Structural Preferential Attachment: Network Organization beyond the Link

Laurent Hébert-Dufresne, Antoine Allard, Vincent Marceau, Pierre-André Noël, and Louis J. Dubé
Département de Physique, de Génie Physique, et d'Optique, Université Laval, Québec (Québec), Canada G1V 0A6
(Received 27 May 2011; published 6 October 2011)

We introduce a mechanism which models the emergence of the universal properties of complex networks, such as scale independence, modularity and self-similarity, and unifies them under a scale-free organization beyond the link. This brings a new perspective on network organization where communities, instead of links, are the fundamental building blocks of complex systems. We show how our simple model can reproduce social and information networks by predicting their community structure and more importantly, how their nodes or communities are interconnected, often in a self-similar manner.

DOI: 10.1103/PhysRevLett.107.158702 PACS numbers: 89.75.Hc, 89.65.Ef, 89.75.Da, 89.75.Fb

A universal matter.—Reducing complex systems to their simplest possible form while retaining their important properties helps model their behavior independently of their nature. Results obtained via these abstract models can then be transferred to other systems sharing a similar simplest form. Such groups of analog systems are called universality classes and are the reason why some models apply just as well to the sizes of earthquakes or solar flares than to the sales number of books or music recordings [1]. That is, their statistical distributions can be reproduced by the same mechanism: preferential attachment. This mechanism has been of special interest to network science [2] because it models the emergence of power-law distributions for the number of links per node. This particular feature is one of the universal properties of network structure [3], alongside modularity [4] and self-similarity [5]. Previous studies have focused on those properties one at a time [3–8], yet a unified point of view is still wanting. In this Letter, we present an overarching model of preferential attachment that unifies the universal properties of network organization under a single principle.

Preferential attachment is one of the most ubiquitous mechanisms describing how elements are distributed within complex systems. More precisely, it predicts the emergence of scale-free (power-law) distributions where the probability P_k of occurrence of an event of order k decreases as an inverse power of k (i.e., $P_k \propto k^{-\gamma}$ with $\gamma > 0$). It was initially introduced outside the realm of network science by Yule [9] as a mathematical model of evolution explaining the power-law distribution of biological genera by number of species. Independently, Gibrat [10] formulated a similar idea as a law governing the growth rate of incomes. Gibrat's law is the sole assumption behind preferential attachment: the growth rates of entities in a system are proportional to their size. Yet, preferential attachment is perhaps better described using Simon's general balls-in-bins process [11].

Simon's model was developed for the distribution of words by their frequency of occurrence in a prose sample [12]. The problem is the following: what is the probability $P_{k+1}(i+1)$ that the $(i+1)$th word of a text is the $(k+1)$th occurrence of one of the $N_k(i)$ words which already appeared k times? By simply stating that $P_{k+1}(i+1) \propto k \cdot N_k(i)$, Simon obtained the desired distribution [Fig. 1(a)]. In this model, the nature of the system is hidden behind a simple logic: the "popularity" of an event is encoded in its number of past occurrences. More clearly, a word used twice is 2 times more likely to reappear next than a word used once. However, before its initial occurrence, a word has appeared exactly zero times, yet it has a certain probability p of appearing for the very first time. Simon's model thus produces systems whose distribution of elements falls as a power law of exponent $\gamma = (2-p)/(1-p)$.

On the matter of networks.—Networks are ensembles of potentially linked elements called nodes. In the late 1990s, it was found that the distribution of links per node (the degree distribution) featured a power-law tail for networks of diverse nature. To model these so-called scale-free networks, Barabási and Albert [3] introduced preferential attachment in network science. In their model, nodes are added to the network and linked to a certain number of existing nodes. The probability that the new node chooses an old one of degree k is proportional to kN_k, where N_k is the number of nodes of degree k. As the system goes to infinity, N_k falls off as k^{-3}.

From the perspective of complex networks, Simon's model may be regarded not as a scheme of throwing balls (e.g., word occurrences) in bins (e.g., unique words), but as an extreme case of scale-free networks where all links are shared within clearly divided structures. Obviously, both Simon's and the Barabási-Albert's (BA) models follow the preferential attachment principle. However, Simon's model creates distinct growing structures, whereas the BA model creates overlapping links of fixed size. By using the same principle, one creates order while the other creates randomness [Fig. 1(b)]. Our approach explores the systems that lie in between.

When structure matters.—The vast majority of natural networks have a modular topology where links are shared within dense subunits [4]. These structures, or communities,

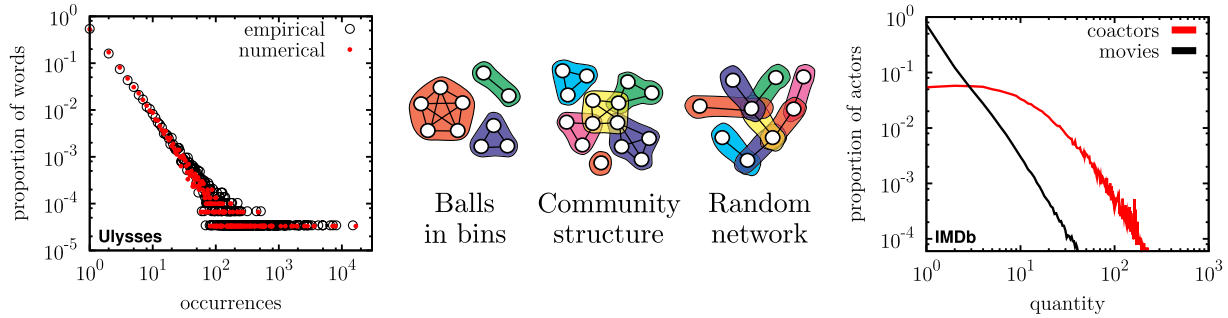


FIG. 1 (color online). (a) The distribution of words by their number of appearances in James Joyce's Ulysses (empirical data). The numerical data was obtained from a single realization of Simon's model with p equal to the ratio of unique words (30 030) on the total word count (267 350). (b) Schematization of the systems considered in this Letter, illustrating how order (Simon's model of balls in bins) and randomness (Barabási-Albert's model of random networks) coexist in a spectrum of complex systems. (c) The distribution of coactors and movies per actor in the Internet Movie Database since 2000. The organization moves closer to a true power law when looking at a higher structural level (i.e., movies versus coactors).

can be identified as social groups, industrial sectors, protein complexes or even semantic fields [13]. They typically overlap with each other by sharing nodes and their number of neighboring structures is called their community degree. This particular topology is often referred to as community structure [Fig. 1(b)]. Because these structures are so important on a global level, they must influence local growth. Consequently, they are at the core of our model.

The use of preferential attachment at a higher structural level is motivated by three observations. First, the number of communities an element belongs to, its membership number, is often a better indicator of its activity level than its total degree. For instance, we judge an actor taking part in many small dramas more active than one of a thousand extras in a single epic movie, just as we may consider a protein part of many complexes more functional than one found in a single big complex.

Second, studies have hinted that Gibrat's law holds true for communities within social networks [14]. The power-law distribution of community sizes recently observed in many systems (e.g., protein interaction, word association and social networks [13] or metabolite and mobile phone networks [15]) supports this hypothesis.

Third, degree distributions can deviate significantly from true power laws, while higher structural levels might be better suited for preferential attachment models [Fig. 1(c)].

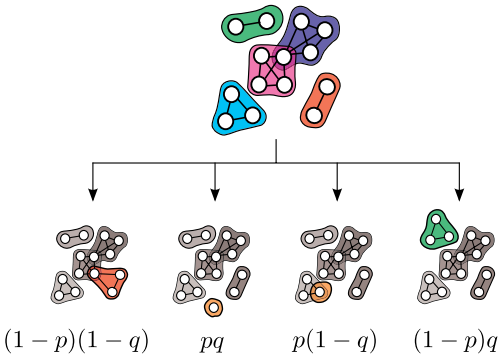
A simple model.—Simon's model assigns elements to structures chosen proportionally to their sizes, while the BA model creates links between elements chosen proportionally to their degree. We thus define structural preferential attachment (SPA), where both elements and structures are chosen according to preferential attachment. Here, links will not be considered as a property of two given nodes, but as part of structures that can grow on the underlying space of nodes and eventually overlap.

Our model can be described as the following stochastic process. At every time step, a node joins a structure. The node is a new one with probability q, or an old one chosen proportionally to its membership number with probability $1 - q$. Moreover, the structure is a new one of size s with probability p, or an old one chosen among existing structures proportionally to their size with probability $1 - p$. These two growth parameters are directly linked to two measurable properties: modularity (p) and connectedness (q) [Fig. 2]. Note that, at this point, no assumption is made on how nodes are linked within structures; our model focuses on the modular organization.

Whenever the structure is a new one, the remaining $s - 1$ elements involved in its creation are once again preferentially chosen among existing nodes. The basic structure size s is called the system base and refers to the smallest structural unit of the system. It is not a parameter of the model *per se*, but depends on the considered system. For instance, the BA model directly creates links, i.e. $s = 2$ (with $p = q = 1$), unlike Simon's model which uses $s = 1$ (with $q = 0$). All the results presented here use a node-based representation ($s = 1$), although they can be reproduced equally well via a link-based representation ($s = 2$). In fact, for sufficiently large systems, the distinction between the two versions seems mainly conceptual (see Supplemental Material for details [16]).

In our process, the growth of structures is not necessarily dependent on the growth of the network (i.e., the creation of nodes). Consequently, we can reproduce statistical properties of real networks without having to consider the large-size limit of the process. This allows our model to naturally include finite size effects (e.g., a distribution cutoff) and increases freedom in the scaling properties. In fact, we can follow S_n and N_m, respectively, the number of structures of size n and of nodes with m memberships, by writing master equations for their time evolution [17]:

$$\dot{S}_n(t) = (1 - p)\frac{(n - 1)S_{n-1}(t) - nS_n(t)}{[1 + p(s - 1)]t} + p\delta_{n,s}; \quad (1)$$



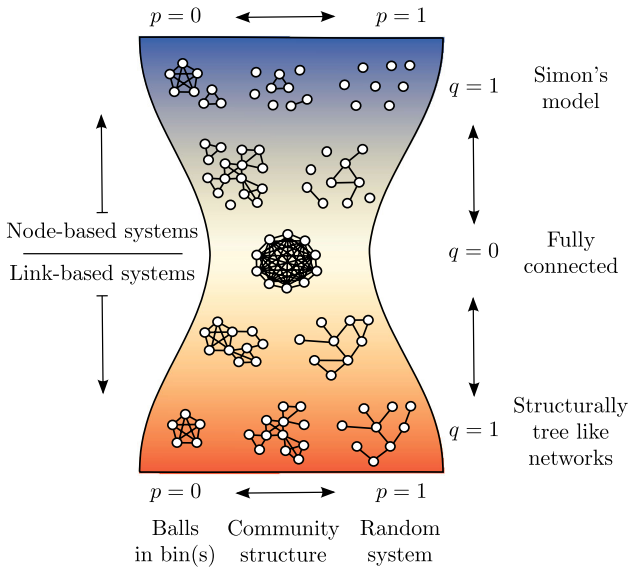


FIG. 2 (color online). (top) Representation of the possible events in a step of node-based SPA; the probability of each event is indicated beneath it. (bottom) A schematization of the spectrum of systems obtainable with SPA. Here, we illustrate the conceptual differences between node-based $s = 1$ and link-based systems $s = 2$: Simon's model ($q = 1$) creates structures of size one (nodes), while the BA model ($p = q = 1$) creates random networks through structures of size two (links).

$$\dot{N}_m(t) = [1 + p(s-1) - q] \times \frac{(m-1)N_{m-1}(t) - mN_m(t)}{[1 + p(s-1)]t} + q\delta_{m,1}. \quad (2)$$

Equations (1) and (2) can be transformed into ordinary differential equations for the evolution of the distribution of nodes per structure and structure per node by normalizing S_n and N_m by the total number of structures and nodes, pt and qt, respectively. One then obtains recursively the following solutions for the normalized distributions at statistical equilibrium, $\{\mathcal{S}_n^*\}$ and $\{\mathcal{N}_m^*\}$:

$$\mathcal{S}_n^* = \frac{\prod_{k=s}^{n-1} k\Omega_s}{\prod_{k=s}^{n}(1 + k\Omega_s)}, \qquad \Omega_s = \frac{1-p}{1 + p(s-1)} \quad (3)$$

TABLE I. Exponents of the power-law distributions of structures per element (membership) and of elements per structure (size) at statistical equilibrium. One easily verifies that the membership scaling of link-based systems with $p = q = 1$ corresponds to that of the BA model ($\gamma_N = 3$), and that node-based systems with $q = 1$ reproduce Simon's model. See Supplemental Material for the derivation [16].

System base s	Membership scaling γ_N	Size scaling γ_S
Node ($s = 1$)	$(2-q)/(1-q)$	$(2-p)/(1-p)$
Link ($s = 2$)	$[2(p+1)-q]/(1+q-p)$	$2/(1-p)$

$$\mathcal{N}_m^* = \frac{\prod_{k=1}^{m-1} k\Gamma_s}{\prod_{k=1}^{m}(1 + k\Gamma_s)}, \qquad \Gamma_s = \frac{1 + p(s-1) - q}{1 + p(s-1)}, \quad (4)$$

which scale as indicated in Table I, $\mathcal{N}_m^* \propto m^{-\gamma_N}$ and $\mathcal{S}_n^* \propto n^{-\gamma_S}$.

Results and discussions.—There are three distributions of interest which can be directly obtained from SPA: the membership, the community size, and the community degree distributions. In systems such as the size of business firms or word frequencies, these distributions suffice to characterize the organization. To obtain them, the SPA parameters, q and p, are fitted to the empirical scaling exponents of the membership and community size distributions. In complex networks, one may also be interested in the degree distribution. Additional assumptions are then needed to determine how nodes are interconnected within communities (specified when required).

The first set of results considered is the community structure of the coautorship network of an electronic preprints archive, the cond-mat arXiv circa 2005 [Fig. 3(a)], whose topology was already characterized using a clique percolation method [13]. Here, the communities are detected using the link community algorithm of Ahn *et al.* [15], confirming previous results.

Using only two parameters, our model can create a system of similar size with an equivalent topology according to the four distributions considered (community sizes, memberships, community degree and node degree). Not only does SPA reproduce the correct density of structures of size 2, 3, 4 or more, but it also correctly predicts how these structures are interconnected via their overlap, i.e., the community degree. This is achieved without imposing any constraints whatsoever for this property. The first portion of the community degree distribution is approximately exponential; a behavior which can be observed in other systems, such as the Internet [Fig. 3(b)] and both a protein interaction and a word-association network [13]. To our knowledge, SPA is the first growth process to reproduce such community structured systems.

Moreover, assuming fully connected structures, SPA correctly produces a similar behavior in the degree distribution of the nodes. Obtaining this distribution alone previously required two parameters and additional

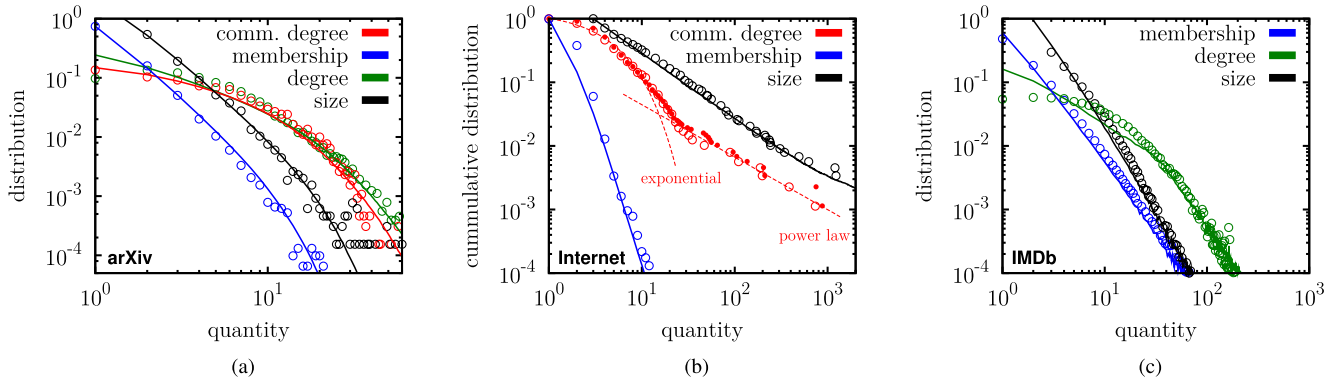


FIG. 3 (color online). Circles: distributions of topological quantities for (a) the cond-mat arXiv circa 2005; (b) Internet at the level of autonomous systems circa 2007; (c) the IMDb network for movies released since 2000. Solid lines: average over multiple realizations of the SPA process with (a) $p = 0.56$ and $q = 0.59$; (b) $p = 0.04$ and $q = 0.66$; (c) $p = 0.47$ and $q = 0.25$. For each realization, iterations are pursued until an equivalent system size is obtained. The Internet data highlights the transition between exponential and scale-free regimes in a typical community degree distribution. It is represented by a single realization of SPA (dots), because averaging masks the transition.

assumptions [7]. In contrast, SPA shows that this is a signature of a scale-free community structure. This is an interesting result in itself, since most observed degree distributions follow a power law only asymptotically. Furthermore, this particular result also illustrates how self-similarity between different structural levels (i.e., node degree and community degree distributions) can emerge from the scale-free organization of communities.

Finally, the Internet Movie Database coacting network is used to illustrate how, for bigger and sparser communities which cannot be considered fully connected, one can still easily approximate the degree distribution. We first observe that the mean density of links in communities of size n approximately behaves as $\log(n)/n$ (see Supplemental Material [16]). Then, using a simple binomial approximation to connect the nodes within communities, it is possible to approximate the correct scaling behavior for the degree distribution [Fig. 3(c)]. This method takes advantage of the fact that communities are, by definition, homogeneous such that their internal organization can be considered random.

Conclusion and perspective.—In this Letter, we have developed a complex network organization model where connections are built through growing communities, whereas past efforts typically tried to arrange random links in a scale-free, modular and/or self-similar manner. Our model shows that these universal properties are a consequence of preferential attachment at the level of communities: the scale-free organization is inherited by the lower structural levels.

Looking at network organization beyond the link is also useful to account for missing links [18] or to help realistic modeling [19,20]. For instance, this new paradigm of scale-free community structure suggests that nodes with the most memberships, i.e., structural hubs, are key elements in propagating epidemics on social networks or viruses on the Internet. These structural hubs connect many different neighborhoods, unlike standard hubs whose links can be redundant if shared within a single community.

There is no denying that communities can interact in more complex ways through time [21]. Yet, from a statistical point of view, those processes can be neglected in the context of a structurally preferential growth. Similarly, even though other theories generating scale-free designs exist [22], they could also benefit from generalizing their point of view to higher levels of organization.

The authors wish to thank Yong-Yeol Ahn *et al.* for their link community algorithm; Gergely Palla *et al.* for providing the CFinder software and the arXiv data set; Mark Newman for the Internet data set; and The Internet Movie Database available at www.imdb.com. This research was funded by CIHR, NSERC and FQRNT.

[1] M. Newman, Contemp. Phys. **46**, 323 (2005).
[2] A.-L. Barabási, Science **325**, 412 (2009).
[3] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[4] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).
[5] C. Song, S. Havlin, and H. Makse, Nature (London) **433**, 392 (2005).
[6] C. Song, S. Havlin, and H. Makse, Nature Phys. **2**, 275 (2006).
[7] R. Albert and A.-L. Barabási, Phys. Rev. Lett. **85**, 5234 (2000).
[8] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas, Phys. Rev. E **68**, 065103 (2003).
[9] G. U. Yule, Phil. Trans. R. Soc. B **213**, 21 (1925).
[10] R. Gibrat, *Les Inégalités Économiques* (Librairie du Recueil Sirey, Paris, 1931).
[11] H. A. Simon, *Models of Man* (John Wiley & Sons, New York, 1961).

[12] G. K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley Press, Cambridge, 1949).
[13] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, Nature (London) **435**, 814 (2005).
[14] D. Rybski, S. V. Buldyrev, S. Havlin, F. Liljeros, and H. A. Makse, Proc. Natl. Acad. Sci. U.S.A. **106**, 12640 (2009).
[15] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, Nature (London) **466**, 761 (2010).
[16] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.107.158702 for additional details, analysis and results.
[17] S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford University Press, New York, 2003).
[18] A. Clauset, C. Moore, and M. E. J. Newman, Nature (London) **453**, 98 (2008).
[19] L. Hébert-Dufresne, P.-A. Noël, V. Marceau, A. Allard, and L. J. Dubé, Phys. Rev. E **82**, 036115 (2010).
[20] B. Karrer and M. E. J. Newman, Phys. Rev. E **82**, 066118 (2010).
[21] G. Palla, A.-L. Barabási, and T. Vicsek, Nature (London) **446**, 664 (2007).
[22] J. Doyle and J. M. Carlson, Phys. Rev. Lett. **84**, 5656 (2000).

Structural preferential attachment: Network organization beyond the link

Supplemental Material

Laurent Hébert-Dufresne, Antoine Allard, Vincent Marceau,
Pierre-André Noël, and Louis J. Dubé

http://dynamica.phy.ulaval.ca/

Département de Physique, de Génie Physique et d'Optique
Université Laval, Québec, Québec, Canada G1V0A6

Human beings, viewed as behaving systems, are quite simple. The apparent complexity of our behavior over time is largely a reflection of the complexity of the environment in which we find ourselves.

— Herbert A. Simon (1916 - 2001)

1 On scale-free distributions and preferential attachment

A distribution is said to follow a power law when the probability of obtaining a particular value decreases as an inverse power of that value (i.e. $P_k \propto k^{-\gamma}$). The systems featuring this property are typically growing, which means that the measured value obtained on a given element will most likely be higher at a later time if given the chance. Examples include the distributions of city or business firm sizes, of written text by length, of computer file sizes, of species by biological genera, of number of telephone messages or airway passengers between pairs of cities and the sales of almost any branded product, even intervals between repetition of notes in Mozart's Bassoon Concerto in B♭ Major, all follow power-law distributions[1,2]. However, it is harder to find a common feature for other systems that also exhibit power-law distributions such as the sizes of earthquakes, moon craters or solar flares[2].

Because this behaviour is typically observed in growing systems, mechanism used to explain the origin of power laws are often based on stochastic growth processes. Of course, other models exist[2], such as the Highly Optimized Tolerance systems of Carlson and Doyle[3,4]. We will focus on one of the most well-known growth processes: preferential attachment (also known as the Yule process[5], Gibrat's law[6], rich-get-richer mechanism and cumulative advantage[7] amongst many names). In essence, preferential attachment is a simple urn scheme where balls are thrown in bins at a rate proportional to the number of balls already in a given bin. The balls can represent money being invested, people taking residence or a new scientific article being published; while bins can respectively represent business firms, cities, scientists and so on. In all cases, the probability P_k that a ball ends up in a bin which already contains k balls is proportional to k (i.e. $P_k = a + bk$), such that the relative growth rate of a structure is (if $a = 0$), or converges toward (if $a \neq 0$), a constant. In our Letter, we generalize preferential attachment by using Simon's model[8] which uses a null ground state ($a = 0$), but introduces the probability p that the bin used in an attachment event (when a ball is thrown) is a new one as yet empty. The absolute number $S_k(t)$ of bins containing k balls after t time steps can be followed by writing:

$$S_k(t+dt) = S_k(t) + dt\left[(1-p)\frac{(k-1)S_{k-1}(t) - kS_k(t)}{t} + p\delta_{k,1}\right] . \tag{1.1}$$

Transforming (1.1) into an equation for the proportion $\mathcal{S}_k(t)$ of bins which contain k balls after t time steps can be done by noting that $\{S_k(t)\}$ is simply $\{\mathcal{S}_k(t)\}$ multiplied by the total number of bins at time t: pt. We can now write

$$p(t+dt)\mathcal{S}_k(t+dt) = pt\mathcal{S}_k(t) + dt\Big\{p\,(1-p)\,[(k-1)\,\mathcal{S}_{k-1}(t) - k\mathcal{S}_k(t)] + p\delta_{k,1}\Big\} \tag{1.2}$$

from which we can obtain a simple ODE of the form:

$$\lim_{dt\to 0}\frac{(t+dt)\mathcal{S}_k(t+dt) - t\mathcal{S}_k(t)}{dt} = \frac{d}{dt}\left[t\mathcal{S}_k(t)\right] = (1-p)\left[(k-1)\,\mathcal{S}_{k-1}(t) - k\mathcal{S}_k(t)\right] + \delta_{k,1} . \tag{1.3}$$

The $\{\mathcal{S}_k^*\}$ ensemble at equilibrium (where $\frac{d}{dt}\mathcal{S}_k(t) = 0$) can be found by the following method:

$$\frac{d}{dt}\left[t\mathcal{S}_k(t)\right] = \mathcal{S}_k(t) + t\frac{d}{dt}\mathcal{S}_k(t) = (1-p)\left[(k-1)\,\mathcal{S}_{k-1}(t) - k\mathcal{S}_k(t)\right] + \delta_{k,1} \tag{1.4}$$

$$\mathcal{S}_k^* = (1-p)\Big[(k-1)\,\mathcal{S}_{k-1}^* - k\mathcal{S}_k^*\Big] + \delta_{k,1} = \frac{(1-p)\,(k-1)\,\mathcal{S}_{k-1}^* + \delta_{k,1}}{(1+k\,(1-p))} \tag{1.5}$$

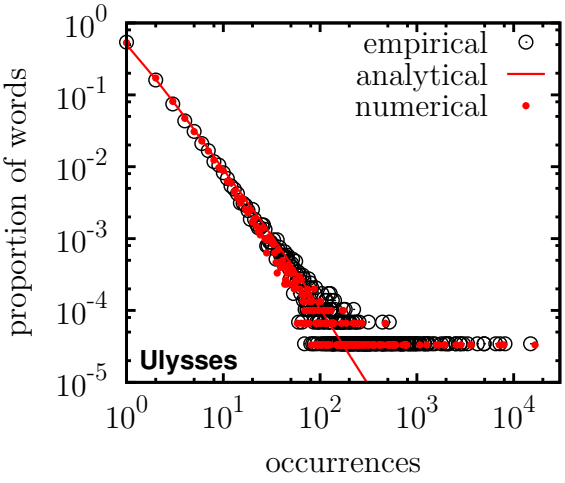


Figure 1: The distribution of words by their number of appearances in James Joyce's Ulysses (empirical data). The numerical data was obtained from a single iteration of Simon's model with an empirical measure of p equal to the ratio of unique words (30 030) on the total word count (267 350). The analytical curve is obtained from Eq. (1.1).

or

$$\mathcal{S}_k^* = \frac{\prod_{m=1}^{k-1} m\,(1-p)}{\prod_{m=1}^{k} [1 + m\,(1-p)]} \quad \forall\ k > 0 \tag{1.6}$$

so that

$$\frac{\mathcal{S}_k^*}{\mathcal{S}_{k-1}^*} = \frac{(k-1)\,(1-p)}{1 + k\,(1-p)} \tag{1.7}$$

which can be shown to yield a power law

$$\lim_{k\to\infty} \frac{\mathcal{S}_k^*}{\mathcal{S}_{k-1}^*} = \left(\frac{k}{k-1}\right)^{-\gamma} \tag{1.8}$$

of exponent

$$\gamma = \lim_{k\to\infty} \left\{ \log\left(\frac{(k-1)(1-p)}{1+k(1-p)}\right) \Big/ \log\left(\frac{k-1}{k}\right) \right\} = \frac{2-p}{1-p}\,. \tag{1.9}$$

Simon's process thus produces systems which feature a scale-free arrangement of elements (balls) by structure (bins) with a scaling exponent $\in [2, \infty)$ so that the distribution is always normalizable.

2 Summary of the model and main results

In Simon's model of preferential attachment, new balls are thrown in bins which are, with probability $(1-p)$, chosen proportionally to their size or, with probability p, new bins as yet empty. We generalize it by allowing balls to be thrown more than once and thus introduce a second probability, q, which is to balls what p is to bins. We just suppose that all balls are tagged with a particular number and that we can guess what number will be on the next ball by using the preferential attachment principle. This idea that balls can belong to more than one bin is fundamental to many systems. Most complex systems are connected systems, i.e. *networks*, such that the bins (structures in a network) used in preferential attachment should be connected

to one another by sharing balls (nodes in a network). For example, sizes of communities on online social networks have been shown to follow a power-law distributions[9], but the same user (ball) can be found in multiple communities (bins).

With this simple model, we are able to reproduce the statistical properties of systems that are neither completely structured, nor completely random. In fact, we show that systems produced according to our *structural preferential attachment* principle have a topology similar to that of real networks featuring *community structure*. More precisely, we reproduce the distributions of balls by bin (sizes), bins by ball (memberships) and of neighbouring bins by bin (community degree) for social (*cond-mat arXiv*, Internet Movie Database) and information (Internet) networks. In doing so, we find that most universal properties of networks are united under a topology best described as a scale-free community structure.

3 Structural preferential attachment (SPA)

This section details the structural preferential attachment model. Using mean-field equations similar to those used in section 1, we then obtain the thermodynamic limit of the process.

3.1 Description

If, as studies have recently implied[10,11], clusters of nodes are the fundamental building blocks of complex networks, then they should influence the local growth of our model. For network growth, this means that nodes should be linked via these structures instead of trying to build these structures by randomly linking nodes. *Structural preferential attachment* (SPA) is thus based on the following idea: instead of only selecting bins according to the preferential attachment principle, we also allow balls to be thrown more than once (assuming they are tagged and come in multiple copies) and thus select them just as we select bins. We then need two parameters: Simon's probability p that the chosen urn is a new one as yet empty, and q which we define as the probability that the chosen ball is a new one.

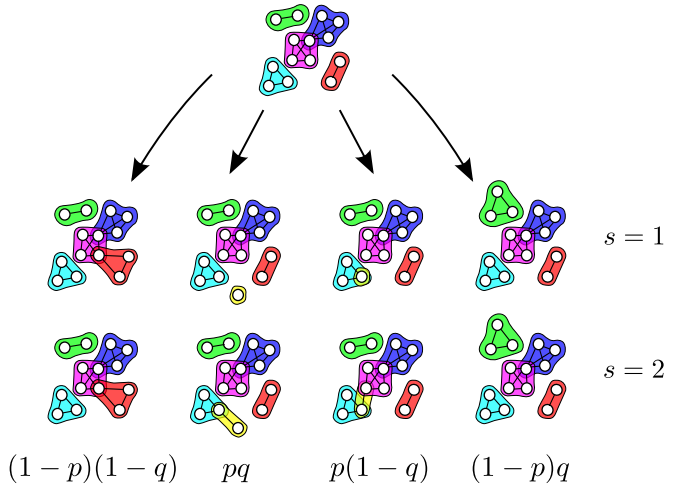


Figure 2: A step of structural preferential attachment. The probability of each type of events is indicated beneath it, with schematization of possible options for that step. When new structures appear, the possible events depend on the nature of the system: node based ($s = 1$) or link based ($s = 2$).

A step of SPA is illustrated on Fig. 2. Whenever we need to choose a structure or a node, this is done

proportionally to its past activity. Remember, this does not imply that the number of memberships of a node (i.e. the number of structures to which it belongs) directly affects its probability of being recruited. Its number of memberships is simply the only information we have and thus encodes whatever property this node might have that makes it popular. Similarly, it is natural to assume that structures probably recruit members at a rate proportional to their current size. This does not mean that a community is more active because it is bigger, but that it is more active because of a hidden property encoded in its size.

SPA encompasses both Simon's model and the well-known model of network growth introduced by Barabási and Albert[12]. Between structured systems and random networks, SPA produces network with the aforementioned scale-free community structure. Noteworthy is the fact that only two parameters are needed in this process (one for the modularity of the system and the other for its connectivity) and that these two parameters are expected to be measurable (or estimable) in growing complex networks.

It is more natural to consider that the most basic structures are of size one (i.e. one node) like in Simon's urn scheme, because structures are simply an extension of the concept of node. However, in most real networks (e.g. World Wide Web), the only information that we have are the links; hence the most basic structures found in such networks are of size two. We will thus consider the most general case with an undefined basic size s for structures. Note that this size is not a parameter of the model, but depends on the nature of the system considered. We will refer to the case $s = 1$ as node-based systems (the fundamental units being independent nodes) and to the case $s = 2$ as link-based systems (links being the basic unit).

3.2 Mathematical description

One can describe this new process in a manner akin to the one used for Simon's simpler model. We shall follow N_m, the number of nodes whose number of memberships is m, and S_n, the absolute number of structures enclosing n nodes. Using the same logic behind equation (1.1), it is straightforward to obtain

$$N_m(t+dt) = N_m(t) + dt\left\{ \left[1 + p(s-1) - q\right] \frac{(m-1)N_{m-1}(t) - mN_m(t)}{\left[1 + p(s-1)\right] t} + q\delta_{m,1} \right\} . \tag{3.1}$$

Similarly, we can write for the structures:

$$S_n(t+dt) = S_n(t) + dt\left\{ (1-p) \frac{(n-1)S_{n-1}(t) - nS_n(t)}{\left[1 + p(s-1)\right] t} + p\delta_{n,s} \right\} . \tag{3.2}$$

We once again transform these equations into time derivatives and switch variables to follow the distribution of nodes by memberships and structures by size, $\mathcal{N}_m$ and $\mathcal{S}_n$ respectively. We simply normalize nodes and structures number by the total number of each entity (i.e. qt nodes and pt structures at time t):

$$q\frac{d}{dt}\left[t\mathcal{N}_m(t)\right] = q\frac{1 + p(s-1) - q}{1 + p(s-1)}\left[(m-1)\mathcal{N}_{m-1}(t) - m\mathcal{N}_m(t)\right] + q\delta_{m,1} \tag{3.3}$$

$$p\frac{d}{dt}\left[t\mathcal{S}_n(t)\right] = p\frac{(1-p)}{1 + p(s-1)}\left[(n-1)\mathcal{S}_{n-1}(t) - n\mathcal{S}_n(t)\right] + p\delta_{n,s} \tag{3.4}$$

From these expressions, we obtain the values at equilibrium:

$$\text{for } m \geq 1 \qquad \mathcal{N}_m^*(s) = \frac{\prod_{k=1}^{m-1} k\Gamma_s}{\prod_{k=1}^{m}(1 + k\Gamma_s)} \quad \text{where} \quad \Gamma_s = \frac{1 + p(s-1) - q}{1 + p(s-1)} , \tag{3.5}$$

$$\text{for } n \geq s \qquad \mathcal{S}_n^*(s) = \frac{\prod_{k=s}^{n-1} k\Omega_s}{\prod_{k=s}^{n}(1 + k\Omega_s)} \quad \text{where} \quad \Omega_s = \frac{1-p}{1 + p(s-1)} , \tag{3.6}$$

whose limit fall with the following scaling exponents (summarized in Table 1):

$$\gamma_N = 1 + \frac{1}{\Gamma_s} \; ; \qquad \gamma_S = 1 + \frac{1}{\Omega_s} \tag{3.7}$$

system base s	membership scaling γ_N	size scaling γ_S
node ($s = 1$)	$(2 - q) / (1 - q)$	$(2 - p) / (1 - p)$
link ($s = 2$)	$[2(p + 1) - q] / (1 + q - p)$	$2/ (1 - p)$

Table 1: **Scaling behaviour.** Exponents of the power-law distributions of structures per element and of elements per structure at statistical equilibrium.

4 The data

This section gives more details on the datasets used in the Letter and on the methods employed to characterize their topology.

4.1 Community detection

Community detection in networks is a challenge in itself. In order to characterize the networks used in this work, two independent and completely different algorithms were used: a link community algorithm[10] and the clique percolation method of CFinder[11]. Results use the link community algorithm, because it proved to be faster and better suited to detect *communities within communities*. When possible, CFinder was used for cross-checking the community partition.

Link communities[10] This algorithm assigns links, instead of nodes, to communities. Groups of links (one or more) are considered as a community depending on how similar the neighbourhoods of their nodes are. The similarity of ensemble of nodes is measured by their Jaccard similarity coefficient. The correct community partition is then selected according to a Jaccard threshold given by the user. A large community can thus be composed of different smaller communities where the similarity of their members' neighbourhoods is higher than in the larger community. The link community algorithm proved to be quite efficient at detecting these nested communities.

CFinder and clique percolation[11] The original clique percolation method used by CFinder is designed to locate the k-clique communities of unweighted, undirected networks. This community definition is based on the observation that a typical member in a community is linked to many other members, but not necessarily to all other nodes in the community. In other words, a community can be interpreted as a union of smaller complete (fully connected) subgraphs that share nodes. Such complete subgraphs in a network are called *k-cliques*, where k refers to the number of nodes in the subgraph, and a *k-clique-community* is defined as the union of all k-cliques that can be reached from each other through a series of adjacent k-cliques. Two k-cliques are said to be adjacent if they share $k - 1$ nodes. CFinder is available at *http://cfinder.org/*.

4.2 Internet Movie Database

The dataset used for the co-acting network of IMDb consists only of movies released after December 31st 1999. Interestingly, the degree distribution is almost identical to that published a decade earlier[12] which consisted of all movies released before the turn of the century. This suggests, since the two networks contain distinct and exclusive ensembles of movies, that the growth parameters of the IMDb network are constant. The network contains 7 665 259 links between 716 463 nodes (actors), where two actors share a link if they are credited alongside another for at least one movie. It was only analysed using the link community algorithm, because of memory issues with CFinder. The organization levels corresponding to actual movies, which is how the dataset was originally compiled, was deemed unsuitable for the study because of the presence of economic (limiting the number of actors in a movie) and artistic (typically requiring a minimal number of characters in a movie) constraints. We believe that a community detection process on the network actually frees the system from these constraints and yield communities of actors linked by genre, time, location, etc.

4.3 arXiv

The cond-mat arXiv database uses articles published at *http://arxiv.org/archive/cond-mat* between April 1998 and February 2004. In this network, an article written by n co-authors contributes to a link of weight $(n-1)$ between every pair of authors. The unweighted network was obtained by deleting all links with a weight under the selected threshold of 0.1; resulting in a network of 125 959 links between 30 561 nodes (authors). This dataset was compiled, analysed and presented in[11].

4.4 Internet

This dataset is a symmetrized snapshot of the structure of the Internet at the level of autonomous systems, reconstructed from BGP tables posted at *archive.routeviews.org*. This snapshot was created by Mark Newman from data for July 22nd 2006 and was not previously published. The network contains 22 962 nodes and 48 436 links.

5 Comparing numerical results with empirical data

In this final section, we detail how SPA can be compared to real systems. Lastly, for systems with larger and sparser communities, we show how to go from a community organization back to a description of how links are distributed within the system.

5.1 Levels of organization

The first step when looking to compare the structure of real networks with systems produced by SPA is to analyse the empirical data. As mentioned earlier, our main algorithm (the link community algorithm[10]) has a single parameter to tune for community detection: its Jaccard threshold. The Jaccard threshold embodies *how similar* the neighbourhoods of the ends of two links must be in order for these links to be considered as part of the same link community. Tuning this parameter, demanding how tightly connected a group of nodes must be in order to be labeled as a community, allows us to look at different levels of organization within the network. If too small, the algorithm will most likely end up with communities corresponding to the connected components of the networks. If too big, significant communities will be broken up into

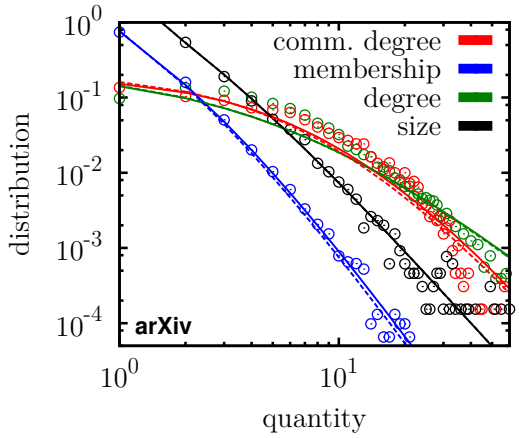


Figure 3: **Comparison between link-based and node-based SPA.** Community structure of the *cond-mat arXiv* as measured by the link community algorithm of Ahn *et al.* [10] (symbols) and analytical description of link-based SPA (continuous lines) or node-based SPA (dotted lines) used to approximate a link-based system by ignoring structures of size one. The degree distributions are obtained by assuming homogeneous mixing [13]. The two black lines perfectly overlap, while the membership distribution of node-based SPA is slightly shifted once we ignore structures of size one. Because node-based SPA results in a network with more than one component, its structure is actually closer to that of the real arXiv system.

different smaller ones. In this paper, we proceeded by sweeping this parameter in order to find the level of scale-free organization.

5.2 A note on node-based and link-based systems

All results presented in this work used a node-based version of SPA. Which means that new structures contain a single node and that they will remain disconnected from the other components of the network until they reach an older node. For the IMDb data, this choice is not even a question as the network contains many such satellites structures (even some of size one) which are disconnected from the giant component. In other systems, like the arXiv network, the choice can be more complicated. One might be tempted to use a link-based system process to reproduce the arXiv, since it is a co-author network and thus cannot contain isolated nodes. However, it does contain some disconnected components, which a link-based process like the Barabási-Albert model[12] is incapable of producing. Hence, it seemed logical to use the node-based process and simply remove the structures of size one (nodes who failed to co-author a paper) from the final system.

As a final point on the subject, it is interesting to note that we have been able to reproduce all results using both the node-based and link-based version of SPA. For example, see Fig. 3 for a comparison between the analytical prediction for link-based SPA and node-based SPA when ignoring structures of size one. In sufficiently large and connected systems, the distinction between the two seems mainly conceptual.

5.3 Results

Figure 4 presents our results for the arXiv network, the Internet and the Internet Movie Database. The arXiv data is completely shown, but the Internet is illustrated for communities of size 3 or bigger (as done by the authors of the detection algorithm[10]) because the algorithm can overestimate the number of communities of

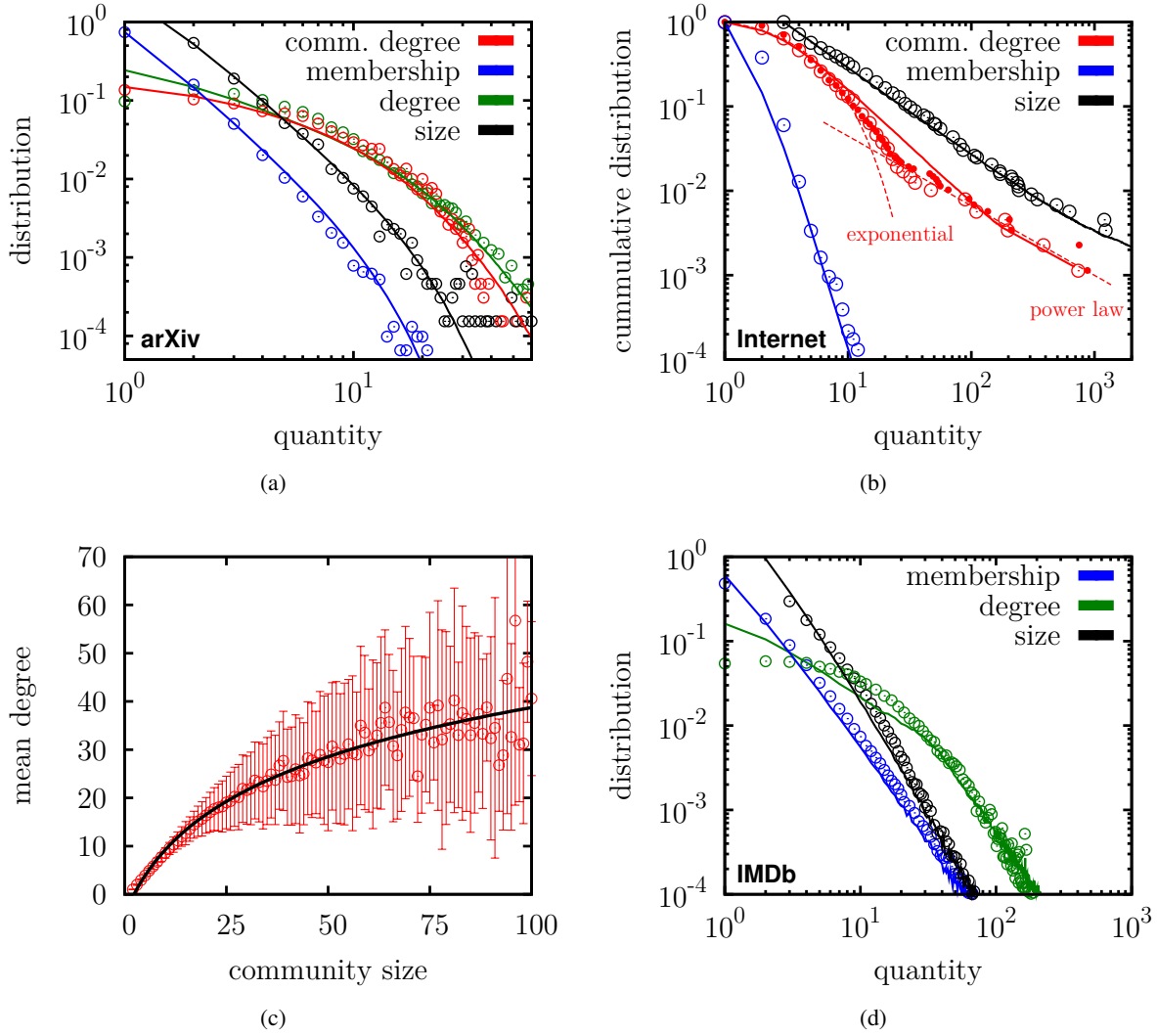


Figure 4: **Community structure and structural preferential attachment (SPA).** ⊙: distributions of topological quantities for the ensemble of (a) the *cond-mat* arXiv circa 2005 and (b) Internet at the level of autonomous system circa 2007. Solid lines: average over multiple realizations of the SPA process with (a) $p = 0.56$ and $q = 0.59$; b) $p = 0.04$ and $q = 0.66$. The empirical networks were analysed using the link community algorithm[10] with Jaccard thresholds of (a) 0.13 and (b) 0.08. (c) The mean number of links per node within a given community as a function of the community size in the IMDb network. The fit is done using a logarithmic function of the form $f(x) = a \cdot log(x + b) - c$. (d) The IMDb network studied with the link community algorithm using a Jaccard threshold of 0.18. The SPA simulation uses $p = 0.47$, $q = 0.25$ and the binomial connection scheme described in section 5.4 with the results of figure 4(c).

size 2 and the goal is here to highlight the connectedness of communities. For the IMDb, the community size distribution is normalized for communities of size 3 or bigger, but the communities of size 2 are considered in the membership and degree distributions. These results highlight how these systems follow a scale-free community structure and how SPA can be used to predict behaviour *outside* of the model's specification. More precisely, the numerical systems predict how the communities are interconnected via their overlap,

reproducing the exponential behaviour and the heavy tail of the community degree distribution.

It is interesting to note that averaging over many iterations of the SPA process highlights the distribution cut-off caused by the finite size of the system. This effect is mostly visible in Fig. 4(a). On the other hand, because the position of the transition between exponential and power-law behaviour observed in the cumulative community degree distribution is highly dependent on the amount of "leading" structures (i.e. the number of structures which are able to break away from the majority and thus have a significantly bigger size), it can differ slightly between two realizations of SPA. In this context, averaging over multiple iterations of the same process partly smooths out the transition. For this reason, a single realization of the model is also presented on Fig. 4(b) to better illustrate the behaviour of community degree in a finite system.

5.4 From communities, back to links

This last subsection presents results which, although preliminary, imply that individuals within a given social community can be approximated as being randomly connected.

The first step in shifting our point of view from communities back to links is to evaluate just how connected the communities of our systems are. Figure 4(c) illustrates the mean number of links per node within a given community as a function of the community size, which is found to grow logarithmically. Using this measure to determine the density of a structure of a given size, we simply throw a dice for each possible link to determine which links actually exist, while respecting the actual density of the network. This allows us to move from a potential degree distribution to an estimated degree distribution. If the binomial approximation (all links within a given community exist only with a certain probability) is correct, this estimated degree distribution should be close to the actual degree distribution of the system we are trying to reproduce. According to Fig. 4(d), this is indeed the case. It is easy to note that the number of nodes of small degree is overestimated by SPA. This is either a consequence of SPA producing too many small satellite components around the main network, or a consequence of IMDb sampling method, where an actor who has only acted in a small scale short film with one or two co-actors is more likely to be absent from the database than actors with hundreds of co-actors.

References

1. Zipf, G. K. *Human Behavior and the Principle of Least Effort* (Addison-Wesley Press, 1949).

2. Newman, M. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* **46**, 323 (2005).

3. Carlson, J. & Doyle, J. Highly optimized tolerance: Robustness and design in complex systems. *Phys. Rev. Lett.* **84**, 2529 (2000).

4. Doyle, J. & Carlson, J. Power laws, highly optimized tolerance, and generalized source coding. *Phys. Rev. Lett.* **84**, 5656 (2000).

5. Yule, G. U. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Phil. Trans. R. Soc. Lond. B* **213**, 21 (1925).

6. Gibrat, R. *Les inégalités économiques* (Librairie du Recueil Sirey, 1931).

7. de Solla Price, D. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science* **27**, 292 (1976).

8. Simon, H. A. *Models of Man* (John Wiley & Sons, 1961).

9. Zhang, Q. & Sornette, D. Predicted and verified deviation from Zipf s law in growing social networks. *arXiv* 1007.2650 (2010).

10. Ahn, Y.-Y., Bagrow, J. P. & Lehmann, S. Link communities reveal multiscale complexity in networks. *Nature* **466**, 761 (2010).

11. Palla, G., Derényi, I., Farkas, I. & Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814 (2005).

12. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509 (1999).

13. Newman, M. E. J. Properties of highly clustered networks. *Phys. Rev. E* **68**, 026121 (2003).