Un modèle de graphes aléatoires croissants pour n'importe quelle distribution des degrés

Thibaud Trolliet, Frédéric Giroire et Stéphane Pérennes

Université Côte d'Azur/CNRS/Inria, France

Les distributions de degrés des réseaux du monde réel sont généralement considérées comme des lois de puissance. Cependant, ce n'est pas le cas pour un grand nombre d'entre eux. Nous proposons donc un nouveau modèle de graphes aléatoires croissants capable de construire des graphes avec presque toute distribution de degrés souhaitée. La distribution des degrés voulue peut être soit théorique, soit extraite d'un réseau du monde réel. L'idée principale est d'inverser l'équation de récurrence couramment utilisée pour calculer la distribution des degrés, afin de trouver une fonction d'attachement adéquate pour le choix des noeuds recevant les nouvelles connexions - généralement choisie comme linéaire. Nous calculons cette fonction d'attachement pour certaines distributions classiques, telles que les distributions de loi de puissance, loi de puissance brisée, et géométrique. Nous utilisons également le modèle sur une version non dirigée du réseau social des suivis de Twitter, pour lequel la distribution des degrés a une forme inhabituelle.

Mots-clefs : Systèmes Complexes, Modélisation, Attachement Préférentiel, Distribution des degrés, Twitter

1 Introduction

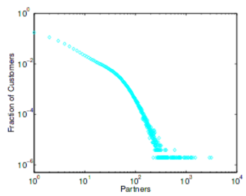
Complex networks appear in the empirical study of real world networks from various domains, such that social, biology, economy, technology, ... Most of those networks exhibit common properties, such as high clustering coefficient, communities, ... Probably the most studied of those properties is the degree distribution (named DD in the rest of the paper), which is often observed as following a power-law distribution.

However, this is common to find real networks with DDs not perfectly following a power-law. Broido and Clauset have already deeply questioned the presence of power-law. In [BC19], they study 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". As a complement, we gathered in Figure 1 DDs from literature which clearly do not follow power-law distributions to show their diversity (see [GPT20] for the references of those figures). 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. 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).

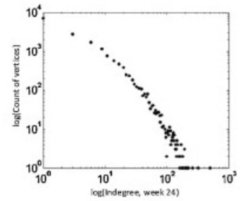
It is yet crucial to build models able to reproduce the properties of real networks. Most of random network models have focused on being able to build graphs exhibiting power-law DDs. Few models permit to build networks with any DD. The most famous one is the configuration model. Goshal and Newman also propose in [GN07] 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.

In this paper [2], 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 $f(i) = i + b$). The theoretical DD of the networks generated by those models is computed using a recurrence equation. The main idea behind this model is to reverse this

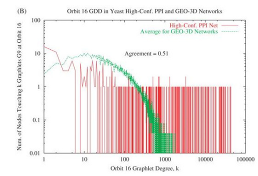
2. The work has been accepted for publication in the proceedings of the conference Complex Networks 2020 [TGP20]. A research report also is available [GPT20].



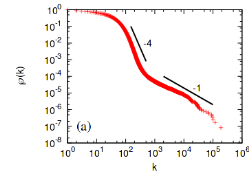
(a) DD of the number of unique callers and callees from a mobile phone operator.



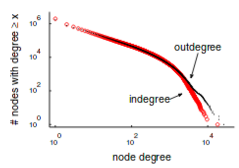
(b) In-DD between shop-to-shop recommendations from an online marketplace.



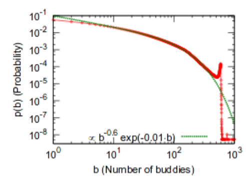
(c) Graphlet DD from a biological model.



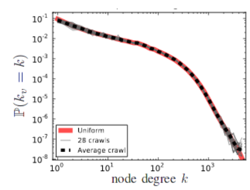
(d) DD of users of Cyworld, the largest online social network of South Korea.



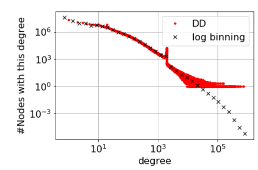
(e) DDs of users of Flickr, an online social network.



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



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



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

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

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 present the model in Section 2. In Section 3.1, we compute the attachment function associated with some classical DD. We also study in Section 3.2 the undirected DD of a Twitter snapshot. 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.

2 Presentation of the model

The proposed model is a generalization of the model introduced by Chung and Lu in [CCG+06]. 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 $\frac{f(deg(v))}{\sum_{w\in V} f(deg(w))}$.

This algorithm has a complexity in $O(t^2)$, since the choice of the random nodes requires looking at the degree of each node. Note that the Chung-lu model is the particular case where $f(i) = i$ 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 $b > -1$.

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. The following

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 [GPT20]	cf [GPT20]

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, $\gamma(a,x)$ the lower incomplete Gamma function.

Theorem is obtained by inverting the recurrence equation of the proposed model such that, if P if given, we can find an associated attachment function f.

Theorem 1. Let P be a probability distribution of mean μ and such that the function $g(i)=\frac{P(k>i+1)}{P(i+1)}-\frac{P(k>i)}{P(i)}$ is bounded. In the proposed model, if p is chosen as $p=1/\mu$ 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.

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.

3 Application to some distributions

3.1 Application to theoretical distributions

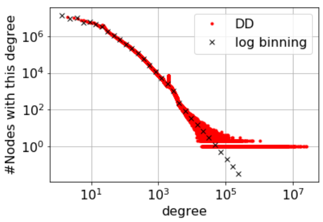
We can apply Equation 1 to compute the attachment function for some classical distributions. Table 1 summarizes the obtained attachment functions for the generalized Chung-Lu, exact power-law, geometric law, and broken power-law distributions.

3.2 Real degree distributions

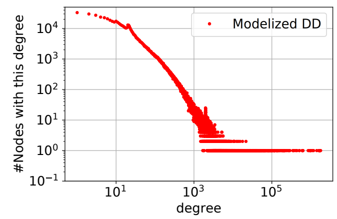
The model can also be applied on empirical DDs. This is a good way to model random networks with an atypical DD. As an example, we apply our model on the DD of an undirected version of Twitter, shown as having atypical behavior due to the Twitter policies.

For this study, we use a Twitter snapshot from 2012, recovered by Gabielkov and Legout [GL12] 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. We look at an undirected version of this graph. Each node corresponds to an account, and an edge exists between two nodes if one of the account follows the other one. The DD of the graph, shown in Figure 2a, exhibits an unusual shape, with spikes at degrees 20 and 2000. Those are due to Twitter's policies to avoid bots. Thus this network cannot be correctly modeled but by a model specifically built for them or by a general model as the one we propose.

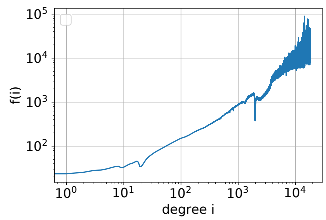
Figure 2c presents the obtained form of the attachment function f computed using Equation 1 with the DD of Twitter. We notice that the overall function is mainly increasing, showing that nodes of higher degrees have a higher chance to connect with new nodes, like in classical preferential attachment 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 2c. 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 that case, f cannot be computed. To get around this difficulty, a possibility is to interpolate the missing values of P - see [GPT20] for more details. Note also that, for the same reason, building graphs with more nodes than the initial one does not have much point.



(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 2c.



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

FIGURE 2: Modelization of the undirected Twitter's graph.

The DD of a random network built with our model is presented in Figure 2b. 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 $d = 2000$.

4 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 can be easily generalized from the presented one.

Références

[BC19] Anna Broido and Aaron Clauset. Scale-free networks are rare. *Nature communications*, 2019.

[CCG+06] Fan Chung, Fan RK Chung, Fan Chung Graham, Linyuan Lu, Kian Fan Chung, et al. *Complex graphs and networks*. American Mathematical Soc., 2006.

[GL12] Maksym Gabielkov and Arnaud Legout. The complete picture of the twitter social graph. In *Proc. on CoNEXT student workshop*, pages 19–20. ACM, 2012.

[GN07] Gourab Ghoshal and MEJ Newman. Growing distributed networks with arbitrary degree distributions. *The European Physical Journal B*, 58(2) :175–184, 2007.

[GPT20] Frédéric Giroire, Stéphane Pérennes, and Thibaud Trolliet. A random growth model with any real or theoretical degree distribution. *arXiv preprint arXiv :2008.03831*, 2020.

[TGP20] Thibaud Trolliet, Frédéric Giroire, and Stéphane Pérennes. A random growth model with any real or theoretical degree distribution. In *COMPLEX NETWORKS*, 2020.