A Random Growth Model with any Real or Theoretical Degree Distribution

Frédéric Giroire, Stéphane Pérennes, Thibaud Trolliet

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 follow a power law distribution. 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, the broken power-law, and the geometric distributions. We also use the model on an undirected version of the Twitter network, for which the degree distribution has an unusual shape. We finally show that the divergence of chosen attachment functions is directly linked to the heavy-tailed property of the obtained degree distributions.

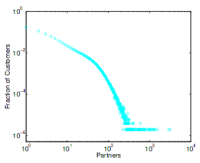
Keywords: Complex Networks, Random Growth Model, Preferential Attachment, Degree Distribution, Random Graphs, Heavy-Tailed Distributions.

1 Acknowledgements

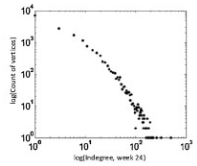
This work has been supported by the French government through the UCA JEDI (ANR-15-IDEX-01) and EUR DS4H (ANR-17-EURE-004) Investments in the Future projects, by the SNIF project, and by Inria associated team EfDyNet.

2 Introduction

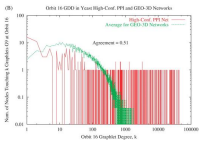
Complex networks appear in the empirical study of real world networks from various domains, such that social, biology, economy, technology, etc. Most of those networks exhibit common properties, such as a high clustering coefficient or the presence of communities. Probably the most studied of those properties 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 Chung-Lu model [7], but also



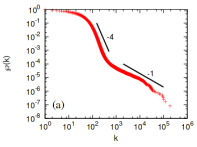
(a) DD of the number of unique callers and callees from a mobile phone operator. [25]



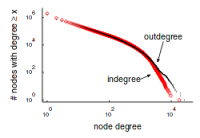
(b) In-DD between shop-to-shop recommendations from an online marketplace. [26]



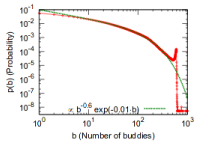
(c) Graphlet DD from a biological model. [22]



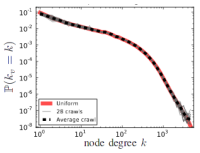
(d) DD of users of Cyworld, the largest online social network of South Korea. [1]



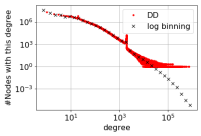
(e) DDs of users of Flickr, an online social network. [6]



(f) DD of the length of the contact list in Microsoft Messenger network. [16]



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



(h) Out-DD of the number of followees on Twitter. [27]

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

models for directed networks [4] or for networks with communities [24]. However, this is common to find real networks with DDs not perfectly following a power-law. For instance for social networks, Facebook DD has been shown to follow a broken power-law[3] [13], while Twitter one only has the distribution tail following a power-law along with some atypical behaviors due to Twitter policies, as we report in Section 6.1.

It is yet crucial to build models able to reproduce the properties of real networks. Indeed, some studies such as the propagation of fake news or the evolution over time of the networks cannot always be done empirically, for technical or ethical reasons. Carrying out simulations with random networks created from 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, i.e., proportionally to the degree of the nodes [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

[3] We call a broken power-law a concatenation of two power-laws, as defined in [15].

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 functions associated with some classical DDs, homogeneous ones such as the geometric distribution, and heterogeneous ones such as exact the power-law and the broken power-law distributions. We also study the undirected DD of a Twitter snapshot of 400 million nodes and 23 billion edges, extracted by Gabielkov et al. [11] and made available by the authors. We discuss its atypical shape, due to Twitter policies. We empirically compute its associated attachment function, and use the model to build random graphs with this DD. Finally, we study some connections between the attachment functions and the associated probability distributions in Section 7. More precisely, we show that in our model, except for some really unusual cases, the probability distribution is heavy-tailed if and only if the attachment function diverges.

The rest of the paper is organized as follows. We first discuss the related work in Section 3. In Section 4, 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 6 we apply our model on a real-world DD, the undirected DD of Twitter. We finally show the link between the divergence of the attachment function and the heavy-tailed property of the probability distribution in Section 7.

3 Related Work

The degree distribution has been computed for a large number of networks, in particular for social networks such as Facebook [13] or Microsoft Messenger [16]. Note that Myers et al. have also studied DDs for Twitter in [19], using a different dataset than the one we use [11]. Most of those studies considered that their DDs follow power-law distributions, and fitted them to find the best suitable parameters.

However, some works question the fact that the best fit of these DDs is a power-law: for instance, Clauset et al. [9] or Lima-Mendez and van Helden [17] 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 the 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 DDs 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. [9] only considered distributions which have a power-law shape when looking at them in a log-log plot, meaning they appear linear. 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, etc. 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 [10], build networks with a homogeneous DD. The observation that a lot of real-world networks follow power-law DDs led Albert and Barabasi to propose their famous model with linear preferential attachment [2]. It has been followed by numerous 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, 20] 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 [12] a model generating non-growing networks (where, at each time-step, a node is added and another one 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.

4 Presentation of the model

The proposed model is a generalization of the model introduced by Chung 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 of the node degree, 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 randomly chosen among 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 randomly chosen among 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 Chung-Lu model is the particular case for which $f(i) = i$ for all $i \geq 1$. We call *generalized Chung-Lu model* the proposed model where $f(i) = i+b$, for all $i \geq 1$ with $b > -1$.

In the following, we note $N(t)$, $E(t)$, and $N(i,t)$ the random variables corresponding to the total number of nodes, edges, and nodes of degree i in the graph at time t, respectively. $P(i) = \lim_{t\to+\infty} \mathbb{E}[\frac{N(i,t)}{N(t)}]$ is the probability that a random node has degree i in the asymptotic DD.

4.1 Inversion of the recurrence equation

A 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 to 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 is given, we can find an associated attachment function f.

Theorem 1. *Let P be a probability distribution with expectation μ such that the following function h is bounded:*

$$h(i) = \frac{P(k > i+1)}{P(i+1)} - \frac{P(k > i)}{P(i)}.$$

In the proposed model, if p is chosen as $p = \frac{1}{\mu}$ and 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.[§]

Remark 1. The condition on p comes from the fact that, by construction of the model, we have $\mathbb{E}[N(t)] = pt$ and $\mathbb{E}(E(t)) = t$. This leads to a mean-degree of $\frac{1}{p}$.

Before proving Theorem 1, we first establish some results on the concentration of $N(t)$ and of $\sum_{j\geq 1} f(j)N(j,t)$. We start with $N(t)$. We need the following lemma:

Lemma 1 (Chernoff bounds, consult Chapter 4.2 in [18]). *Let $X_1, X_2, \ldots, X_t$ be independent indicator random variables with $\mathbb{P}[X_i = 1] = p_i$ and $\mathbb{P}[X_i = 0] = 1 - p_i$. Let $X = \sum_{i=1}^{t} X_i$ and $\mu = \mathbb{E}[X] = \sum_{i=1}^{t} p_i$. Then, we have*

$$\mathbb{P}[|X - \mu| \geqslant \delta\mu] \leq 2e^{-\mu\delta^2/3}.$$

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

$N(t)$ is a random variable following a binomial distribution $N(t) \sim B(t,p) + n_0$, with n_0 the number of nodes in the initial graph. We can thus use Lemma 1 on $N(t)$. Since $\mathbb{E}[N(t)] = pt$, setting $\delta = \sqrt{\frac{9 \ln t}{pt}}$ we get:

Corollary 1.

$$\mathbb{P}[|N(t) - pt| \geqslant \sqrt{9pt \ln t}] \leq 2/t^3. \tag{2}$$

We also have the following result on P:

Lemma 2. $P(i) \underset{t \to +\infty}{\sim} \frac{\mathbb{E}[N(i,t)]}{pt}$

Proof. For more clarity in this proof let us denote $N(t)$ as N_t and $N(i,t)$ as $N_{i,t}$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space on which random variables $N_{i,t}$ and N_t are defined. Thus $N_{i,t} : \Omega \to \mathbb{R}$ and $N_t : \Omega \to \mathbb{R}$. Let $\Omega_1 \subseteq \Omega$ denote the set of all $\omega \in \Omega$ such that $N_t(\omega) \in (\mathbb{E}[N_t] - \sqrt{9pt \ln t}, \mathbb{E}[N_t] + \sqrt{9pt \ln t})$. By Corollary 1 we know that $\sum_{\omega \in \Omega \setminus \Omega_1} \mathbb{P}[\omega] \leqslant 2/t^3$. Using the fact that $\mathbb{E}[N_t] = pt$ and that, for each ω, we have $\frac{N_{i,t}(\omega)}{N_t(\omega)} \leqslant 1$ and

$$\begin{aligned}
\mathbb{E}\left[\frac{N_{i,t}}{N_t}\right] &= \sum_{\omega \in \Omega} \frac{N_{i,t}(\omega)}{N_t(\omega)} \mathbb{P}[\omega] = \sum_{\omega \in \Omega_1} \frac{N_{i,t}(\omega)}{N_t(\omega)} \mathbb{P}[\omega] + \sum_{\omega \in \Omega \setminus \Omega_1} \frac{N_{i,t}(\omega)}{N_t(\omega)} \mathbb{P}[\omega] \\
&\leqslant \sum_{\omega \in \Omega} \frac{N_{i,t}(\omega)}{\mathbb{E}[N_t] - \sqrt{9pt \ln t}} \mathbb{P}[\omega] + \sum_{\omega \in \Omega \setminus \Omega_1} 1 \cdot \mathbb{P}[\omega] \\
&\leqslant \frac{\mathbb{E}[N_{i,t}]}{\mathbb{E}[N_t] - \sqrt{9pt \ln t}} + 2/t^3 \sim \frac{\mathbb{E}[N_{i,t}]}{pt}.
\end{aligned}$$

On the other hand, since $N_{i,t} \leq t$,

$$\begin{aligned}
\mathbb{E}\left[\frac{N_{i,t}}{N_t}\right] &\geqslant \sum_{\omega \in \Omega_1} \frac{N_{i,t}(\omega)}{N_t(\omega)} \mathbb{P}[\omega] \geqslant \sum_{\omega \in \Omega_1} \frac{N_{i,t}(\omega)}{\mathbb{E}[N_t] + \sqrt{9pt \ln t}} \mathbb{P}[\omega] \\
&= \frac{1}{\mathbb{E}[N_t] + \sqrt{9pt \ln t}} \left(\mathbb{E}[N_{i,t}] - \sum_{\omega \in \Omega \setminus \Omega_1} N_{i,t}(\omega) \mathbb{P}[\omega] \right) \\
&\geqslant \frac{1}{\mathbb{E}[N_t] + \sqrt{9pt \ln t}} \left(\mathbb{E}[N_{i,t}] - \sum_{\omega \in \Omega \setminus \Omega_1} t \cdot \mathbb{P}[\omega] \right) \\
&\geqslant \frac{\mathbb{E}[N_{i,t}]}{\mathbb{E}[N_t] + \sqrt{9pt \ln t}} - \frac{t \cdot 2/t^3}{\mathbb{E}[N_t] + \sqrt{9pt \ln t}} \\
&\sim \frac{\mathbb{E}[N_{i,t}]}{pt}.
\end{aligned}$$

□

We now discuss the concentration of $\sum_{j \geq 1} f(j) N(j,t)$. Let us define

$$Z_t = \sum_{j \geq 1} f(j) N(j,t).$$

Using the following lemma from [14]:

Lemma 3 (Hoeffding's inequality, [14]). *Let $X_1, X_2, \ldots, X_t$ be independent random variables such that $\mathbb{P}[X_k \in [a_k, b_k]] = 1$. Let $X = \sum_{k=1}^{t} X_k$. Then*

$$\mathbb{P}[|X - \mathbb{E}[X]| \geqslant \delta] \leqslant 2\exp\left\{-\frac{2\delta^2}{\sum_{k=1}^{t}(a_k - b_k)^2}\right\}.$$

We can show that:

Lemma 4. *If the following condition is satisfied:*

$$\exists K \text{ such that } \forall i \geqslant 1, |f(i+1) - f(i)| \leqslant K$$

Then:

$$\mathbb{P}[|Z_t - \mathbb{E}[Z_t]| \geqslant \sqrt{32K^2 t \ln t}] = \mathcal{O}\left(\frac{1}{t^4}\right).$$

Proof. First, recall that Z_t can either be expressed as $Z_t = \sum_{j\geq 1} f(j)N(j,t)$ or $Z_t = \sum_{u\in V_t} f(deg_t(u))$, with $deg_t(u)$ being the degree of node u at time t. Now Z_t can also be express as the sum of independent random variables $X_1 + X_2 + ... + X_t$, where X_k is the variation of Z_k during the time step k, i.e., $X_k = Z_k - Z_{k-1}$. In practice, X_k can take the following different values:

- With probability p, a node and an edge are added to the graph, and $X_k = f(deg_k(u)+1) - f(deg_k(u)) + f(1)$, where u is the chosen node at time step k;
- With probability $(1-p)$, an edge is added between two existing nodes, and $X_k = f(deg_k(u)+1) - f(deg_k(u)) + f(deg_k(v)+1) - f(deg_k(v))$, where u and v are the chosen nodes.

Using the condition on f, we bound X_k by $-2K \leqslant X_k \leqslant 2K$.
We can thus apply Lemma 3 with $X = \sum_{k=1}^{t} X_k = Z_t$, $a_i = -2K$ and $b_i = 2K$ to obtain:

$$\mathbb{P}[|Z_t - \mathbb{E}[Z_t]| \geqslant \delta] \leqslant 2\exp\left\{-\frac{2\delta^2}{t(4K)^2}\right\}. \tag{3}$$

Now, setting $\delta = \sqrt{32K^2 t \ln t}$ we get:

$$\mathbb{P}[|Z_t - \mathbb{E}[Z_t]| \geqslant \sqrt{32K^2 t \ln t}] \leqslant 2\exp\left\{-\frac{2 \cdot 32K^2 t \ln t}{t(4K)^2}\right\} = \mathcal{O}\left(\frac{1}{t^4}\right).$$

□

We finally need the following lemma from [8]:

Lemma 5 (Compare Chapter 3.3 in [8]). *Let (a_t), (b_t), (c_t) be three sequences such that $a_{t+1} = (1 - \frac{b_t}{t})a_t + c_t$, $\lim_{t\to+\infty} b_t = b > 0$, and $\lim_{t\to+\infty} c_t = c$. Then, $\lim_{t\to+\infty} \frac{a_t}{t}$ exists and equals $\frac{c}{1+b}$.*

We are now ready to prove Theorem 1.

Proof (Proof of Theorem 1). During the proof, we assume that the following conditions as true:

C1) $\exists K$ such that $\forall i \geqslant 1, |f(i+1) - f(i)| \leqslant K$,
C2) $\sum_{j\geq 1} f(j)P(j) = \mu$, $\mu \in \mathbb{R}_+^*$.

We remind the reader that P is defined as $P(i) = \lim_{t\to+\infty} \mathbb{E}[\frac{N(i,t)}{N(t)}]$. We will verify at the end of the proof that both conditions are indeed satisfied for the chosen f.

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 of degree $i-1$ may gain an edge and, thus, increases by 1 the number of nodes of degree i. This happens with probability $p+2(1-p)$ (the mean number of half-edges connected to existing nodes during a time step) $\times \frac{f(i-1)}{\sum_{j\geq 1} f(j)N(j,t)}$ (the probability for this particular node of degree $i-1$ to be chosen). Since it is the same for all nodes of degree $i-1$, the number of nodes whose degree increases from $i-1$ to i during a time step is $\big(p+2(1-p)\big) \times \frac{f(i-1)}{\sum_{j\geq 1} f(j)N(j,t)} \times N(i-1,t)$. In the same way, a node of degree i may be connected to a new edge, thus, becoming a node of degree $i+1$. The number of nodes of degree i decreases by one in this case. Finally, a node of degree 1 is added with probability p. Gathering those contributions gives the following equation:

$$\begin{aligned} &N(i,t+1) - N(i,t) = \\ &p\delta_{i,1} + (2-p)\frac{f(i-1)}{\sum\limits_{j\geq 1} f(j)N(j,t)}N(i-1,t) - (2-p)\frac{f(i)}{\sum\limits_{j\geq 1} f(j)N(j,t)}N(i,t) \end{aligned} \tag{4}$$

where $\delta_{i,j}$ is the Kronecker delta. The first term of the right hand side is the probability that a new node is added. 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).

We take the expectation on both sides and use Lemma 4 to obtain:

$$\begin{aligned} \mathbb{E}[N(i,t+1)] - \mathbb{E}[N(i,t)] &= p\delta_{i,1} \\ &+ (2-p)\frac{f(i-1)}{\sum\limits_{j\geq 1} f(j)\mathbb{E}[N(j,t)] + \mathcal{O}\left(\sqrt{t\ln t}\right)}\mathbb{E}[N(i-1,t)] \\ &- (2-p)\frac{f(i)}{\sum\limits_{j\geq 1} f(j)\mathbb{E}[N(j,t)] + \mathcal{O}\left(\sqrt{t\ln t}\right)}\mathbb{E}[N(i,t)]. \end{aligned} \tag{5}$$

We introduce the new notation $g(i) = \frac{2-p}{p} \frac{f(i)}{\sum_{j\geq 1} f(j)P(j)}$. We first prove that $g(i) = \frac{1}{P(i)} \sum_{k=i+1}^{\infty} P(k)$. We then show that we can choose $f = g$.

For $i = 1$, Equation 5 becomes:

$$\mathbb{E}[N(1,t+1)] - \mathbb{E}[N(1,t)] = p - (2-p) \frac{f(1)}{\sum_{j\geq 1} f(j)\mathbb{E}[N(j,t)] + \mathcal{O}\left(\sqrt{t \ln t}\right)} \mathbb{E}[N(1,t)]. \tag{6}$$

Taking:

$$a_t = \frac{\mathbb{E}[N(1,t)]}{p},$$

$$b_t = \frac{(2-p)f(1)}{p \sum_{j\geq 1} f(j) \frac{\mathbb{E}[N(j,t)]}{pt} + \mathcal{O}\left(\sqrt{\frac{\ln t}{t}}\right)}, \text{ and}$$

$$c_t = 1,$$

we have $\lim_{t\to+\infty} b_t = g(1) > 0$ and $\lim_{t\to+\infty} c_t = 1$. We can thus apply Lemma 5 (and use Lemma 2 to recognize $P(1)$):

$$\lim_{t\to+\infty} \frac{\mathbb{E}[N(1,t)]}{pt} = P(1) = \frac{1}{1+g(1)}. \tag{7}$$

Now, $\forall i \geq 2$, taking:

$$a_t = \frac{\mathbb{E}[N(i,t)]}{p},$$

$$b_t = \frac{(2-p)f(i)}{p \sum_{j\geq 1} f(j) \frac{\mathbb{E}[N(j,t)]}{pt} + \mathcal{O}\left(\sqrt{\frac{\ln t}{t}}\right)}, \text{ and}$$

$$c_t = \frac{(2-p)f(i-1)}{p \sum_{j\geq 1} f(j) \frac{\mathbb{E}[N(j,t)]}{pt} + \mathcal{O}\left(\sqrt{\frac{\ln t}{t}}\right)} \frac{\mathbb{E}[N(i-1,t)]}{pt},$$

we have $\lim_{t\to+\infty} b_t = g(i) > 0$ and $\lim_{t\to+\infty} c_t = g(i-1)P(i-1)$. Applying Lemma 5 and Lemma 2 give:

$$\lim_{t\to+\infty} \frac{\mathbb{E}[N(i,t)]}{pt} = P(i) = \frac{g(i-1)P(i-1)}{1+g(i)}. \tag{8}$$

Iterating over Equation 8, we express g as a function of P:

$$\begin{aligned} g(i)P(i) &= g(i-1)P(i-1) - P(i) \\ &= g(i-2)P(i-2) - P(i-1) - P(i) \\ &= \cdots \\ &= g(1)P(1) - \sum_{k=2}^{i} P(k) \\ &= 1 - \sum_{k=1}^{i} P(k) \end{aligned}$$

Note that we used Equation 7 to replace $g(1)P(1)$. We thus obtain:

$$g(i) = \frac{1}{P(i)} \sum_{k=i+1}^{\infty} P(k). \tag{9}$$

Now, notice that:

$$\sum_{k=1}^{\infty} g(k)P(k) = \sum_{k=1}^{\infty} \frac{2-p}{p} \frac{f(k)}{\sum_{k'=1}^{\infty} f(k')P(k')} P(k) = \frac{(2-p)}{p}. \tag{10}$$

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 $f = g$.
We finally have to verify that the conditions we put at the beginning of the proof are true. The first condition is equivalent to the one of the theorem for the given f. The second condition is given by Equation 10, concluding the proof. □

5 Application to some distributions

We now apply Equation (1) of Theorem 1 to compute the attachment functions for some classical distributions. We start from the distribution obtained with the generalized Chung-Lu model in Section 5.1 and show that we find a linear dependence with the degree, as expected. We then compute the associated attachment functions of the broken power-law distribution in Section 5.3, of the exact power-law in Section 5.2, and of the geometric law in Section 5.4. Table 1 summarizes those results.

5.1 Preliminary: Generalized Chung-Lu model

As a first example and to present our method, we take a model, the generalized Chung-Lu model, for which we know its DD (power law) and its attachment

Name	P(i)	f(i)	Condition
Generalized Chung-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	$q(1-q)^{i-1}$	$\frac{1-q}{q}$	$p=q$
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. 19& 20	cf. eq. 18

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

function (linear). Applying Theorem 1 should thus lead to a linear attachment function in this case.

In fact, in the generalized Chung-Lu model, we can show that the real DD is not an exact power-law but a fraction of Gamma functions —equivalent to a power-law for high degrees— of the form:

$$\forall i \geq 1, P(i) = C\frac{\Gamma(i+b)}{\Gamma(i+b+\alpha)} \underset{i\gg 1}{\sim} i^{-\alpha}, \tag{11}$$

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 one of b fixes the mean-degree of the graph.

Constraint on p: The condition on p from Theorem 1 gives:

$$\frac{1}{p}=\sum_{k=1}^{\infty}kP(k)=(\alpha-1)\frac{\Gamma(b+\alpha)}{\Gamma(b+1)}\times\frac{\alpha^2+\alpha(2b-1)+b(b-1)}{(\alpha-2)(\alpha-1)}\frac{\Gamma(b+1)}{\Gamma(\alpha+b+1)}$$
$$\implies p=\frac{(\alpha-2)}{\alpha+b-1}. \tag{12}$$

Attachment function f: Using Theorem 1, we get:

$$f(i)=\frac{1}{P(i)}\sum_{k\geq i+1}P(k)=\frac{\Gamma(i+b+\alpha)}{\Gamma(i+b)}\frac{\Gamma(i+b+1)}{(\alpha-1)\Gamma(i+\alpha+b)} \tag{13}$$

$$\implies f(i)=\frac{1}{\alpha-1}i+\frac{b}{\alpha-1}. \tag{14}$$

As expected, we find a linear attachment function. To create a graph with a DD slope α and with a mean-degree p^{-1}, one only has to choose α as the wanted slope and b following Equation (12). In the particular case for which $b=0$, we recover the Chung-Lu model of [7], with a slope of $\alpha=2+\frac{p}{2-p}$ as expected.

5.2 Broken Power-law

We now study the case of a broken power-law, corresponding to the DD of some real world complex networks, as discussed in Section 3, and which was the one we were interested in initially. We consider a distribution of the form:

$$P(i) = \begin{cases} C\frac{\Gamma(i+b_1)}{\Gamma(i+b_1+\alpha_1)} & \text{if } i \leq d \\ C\gamma\frac{\Gamma(i+b_2)}{\Gamma(i+b_2+\alpha_2)} & \text{if } i > d \end{cases} \tag{15}$$

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 is a normalisation constant, and γ is chosen in order to obtain continuity for $i = d$ (see Equation (16)). As seen in Section 5.1, the ratio of such 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 with a switch between the two at the value d.

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

$$\frac{\Gamma(d+b_1)}{\Gamma(d+b_1+\alpha_1)} = \gamma\frac{\Gamma(d+b_2)}{\Gamma(d+b_2+\alpha_2)} \implies \gamma = \frac{\Gamma(d+b_1)\Gamma(d+b_2+\alpha_2)}{\Gamma(d+b_1+\alpha_1)\Gamma(d+b_2)}. \tag{16}$$

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

$$C = \Big(\sum_{k=1}^{\infty} P(k)\Big)^{-1} = \Big(\frac{1}{\alpha_1 - 1}\frac{\Gamma(b_1+1)}{\Gamma(\alpha_1+b_1)} + \frac{\Gamma(b_1+d)}{\Gamma(\alpha_1+b_1+d)}\big(\frac{b_2+d}{\alpha_2-1} - \frac{b_1+d}{\alpha_1-1}\big)\Big)^{-1} \tag{17}$$

Using the condition in Theorem 1, p is defined by the following equation:

$$\begin{aligned} \frac{1}{pC} &= \sum_{k=1}^{d} k\frac{\Gamma(k+b_1)}{\Gamma(k+b_1+\alpha_1)} + \gamma\sum_{k=d+1}^{\infty} k\frac{\Gamma(k+b_2)}{\Gamma(k+b_2+\alpha_2)} \\ &= \frac{\alpha_1^2+\alpha_1(2b_1-1)+b_1(b_1-1)}{(\alpha_1-2)(\alpha_1-1)}\frac{\Gamma(b_1+1)}{\Gamma(\alpha_1+b_1+1)} \\ &- \frac{\alpha_1^2(d+1)+\alpha_1(b_1(d+2)+d^2-1)+b_1(b_1-1)-d(d+1)}{(\alpha_1-2)(\alpha_1-1)}\frac{\Gamma(b_1+d+1)}{\Gamma(\alpha_1+b_1+d+1)} \\ &+ \gamma\frac{\alpha_2^2(d+1)+\alpha_2(b_2(d+2)+d^2-1)+b_2(b_2-1)-d(d+1)}{(\alpha_2-2)(\alpha_2-1)}\frac{\Gamma(b_2+d+1)}{\Gamma(\alpha_2+b_2+d+1)}. \end{aligned} \tag{18}$$

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

Case 1: $i \geq d$

$$f(i) = \frac{1}{P(i)} \sum_{k=i+1}^{\infty} P(k) = \frac{\Gamma(i+b_2+\alpha_2)}{\Gamma(i+b_2)} \sum_{k=i+1}^{\infty} \frac{\Gamma(k+b_2)}{\Gamma(k+b_2+\alpha_2)}$$
$$= \frac{\Gamma(i+b_2+\alpha_2)}{\Gamma(i+b_2)} \frac{1}{\alpha_2 - 1} \frac{\Gamma(i+b_2+1)}{\Gamma(i+b_2+\alpha_2)}$$
$$\implies f(i) = \frac{1}{\alpha_2 - 1} i + \frac{b_2}{\alpha_2 - 1}. \tag{19}$$

We find a linear attachment function. Indeed, for $i > d$, we only take into account the second power-law. Hence, we find the same result than in Section 5.1.

Case 2: $i < d$

$$f(i) = \frac{\Gamma(i+b_1+\alpha_1)}{\Gamma(i+b_1)} \left(\sum_{k=i+1}^{d} \frac{\Gamma(k+b_1)}{\Gamma(k+b_1+\alpha_1)} + \gamma \sum_{k=d+1}^{\infty} \frac{\Gamma(k+b_2)}{\Gamma(k+b_2+\alpha_2)} \right)$$
$$= \frac{\Gamma(i+b_1+\alpha_1)}{\Gamma(i+b_1)} \left(\frac{1}{\alpha_1 - 1} \left(\frac{\Gamma(i+b_1+1)}{\Gamma(i+\alpha_1+b_1)} - \frac{\Gamma(b_1+d+1)}{\Gamma(b_1+\alpha_1+d)} \right) + \frac{\gamma}{\alpha_2 - 1} \frac{\Gamma(b_2+d+1)}{\Gamma(b_2+\alpha_1+d)} \right)$$
$$= \frac{i+b_1}{\alpha_1 - 1} + \frac{\Gamma(i+b_1+\alpha_1)}{\Gamma(i+b_1)} \left(\frac{d+b_2}{\alpha_2 - 1} \frac{\Gamma(b_1+d)}{\Gamma(b_1+\alpha_1+d)} - \frac{1}{\alpha_1 - 1} \frac{\Gamma(b_1+d+1)}{\Gamma(b_1+\alpha_1+d)} \right)$$
$$f(i) = \frac{i+b_1}{\alpha_1 - 1} + \frac{\Gamma(i+b_1+\alpha_1)\Gamma(d+b_1)}{\Gamma(i+b_1)\Gamma(d+b_1+\alpha_1)} \left(\frac{b_2+d}{\alpha_2 - 1} - \frac{b_1+d}{\alpha_1 - 1} \right). \tag{20}$$

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 observe that, while the second part is linear as discussed before, the first part is sub-linear.

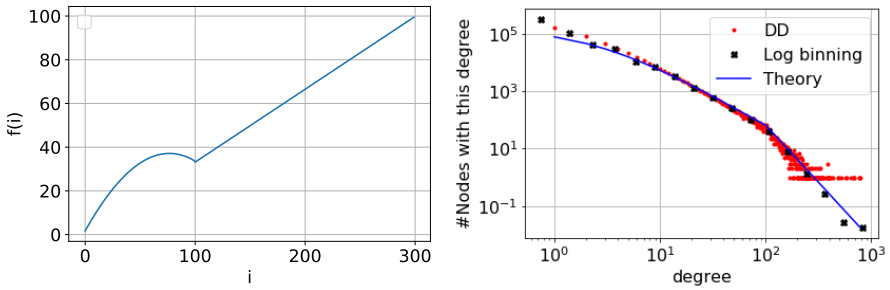
We used this attachment function to build a network using our model. Its DD is shown in Figure 2b. The experiment confirms that, as desired, the random network has a broken power-law distribution.

5.3 Exact power-law degree distribution

The DD obtained with the Chun-Lu model -and most of other classical models- gives a power-law only for high degrees. We may ask ourselves what would be the attachment function associated with an exact power-law degree distribution of the form $P(i) = \frac{i^{-\alpha}}{\zeta(\alpha)}$, where $\zeta(s) = \sum_{k \geq 1} \frac{1}{k^s}$ is the Riemann zeta function.

Constraints on C and p We have the following equation for p:

$$\frac{1}{p} = \frac{1}{\zeta(\alpha)} \sum_{k=1}^{\infty} k^{1-\alpha} = \frac{\zeta(\alpha - 1)}{\zeta(\alpha)}$$



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

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

$$\implies p = \frac{\zeta(\alpha)}{\zeta(\alpha - 1)}. \tag{21}$$

Attachment function Theorem 1 immediately gives:

$$f(i) = \frac{1}{P(i)} \sum_{k=i+1}^{\infty} P(k) = \frac{\zeta(\alpha, i+1)}{i^{-\alpha}}. \tag{22}$$

5.4 Geometric law

We now study the geometric distribution:

$$\forall i \geq 1, P(i) = q(1-q)^{i-1}. \tag{23}$$

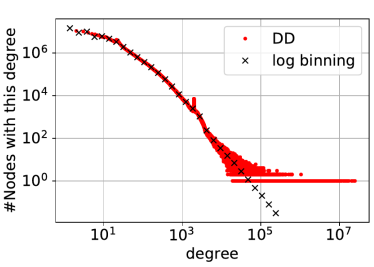
Constraints on p We have:

$$\frac{1}{p} = \sum_{k \geq 1} kq(1-q)^{k-1} = \frac{q}{(1-q)} \frac{(1-q)}{q^2} = \frac{1}{q} \tag{24}$$

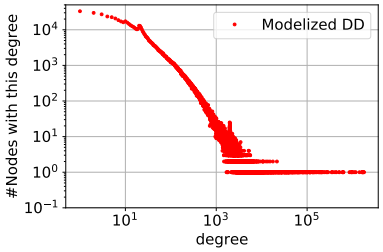
$$\implies p = q. \tag{25}$$

Attachment function The attachment function is easy to compute:

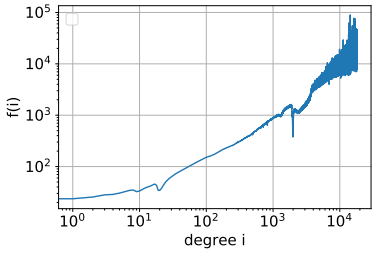
$$f(i) = \frac{1}{q(1-q)^{i-1}} \sum_{k \geq i+1} q(1-q)^{k-1} = \frac{1}{(1-q)^i} \frac{(1-q)^{i+1}}{q} = \frac{1-q}{q}. \tag{26}$$



(a) DD of the Twitter 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 graph.

6 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 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 to the DD of an undirected version of Twitter, shown as having an atypical behavior due to the Twitter policies. We start with a presentation of this distribution, then apply our model to build a random graph with this DD.

6.1 Undirected DD of Twitter

For this study, we use a Twitter snapshot from 2012, recovered by Gabielkov and Legout [11] 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. An arc (u, v) exists if the account u follows the account v. The in- and out-DDs are presented in [28].

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 due to a social phenomena or to the recommendation system). The second spike is explained by a specificity of Twitter: until 2015, to avoid bots 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 was allowed to follow more than 2000 people only if he was also followed by more than 2000 people. This led to a large number of accounts with around 2000 followings. This highlights the fact that some networks have their own characteristics, sometimes due to intern policies, which cannot be modeled but by a model specifically built for them.

6.2 Modeling

Figure 3c presents 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 spikes on the DD for the same degrees. Indeed, to increase the amount of nodes with those degrees, the attachment function for them has to be smaller. So nodes with this degree have less chance of gaining new edges.

We finally use our model with the empirical attachment function of Figure 3c to build random networks with the same DD as the one of Twitter. 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 such 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 larger 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 missing values only occur in the tail of the distribution, which looks like a straight line, and since we interpolate between each pair of closest 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 computation time 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 DD: in particular, we observe the spikes around $d = 20$ and $d = 2000$.

7 Link between the attachment function and heavy-tailed distributions

In this section, we show a relation between the shape of the attachment function f and the tail of the probability function P. More precisely, we show that

(under some conditions on f), if f verifies $\lim_{i\to+\infty} f(i) = +\infty$, then, the associated distribution P is heavy-tailed, and, if f is bounded from above, then, the associated distribution P is not heavy-tailed.

The heavy-tailed feature of DDs is an interesting property of networks: most of the time, real-world networks exhibit heavy-tailed DDs, while pure randomness (as we find in the Erdos-Reyni model) builds networks with homogeneous DDs. The particular case of linear preferential attachment is known to build networks with heavy-tailed DDs.

To the best of our knowledge, our result is the first to make such a general relation between the attachment function of random growing models and the heavy-tailed feature of the DD. Moreover, if the results presented here only apply to the model proposed in Section 4, we believe the proofs can be extended to almost any other random growing models to show similar results.

Note that we now impose an attachment function f and we study the shape of the corresponding DD (instead of imposing a probability distribution and studying the attachment function, as we have made until now.)

7.1 Conditions on f

First of all, f has to verify some conditions in order to give a coherent probability distribution. For instance, choosing $f(i) = i^\alpha$ with $\alpha > 1$ builds a graph in which a dominant vertex emerges such that after n time steps, the degree of this node is of order n, while the degrees of all other vertices are bounded [21]. The DD associated with this attachment function thus is not well-defined. We first express the conditions on f. It can be summed up by:

Condition 1 *In order to obtain a distribution P for the DD verifying $\sum_{k\geq 1} P(k) = 1$ and $\sum_{k\geq 1} kP(k) = \mu$, $\mu \in \mathrm{R}_+^*$, the attachment function f has to verify:*

- *If f converges, $\sum_{i=1}^{+\infty} \frac{(1+\frac{1}{c})^{-i+1}}{f(i)}$ is finite, where $c = \max_{i\geq 1}\Big(f(i)\Big)$;*
- *If f diverges, $\sum_{i=1}^{+\infty} \exp\Big(-\sum_{k=1}^{i} \frac{1}{f(k)}\Big)$ is finite.*

Proof. First, we express the condition $\sum_{k\geq 1} kP(k)$ in an interesting form:

Lemma 6.

$$\sum_{k=1}^{+\infty} f(k)P(k) = \sum_{k=1}^{+\infty} kP(k). \tag{27}$$

Proof. Using Equation 1, we have:

$$\sum_{k=1}^{+\infty} f(k)P(k) = \sum_{k=1}^{+\infty} \sum_{k'=k+1}^{+\infty} P(k') \tag{28}$$

$$= \sum_{k=1}^{+\infty} kP(k). \tag{29}$$

□

We believe this surprising equality between the two sums might lead to some deeper understandings of the links between P and f. We keep this exploration for future works.

We are now left with the study of the convergence of $\sum_{k=1}^{+\infty} P(k)$ and of $\sum_{k=1}^{+\infty} f(k)P(k)$.
Iterating over Equation 8 to express P as a function of f gives:

$$P(i) = P(1) \prod_{k=2}^{i} \frac{f(k-1)}{1+f(k)}. \tag{30}$$

We can rewrite this expression as:

$$P(i) = P(1)\frac{f(1)}{f(i)} \prod_{k=2}^{i} \frac{f(k)}{1+f(k)} \tag{31}$$

$$= P(1)\frac{f(1)}{f(i)} \exp\Big(\ln(\prod_{k=2}^{i} \frac{f(k)}{1+f(k)})\Big) \tag{32}$$

$$= P(1)\frac{f(1)}{f(i)} \exp\Big(-\sum_{k=2}^{i} \ln(1+\frac{1}{f(k)})\Big). \tag{33}$$

From now on we distinguish two cases:

1) $\underline{f \text{ converges}}$:
In this case, $\exists c > 0$ such that $\forall i \geq 1, f(i) \leq c$. We have:

$$P(i) \leq P(1)\frac{f(1)}{f(i)} \exp\Big(-\sum_{k=2}^{i} \ln(1+\frac{1}{c})\Big) \tag{34}$$

$$\leq P(1)f(1)\frac{(1+\frac{1}{c})^{-i+1}}{f(i)}. \tag{35}$$

So, if f converges, $\sum_{k=1}^{+\infty} f(k)P(k)$ always converges, and, by Lemma 6, the mean of P is finite. The condition on $\sum_{k=1}^{+\infty} P(k)$ gives that $\sum_{k\geq 1} \frac{(1+\frac{1}{c})^{-k}}{f(k)}$ has to be finite.

2) $\underline{f \text{ diverges}}$:

Then, we can find i_0 such that $\sum_{k=2}^{i} \ln(1+\frac{1}{f(k)}) \underset{i\to+\infty}{\sim} \sum_{k=2}^{i_0} \ln(1+\frac{1}{f(k)}) + \sum_{k=i_0}^{i} \frac{1}{f(k)}$. We can rewrite Equation (33) as:

$$P(i) \sim P(1)\frac{f(1)}{f(i)} \exp\Big(-\sum_{k=2}^{i_0} \ln(1+\frac{1}{f(k)}) + \sum_{k=1}^{i_0-1} \frac{1}{f(k)} - \sum_{k=1}^{i} \frac{1}{f(k)}\Big) \tag{36}$$

$$\sim K_{f,i_0}\frac{1}{f(i)} \exp\Big(-\sum_{k=1}^{i} \frac{1}{f(k)}\Big), \tag{37}$$

with K_{f,i_0} a constant depending on f and i_0. Thus, by Lemma 6, the mean of P is finite if and only if the following quantity is finite:

$$\sum_{i=1}^{+\infty} \exp\Big(-\sum_{k=1}^{i} \frac{1}{f(k)}\Big).$$

Note that the other condition, i.e., the convergence of $\sum_{i=1}^{+\infty} \frac{1}{f(i)} \exp\Big(-\sum_{k=1}^{i} \frac{1}{f(k)}\Big)$, is included in the first one. Indeed, since f diverges, there exists a constant i_0 such that $\forall i \geq i_0, \frac{1}{f(i)} \leq 1$, and the second condition can be bounded by the first one.

□

It is interesting to note that, for $f(i) \propto i^\alpha$, $\alpha = 1$ is the limit case for which Condition 1 holds, as expected from the results of [21].

7.2 Link between the limit of f and heavy-tailed DDs

Definition 1. *[23] We say that a distribution P is heavy-tailed if it decays more slowly than an exponential, i.e.:*

$$\forall t > 0, e^{ti} P(X > i) \underset{i \to +\infty}{\to} +\infty.$$

We show the two following theorems:

Theorem 2. *Let f be an attachment function verifying Condition 1 and such that $\lim_{i \to +\infty} f(i) = +\infty$. Then, the associated distribution P is heavy-tailed.*

Theorem 3. *Let f be an attachment function verifying Condition 1 and such that f is bounded from above by $M > 0$. Then, the associated distribution P is not heavy-tailed.*

To prove those theorems, we use the following lemma:

Lemma 7. *P is heavy-tailed if and only if*

$$\forall t > 0, \exists i_0 > 0 \text{ such that } \lim_{i \to +\infty} g_{t,i_0}(i) = +\infty,$$

where $g_{t,i_0}(i) = ti + \log(f(i_0)) - \sum_{k=i_0}^{i-1} \log(1 + \frac{1}{f(k+1)})$.

Proof. We recall that $P(i) = P(1) \prod_{k=1}^{i-1} \frac{f(k)}{1+f(k+1)}$ and $f(i) = \frac{1}{P(i)} \sum_{k=i+1}^{\infty} P(k)$. It implies

$$P(X > i) = \sum_{k=i+1}^{\infty} P(k) = f(i)P(i) = f(i)P(1) \prod_{k=1}^{i-1} \frac{f(k)}{1 + f(k+1)}. \tag{38}$$

Let $t > 0$, $i_0 > 0$. We have:

$$e^{ti}P(X > i) = e^{ti}f(i)P(1)\prod_{k=1}^{i-1}\frac{f(k)}{1+f(k+1)} \tag{39}$$

$$= e^{ti}e^{\log(f(i))}P(1)\prod_{k=1}^{i_0-1}\frac{f(k)}{1+f(k+1)}\prod_{k=i_0}^{i-1}e^{\log(\frac{f(k)}{1+f(k+1)})} \tag{40}$$

$$= P(1)\prod_{k=1}^{i_0-1}\left(\frac{f(k)}{1+f(k+1)}\right) \times e^{ti+\log(f(i))+\sum_{k=i_0}^{i-1}\log\left(\frac{f(k)}{1+f(k+1)}\right)}. \tag{41}$$

We call $g_{t,i_0}(i) = ti+\log(f(i))+\sum_{k=i_0}^{i-1}\log(\frac{f(k)}{1+f(k+1)})$. P is heavy-tailed if and only if $\lim_{i\to+\infty} g_{t,i_0}(i) = +\infty$. But g_{t,i_0} can also be expressed as:

$$g_{t,i_0}(i) = ti + \log(f(i)) - \sum_{k=i_0}^{i-1}\log\left(\frac{1+f(k+1)}{f(k)}\right) \tag{42}$$

$$= ti + \log(f(i)) - \sum_{k=i_0}^{i-1}\log(f(k+1)) + \sum_{k=i_0}^{i-1}\log(f(k)) - \sum_{k=i_0}^{i-1}\log\left(1+\frac{1}{f(k+1)}\right) \tag{43}$$

$$= ti + \log(f(i)) - \log(f(i)) + \log(f(i_0)) - \sum_{k=i_0}^{i-1}\log\left(1+\frac{1}{f(k+1)}\right) \tag{44}$$

$$= ti + \log(f(i_0)) - \sum_{k=i_0}^{i-1}\log\left(1+\frac{1}{f(k+1)}\right). \tag{45}$$

□

Proof of Theorem 2.
Let $t > 0$. By definition of the limit, $\exists i_0$ such that $\forall i > i_0, f(i) > \frac{1}{e^{t/2}-1}$. So:

$$g_{t,i_0}(i) = ti + \log(f(i_0)) - \sum_{k=i_0}^{i-1}\log\left(1+\frac{1}{f(k+1)}\right) \tag{46}$$

$$> ti + \log(f(i_0)) - \sum_{k=i_0}^{i-1}\log\left(1+\frac{1}{(\frac{1}{e^{t/2}-1})}\right) \tag{47}$$

$$= ti + \log(f(i_0)) - (i-i_0-1)\frac{t}{2} \tag{48}$$

$$= \frac{1}{2}i + \log(f(i_0)) + (i_0+1)\frac{t}{2} \tag{49}$$

$$\underset{i\to+\infty}{\to} +\infty. \tag{50}$$

□

Proof of Theorem 3.

$$g_{t,i_0}(i) = ti + \log(f(i_0)) - \sum_{k=i_0}^{i-1} \log\left(1 + \frac{1}{f(k+1)}\right) \tag{51}$$

$$< ti + \log(f(i_0)) - \sum_{k=i_0}^{i-1} \log\left(1 + \frac{1}{M}\right) \tag{52}$$

$$= ti + \log(f(i_0)) - (i - i_0 - 1)\log\left(1 + \frac{1}{M}\right). \tag{53}$$

Let $t = \frac{1}{2}\log(1 + \frac{1}{M})$.

$$g_{t,i_0}(i) = -\frac{1}{2}\log\left(1 + \frac{1}{M}\right) i + \log\left(f(i_0)) + (i_0 + 1)\log(1 + \frac{1}{M}\right) \tag{54}$$

$$\underset{i \to +\infty}{\to} -\infty. \tag{55}$$

There exists a value of $t > 0$ such that the limit of g_{t,i_0} goes to $-\infty$, hence P is not heavy-tailed. □

Remark 2. The set of *preferential attachment functions* (i.e., increasing functions) is not included nor it contains any of previous cases. Indeed, we can have a preferential attachment (or a non preferential attachment) function in the first case, as well as in the second case.

Remark 3. Not all functions are included in the previous cases. It remains the cases in which the limit of f is not infinite but f is not bounded either (for instance, $f(i) = 1$ if i is pair, $f(i) = i$ otherwise). However, we believe those cases are hardly encountered in practice.

8 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, called the attachment function. We expressed f as a function of any degree distribution P, leading to the possibility to build a random network with *any* wanted degree distribution. We computed f for some classical distributions, as well as for a snapshot of Twitter with 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. We also believe this model enlightens the relations between the degree distributions of networks and the attachment function behind them, both in random growth models as well as in real-world networks.

To take a step in that direction, we show that, in our model, the limit of the attachment function f is sufficient to determine if the probability distribution of the graphs is heavy-tailed or not. We believe this result can be extended to other models, and hopefully lead to interesting studies on real-world networks.

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. Fan Chung and Linyuan Lu. *Complex Graphs and Networks*. American Mathematical Society, 2006.
9. Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.
10. Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.
11. Maksym Gabielkov and Arnaud Legout. The complete picture of the twitter social graph. In *Proc. on CoNEXT student workshop*, pages 19–20. ACM, 2012.
12. Gourab Ghoshal and MEJ Newman. Growing distributed networks with arbitrary degree distributions. *The European Physical Journal B*, 58(2):175–184, 2007.
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. Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301), 1963.
15. 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.
16. 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.
17. 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.
18. Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge University Press, USA, 2nd edition, 2017.
19. 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.

20. Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. Random graphs with arbitrary degree distributions and their applications. *Physical review E*, 2001.
21. Roberto Oliveira and Joel Spencer. Connectivity transitions in networks with super-linear preferential attachment. *Internet Mathematics*, 2(2):121–163, 2005.
22. Nataša Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007.
23. Tomasz Rolski, Hanspeter Schmidli, Volker Schmidt, and Jozef L Teugels. *Stochastic processes for insurance and finance*, volume 505. John Wiley & Sons, 2009.
24. 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.
25. 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.
26. Andrew T Stephen and Olivier Toubia. Explaining the power-law degree distribution in a social commerce network. *Social Networks*, 31(4):262–270, 2009.
27. 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.
28. Thibaud Trolliet, Nathann Cohen, Frédéric Giroire, Luc Hogie, and Stéphane Pérennes. Interest clustering coefficient: a new metric for directed networks like twitter. *Journal of Complex Networks*, 10(1):cnab030, 2022.