model studies lead to the basic conclusion that *only* linear attachment is able to generate power laws.

This is an interesting and strong conclusion, because — if correct — it seems to settle the matter as to how power laws emerge. The implied and pretty attractive argument goes as follows:

1. Power-law degree distributions are empirically observed in many networks (including the Web).
2. The generative network model studies generally show that linear attachment is able to produce power laws in networks.
3. The studies of Krapivsky *et al.* offer grounds to believe that nonlinear attachment functions *do not and can not* lead to power-law behaviour.
4. Therefore, the hypothesis of linear preferential attachment must be correct.

However, we will show now that point 3 is an overstatement and in part invalid, so that nonlinear attachment cannot be ruled out.

Krapivsky *et al.* consider growing networks (one node added per unit of time, only link creation). Their basic model is specified by a rate equation for the number of nodes with k links $N_k(t)$:

$$\tfrac{d}{dt} N_k = \frac{A_{k-1} \cdot N_{k-1} - A_k \cdot N_k}{A} + \delta_{k,1}, \qquad (16)$$

where $A_k = k^\gamma$, and $A = \sum_k A_k N_k$. which is a normalization factor: $A_{k-1} \cdot N_{k-1}$ is the rate at which link creation occurs and the factor A turns this into a normalized probability.

These authors then show, by summing Eq. (16) over k, that $N_k(t) = n_k \cdot t$: the number of nodes with k links ultimately increases linearly in time for all k, and also $A(t) = M_\gamma \cdot t$. Substituting this into Eq. (16), one sees that the time dependence cancels, and one is left with a static equation for n_k, where n_k is the degree distribution in the long-time limit $t \longrightarrow \infty$:

$$n_k = \frac{A_{k-1} \cdot n_{k-1} - A_k \cdot n_k}{M_\gamma} + \delta_{k,1}. \qquad (17)$$

In fact, the degree distribution n_k equals what we call the stationary solution q_k^{stat} in this paper and, following the definition of A, M_γ is in fact the γ-th moment of the degree distribution. For linear attachment, it is the first moment of the degree distribution, which is equal to the average degree (which equals one *by construction* in the model of Krapivsky *et al.*). Generally, for this model Eq. (17) is the equation for the stationary solution of the degree distribution if we have nonlinear attachment.

Now let us compare this with our exciton model of Eq. (1). If we assume the same growing network case (no link and node removal, i.e., $\lambda_k^- = w_0 = 0$) and look for the stationary solution by setting the left-hand side equal to zero we get:

$$0 = \lambda_{k-1}^+ \cdot n_{k-1} - [\lambda_k^+ + f_0] \cdot n_k + f_0 \cdot \delta_{k,0}. \qquad (18)$$

It is easy to see that the two models represented by Eq. (17) and Eq. (18) will essentially predict the same degree distribution if we set:

$$\lambda_k^+ = A_k = k^\gamma \text{ and } f_0 = M_\gamma. \qquad (19)$$

Hence, what the model by Krapivsky *et al.* represents is indeed modelling a nonlinear (broken-power) link creation rate, but through its assumed normalization factor it effectively changes the node addition rate. Rather than the node addition rate being one (the standard assumption in generative network model studies) it equals M_γ.

Accordingly, the comparison by Krapivsky *et al.* between different nonlinear attachment kernels is not on the same basis, because M_γ quickly monotonically increases as a function of γ: it is easy to generally prove that if $\gamma_2 > \gamma_1$ then $M_{\gamma_2} > M_{\gamma_1}$. Rather than staying constant at one (which it is for the linear case), the node addition rate quickly grows for superlinear attachment. So the model by these authors has (implicitly) different node addition rates for each different value of γ. This is not proper: for the study of nonlinear forms of attachment, one should vary only the parameter γ, but certainly not the node addition rates at the same time.

This explains for example why our results in this paper for quadratic attachment (Sec. 3) are very different from those of Krapivsky *et al.*. It is in fact no surprise that no power laws emerge in their model: recall that for quadratic attachment a power law will have a second moment M_2 that is infinite. Then the node addition rate has to be also infinite in the model by Krapivsky *et al.* which is clearly absurd. If we set the node addition rate to a finite value that is different from the second moment (as we do in our model) then in the quadratic attachment case power laws can occur with realistic exponents (as demonstrated in Sec. 3, also for the general case including link removal and network churn).

This whole argument also applies to other values of γ. Our model predicts that for any superlinear form of attachment power laws may emerge (although the exponent may not necessarily assume values that are in the range of the empirical observations; the predicted exponent in our model generally is γ ± additional terms coming from link removal and network churn, roughly analogous to what we see in Eq. (8) for the quadratic case. For sublinear attachment it predicts that they generally don't (a proof of this is outside size and scope of this paper).

The general and essential point is that in a continuous-time model (expressed by ordinary differential equations) rates rather than probabilities occur, and the former do not require normalization. Rates occur because otherwise the dimensions (one over time, t^{-1}) at both sides in the dynamic equations can not be right. Even if rates occur (as in our model), it can be formally demonstrated that the degree distribution $q_k(t)$ is properly normalized for any time t, and therefore this holds also for the long-time limit n_k or q_k^{stat} (to prove this requires some matrix algebra).

The conclusion of no power laws for nonlinear attachment is therefore basically an artefact of an improper way of normalization. It is possible to rewrite continuous-time models in terms of transition probabilities that should be normalized (although this requires transform theory and convolution equations). But then the evolution equations do not take the shape of ordinary differential equations but of random-walk discrete-time equations that are event-driven. The connection of transition rates and properly normalized transition probabilities is given by:

$$p_k^{\pm} = \frac{\lambda_k^{\pm}}{\Lambda_k}, \text{ with } \Lambda_k = [\lambda_k^{+} + \lambda_k^{-} + w_0 + f_0], \tag{20}$$

where $p_k^{\pm}$ are the properly normalized transition probabilities for link attachment and removal. If we include node addition and removal, the total transition probability is properly normalized to unity. Without going into further detail, it is clear that this is fundamentally different from the normalization procedure followed by Krapivsky *et al.*.

Overall, the conclusion of Krapivsky *et al.* expressed in point 3 above that nonlinear attachment can not lead to power laws is generally unwarranted. Consequently, the hypothesis of nonlinear attachment cannot be ruled out.