



**HAL**  
open science

## Growth model for collaboration networks

Ghislain Romaric Meleu, Paulin Melatagia Yonta

► **To cite this version:**

Ghislain Romaric Meleu, Paulin Melatagia Yonta. Growth model for collaboration networks. 2016. hal-01304882

**HAL Id: hal-01304882**

**<https://hal.science/hal-01304882>**

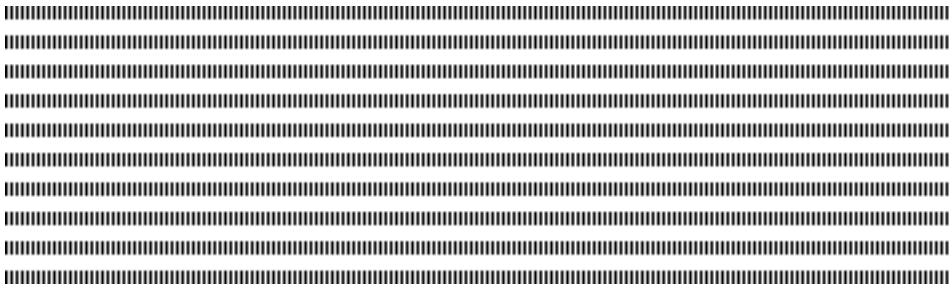
Preprint submitted on 20 Apr 2016

**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.



Distributed under a Creative Commons Attribution 4.0 International License



## Growth model for collaboration networks

Ghislain Romaric MELEU<sup>1,2,3</sup> — Paulin MELATAGIA YONTA<sup>1,2</sup>

<sup>1</sup> UMI 209 UMMISCO, Université de Yaoundé I,  
B.P. 337 Yaoundé, Cameroun

<sup>2</sup> LIRIMA, Equipe IDASCO, Faculté des Sciences,  
Département d'Informatique, Université de Yaoundé I  
B.P. 812 Yaoundé, Cameroun

<sup>3</sup> Département de Mathématiques et Informatique,  
Faculté des Sciences, Université de Ngaoundéré  
B.P. 454 Ngaoundéré-Cameroun.



**ABSTRACT.** We propose a model of growing networks based on cliques formation. A clique is used to illustrate for example co-authorship in co-publication networks, co-cocurrence of words or collaboration between actors of the same movie. Our model is iterative and at each step, a clique of  $\lambda\eta$  existing vertices and  $(1 - \lambda)\eta$  new vertices is created and added in the network;  $\eta$  is the mean of vertices per clique and  $\lambda$  is the proportion of old vertices per clique. The old vertices are selected according preferential attachment. We show that the degree distribution of the generated networks follow Power Law of parameter  $1 + 1/\lambda$ ; those networks are ultra-small world networks whith high clustering coefficient and weak density. Moreover, the networks generated by the proposed model match some real co-publication networks as CARI, EGC and HepTh.

**RÉSUMÉ.** Nous proposons un modèle de croissance de graphe basé sur la formation de clique. Une clique peut par exemple illustrer la collaboration entre auteurs dans un réseau de co-publication, les relations de co-occurrence des mots dans une phrase ou les relations entre acteurs d'un film. C'est un modèle itératif qui à chaque étape crée une clique de  $\lambda\eta$  anciens sommets et  $(1 - \lambda)\eta$  nouveaux sommets et l'insère dans le graphe.  $\eta$  est le nombre moyen de sommets dans une clique et  $\lambda$  la proportion moyenne d'anciens sommets dans une clique. La distribution des degrés des réseaux générés suit la Loi de Puissance de paramètre  $1 + 1/\lambda$  et par conséquent ce sont des réseaux petit-mondes qui présentent un coefficient de clustering élevé et une faible densité. En outre, les réseaux générés par le modèle proposé correspondent à des réseaux de terrains à l'instar des réseaux de co-publication du CARI, de EGC et de HepTh.

**KEYWORDS :** Social Networks Analysis, Collaborative Network, Random graph, Preferential Attachment, Structural property.

**MOTS-CLÉS :** Analyse des réseaux sociaux, Réseau de collaboration, Graphe aléatoire, Attachement préférentiel, Propriété structurelle.



---

## 1. Introduction

In many application contexts, we encounter large graphs with no apparent simple structure called real networks. Examples are Internet topology, web graphs and social networks, biological or linguistic networks. A social network is a set of people or groups of people with some pattern of contacts or interactions between them. It appeared that the classical random graph model used to represent real-world complex networks does not capture their main properties[17]. In particular, real networks have a very low density, an average short distance, degree distribution in Power Law, and high clustering coefficient[17, 22].

Inspired by empirical studies of networked systems such as the Internet, social networks, and biological networks, researchers have in recent years developed a variety of models to help us understand or predict the behavior of these systems[22]. The classical random graph reproduces well the low average distance. However almost all other properties of the random graphs do not match those of real world networks. These random graphs have a low clustering coefficient and a Poisson degree distribution. The model based on preferential attachment [12, 13] reproduces well Power Law distribution. However, the generated network have low clustering coefficient. Some models as Watts and Strogatz model [6] capture the high clustering coefficient, but not the distribution in Power Law. In our knowledge, despite numerous attempts, the definition of models that generate networks with the following main characteristics (short average distance, low density, Power law degree distribution and high clustering coefficient) remains an open problem.

Real networks as co-publication networks have short average distance, low density, Power law distribution and high clustering coefficient. We propose in this paper a new model of growing networks that reproduce graphs with such characteristics. The proposed model is based on small cliques formation. A clique is used to illustrate for example co-authorship in co-publication network, co-occurrence of words or collaboration between actors of the same movie. We show that the degree distribution of the generated networks follow Power Law of parameter  $1 + 1/\lambda$ ; they are ultra-small network with high clustering coefficient and weak density.  $\lambda$  is the proportion of old vertices per clique.

The remainder of the paper is organized as follows: in Section 2 we present a brief state of the art on networks generation models. In section 3 we present the collaboration networks and their generation models. In Section 4 we present a brief analysis of the networks that are used in this paper to validate our model. In Section 5 we present our model and an analysis of its properties. Section 6 provides a validation of the model on real datasets. The article ends with a conclusion.

---

## 2. Networks generation models

Many real world networks exhibit the small world property, i.e. short average distance[6, 10]. This concept has been born from the famous experience made by Milgram[11]. Another property of many real world networks is the presence of high average clustering coefficient i.e. if a vertex  $i$  is connected to vertices  $j$  and  $k$ , there is a high probability of vertices  $j$  and  $k$  of being connected.

A number of models of random graphs have been proposed to explain the dynamic of real word networks. The random graph model developed by Rapoport [26, 27, 28] and independently by Erdos and Rényi [8, 9](ER) can be considered as the most basic model of complex networks. The networks generated by these models have a degree distribution

that follows Poisson law, the small world property and a small average clustering coefficient. The most popular model of random networks (WS) that reproduce short average distance and high clustering coefficient was developed by Watts and Strogatz [6].

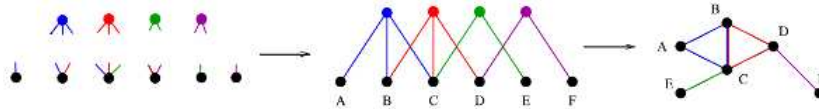
Barabási and Albert [12] (AB) shown that the degree distribution of many real systems is characterized by a degree distribution that follow a Power Law. More specifically, the degree distribution has been found for large  $k$ ,  $P(k) \approx k^{-\lambda}$ . Those networks are called scale-free networks. The Barabási-Albert network model is based on two basic rules: growth and preferential attachment which mean that the probability of a new vertex to be connected to an existing vertex  $j$  is proportional to the degree  $k_j$  of  $j$ .

Price [13] was the first to introduce preferential attachment. Many variants of Barabási model was proposed [24, 14, 23, 15, 16]. Dorogovtsev et al. [24] and Krapivsky and Redner [14, 23] studied the model of preferential attachment in which the probability of attachment to a vertex of degree  $k$  is proportional to  $k + k_0$ . They established that under these conditions, the degree distribution follows a power Law of parameter

$$\lambda = 3 + \frac{k_0}{m}$$

Bianconi and Barabási [15] and Ergun and Rodgers [16] proposed an extension of Barabási and Albert model in which for a new vertex  $i$ , the model assigns a coefficient  $\eta_i$  following a distribution  $\rho(n)$  which represents its attractiveness i.e, its ability to build new relationships. The edge is formed with a vertex with a probability proportional to the product  $\eta_i k_i$ . Depending on the shape of the distribution  $\rho(n)$  the model has two driving schemes model [16]. If the size of the distribution  $\rho$  is finite, then the network shows a distribution of degrees with Power Law, as in the original Barabási-Albert model. However, if the distribution has an infinite size, then the vertex which as highest attraction ability attracts most of the relationship in the graph.

Jean-Loup Guillaume and Matthieu Latapy proposed a bipartite random network model [17] (JM) to generate real world networks. They have shown that all complex networks can be considered as a bipartite graph with specific characteristics [18] and that their main properties can be considered as consequences of the underlying bipartite structure. This model has the merit to reproduce graphs with degree distribution in Power Law, low average distance and high average clustering coefficient. However, its randomness can be considered as a limit since it is not related to the evolution of real word networks.



**Figure 1.** Illustration of Jean-Loup Guillaume and Matthieu Latapy model

Ultimately, there are many models, but almost all fail to reproduce simultaneously the degree distribution in Power Law, high average clustering coefficient, low average distance and low transitivity.

We can consider our model as the uni-partite version of growing model of Jean-Loup Guillaume and Matthieu Latapy. We give several properties of the model which depend on the average number of the vertices and the proportion of the old vertices involved in a collaboration. Specifically, we show that the degree distribution follows a Power Law with parameter  $\gamma = 1 + \frac{1}{\lambda}$  and therefore, the average distance is always logarithmic to the number of vertices.

---

### 3. Collaboration networks

There have been considerable of interest in the study of a special class of social networks, called social collaboration networks[1]. These include movie actor collaboration networks and scientist collaboration networks. This kind of networks can be described using bipartite graphs[1, 25]. One type of nodes can be called 'actor' such as movie actors or scientists and the other can be called 'act' or 'collaboration' such as movies or scientific papers. In these graphs, only undirected edges between different types of vertices are considered. An edge represents an actor taking part in an act or collaboration. If we consider one type of nodes only, two edges sharing a common vertex in the bigraph are projected onto an edge between the two nodes of the same type. Take, for example, a movie actor collaboration network. Sometimes, we need to consider only the collaboration between actors. In this situation, an edge between two actor's shows their collaboration in the same movie. On the other hand, we can define an edge between two movies, which indicates that the same common actor takes part in both movies. If we have to consider how many actors are taking part in movie, we can define a quantity  $T$ , 'act-size', which stands for the number of actors in act; these  $T$  nodes form a complete graph in the down-projected graph consisting of only  $T$  nodes. Each node has a degree value  $T - 1$ . Of course, two complete graphs may share one or more edges in the, down-projected graph. It is easy to verify that such a down-projected network is still a set of complete graphs. We present in the following paragraphs, the model of collaboration networks that are similar to our work.

The model of Pei-Pei Zhang and al.[25] suppose that there are  $m_0$  nodes at  $t = 0$ , which are connected and form some complete graphs representing a number of acts. In each time step a new node is added. It connects to  $T - 1$  old nodes selected according to a specified rule; a complete graph is formed consisting of these  $T - 1$  old nodes and the new node. Considering the rule of selecting  $T - 1$  old nodes ( $T$  is a constant) with a probability proportional to the act-degree  $h_i$  of each old node  $i$ . This is the 'act-degree linear preference rule', which means that, in the case of a network of movie actors, selecting a movie actor according to how many movies he has acted in. The act-degree distribution follows a Power law with the scaling exponent,  $\gamma$  equals

$$\frac{2T - 1}{T - 1} = 2 + \frac{1}{T - 1}$$

$\gamma$  decreases as the act-size,  $T$ , increases. It tends with limit 2. Because the degree  $k_i = h_i(T - 1)$  when considering multiple edges; they obtain the degree distribution (with multiple edges counted) as  $P(k) = k^{-\gamma}$ . Thus the degree distribution  $P(k)$  and the act-degree distribution  $P(h)$  are both exact power functions with the same scaling exponent.

The main difference between this model and our model is that, this model use bipartite graph while our model use simple graph and for new collaboration, author consider only one new vertex while we define a parameter  $\lambda$  that control the proportion of new vertices.

We can also consider that projected graph dynamics is characterized by the arrival of new vertices in the networks (authors or actor) and the addition of clique on the network. Clique is used to illustrate for example co-authorship in co-publication network or collaboration between actors of the same movie. New vertices are working with olds for collaboration. So we can deduce an average proportion of old vertex per collaboration. Our objectives in this work is to offer and deduce the properties of a model of growing collaboration networks based on adding some new cliques using directly simple graphe.

---

## 4. Datasets

The datasets used in this paper consist are co-authorship networks and producers network from Internet Movie Database.

### 4.1. Co-authorship networks

We used:

1) CARI co-citation network [20] (CARI) collected from all the articles of the proceedings of CARI'92 to CARI'10 (except that of CARI'00). This dataset contain 646 articles and 1070 authors.

2) EGC co-citation network (EGC) obtained from all the articles published in conference EGC<sup>1</sup> since 2001. The dataset contain 1921 papers and 2741 authors.

3) High energy physics theory co-citation network[21] (HepTh). It is obtained from the e-print arXiv and covers all the co-citations content on papers meta information obtained from the project site of Stanford Network Analysis Project (SNAP)<sup>2</sup>. The data covers papers in the period from January 1992 to April 2003. This dataset contain 29554 and 11913 authors.

Since we analyse the growing of the collaboration networks, the data of a year of the dataset is added to the data of the previous years in such a way that, a vertex of the current network is a researcher who has published at least one paper at the current or previous years. A link between two vertices means that the associated researchers have co-authored at least one paper at the current year or previous years. If a paper is co-authored by  $k$  authors this generates a clique of  $k$  vertices. The edges are not weighted in the networks.

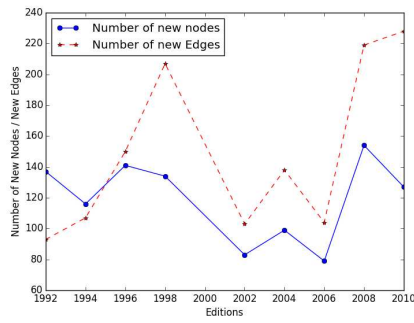
We observed that the papers in the datasets have a mean of 2.38, 2.41 and 1.68 authors per paper, respectively for CARI, EGC and HepTh. In these datasets respectively, there is proportion of 0.3, 0.4 and 0.7 old authors per paper. This implies at each new edition of the CARI and EGC, the publications involve more authors who have not yet published in the conference than author who have already published. At the contrary, HepTh publications involve more authors who have already published in this field.

We study the dynamic behavior of new vertices and its impact on new edges in networks. We note that on the networks of CARI and EGC, new edges and new vertices have the same variation on several points. This leads us to understand that the new edges are mainly generated by the arrival of new vertices that adds both relationships with old and new authors. The low proportion of older authors per paper can help to explain this. In contrast, the variation of new authors and new edges are opposed in the HepTh network; while the number of vertices of the network arriving gradually decrease the number of new edges meanwhile is growing. This implies that new edges are formed mainly between the old vertices and their number is not so much linked to the arrival of new vertices.

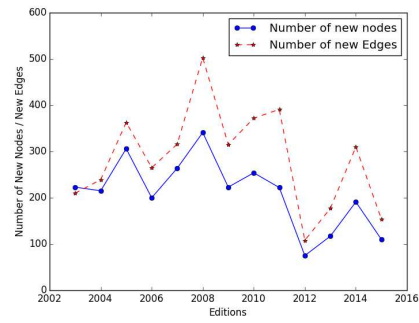
We find that co-authorship networks consist of larger proportion of connected components of size  $< 6$ . Those Components are always complete sub-graphs and are surely obtain by isolated publication. The mains components have respectively 13%, 34% and

---

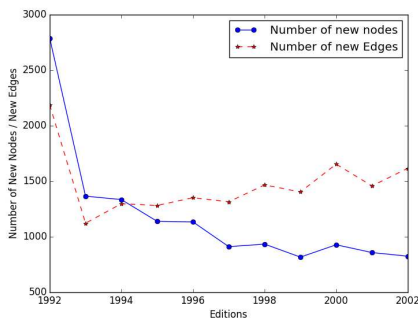
1. [http://editions-rnti.fr/files/EGC\\_articles\\_20150204.txt.zip](http://editions-rnti.fr/files/EGC_articles_20150204.txt.zip)  
2. <http://snap.stanford.edu/data/cit-HepTh-abstracts.tar.gz>



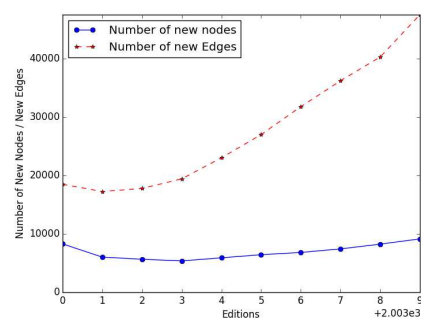
*CARI*



*EGC*



*HepTh*



*Producers*

**Figure 2.** Timegraphs of the numbers of new vertices and new edges.

50.7% of number of vertices on CARI, EGC, and HepTh. Components of size  $> 6$ , are formed by a complete connected core of 1 – 4 vertices with leaves. They are formed following the fusion of singletons with small complete components.

The mains components are formed by the small highly connected components links between them through a small number (1, 2 or 3) of authors. The main component of CARI has a low transitivity, a high clustering coefficient, a high average distance and a degree distribution which follows Poisson law. This structure is different from the one found by Newman [4, 5] and the structure of two others dataset; it is possible that the small size of the studied network partly explains this difference. The main component of EGC and HepTh are small word network. They have a low transitivity, a high clustering coefficient, a low average distance and a degree distribution which follows Power law as shown in the following table.

	n	m	l	CC	T	$d$	$\delta$
CARI	140	269	6.38	0.77	0.52	3.8	$2.10^{-2}$
EGC	957	1842	7.76	0.77	0.25	3.85	$4.10^{-3}$
HepTh	6063	12073	7.50	0.5	0.2	3.98	$6.510^{-4}$
Producers	49340	254118	6.1	0.72	0.3	10.30	$210^{-4}$

**Table 1.** Properties of main component : total number of vertices  $n$ ; total number of edges  $m$  ; mean degree  $d$  ; mean distance  $l$  ;clustering coefficient  $CC$ ; Transitivity  $T$ ; density  $\delta$ .

## 4.2. Producers network from Internet Movie Database(IMDB)

In these social networks, two producers are connected if they have produced a movie together. For this network, we used movies produced between 1990 and 1999. It consists of 181692 movies, 69241 producers and 278446 edges. This graph of IMDB is widely studied for many reasons: it is very large, well representative of social networks, evolving with each new movie produced, and easily available through the Internet Movie Database.

We used the same methodology yield used to build co-authorship networks to build IMDB collaboration network. We observed that, the movie have mean of 3.5 producers and average proportion of 0.71 old producers. This implies that in IMDB, movies involve more producers who have already produced movies.

The network of producers consists of many components. Like for the co-authorship networks, we find that producers networks consist of larger proportion of connected components of size  $\leq 6$ . Those components are always complete sub-graphs and are surely obtained by isolated production of movies. The largest component contains 27% of the vertices in 1990 (this network contains only movies produced in year 1990); it grows rapidly and contains 71% of the vertices in 1999 (this network consist of movies produced between 1990 and 1999). In 1999, the main component of the producers network have 0.72 of average clustering coefficient, 0.3 of transitivity,  $210^{-4}$  of density and 6.1 of average distance. Also, the degree distribution of the main component follow the Power Law. It is a small-world network.

Based on the above observations, we can assume that the dynamics of the structural properties of the studied network are based on three processes that can explain the observed properties: collaboration between old and new vertices, the creation of clique between the vertices and preferential attachment.

- Collaboration between old and new vertices generate the growth of the network and the creation of components.

- The high clustering coefficient can be explained by the explicit process of creation of clique that include the creation of triangles in the graph.

- The degree distribution in Power Law of the datasets supposes that the collaborations between vertices are made according preferential attachment.

We propose to use these elements to produce a generic model of growing collaboration networks. Each collaboration is started by defining its participants. A collaboration contains a variable number of participants, we will assume to have a distribution of numbers of participants per collaboration i.e the distribution of the size of cliques in a network. To define the participants in a collaboration, we will choose between participants already present in the network and new participants. We use a proportion of old vertices by collaboration to create a new vertex or select an old ones. To reproduce the preferential



attachment we suppose that the probability of an old vertex to participate in a collaboration is proportional to its degree.

## 5. The proposed model

### 5.1. Description

We propose a growth model for the collaborative network from random collaborations. It is an iterative model that simulate at each step an collaboration and create relationships in networks. At each step, the model begins by defining the number of vertices, then selects or creates the vertices involving in a collaboration, and finally creates the relationships between these vertices. The selection of old vertices is made according to preferential attachment. The model parameters are listed in Table 5.1 and the algorithm of generation of the random collaborations is given by Algorithm 1.

Designation	Description
$N_a$	Number of collaboration to generate
$P(x = i)$	The distribution of the number of vertices per collaboration
$\lambda$	Proportion of old vertex by collaboration

**Table 2.** Parameters of the model

```

for  $t = 1$  to  $N_a$  do
   $n \leftarrow nb\_vertices(P)$ ;
  for  $i = 1$  to  $n$  do
    Select old vertice with probability  $\lambda$  using preferential attachment or
    create a new vertice with probability  $1 - \lambda$ 
  end
  Create a clique between the  $n$  vertices created and/or selected
end

```

**Algorithm 1:** Collaborative Scale-free model(CSFM)

Using randomness for the selection of an old vertex offers several advantages: it allows to generate collaborations consisting only by vertices present in the network, collaborations that consist of old and new vertices and collaborations that consist only of new vertices. In the latter case it promotes the creation of new components in the network. It allows to manage the existence of many components in the generated network as observed on real word networks.

The properties we study in the following section are the average properties. We assume that at each step, the number of vertices is constant and equals to  $\eta = \sum i P(x = i)$  i.e. the average number of vertices per collaboration.

### 5.2. Properties of the generated networks

Let a collaborative scale-free model (CSFM) with  $\eta$  the average number of vertices by collaboration and  $\lambda$  the proportion of old vertex by collaboration. Let  $t$  the number of iterations done by the CSFM algorithm, for  $t \gg 1$ , we can deduce the following:

**Proposition 1.** *The number of vertices of the networks is:*

$$n_t = t(1 - \lambda)\eta. \quad [1]$$

Proof

At each step we have  $(1 - \lambda)\eta$  new vertices where  $(1 - \lambda)$  is the probability to create a new vertex and  $\eta$  the number of vertex possible. So at time  $t$  we have create :  $n_t = t(1 - \lambda)\eta$ .

**Proposition 2.** *Neglecting the existence of multiple relationships, the number of edges in the network is:*

$$m_t = \frac{t}{2}(\eta - 1)\eta \quad [2]$$

Proof

We approximate the number of edges at each step by the maximum number that can be created. Indeed, selecting  $\lambda\eta$  old actors, it is possible that some of them already have relationships; we neglect this fact. The number of edges created is then:

$$\frac{1}{2}\eta(\eta - 1)$$

because we created a complete graph with all  $\eta$  selected and/or created vertices. So at a given iteration  $t$ , we can consider that the number of edges in the network is :  $m_t = \frac{t}{2}(\eta - 1)\eta$

**Proposition 3.** *The density of the network is :*

$$\delta_t = \frac{(\eta - 1)}{(1 - \lambda)(n_t - 1)} \quad [3]$$

Proof

By definition,  $\delta_t = \frac{2m_t}{n_t(n_t - 1)}$

According to Eq. [1] and Eq.[2] we deduce that :  $\delta_t = \frac{t(\eta - 1)\eta}{t(1 - \lambda)\eta(t(1 - \lambda)\eta - 1)} = \frac{(\eta - 1)}{(1 - \lambda)(n_t - 1)}$

**Lemma 1.** *The average degree of the network is:*

$$\bar{d} = \frac{(\eta - 1)}{(1 - \lambda)} \quad [4]$$

Proof

By definition,  $\delta_t = \frac{\bar{d}}{n_t - 1}$

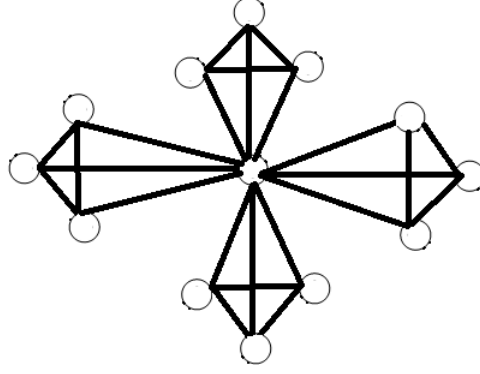
From Eq. [3] we deduce that :  $\bar{d} = \frac{(\eta - 1)}{(1 - \lambda)}$

**Lemma 2.** *The clustering coefficient of a vertex of degree  $k$  is:*

$$C_k \geq \frac{\eta - 2}{k - 1} \quad [5]$$

Proof

An actor may have multiple collaborations with another one or not. In the case it is not only one collaboration per another actor, the structure of the graph that summarizes the collaborations of the vertex form a star (see Fig 3). Each collaboration generates an average increase of the degree of  $\eta - 1$ .



**Figure 3.** Star structure generate by more collaboration

In the other case where an actor collaborate with the same others actors in all its collaborations, the structure of the graph that summarizes collaborations is a complete graph. These two cases are extremes. Let call **A** the first one and **B** the second.

Let  $E_k$ , number of links between the neighbors of a vertex of degree  $k$ . Remember that the clustering coefficient or local density of a vertex is given by:

$$C_k = 2 \frac{E_k}{(k-1)(k-2)} \quad [6]$$

The number of collaboration in case **A** for vertex of degree  $k$  is :

$$n_k = \frac{k}{\eta - 1} \quad [7]$$

The other  $\eta - 1$  vertices of each collaboration form a complete graph. It appears that, the number of edges between the neighbors of considerate vertex is :

$$E_k = \frac{1}{2} k (\eta - 2) \quad [8]$$

In case case **B** we find :

$$E_k = \frac{1}{2} k (k - 1) \quad [9]$$

Hence,

$$\forall k \geq \eta - 1, \frac{\eta - 2}{k - 1} \leq C_k \leq 1 \quad [10]$$

**Theorem 1.** The degree distribution is:

$$p_k \approx \left( \frac{k}{\eta - 1} \right)^{-(1+\frac{1}{\eta})} \approx k^{-(1+\frac{1}{\eta})} \quad [11]$$

Proof

Considering that there is no vertex of degree 0 generated by **CSFM** algorithm. At the step  $t$ , the probability to choose an old vertex of degree  $k$  to participate in the collaboration using preferential attachment, according to [2, 12, 13] is :

$$\frac{k}{\sum x p_x} p_{k,t} \quad [12]$$

where  $p_{k,t}$  is density of vertices of degree  $k$  at step  $t$ .

It follows that, the mean number of vertices of degree  $k$  at step  $t$  that gain an edge when the algorithm creates a new collaboration is :

$$\lambda\eta \frac{k}{\sum x p_x} p_{k,t}$$

Let  $n_t$  the number of vertices at  $t$  step of **CSFM** algorithm;  $n_t p_{k,t}$ , the number of vertices of degree  $k$  at  $t$  step will decrease by  $\lambda\eta \frac{k}{\sum x p_x} p_{k,t}$ . Since this number of vertices will be chosen for the new collaboration, their degree will increase from  $k$  to  $k + \eta - 1$ . In the same time some existing vertices will establish new links and their degree will increase to  $k$  for some of them. These last vertices are those of degree  $k - \eta + 1$  at step  $t$ . i.e  $\lambda\eta \frac{k-\eta+1}{\sum x p_x} p_{k-\eta+1,t}$  vertices.

Let us remember that when we express the number of edges in Eq. [2], we have neglected the existence of an edge between two old vertices at each step. Therefore, every vertex selected and/or created generates  $\eta - 1$  relationships. As a consequence the degree of each vertex is a multiple of  $\eta - 1$ .

When a new collaboration is added in the network, at step  $t$ , since the number of new vertices is  $(1 - \lambda)\eta$ , the variation of the number of vertices of degree  $k$  is then :

$$(n + (1 - \lambda)\eta) p_{k,t+1} - n_t p_{k,t} = \frac{\lambda\eta}{\sum x p_x} [(k - \eta + 1) p_{k-\eta+1,t} - k p_{k,t}] \quad [13]$$

Looking for a stationary state  $p_{k,t+1} = p_{k,t} = p_k$  as

$$(1 - \lambda) p_k = \frac{\lambda}{\sum x p_x} [(k - \eta + 1) p_{k-\eta+1} - k p_k] \quad \forall k > \eta - 1 \quad [14]$$

in this state, the variation of the number of vertices of degree  $\eta - 1$  is :

$$\begin{aligned} (1 - \lambda) p_{\eta-1} &= (1 - \lambda) - \frac{\lambda(\eta-1)}{\sum x p_x} p_{\eta-1} \\ \Leftrightarrow \left[ (1 - \lambda) + \frac{\lambda(\eta-1)}{\sum x p_x} \right] p_{\eta-1} &= (1 - \lambda) \\ \Leftrightarrow \left[ (1 - \lambda) + \frac{\lambda(\eta-1)}{d} \right] p_{\eta-1} &= (1 - \lambda) \\ \Leftrightarrow \left[ (1 - \lambda) + \frac{\lambda(\eta-1)}{\frac{(\eta-1)}{(1-\lambda)}} \right] p_{\eta-1} &= (1 - \lambda) \\ \Leftrightarrow [(1 - \lambda) + \lambda(\eta - 1)] p_{\eta-1} &= (1 - \lambda) \\ \Leftrightarrow p_{\eta-1} &= \frac{1}{1+\lambda} \end{aligned} \quad [15]$$

From Eq. [14] we deduce

$$p_k = \frac{k-\eta+1}{k+\frac{1}{\lambda}(\eta-1)} p_{k-\eta+1} \quad [16]$$

Since the degree of connections of each vertex is a multiple of  $\eta - 1$ , it follows that:

$$\begin{aligned}
p_k &= \frac{\left(\frac{k}{\eta-1}-1\right)\dots 1}{\left(\frac{k}{\eta-1}+\frac{1}{\lambda}\right)\dots\left(2+\frac{1}{\lambda}\right)} \cdot \frac{1}{1+\lambda} \\
&= \frac{\Gamma\left(\frac{k}{\eta-1}\right)\Gamma\left(1+\frac{1}{\lambda}\right)}{\Gamma\left(\frac{k}{\eta-1}+1+\frac{1}{\lambda}\right)} \\
&= B\left(\frac{k}{\eta-1}, 1+\frac{1}{\lambda}\right)
\end{aligned} \tag{17}$$

where  $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$  is Legendre's beta-function, which goes asymptotically as  $a^{-b}$  for large  $a$  and fixed  $b$ , and hence

$$p_k \approx \left(\frac{k}{\eta-1}\right)^{-\left(1+\frac{1}{\lambda}\right)}$$

**Corollary 1.** *Cohen and Havlin [19] shown that scale free networks with parameter  $2 < \gamma < 3$  have a much smaller diameter  $d \approx \ln \ln n$  for network with  $n$  vertices. For  $\gamma = 3$ ,  $d \approx \ln n / \ln \ln n$  while for  $\gamma > 3$ ,  $d \approx \ln n$ . The networks generated by our model have parameter  $\gamma = 1 + \frac{1}{\lambda}$  for the degree distribution, so  $\gamma > 2$ . Then we can deduce that the diameter of the networks are:*

$$d \approx \ln n \tag{18}$$

*In particular, if  $\lambda \geq 1/2$  the proposed algorithm generates ultra-small word networks and*

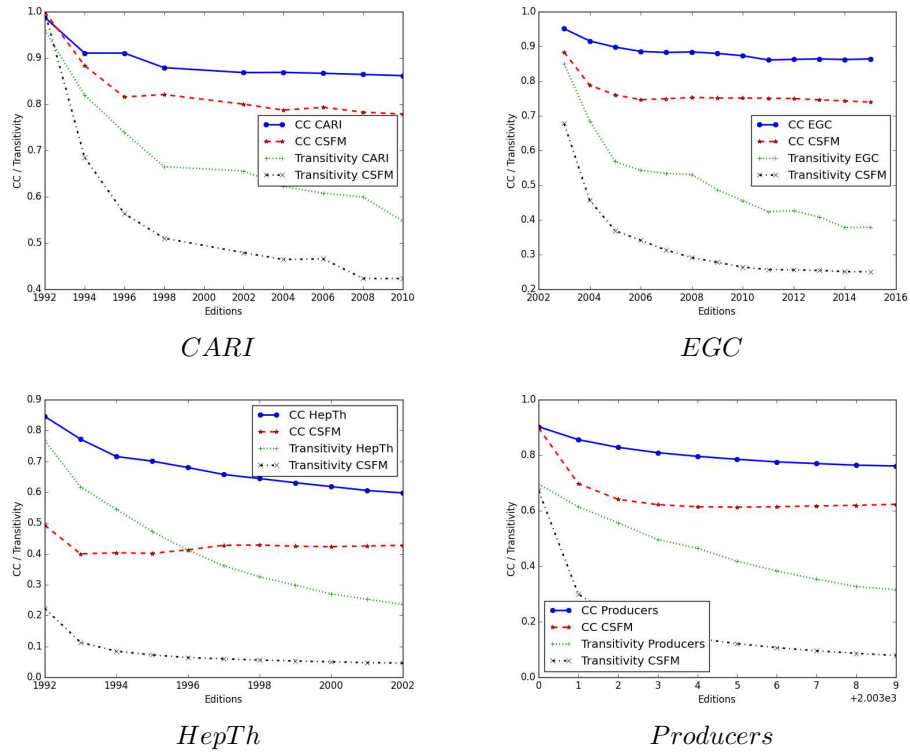
$$d \approx \ln \ln n \tag{19}$$

---

## 6. Simulations

To generate the networks, we have extracted parameters from different datasets. We also extracted the number of collaborations and generate that match at each step (a year) the same number of collaboration.

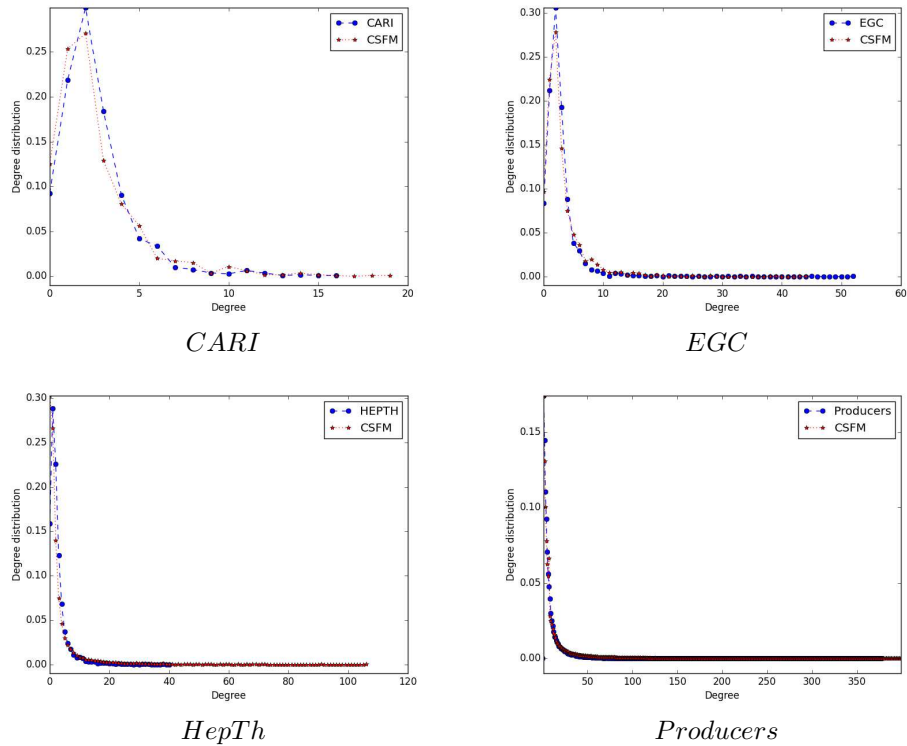
From simulations and in accordance with the theoretical results, we find that the proposed algorithm perfectly reproduces the observed distributions degrees. This is the result of the preferential attachment used for the selection of older vertices in collaborations. This result provides confirmation to the hypothesis that we have formulated on the networks of CARI. Indeed, we assumed that the distribution in Poisson Low was probably due to the small size of the dataset. We can therefore conclude that the creation of the links in the networks is based on preferential attachment. The simulated networks also have very high clustering coefficient. This is due to the creation of complete graphs for each collaboration. Furthermore, the networks have very low average distances, low transivities and similar densities than those observed for the different datasets(see Table 6).



**Figure 4.** Clustering Coefficient and Transitivity of real world network and generated network

	n	m	l	CC	T	d	$\delta$
CARI	1070	1349	5.27	0.86	0.54	2.52	$2 \cdot 10^{-3}$
CSFM CARI	1102	1395	6.71	0.78	0.42	2.53	$2 \cdot 10^{-3}$
EGC	2741	3723	7.7	0.86	0.37	2.71	$10^{-3}$
CSFM EGC	2866	4411	5.65	0.74	0.25	3.07	$2 \cdot 10^{-3}$
HepTh	11913	15509	7.50	0.59	0.23	2.6	$2 \cdot 10^{-4}$
CSFM HepTh	14868	25275	4.41	0.43	0.05	3.4	$2 \cdot 10^{-4}$
Producers	69241	278446	5.7	0.76	0.31	8.04	$1 \cdot 10^{-4}$
CSFM Producers	193684	1098364	5.1	0.62	0.07	11.34	$5,85 \cdot 10^{-4}$

**Table 3.** Comparison between global properties of real networks and generated network : total number of vertices  $n$ ; total number of edges  $m$ ; mean degree  $d$ ; mean distance  $l$ ; clustering coefficient  $CC$ ; Transitivity  $T$ ; density  $\delta$ .



**Figure 5.** The degree distribution of real world network and generated network

## 7. Conclusion

We have presented in this paper a collaborative model of growing graphs. It is an iterative model that simulates at each step a collaboration and creates relationships in networks. The collaboration involves several old and new vertices. The model is set by the distribution of the number of vertices and proportion of old vertices by collaboration.

We conducted a theoretical analysis of the model and the result of simulations were compared with four real datasets. The parameters for the simulations were extracted from those datasets. It appears that the generated networks have the distributions that follow Power Law, low average distance, a high clustering coefficient, low transitivity and low density. Therefore, we can say that the proposed model reproduces random networks with characteristic similar to real-world networks.

However, after analyzing these basic properties, the future prospect of this work may be to study more complex properties. For example one can analyze the structure and dynamics of communities in these graphs related to other models on one hand, and on the other hand to the real-world networks. Indeed, the high clustering coefficient and low transitivity in these graphs suggest that the existence of many communities.

---

## 8. References

- [1] S. WASSERMAN, K. FAUST, “ Social networks analysis : methods and applications”, *Cambridge university press*, vol. 8, 1994.
- [2] A.L. BARABÁSI, H. JEONG, NÉDA, E. RAVASZ, A. SCHUBERT, T. VICSEK, “Evolution of the social network of scientific collaborations”, *Physica A*, vol. 311, num. 3, 2002.
- [3] M. E. J. NEWMAN, “The structure of scientific collaboration networks”, *Proc. Natl. Acad. Sci. USA* vol. 98, num. 2, 2000.
- [4] M. E. J. NEWMAN , “Scientific collaboration networks. I. Network construction and fundamental results”, *Physical Review E* vol. 64, num. 1, 2001.
- [5] M. E. J. NEWMAN, “ Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality”, *Physical Review* , vol. 64, num. 1, 2001.
- [6] D. J. WATTS, S. H. STROGATZ, “ Collective dynamics of small-world networks ”, *Nature*, vol. 393, num. 6684, 1998.
- [7] R. FERRER, R. V. SOLÉ , “ The small world of human language. Proceedings of the Royal Society of London”, *Series B: Biological Sciences*, vol. 268, num. 1482, 2001.
- [8] P. ERDOS, A. RÉNYI, “On random graphs”, *Publicationes Mathematicae*, vol. 6, 1959.
- [9] P. ERDOS, A. RÉNYI, “ On the evolution of random graphs”, *Publ. Math. Inst. Hungar. Acad. Sci.*, vol. 5, 1960.
- [10] D. J. WATTS, “Small worlds : the dynamics of networks between order and randomness”, *Princeton University Press*, 1999.
- [11] S. MILGRAN, “ The small world problem”, *Psychology Today*, vol. 2, num. 1, 1967.
- [12] A. L. BARABÁSI, R. ALBERT , “ Emergence of scaling in a random networks”, *Science* vol. 286, num. 509, 1999.
- [13] D. J. DE S. PRICE, “ A general theory of bibliometric and other cumulative advantage processes”, *J. Amer. Soc. In-form. Sci.*, vol. 27, num. 5, 1976.
- [14] P. L. KRAPIVSKY, G. J. RODGERS, S. REDNER, “Degree distributions of growing networks”, *Physical Review Letters*, vol. 86, num. 23, 2001.
- [15] G. BIANCONI, A. L. BARABÁSI, “Competition and multiscaling in evolving networks”, *Europhysics Letters*, vol. 54, num. 4, 2001.
- [16] G. ERGUN , G. J. RODGERS, “Growing random networks with fitness”, *Physica A: Statistical Mechanics and its Applications*, vol. 303, num. 1, 2002.
- [17] J. L. GUILLAUME, M. LATAPY, “Bipartite Graphs As Models Of Complex Networks”, *Physica A: Statistical Mechanics and its Applications*, vol. 371, num. 2, 2006.
- [18] J. L. GUILLAUME , M. LATAPY , “Bipartite structure of all complex networks”, *Information processing letters*, vol. 90, num. 5, 2004.
- [19] R. COHEN , S. HAVLIN , “Scale-free networks are ultrasmall”, *Physical review letters*, vol. 90, num. 55, 2003.
- [20] R. MELEU , P.Y.MELATAGIA , “ Analyse et modélisation du CARI : Croissance de la communauté des chercheurs du CARI”, *In proceeding of CRI 2013, Yaoundé, Cameroun* vol. 1, 2013.
- [21] J. LESKOVEC, J. KLEINBERG, C. FALOUTSOS, “ Graph Evolution: Densification and Shrinking Diameters”, *ACM Transactions on Knowledge Discovery from Data (ACM TKDD)*, vol. 1, num. 1, 2007.
- [22] M. E. J. NEWMAN, “ The structure and function of complex networks”, *SIAM Review* , vol. 45, num. 2, 2003.



- [23] P. L. KRAPIVSKY , S. REDNER , “ Organization of growing random networks”, *Physical Review E* , vol. 63, num. 6, 2001.
- [24] S. N. DOROGOVTSSEV, A. V. GOLTSEV , J. F. F. MENDES , “ Pseudofractal scale-free web”, *Physical Review E* , vol. 65, num. 6, 2002.
- [25] PEI-PEI ZHANG, KAN CHEN, YUE HE, TAO ZHOU, BEI-BEI SU, YINGDI JIN, HUI CHANG, YUE-PING ZHOU, LI-CHENG SUND, BING-HONG WANG, DA-REN HE , “ Model and empirical study on some collaboration networks”, *Physica A: Statistical Mechanics and its applications* , vol. 360, num. 2, 2006.
- [26] A. RAPOPORT , “ Nets with distance bias”, *Bulletin of Mathematical Biophysics* , vol. 13, 1951.
- [27] A. RAPOPORT , “ Spread of information through a population with sociostructural bias: I. Assumption of transitivity”, *Bulletin of Mathematical Biophysics* , vol. 15, 1953.
- [28] A. RAPOPORT , “ Contribution to the theory of random and biased nets”, *Bulletin of Mathematical Biophysics*, vol. 19, 1957.