HAL
archives-ouvertes.fr

A Random Growth Model with any Real or Theoretical Degree Distribution

Frédéric Giroire, Stéphane Pérennes, Thibaud Trolliet

▶ **To cite this version:**

Frédéric Giroire, Stéphane Pérennes, Thibaud Trolliet. A Random Growth Model with any Real or Theoretical Degree Distribution. COMPLEX NETWORKS 2020 - 9th International Conference on Complex Networks and their Applications, Dec 2020, Madrid / Virtual, Spain. hal-03052144

HAL Id: hal-03052144
https://hal.inria.fr/hal-03052144

Submitted on 10 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Random Growth Model with any Real or Theoretical Degree Distribution*

Frédéric Giroire[1], Stéphane Pérennes[1], and Thibaud Trolliet[2]

[1] Université Côte d'Azur/CNRS, France
[2] INRIA Sophia-Antipolis, France

Abstract. The degree distributions of complex networks are usually considered to be power law. However, it is not the case for a large number of them. We thus propose a new model able to build random growing networks with (almost) any wanted degree distribution. The degree distribution can either be theoretical or extracted from a real-world network. The main idea is to invert the recurrence equation commonly used to compute the degree distribution in order to find a convenient attachment function for node connections - commonly chosen as linear. We compute this attachment function for some classical distributions, as the power-law, broken power-law, geometric and Poisson distributions. We also use the model on an undirected version of the Twitter network, for which the degree distribution has an unusual shape.

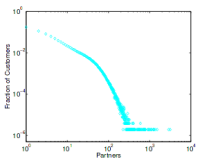
Keywords: Complex Networks, Random Growth Model, Preferential Attachment, Degree Distribution, Twitter

1 Introduction

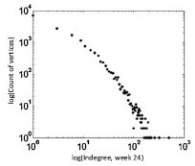
Complex networks appear in the empirical study of real world networks from various domains, such that social, biology, economy, technology, ... Most of those networks exhibit common properties, such as high clustering coefficient, communities, ... Probably the most studied of those properties is the degree distribution (named DD in the rest of the paper), which is often observed as following a power-law distribution. Random network models have thus focused on being able to build graphs exhibiting power-law DDs, such as the well-known Barabasi-Albert model [2] or the Chun-Lu model [7], but also models for directed networks [4] or for networks with communities [20]. However, this is common to find real networks with DDs not perfectly following a power-law. For instance for social networks, Facebook has been shown to follow a broken power-law[3] [13], while Twitter only has the distribution tail following a power-law and some atypical behaviors due to Twitter's policies, as we report in Section 5.1.

*This work has been supported by the French government through the UCA JEDI (ANR-15-IDEX-01) and EUR DS4H (ANR-17-EURE-004) Investments in the Future projects, by the SNIF project, and by Inria associated team EfDyNet.

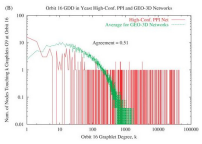
[3] We call a broken power-law a concatenation of two power-laws, as defined in [14].



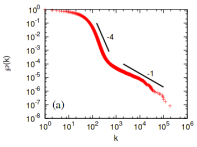
(a) DD of the number of unique callers and callees from a mobile phone operator. [21]



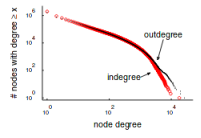
(b) In-DD between shop-to-shop recommendations from an online marketplace. [22]



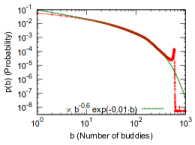
(c) Graphlet DD from a biological model. [19]



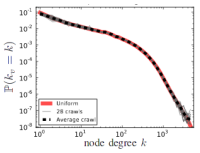
(d) DD of users of Cyworld, the largest online social network of South Korea. [1]



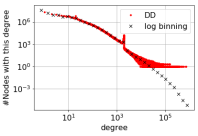
(e) DDs of users of Flickr, an online social network. [6]



(f) DD of the length of the contact list in Microsoft Messenger network. [15]



(g) DD of the number of friends from FaceBook, a social network. [13]



(h) Out-DD of the number of followees on Twitter. [23]

Fig. 1: DDs extracted from different seminal papers studying networks from various domains.

It is yet crucial to build models able to reproduce the properties of real networks. Indeed, some studies such as fake news propagation or evolution over time of the networks cannot always be done empirically, for technical or ethical reasons. Carrying out simulations with random networks created with well-built models is a solution to study real networks without directly experimenting on them. Those models have to create networks with similar properties as real ones, while staying as simple as possible.

In this paper, we propose a random growth model able to create graphs with almost any (under some conditions) given DD. Classical models usually choose the nodes receiving new edges proportionally to a linear attachment function $f(i) = i$ (or $f(i) = i + b$) [2, 4]. The theoretical DD of the networks generated by those models is computed using a recurrence equation. The main idea of this paper is to reverse this recurrence equation to express the attachment function f as a function of the DD. This way, for a given DD, we can compute the associated attachment function, and use it in a proposed random growth model to create graphs with the wanted DD. The given DD can either be theoretical, or extracted from a real network.

We compute the attachment function associated with some classical DD, homogeneous ones such as Poisson or geometric distributions, and heterogeneous ones such as exact power-law and broken power-law. We also study the undirected DD of a Twitter snapshot of 400 million nodes and 23 billion edges, extracted by Gabielkov et al. [10] and made available by the authors. We notice it has an atypical shape, due to Twitter's policies. We compute empirically the

associated attachment function, and use the model to build random graphs with this DD. A necessary condition is that the given DD must be defined for all degrees under the (arbitrary chosen) maximum value. However this condition can be circumvented doing an interpolation between existing points to estimate the missing ones, as discussed in Section 5.

The rest of the paper is organized as follows. We first discuss the related work in Section 2. In Section 3, we present the new model, and invert the recurrence equation to find the relation between the attachment function and the DD. We apply this relation to compute the attachment function associated to a power-law DD, a broken-power law DD, and other theoretical distributions. In Section 5 we apply our model on a real-world DD, the undirected DD of Twitter.

2 Related Work

The degree distribution has been computed for a lot of networks, in particular for social networks such as Facebook [13] or Microsoft Messenger [15]. Note that Myers et al. have also studied DDs for Twitter in [17], using a different dataset than the one of [10].

Questioning the relevance of power-law fits is not new: for instance, Clauset et al. [8] or Lima-Mendez and van Helden [16] have already deeply questioned the myth of power-law -as Lima-Mendez and van Helden call it-, and develop tools to verify if a distribution can be considered as a power-law or not. Clauset et al. apply the developed tools on 24 distributions extracted from various domains of literature, which have all been considered to be power-laws. Among them, "17 of the 24 data sets are consistent with a power-law distribution", and "there is only one case in which the power law appears to be truly convincing, in the sense that it is an excellent fit to the data and none of the alternatives carries any weight". In the continuity of this work, Broido and Clauset study in [5] the DD of nearly 1000 networks from various domains, and conclude that "fewer than 36 networks (4%) exhibit the strongest level of evidence for scale-free structure".

The study of Clauset et al. [8] only considered distributions which have a power-law shape when looking at the distribution in log-log. As a complement, we gathered DDs from literature which clearly do not follow power-law distributions to show their diversity. We extracted from literature DDs of networks from various domains: biology, economy, computer science, ... Each presented DD comes from a seminal well cited paper of the respective domains. They are gathered in Figure 1. Various shapes can be observed from those DDs, which could (by eyes) be associated with exponential (Fig. 1b, 1c), broken power-law (Fig. 1a, 1e, 1g), or even some kind of inverted broken power-law (Fig 1d). We also observe DDs with specific behaviors (Fig. 1f, 1h).

The first proposed models of random networks, such as the Erdős–Rényi model [9], build networks with a homogeneous DD. The observation that a lot of real-world networks follow power-law DDs lead Albert and Barabasi to propose their famous model with linear preferential attachment [2]. It has been followed by a lot of random growth models, e.g. [4, 7] also giving a DD in power-law. A few models permit to build networks with any DD: for instance, the configuration model [3, 18] takes as parameter a DD P and a number of nodes n, creates n

nodes with a degree randomly picked following P, then randomly connects the half-edges of every node. Goshal and Newman propose in [11] a model generating non-growing networks (where, at each time-step, a node is added and another is deleted) which can achieve any DD, using a method close to the one proposed in this paper. However, both of those models generate non-growing networks, while most real-world networks are constantly growing.

3 Presentation of the model

The proposed model is a generalization of the model introduced by Chun and Lu in [7]. At each time step, we have either a node event or an edge event. During a node event, a node is added with an edge attached to it; during an edge event, an edge is added between two existing nodes. Each node to which the edge is connected is randomly chosen among all nodes with a probability proportional to a given function f, called the *attachment function*. The model is as follows:

- ▷ We start with an initial graph G_0.
- ▷ At each time step t:
  - With probability p: we add a node u, and an edge (u, v) where the node v is chosen randomly between all existing nodes with a probability $\frac{f(deg(v))}{\sum_{w\in V} f(deg(w))}$;
  - With probability $(1-p)$: we add an edge (u, v) where the nodes u and v are chosen randomly between all existing nodes with a probability $\frac{f(deg(u))}{\sum_{w\in V} f(deg(w))}$ and $\frac{f(deg(v))}{\sum_{w\in V} f(deg(w))}$.

Note that the Chun-lu model is the particular case where $f(i) = i$ for all $i \geq 1$. We call *generalized Chun-Lu model* the proposed model where $f(i) = i + b$, for all $i \geq 1$ with $b > -1$.

3.1 Inversion of the recurrence equation

The common way to find the DD of classical random growth models is to study the recurrence equation of the evolution of the number of nodes with degree i between two time steps. This equation can sometimes be easily solved, sometimes not. But what matters for us is that the common process is to start from a given model -thus an attachment function f-, and use the recurrence equation to find the DD P. In this section, we show that the recurrence equation of the proposed model can be reversed such that, if P if given, we can find an associated attachment function f.

Theorem 1. *In the proposed model, if the attachment function is chosen as:*

$$\forall i \geq 1, f(i) = \frac{1}{P(i)} \sum_{k=i+1}^{\infty} P(k), \tag{1}$$

then the DD of the created graph is distributed according to P.[§]

[§]Note that Equation 1 can also be expressed as $f(i) = \frac{P(k>i)}{P(i)}$.

Proof. We consider the variation of the number of nodes of degree i $N(i,t)$ between a time step from t to (t+1). During this time step, a node with degree i may gain a degree and thus diminishes by 1 the number of nodes of degree i. This happens with a probability $p + 2(1-p)$ (the mean number of half-edges connected to existing nodes during a time step) $\times \frac{f(i)}{\sum_{j\geq 1} f(j)N(j,t)}$ (the probability for this particular node of degree i to be chosen). Since it is the same for all nodes of degree i, the number of nodes going from degree i to $i+1$ during a time step is $\big(p + 2(1-p)\big) \times \frac{f(i)}{\sum_{j\geq 1} f(j)N(j,t)} \times N(i,t)$. In the same way, some nodes with degree $i-1$ may be connected to an edge and increase the number of nodes of degree i. Finally, with probability p, a node of degree 1 is added. Gathering those contributions, taking the expectation, and using concentration results give the following equation:

$$\mathbb{E}[N(i,t+1)] - \mathbb{E}[N(i,t)] = \tag{2}$$
$$p\delta_{i,1} + (2-p)\frac{f(i-1)}{\sum\limits_{j\geq 1} f(j)\mathbb{E}[N(j,t)]}\mathbb{E}[N(i-1,t)] - (2-p)\frac{f(i)}{\sum\limits_{j\geq 1} f(j)\mathbb{E}[N(j,t)]}\mathbb{E}[N(i,t)]$$

where $\delta_{i,j}$ is the Kronecker delta. The first term of the right hand is the probability of addition of a node. The second (resp. third) term is the probability that a node of degree $i-1$ (resp. i) gets chosen to be the end of an edge. The factor $(2-p) = p + 2(1-p)$ comes from the fact that this happens with probability p during a node event (connection of a single half-edge) and with probability $2(1-p)$ during an edge event (possible connection of 2 half-edges).

Let $P(i) = \lim\limits_{t\to+\infty} \frac{\mathbb{E}[N(i,t)]}{pt}$ (the p in the denominator comes from the fact that $\mathbb{E}[N(t)] = pt$). We denote $g(i) = \frac{2-p}{p}\frac{f(i)}{\sum_{j\geq 1} f(j)P(j)}$. We first show that $g(i) = \frac{1}{P(i)} \sum\limits_{k=i+1}^{\infty} P(k)$. We will then show that we can choose $f = g$.

We use the following lemma from [7]:

Lemma 1. *Let (a_t), (b_t), (c_t) be three sequences such that $a_{t+1} = (1-\frac{b_t}{t})a_t + c_t$, $\lim\limits_{t\to+\infty} b_t = b > 0$, and $\lim\limits_{t\to+\infty} c_t = c$. Then $\lim\limits_{t\to+\infty} \frac{a_t}{t}$ exists and equals $\frac{c}{1+b}$.*

For $i = 1$, the equation becomes:

$$\mathbb{E}[N(1,t+1)] - \mathbb{E}[N(1,t)] = p - (2-p)\frac{f(1)}{\sum\limits_{j\geq 1} f(j)\mathbb{E}[N(j,t)]}\mathbb{E}[N(1,t)]. \tag{3}$$

Taking $a_t = \frac{\mathbb{E}[N(1,t)]}{p}$, $b_t = \frac{(2-p)f(1)}{p\sum_{j\geq 1} f(j)\frac{\mathbb{E}[N(j,t)]}{pt}}$, and $c_t = 1$, we have $\lim\limits_{t\to+\infty} b_t = g(1) > 0$ and $\lim\limits_{t\to+\infty} c_t = 1$. We can thus apply Lemma 1:

$$\lim_{t\to+\infty} \frac{\mathbb{E}[N(1,t)]}{pt} = P(1) = \frac{1}{1+g(1)}. \tag{4}$$

Now, $\forall i \geq 2$, taking $a_t = \frac{\mathbb{E}[N(i,t)]}{p}$, $b_t = \frac{(2-p)f(i)}{p\sum_{j\geq 1} f(j)\frac{\mathbb{E}[N(j,t)]}{pt}}$, and $c_t = \frac{(2-p)f(i-1)}{p\sum_{j\geq 1} f(j)\frac{\mathbb{E}[N(j,t)]}{pt}} \frac{\mathbb{E}[N(i-1,t)]}{pt}$, we have $\lim_{t\to+\infty} b_t = g(i) > 0$ and $\lim_{t\to+\infty} c_t = g(i-1)P(i-1)$. Lemma 1 gives:

$$\lim_{t\to+\infty} \frac{\mathbb{E}[N(i,t)]}{pt} = P(i) = \frac{g(i-1)P(i-1)}{1+g(i)}. \tag{5}$$

Iterating over Equation 5, we express g as a function of P:

$$g(i)P(i) = g(i-1)P(i-1) - P(i) = g(1)P(1) - \sum_{k=2}^{i} P(k) = 1 - \sum_{k=1}^{i} P(k)$$

$$\implies g(i) = \frac{1}{P(i)} \sum_{k=i+1}^{\infty} P(k) \tag{6}$$

Now, notice that:

$$\sum_{k=1}^{\infty} g(k)P(k) = \sum_{k=1}^{\infty} \frac{2-p}{p} \frac{f(k)}{\sum_{k'=1}^{\infty} f(k')P(k')} P(k) = \frac{(2-p)}{p}. \tag{7}$$

So $g(i)$ satisfies $g(i) = \frac{2-p}{p} \frac{g(i)}{\sum_{k=1}^{\infty} g(k)P(k)}$. Hence the attachment function can be chosen as $f = g$, which concludes the proof. □

For a given probability law, Theorem 1 can be used to compute the attachment function which, when used in the model, will give this probability law as DD.

With the presented model, we also have an implicit constraint between the mean degree and the parameter p. Indeed by construction, we have $\mathbb{E}[N(t)] = pt$ and $\mathbb{E}(|E|(t)) = t$ with $|E|(t)$ the number of edges at time t, leading to a mean-degree of $\frac{1}{p}$. But the mean-degree can also be expressed as $\sum_{k\geq 1} kP(k)$.

Condition 1 *The parameter p has to satisfy:*

$$\frac{1}{p} =< k > \tag{8}$$

We can finally combine the previous results and present the method to build a random network with a fixed DD:

1) Use Equation 1 to compute f from P;
2) Compute p using Condition 1;
2) Build the graph with the proposed model, given (f, p) as parameters.

Name	P(i)	f(i)	Condition
Generalized Chun-Lu	$C\frac{\Gamma(i+b)}{\Gamma(i+b+\alpha)}$	$\frac{1}{\alpha-1}i+\frac{b}{\alpha-1}$	$p=\frac{\alpha-2}{\alpha+b-1}$
Exact Power-Law	$\frac{i^{-\alpha}}{\zeta(\alpha)}$	$\frac{\zeta(\alpha,i+1)}{i^{-\alpha}}$	$p=\frac{\zeta(\alpha)}{\zeta(\alpha-1)}$
Geometric Law	$q(1-q)^{i-1}$	$\frac{1-q}{q}$	$p=q$
Poisson Law	$\frac{1}{e^{\lambda}-1}\frac{\lambda^i}{i!}$	$e^{\lambda}\frac{\gamma(i+1,\lambda)}{\lambda^i}$	$p=\frac{1-e^{-\lambda}}{\lambda}$
Broken Power-Law	$\begin{cases} C\frac{\Gamma(i+b_1)}{\Gamma(i+b_1+\alpha_1)} & \text{if } i\leq d \\ C\gamma\frac{\Gamma(i+b_2)}{\Gamma(i+b_2+\alpha_2)} & \text{if } i>d \end{cases}$	cf eq. 17& 18	cf eq. 16

Table 1: Attachment functions f and conditions on p for some classical probability distributions P. $\zeta(s)$ is the Riemann zeta function, $\zeta(s,q)$ the Hurwitz zeta function, and $\gamma(a,x)$ is the lower incomplete Gamma function.

4 Application to some distributions

We now apply Equation 1 to compute the attachment function for some classical distributions. We first start in Section 4.1 from the distribution obtained with the generalized Chun-Lu model to show we find a linear dependence, as expected. We then compute in Section 4.2 the associated attachment function of the broken power-law distribution. Using similar computations (which can be found in Report [12]), we computed the attachment function of other classical distributions. Table 1 summarizes those results.

4.1 Preliminary: Generalized Chun-Lu model

As a first example, by taking a power-law DD, we should be able to find a linear probability distribution for the generalized Chun-Lu model.

In the general Chun-Lu model, we can show that the real DD is not an exact power-law but a fraction of Gamma function -equivalent to a power-law for high degrees- of the form:

$$\forall i \geq 1, P(i) = C\frac{\Gamma(i+b)}{\Gamma(i+b+\alpha)} \underset{i\gg 1}{\sim} i^{-\alpha} \tag{9}$$

where $C=(\alpha-1)\frac{\Gamma(b+\alpha)}{\Gamma(b+1)}$, and $\alpha>2$. The choice of α determines the slope of the DD, while the choice of b determines the mean-degree of the graph.

Constraint on p: Condition 1 gives:

$$\frac{1}{p}=\sum_{k=1}^{\infty}kP(k)=(\alpha-1)\frac{\Gamma(b+\alpha)}{\Gamma(b+1)}\times\frac{\alpha^2+\alpha(2b-1)+b(b-1)}{(\alpha-2)(\alpha-1)}\frac{\Gamma(b+1)}{\Gamma(\alpha+b+1)}$$

$$\implies p=\frac{(\alpha-2)}{\alpha+b-1} \tag{10}$$

Attachment function f: Using Theorem 1:

$$f(i) = \frac{1}{P(i)} \sum_{k \geq i+1} P(k) = \frac{\Gamma(i+b+\alpha)}{\Gamma(i+b)} \frac{\Gamma(i+b+1)}{(\alpha-1)\Gamma(i+\alpha+b)} \tag{11}$$

$$\implies f(i) = \frac{1}{\alpha-1} i + \frac{b}{\alpha-1} \tag{12}$$

As expected, we find a linear attachment function. To create a graph with a wanted slope α and mean-degree p^{-1}, one only has to choose α as the wanted slope and b following equation 10. In the particular case $b = 0$, we recover the Chun-Lu model of [7], with a slope of $\alpha = 2 + \frac{p}{2-p}$ as expected.

4.2 Broken Power-law

We now study the case of a broken power-law, corresponding to the DD of real world complex networks, as discussed in Section 2. which was the one we were interested in initially. We consider a distribution of the form:

$$P(i) = \begin{cases} C \frac{\Gamma(i+b_1)}{\Gamma(i+b_1+\alpha_1)} & \text{if } i \leq d \\ C\gamma \frac{\Gamma(i+b_2)}{\Gamma(i+b_2+\alpha_2)} & \text{if } i > d \end{cases} \tag{13}$$

where d, b_1, α_1, b_2, and α_2 are parameters of our distribution such that $\alpha_1 > 2$, $\alpha_2 > 2$, C a normalisation constant, and γ chosen in order to obtain continuity for $i = d$. As seen in section 4.1, the ratio of gamma functions is close to a power-law as soon as i gets large. Hence, this distribution corresponds to two powers-laws, with different slopes, and a switch between the two at the value d.

We can easily find the continuity constant γ, since it verifies:

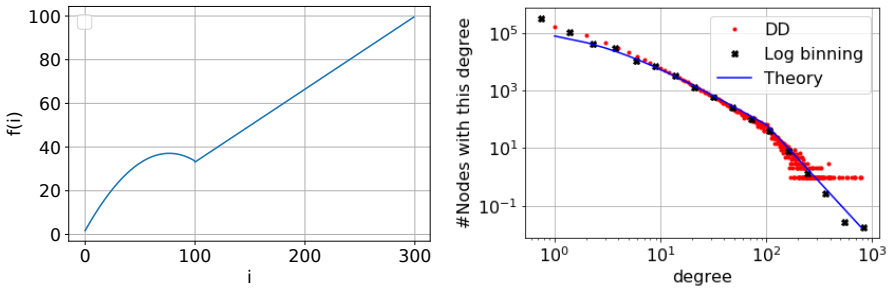
$$\frac{\Gamma(d+b_1)}{\Gamma(d+b_1+\alpha_1)} = \gamma \frac{\Gamma(d+b_2)}{\Gamma(d+b_2+\alpha_2)} \implies \gamma = \frac{\Gamma(d+b_1)\Gamma(d+b_2+\alpha_2)}{\Gamma(d+b_1+\alpha_1)\Gamma(d+b_2)}. \tag{14}$$

Constraints on C and p: The value of C can be computed by summing over all degrees:

$$C = \Big(\sum_{k=1}^{\infty} P(k) \Big)^{-1} = \Big(\frac{1}{\alpha_1 - 1} \frac{\Gamma(b_1+1)}{\Gamma(\alpha_1+b_1)} + \frac{\Gamma(b_1+d)}{\Gamma(\alpha_1+b_1+d)} \big(\frac{b_2+d}{\alpha_2-1} - \frac{b_1+d}{\alpha_1-1} \big) \Big)^{-1} \tag{15}$$

Using Condition 1, p is defined by the following equation:

$$\begin{aligned} \frac{1}{pC} &= \sum_{k=1}^{d} k \frac{\Gamma(k+b_1)}{\Gamma(k+b_1+\alpha_1)} + \gamma \sum_{k=d+1}^{\infty} k \frac{\Gamma(k+b_2)}{\Gamma(k+b_2+\alpha_2)} \\ &= \frac{\alpha_1^2 + \alpha_1(2b_1-1) + b_1(b_1-1)}{(\alpha_1-2)(\alpha_1-1)} \frac{\Gamma(b_1+1)}{\Gamma(\alpha_1+b_1+1)} \\ &- \frac{\alpha_1^2(d+1) + \alpha_1(b_1(d+2)+d^2-1) + b_1(b_1-1) - d(d+1)}{(\alpha_1-2)(\alpha_1-1)} \frac{\Gamma(b_1+d+1)}{\Gamma(\alpha_1+b_1+d+1)} \\ &+ \gamma \frac{\alpha_2^2(d+1) + \alpha_2(b_2(d+2)+d^2-1) + b_2(b_2-1) - d(d+1)}{(\alpha_2-2)(\alpha_2-1)} \frac{\Gamma(b_2+d+1)}{\Gamma(\alpha_2+b_2+d+1)} \end{aligned} \tag{16}$$



(a) Theoretical attachment function f (b) DD of a random network

Fig. 2: Theoretical attachment function f and degree distribution of a random network for the broken power-law distribution. Parameters are $N = 5 \cdot 10^5$, $b_1 = b_2 = 1$, $\alpha_1 = 2.1$, $\alpha_2 = 4$ and $d = 100$.

Attachment function f: For the computation of the attachment function, we have to distinguish two cases:

Case 1: $i \geq d$

$$f(i) = \frac{\Gamma(i + b_2 + \alpha_2)}{\Gamma(i + b_2)} \frac{1}{\alpha_2 - 1} \frac{\Gamma(i + b_2 + 1)}{\Gamma(i + b_2 + \alpha_2)} = \frac{1}{\alpha_2 - 1} i + \frac{b_2}{\alpha_2 - 1} \tag{17}$$

We find a linear attachment function: indeed for $i > d$, we only take into account the second power-law, hence we expect to find the same result than in section 4.1.

Case 2: $i < d$

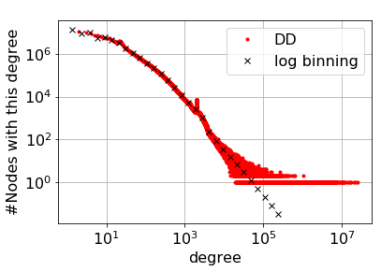
$$\begin{aligned} f(i) &= \frac{\Gamma(i + b_1 + \alpha_1)}{\Gamma(i + b_1)} \left(\sum_{k=i+1}^{d} \frac{\Gamma(k + b_1)}{\Gamma(k + b_1 + \alpha_1)} + \gamma \sum_{k=d+1}^{\infty} \frac{\Gamma(k + b_2)}{\Gamma(k + b_2 + \alpha_2)} \right) \\ &= \frac{i + b_1}{\alpha_1 - 1} + \frac{\Gamma(i + b_1 + \alpha_1)\Gamma(d + b_1)}{\Gamma(i + b_1)\Gamma(d + b_1 + \alpha_1)} \left(\frac{b_2 + d}{\alpha_2 - 1} - \frac{b_1 + d}{\alpha_1 - 1} \right) \end{aligned} \tag{18}$$

In this second case, we have a linear part, in addition to a more complicated part. Note that, for $(\alpha_1, b_1) = (\alpha_2, b_2)$, i.e., when the two power-laws are equals, this second term vanishes, letting as expected only the linear part. Figure 2a shows the shape of f. We see that, while the second part is linear as discussed before, the first part is sub-linear.

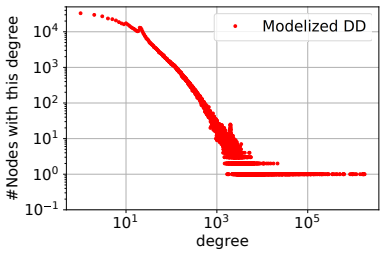
We used this attachment function to build a network using our model. The DD is shown in Figure 2b: we see we built a random network with a broken power-law distribution as wanted.

5 Real degree distributions

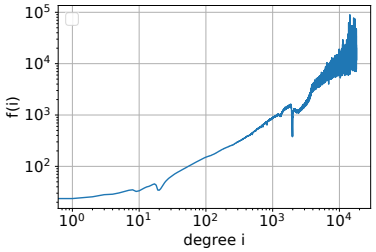
The model can also be applied to an empirical DD. Indeed, we observe in Theorem 1 that $f(i)$ only depends on the values $P(i)$ which can be arbitrary, that



(a) DD of the Twitter's undirected network.



(b) DD of a random network with $8 \cdot 10^5$ nodes using the attachment function of Figure 3c.



(c) Attachment function f resulting from the undirected DD of Twitter.

Fig. 3: Modelization of the undirected Twitter's graph.

is not following any classical function. This is a good way to model random networks with an atypical DD. As an example, we apply our model on the DD of an undirected version of Twitter, shown as having atypical behavior due to the Twitter policies. We start with a presentation of this DD, then apply our model to build a random graph with this distribution.

5.1 Undirected DD of Twitter

For this study, we use a Twitter snapshot from 2012, recovered by Gabielkov and Legout [10] and made available by the authors. This network contains 505 million nodes and 23 billion edges, making it one of the biggest social graph available nowadays. Each node corresponds to an account, and an arc (u, v) exists if the account u follows the account v. The in- and out-DDs are presented in [23].

In our case, we look at an undirected version of the Twitter snapshot. We consider the degree of each node as being the sum of its in- and out-degrees. The distribution of this undirected graph is presented in Figure 3a. We notice two spikes, around $d = 20$ and $d = 2000$. We do not know the reason of the first one (which could be social, or due to recommendation system). The second spike is explained by a specificity of Twitter: until 2015, to avoid bots which were following a very large number of users, Twitter limited the number of possible followings to $\max(2000, \text{number of followers})$. In other words, a user is allowed to follow more than 2000 people only if he is also followed by more than 2000 people.

This leads to a lot of accounts with around 2000 followings. This highlights the fact that some networks have their own specificities, sometimes due to intern policies, which cannot be modeled but by a model specifically built for them.

5.2 Modelization

Figure 3c presents the obtained form of the attachment function f computed using Equation 1 with the DD of Twitter. We notice that the overall function is mainly increasing, showing that nodes of higher degrees have a higher chance to connect with new nodes, like in classical preferential attachment models. We also notice two drops, around 20 and 2000. They are associated with the risings on the DD on the same degrees: to increase the amount of nodes with those degrees, the attachment function has to be smaller, so nodes with this degree have less chance to gain new edges.

We finally use our model with the empirical attachment function of Figure 3c. Note that, in an empirical study, P can be equal to zero for some degrees, for which no node has this degree in the network. In Twitter, the smallest of those degrees occurs around 18.000. In that case, f cannot be computed. To get around this difficulty, we interpolate the missing values of P, using the two closest smaller and bigger degrees of the missing points. Since we observe the probability distribution on a log-log scale, we interpolate between the two points as a straight line on a log-log scale, i.e., as a power-law function. We believe this is a fair choice since we only look at the tail of the distribution, which looks like a straight line, and since we interpolate between each pair of closest two points only, instead of fitting on the whole tail of the distribution.

The DD of a random network built with our model is presented in Figure 3b. For time computation reasons, the built network only has $N = 2 \cdot 10^5$ nodes, to be compared to the $5 \cdot 10^8$ nodes of Twitter. However, it is enough to verify that its DD shape follows the one of the real Twitter's DD: in particular we recognize the spikes around $d = 20$ and $d = 2000$.

6 Conclusion

In this paper, we proposed a new random growth model picking the nodes to be connected together in the graph with a flexible probability f. We expressed this f as a function of any distribution P, leading to the possibility to build a random network with any wanted degree distribution. We computed f for some classical distributions, as much as for a snapshot of Twitter of 505 million nodes and 23 billion edges. We believe this model is useful for anyone studying networks with atypical degree distributions, regardless of the domain. If the presented model is undirected, we also believe a directed version of it, based on the Bollobás et al. model [4], can be easily generalized from the presented one.

References

1. Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. Analysis of topological characteristics of huge online social networking services. In *Proceedings of the 16th int. conference on World Wide Web*, pages 835–844, 2007.

2. Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.
3. Béla Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European Journal of Combinatorics*, 1(4):311–316, 1980.
4. Béla Bollobás, Christian Borgs, Jennifer T Chayes, and Oliver Riordan. Directed scale-free graphs. In *SODA*, volume 3, pages 132–139, 2003.
5. Anna D Broido and Aaron Clauset. Scale-free networks are rare. *Nature communications*, 10(1):1–10, 2019.
6. Meeyoung Cha, Alan Mislove, and Krishna P Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World wide web*, pages 721–730, 2009.
7. Fan Chung, Fan RK Chung, Fan Chung Graham, Linyuan Lu, Kian Fan Chung, et al. *Complex graphs and networks*. American Mathematical Soc., 2006.
8. Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.
9. Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.
10. Maksym Gabielkov and Arnaud Legout. The complete picture of the twitter social graph. In *Proc. on CoNEXT student workshop*, pages 19–20. ACM, 2012.
11. Gourab Ghoshal and MEJ Newman. Growing distributed networks with arbitrary degree distributions. *The European Physical Journal B*, 58(2):175–184, 2007.
12. Frédéric Giroire, Stéphane Pérennes, and Thibaud Trolliet. A random growth model with any real or theoretical degree distribution. *arXiv preprint arXiv:2008.03831*, 2020.
13. Minas Gjoka, Maciej Kurant, Carter T Butts, and Athina Markopoulou. Walking in facebook: A case study of unbiased sampling of osns. In *IEEE INFOCOM*, 2010.
14. Gudlaugur Jóhannesson, Gunnlaugur Björnsson, and Einar H Gudmundsson. Afterglow light curves and broken power laws: a statistical study. *The Astrophysical Journal Letters*, 640(1):L5, 2006.
15. Jure Leskovec and Eric Horvitz. Planetary-scale views on a large instant-messaging network. In *Proc. of the 17th international conference on World Wide Web*, 2008.
16. Gipsi Lima-Mendez and Jacques van Helden. The powerful law of the power law and other myths in network biology. *Molecular BioSystems*, 5(12):1482–1493, 2009.
17. Seth A Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. Information network or social network?: the structure of the twitter follow graph. In *Proceedings of the 23rd Int. Conference on World Wide Web*, pages 493–498. ACM, 2014.
18. Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. Random graphs with arbitrary degree distributions and their applications. *Physical review E*, 2001.
19. Nataša Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007.
20. Arnaud Sallaberry, Faraz Zaidi, and Guy Melançon. Model for generating artificial social networks having community structures with small-world and scale-free properties. *Social Network Analysis and Mining*, 3(3):597–609, 2013.
21. Mukund Seshadri, Sridhar Machiraju, Ashwin Sridharan, Jean Bolot, Christos Faloutsos, and Jure Leskove. Mobile call graphs: beyond power-law and lognormal distributions. In *ACM SIGKDD*, pages 596–604, 2008.
22. Andrew T Stephen and Olivier Toubia. Explaining the power-law degree distribution in a social commerce network. *Social Networks*, 31(4):262–270, 2009.
23. Thibaud Trolliet, Nathann Cohen, Frédéric Giroire, Luc Hogie, and Stéphane Pérennes. Interest clustering coefficient: a new metric for directed networks like twitter. *arXiv preprint arXiv:2008.00517*, 2020.