Growth model for collaboration networks

Ghislain Romaric Meleu, Paulin Melatagia Yonta

Growth model for collaboration networks

Ghislain Romaric MELEU[1,2,3] — Paulin MELATAGIA YONTA[1,2]

[1] UMI 209 UMMISCO, Université de Yaoundé I,
B.P. 337 Yaoundé, Cameroun

[2] LIRIMA, Equipe IDASCO, Faculté des Sciences,
Département d'Informatique, Université de Yaoundé I
B.P. 812 Yaoundé, Cameroun

[3] 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 the formation of cliques. A clique is used to illustrate for example co-authorship in co-publication networks, co-occurence 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; η is the mean of vertices per clique and λ is the proportion of old vertices per clique. The old vertices are selected according to preferential attachment. We show that the degree distribution of the generated networks follows Power Law of parameter $1+1/\lambda$. The latter are ultra small-world networks with high clustering coefficient and low density. Moreover, the networks generated by the proposed model match with some real co-publication networks such 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. η est le nombre moyen de sommets dans une clique et λ 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 reproduisent la structure 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 préféntiel, 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 [12]. In particular, real networks have a very low density, an average short distance, have a degree distribution that follows the Power Law, high clustering coefficient and high transitivity [12, 17]. The transitivity is the probability that if vertex A is connected to vertex B and vertex B to vertex C, then the vertex A will also be connected to vertex C [16]. An alternative definition of the transitivity is clustering coefficient,wich has been given by Watts and Strogatz [6], who proposed to define a local value of transitivity in each vertex; the clustering coefficient for the whole network is the average of those local transitivity. Leskovec et al.[13] made two stunning empirical observations: they reported that real-world networks became denser over time (super-constant average degree), and their diameters effectively decreased over time!

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 [17]. 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 [2, 7] reproduces the Power Law distribution efficiently. However, the generated network has low clustering coefficient. Some models such as the Watts and Strogatz model [6] capture the high clustering coefficient, but not the distribution in Power Law. Some models, among those described in the state of the art, generate networks with the following main characteristics : average short distance, low density, degree in Power Law, High transitivity and high clustering coefficient. We propose in this paper a new way, both simple and realistic, for reproducing these characteristics.

Real networks such as co-publication networks have short average distances, low densities, Power law distribution and high clustering coefficients. We propose in this paper a new model of growing networks that reproduce graphs with such characteristics. The proposed model is based on the formation of small cliques. A clique is used to illustrate for example co-authorship in co-publication network, co-coccurence 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 $1 + 1/\ \lambda$; the latter are ultra small-world networks with a high clustering coefficient and low density. λ 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 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, 8] . This concept stems from the famous experience made by Milgram [28]. In particular, if the average distance $d \approx \ln \ln n$ we say that, the networks are ultra small-word networks [23]. 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 being connected.

A number of models of random graphs have been proposed to explain the dynamics of real word networks. The random graph model developed by Rapoport [3, 1, 4] and independently by Erdos and Rényi [18, 19] can be considered as the most basic model of complex networks. The networks generated by these models have a degree distribution that follows a Poisson law, the small world property and a small average clustering coefficient. The most popular model of random networks that reproduces short average distance and high clustering coefficient was developed by Watts and Strogatz [6].

Barabási and Albert [2] showed that the degree distribution of many real systems is characterized by a degree distribution that follows 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 [7] was the first to introduce preferential attachment. Many variants of Barabàsi model were proposed [29, 20, 21, 9, 10]. Dorogovtsev et al. [29] and Krapivsky and Redner [20, 21] studied the model of preferential attachement 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 [9] and Ergun and Rodgers [10] proposed an extension of Barabási and Albert's model in which for a new vertex i, the model assigns a coefficient η_i following a distribution $\rho(n)$ which represents its attractiveness i.e, its ability to build new relationships. An 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 scheme [10]. If the size of the distribution ρ 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 [12] to generate real world networks. They have showed that all complex networks can be considered as a bipartite graph with specific characteristics [11] and that their main properties can be considered as consequences of the underlying bipartite structure. This model consists in sampling a random bipartite graph with prescribed (top and bottom) degree distributions as follows (see Figure 1):

1) generate both top and bottom vertices and assign to each vertex a degree drawn from the given distributions,

2) create for each vertex as many connection points as its degree,

3) link top and bottom connection points randomly,

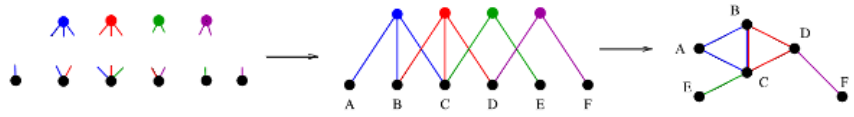


Figure 1. *Illustration of Jean-Loup Guillaume and Matthieu Latapy model*

4) the desired network is obtained by projecting the bi-partite graph on top or bottom.

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.

Copying is another mechanism that may be observed in real-world networks. The basic idea of copying comes from the fact that a new web page is often made by copying an old one. A kind of copying models was proposed in Kumar et al. [25] to explain the emergence of the degree power laws in the web graphs. These models are parameterized by a copy factor $\alpha \in (0, 1)$ and a constant out-degree $d \geq 1$. It is proved in [25] that the copying models possess a power law degree sequence as

$$p_k \approx ck^{-\frac{2-\alpha}{1-\alpha}} \qquad \texttt{where } c \texttt{ is a constant}$$

.

More recently, Silvio Lattanzi et al. [27] presented a model where two graphs evolve at the same time. The first one is a simple bipartite graph that represents the affiliation network and the second one is the social network of actors. The underlying idea behind this model is that in social networks there are two types of entities (actors and societies) that are related by affiliation of the former in the latter. The social network among the actors that results from the bipartite graph is obtained by "folding" the graph, that is, replacing paths of length two in the bipartite graph among actors by an (undirected) edge. The central thesis in developing this social network as a folded affiliation network is that acquaintanceships among people often stem from one or more common or shared affiliations (living on the same street, working at the same place, being fans of the same football club, having coauthored a paper together, etc). This model incorporates elements of preferential attachment and edge copying in fairly natural ways. Silvio Lattanzi et al. show that when an affiliation network B is generated according this model and its folding G on n vertices is produced, the resulting graphs satisfy the following properties:

1) B has a power-law distribution, and G has a heavy-tailed degree distribution as well, and all but $o(n)$ vertices of G have bounded degree;

2) under a mild condition on the ratio of the expected degree of actor nodes and society nodes in B, the graph G has superlinear number of edges;

3) under the same condition, the effective diameter of G stabilizes to a constant.

3. Collaboration networks

There have been considerable interest in the study of a special class of social networks, called social collaboration networks [30]. These include movie actor collaboration networks and scientist collaboration networks. This kind of networks can be described using bipartite graphs [30, 22]. 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 an 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, a model of collaboration networks that are similar to one proposed on this paper.

The model of Zhang et al. [22] supposes 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 γ equals to

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

γ 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 model proposed in this paper is similar to that of Zhang et al. [22]; the major difference is that for a new collaboration, they consider only one new vertex while the model proposed on this paper define a parameter λ that controls the proportion of new vertices. We can thus consider that our model is a generalization of the model of Zhang et al.

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 cliques on the network. A 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 objective in this work is to offer and deduce the properties of a model of growing collaboration networks based on adding some new cliques in a network.

4. Datasets

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

1) CARI co-citation network [26] (CARI) collected from all the articles of the proceedings of CARI'92 to CARI'10 (except that of CARI'00). This dataset contains 646 articles and 1070 authors.

2) EGC co-citation network (EGC) obtained from all the articles published in conference EGC [1] since 2001. The dataset contains 1921 papers and 2741 authors.

3) High energy physics theory co-citation network [13] (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) [2]. The data covers papers in the period from January 1992 to April 2003. This dataset contains 29554 and 11913 authors.

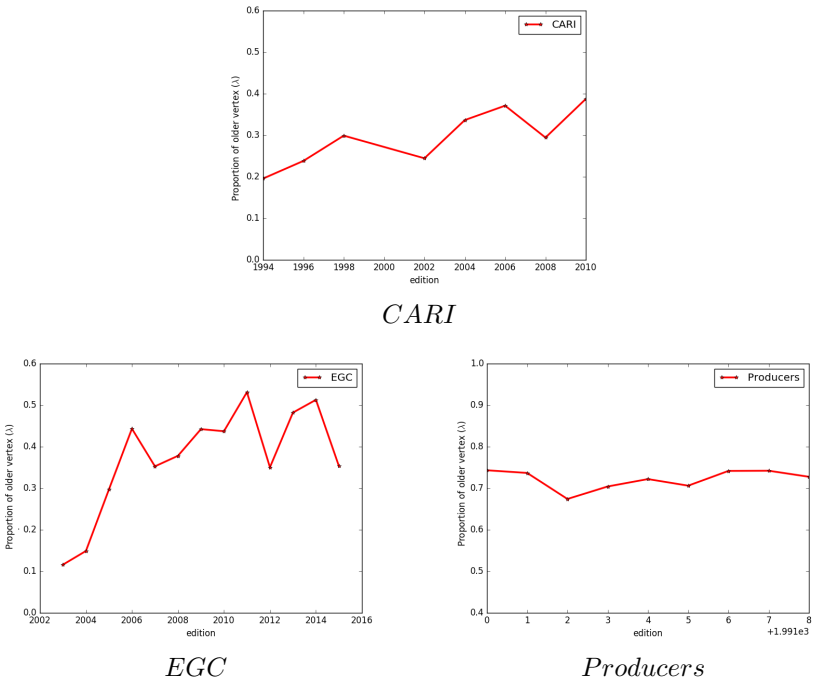
4) Producers network from Internet Movie Database(IMDB) : in these socials networks, two producers are connected if they have produced a movie together. We used movies produced between 1990 and 1999. It consist of 181692 movies, 69241 producers and 278446 edges. Graphs of IMDB are widely studied for many reasons: they are very large, well representative of social networks, evolving with each new movie produced, and easily available through the Internet Movie Database.

Since we analyse the growth 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 (resp. producer) who has published (resp. produced) at least one paper (resp. one movie) during the current or previous years. A link between two vertices means that the associated researchers (resp. associated producers) co-authored (resp. co-produced) at least one paper (resp. one movie) during the current year or previous years. If a paper (resp. movie) is co-authored (co-produced) by k authors (resp. producers) 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 a 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 authors who have already published. On the contrary, HepTh publications involve more authors who have already published in this field. The movie graph has 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 (see Figure 2).

We studied the dynamic behavior of new vertices and its impact on new edges in networks (see the left part of Figure 9). 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 add both relationships with old and new authors. The low proportion of older authors per paper can help to explain this. However, 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 grows. 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.

1. http://editions-rnti.fr/files/EGC_articles_20150204.txt.zip
2. http://snap.stanford.edu/data/cit-HepTh-abstracts.tar.gz



CARI

EGC *Producers*

Figure 2. *Timegraphs of proportion of old authors by collaboration (λ).*

We find that all networks consist of a larger proportion of connected components of size < 6. These components are always complete sub-graphs and are obtained from isolated publications or movies. The components of size > 6 are formed following the fusion of a small complete components (see Figure 3). The main component have respectively 13%, 34% and 50.7% of number of vertices on CARI, EGC, and HepTh. In producers networks, the largest component contains 27% of the vertices in 1990 (this network contains only movies produced in 1990); it grows rapidly and contains 71% of the vertices in 1999 (this network consists of movies produced between 1990 and 1999).

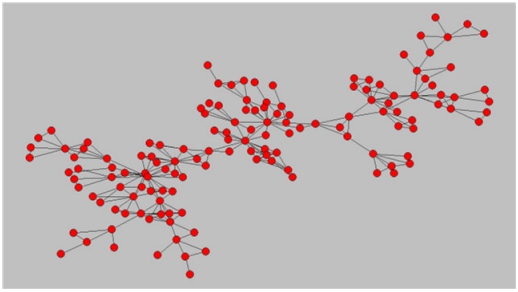


Figure 3. *Example of a component of size* > 6

The main component is 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 high transitivity, a high clustering coefficient, a high average distance and a degree distribution which follows the Power Law. The main component of EGC and HepTh are

small-word networks. They have a high transitivity, a high clustering coefficient, a low average distance and a degree distribution which follows a Power Law as shown in the following table. This structure is the same as those found by Newman [14, 15] for the same type of network. The main component of the producers network has a clustering coefficient of 0.72, transitivity equals 0.3, a density of 0.00021, and an average distance of 6.1. Also, the degree distribution of the main component follows the Power Low. It is a small-word network.

	n	m	l	CC	T	d	δ
CARI	140	269	6.38	0.77	0.52	3.8	0.02
EGC	957	1842	7.76	0.77	0.25	3.85	0.004
HepTh	6063	12073	7.50	0.5	0.2	3.98	0.00065
Producers	49340	254118	6.1	0.72	0.3	10.30	0.0002

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

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.

– The collaboration between old and new vertices generates the growth of the network and the creation of components.

– The high clustering coefficient can be explained by the explicit process of creation of cliques 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 for 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 simulates at each step a collaboration and create relationships in networks. At each step, the model begins by defining the number of vertices, then selects or creates the vertices involved in a collaboration, and finally creates the relationships between these vertices. The selection of old vertices is made according to preferential at-

tachment. 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 collaborations to generate
P	The distribution of the number of vertices per collaboration P_i = probability to have i actors in a collaboration
λ	Proportion of old vertices 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 vertex with probability λ using preferential attachment or create a new vertex with probability $1 - \lambda$
  end
  Create a clique between the n vertices
end

Algorithm 1: Collaboration Model (CM)

Using randomness for the selection of an old vertex offers several advantages: it allows to generate collaborations consisting only of 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 also allows to manage the existence of many components in the generated network as observed on real word networks.

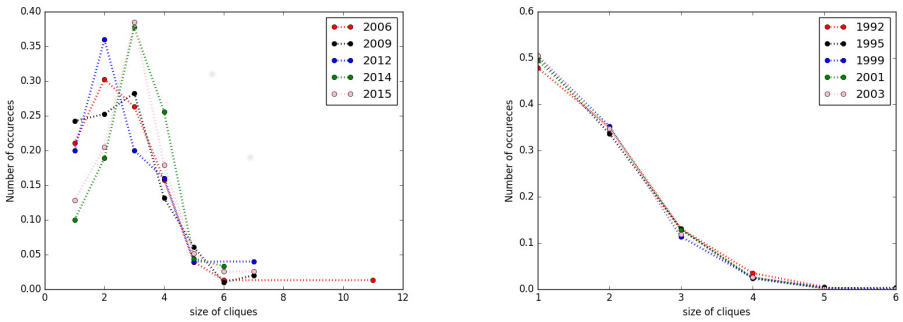
We can consider our model as the dynamic version of the model of Jean-Loup Guillaume et al. [12]. The distribution of the number of actors by collaboration representing the fixed distribution of the top part of the bipartite graph. We deduce the formal properties as the consequence of network dynamics while Jean-Loup Guillaume et al. deduce the properties of the random model as the consequence of fixed distributions on input of the random model and do not address the formalization of the dynamics of the graph.

We give several properties of the genereted networks which depend on the mathematical expectation of distribution P ($\eta = \sum iP(x = i)$) 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.

The properties we will study in the following section are the average properties. We assume that at each step, the number of vertices in a collaboration is equals to $\eta = \sum iP(x = i)$ i.e. the mathematical expectation of distribution P (see Figure 4).

5.2. Properties of the generated networks

For simplification in theorical analyses, we will use the mathematical expectation of distribution P as the number of vertices per collaboration ($\eta = \sum iP(x = i)$) in our demonstrations.



EGC $HepTh$

Figure 4. *Distribution P*

Let **CM** be a collaboration model with $\eta = \sum iP_i$ the mathematical expectation of distribution P and λ the proportion of old vertices by collaboration. Let t be the current number of iterations of the **CM** algorithm , for $t >> 1$, we can deduce the following:

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

$$n_t = t(1-\lambda)\sum iP_i = t(1-\lambda)\eta. \qquad [1]$$

Proof

At each step we have $(1-\lambda)\sum iP_i = (1-\lambda)\eta$ new vertices where $(1-\lambda)$ is the probability to create a new vertex and η the mathematical expectation of distribution P. So at time t we have create : $n_t = t(1-\lambda)\eta$.

Proposition 2. *The number of edges m_t in the network is:*

$$\frac{t}{2}(1-\lambda)\eta[(1-\lambda)\eta - 1] + t\lambda(1-\lambda)\eta^2 \leq m_t \leq \frac{t}{2}(\eta-1)\eta \qquad [2]$$

Proof

Selecting $\lambda\eta$ old actors, it is possible that some of them already have relationships. If we neglect this fact, the maximum number of edges created is then:

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

In other hand, if all $\lambda\eta$ old actors, already have relationships between them, the number of edges created is then

$$\frac{1}{2}(1-\lambda)\eta[(1-\lambda)\eta - 1] + \lambda(1-\lambda)\eta^2$$

Indeed, the number of edges created is obtained by the clique constituted by $(1-\lambda)\eta$ new actors and edges between thoses new actors and old $\lambda\eta$ actors. So at a given iteration t, we can consider that number of edges in the network is framed by:

$$\frac{t}{2}(1-\lambda)\eta[(1-\lambda)\eta - 1] + t\lambda(1-\lambda)\eta^2 \leq m_t \leq \frac{t}{2}(\eta-1)\eta$$

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

$$\frac{(1+\lambda)\eta - 1}{n_t - 1} \leq \delta_t \leq \frac{(\eta - 1)}{(1-\lambda)\,(n_t - 1)} \qquad [3]$$

Proof
By definition :

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

According to Eq. [1] and Eq.[2] we deduce that :

$$\frac{t(1-\lambda)\eta[(1-\lambda)\eta - 1] + 2t\lambda(1-\lambda)\eta^2}{t(1-\lambda)\eta\,(t(1-\lambda)\eta - 1)} \leq \delta_t \leq \frac{t(\eta-1)\eta}{t(1-\lambda)\eta\,(t(1-\lambda)\eta - 1)}$$

So

$$\frac{(1+\lambda)\eta - 1}{n_t - 1} \leq \delta_t \leq \frac{(\eta - 1)}{(1-\lambda)\,(n_t - 1)}$$

Lemma 1. *The average degree $\bar{d}$ of the network is:*

$$(1+\lambda)\eta - 1 \leq \bar{d} \leq \frac{\eta - 1}{1-\lambda} \qquad [4]$$

Proof
By definition :

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

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

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

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

Proof

An actor can have zero or multiple collaborations with other actors. In the case where each actor collaborates only once with another actor, the structure of the graph that summarizes the collaborations of the vertices form a star (see Fig 5). Each collaboration generates an average increase of the degree of $\eta - 1$.

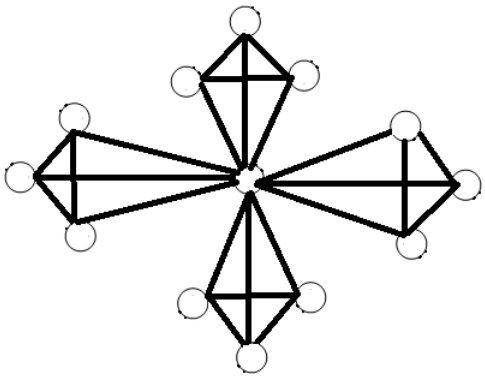


Figure 5. *Star structure generated by collaborations*

In the other case where an actor collaborates with the same actors in all his collaborations, the structure of the graph that summarizes collaborations is a complete graph. These two cases are extremes. Let's call **A** the first case 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 collaborations in case **A** for a 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 the considerated vertex is :

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

In case **B** we find :

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

Hence,

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

Theorem 1. *The degree distribution is:*

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

Proof

Consider that there is no vertex of degree 0 generated by the **CM** 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 [5, 2, 7] is :

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

where $p_{k,t}$ is the 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 :

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

Let n_t be the number of vertices after t step of the **CM** algorithm; $n_t p_{k,t}$, the number of vertices of degree k at step t will decrease by $\lambda\eta\frac{k}{\sum xp_x}p_{k,t}$. Since this number of vertices will be choosen for the new collaboration, their degree will increase from k to $k+\eta-1$. At 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 $\lambda\eta\frac{k-\eta+1}{\sum xp_x}p_{k-\eta+1,t}$ 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 :

$$(n+(1-\lambda)\eta)p_{k,t+1}-n_t p_{k,t}=\frac{\lambda\eta}{\sum xp_x}\left[(k-\eta+1)p_{k-\eta+1,t}-kp_{k,t}\right] \qquad [13]$$

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

$$(1-\lambda)p_k=\frac{\lambda}{\sum xp_x}\left[(k-\eta+1)p_{k-\eta+1}-kp_k\right] \qquad \forall\, k>\eta-1 \qquad [14]$$

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

$$\begin{aligned}
& (1-\lambda)p_{\eta-1} && = && (1-\lambda)-\tfrac{\lambda(\eta-1)}{\sum xp_x}p_{\eta-1} \\
\Leftrightarrow\quad & \left[(1-\lambda)+\tfrac{\lambda(\eta-1)}{\sum xp_x}\right]p_{\eta-1} && = && (1-\lambda) \\
\Leftrightarrow\quad & \left[(1-\lambda)+\tfrac{\lambda(\eta-1)}{\bar{d}}\right]p_{\eta-1} && = && (1-\lambda)
\end{aligned} \qquad [15]$$

By negleting the possible existence of multiple collaboration between actors for simplification purpose, we have:

$$\begin{aligned}
\Leftrightarrow\quad & \left[(1-\lambda)+\tfrac{\lambda(\eta-1)}{\frac{(\eta-1)}{(1-\lambda)}}\right]p_{\eta-1} && = && (1-\lambda)(\textit{using Lemma 1}) \\
\Leftrightarrow\quad & \left[(1-\lambda)+\lambda(1-\lambda)\right]p_{\eta-1} && = && (1-\lambda) \\
\Leftrightarrow\quad & p_{\eta-1}=\tfrac{1}{1+\lambda}
\end{aligned} \qquad [16]$$

From Eq. [14] we deduce

$$p_k \quad = \quad \tfrac{k-\eta+1}{k+\frac{1}{\lambda}(\eta-1)}p_{k-\eta+1} \qquad [17]$$

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

$$\begin{aligned} p_k &= \frac{(\frac{k}{\eta-1}-1)\ldots 1}{(\frac{k}{\eta-1}+\frac{1}{\lambda})\ldots(2+\frac{1}{\lambda})}\cdot\frac{1}{1+\lambda} \\ &= \frac{\Gamma\left(\frac{k}{\eta-1}\right)\Gamma\left(1+\frac{1}{\lambda}\right)}{\Gamma\left(\frac{k}{\eta-1}+1+\frac{1}{\lambda}\right)} \\ &= B\left(\frac{k}{\eta-1}, 1+\frac{1}{\lambda}\right) \end{aligned} \quad [18]$$

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

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

Corrolary 1. *The networks generated by this model are small-word networks.*

Proof

Cohen and Havlin [23] shown that scale free networks with parameter $2 < \gamma < 3$ have a much smaller diameter $d \approx \ln\ln n$ for networks 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 $\gamma > 2$. Thus we can deduce that the diameter of the networks is:

$$d \approx \ln n \quad [19]$$

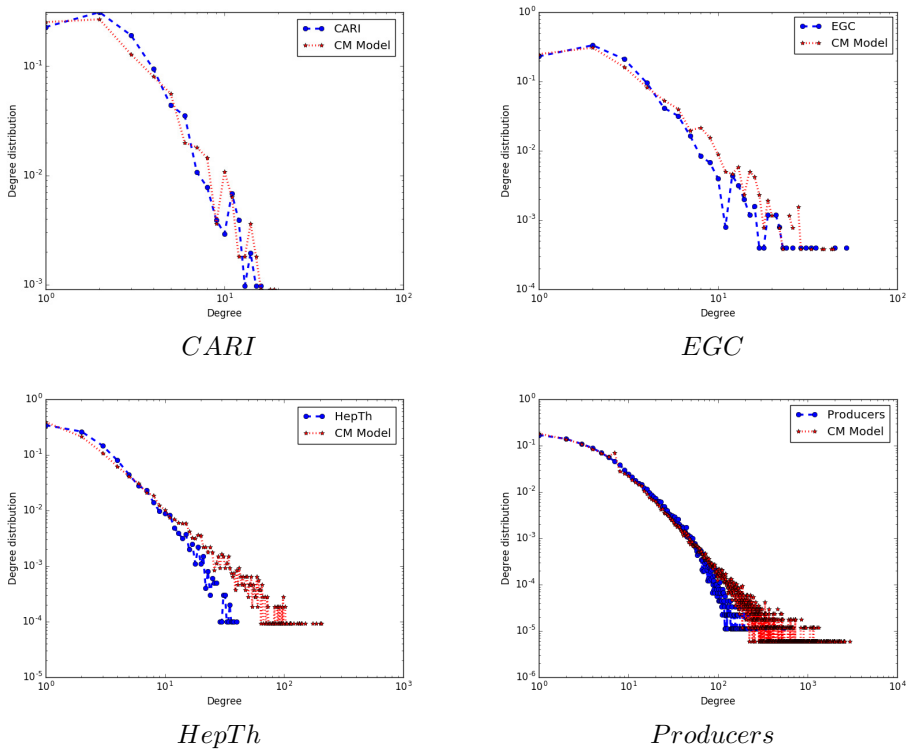
In particular, if $\lambda \geq 1/2$ the proposed algorithm generates ultra-small world networks and

$$d \approx \ln\ln n \quad [20]$$

6. Simulations

To generate the networks, we extracted parameters from different datasets. We also extracted the number of collaborations at each edition and generated the networks accordingly.

From simulations and in accordance with the theorical results, we find that the proposed algorithm reproduces perfectly the observed distributions degrees (see Figure 6). This is the result of the preferential attachment rule used for the selection of older vertices in collaborations.



$CARI$ EGC $HepTh$ $Producers$

Figure 6. *The degree distribution of real world networks and generated networks*

The simulated networks have very high clustering coefficients and high transitivities (see Figuere 7). This is due to the creation of complete graphs for each collaboration. We also compared the distribution of the clustering coefficient and the correlation of degree and clustering coefficient (see Figure 8) of generated networks and real networks. Based on this, we are therefore able to say that our model reproduces distribution of clustering coefficient and the correlation observed in practice between the degree of a node and its clustering coefficient.

The variation of the new vertices and edges in the generated network is the same as that of real networks (see Figure 9). As explained in section 3, it is the consequence of the values of the proportions of old vertices per collaboration. For small values of λ (CARI and EGC) new edges and new vertices vary in the same direction because, the edges are mainly created by the addition of news vertices. For large values of λ (HepTH and Producers), the number of new edges does not depend on the number of news vertices. In fact, the edges are created mainly by the old vertices which are predominantly present in collaborations. The results would certainly be better if the values of λ follows those obtained on the real networks (see Figure 4).

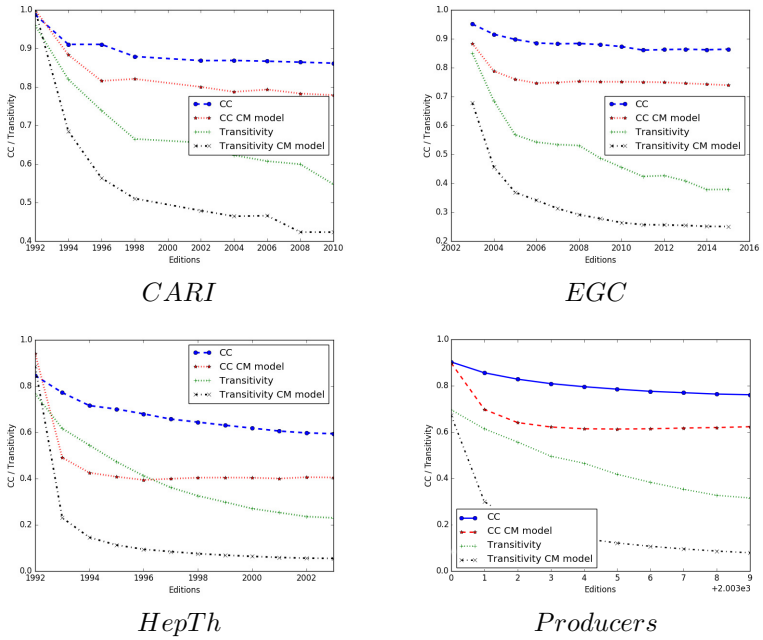


Figure 7. *Clustering Coefficent and Transitivity of real world networks and generated networks*

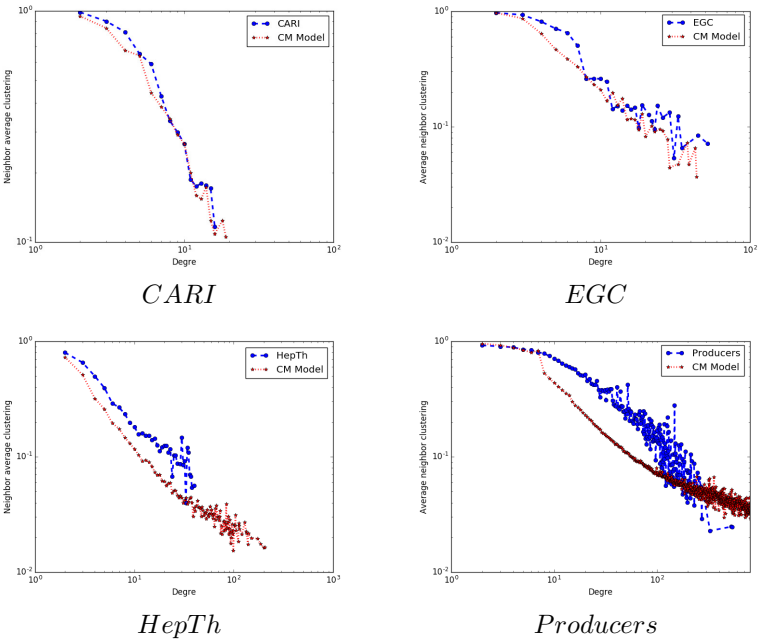
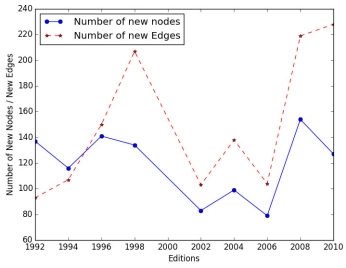
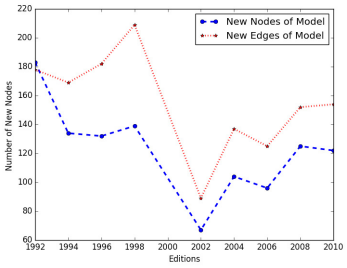


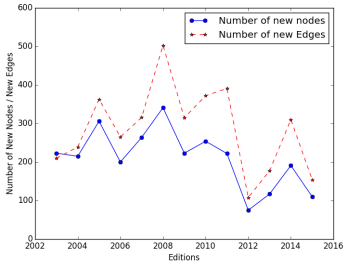
Figure 8. *Corelation Degre / Clustering Coefficent of real world networks and generated networks*



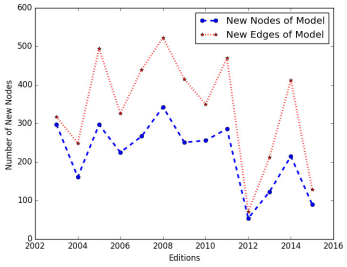
CARI



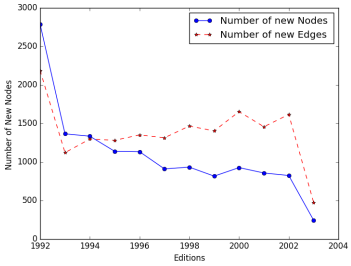
CM model of CARI



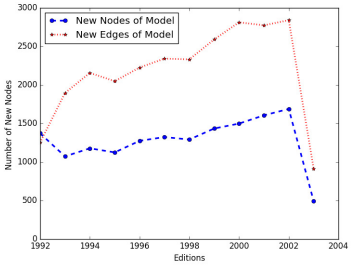
EGC



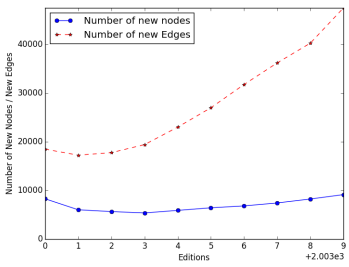
CM model of EGC



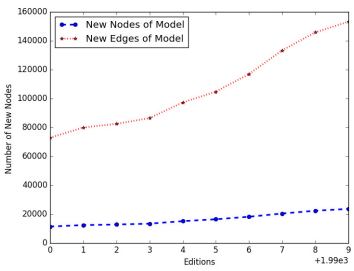
Hepth



CM model of Hepth



Producers



CM model of Producers

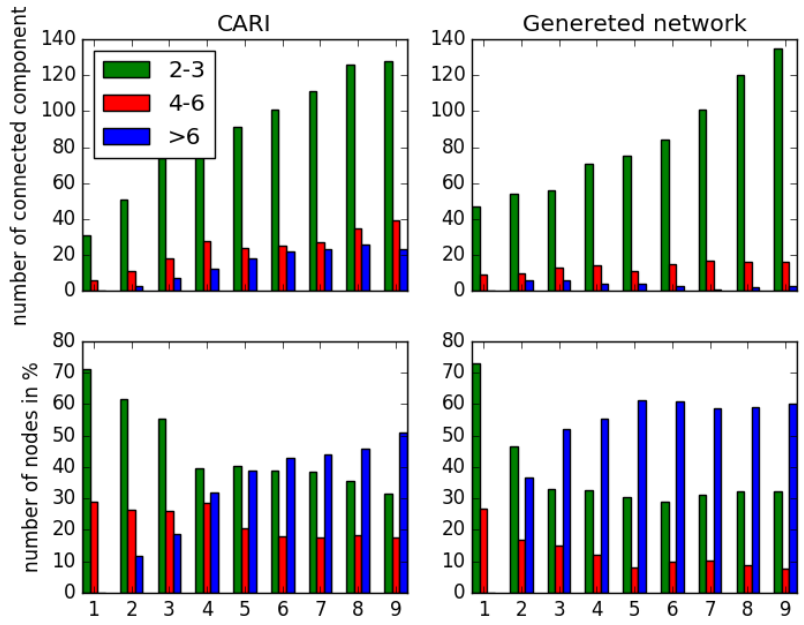
Figure 9. *Timegraphs of the numbers of new vertices and new edges: at the left the timegraphs for reals networks and at right the timegraph for the genereted networks*

The generated networks consist of larger proportions of connected components of size < 6 as real networks (see Figure 10). The main difference resides on the number of large component (size > 6). This difference can be explained by the nature of the analyzed networks. Indeed, in the networks of co-publication (or production) links are created between the vertices belonging to the same discipline or the same research area; this decelerates fusion of components in the networks. The fusion of the components is facilitated by researchers involved in several disciplines or more domains. Contrariwise, this constraint is not implemented in our model. This explains the rapid growth of the giant component, reducing all a large component into a giant component and rapidly increasing the proportion of vertices in this giant component.

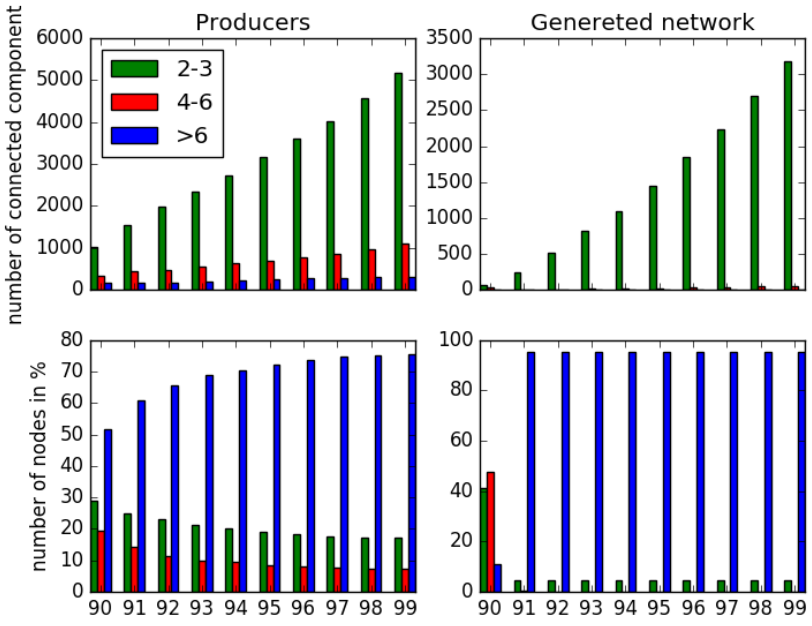
Furthermore, the networks have similar densities than those observed for the different datasets (see Table 3).

	n	m	l	CC	T	d	δ
CARI	1070	1349	5.27	0.86	0.54	2.52	0.002
CM model of CARI	1102	1395	6.71	0.78	0.42	2.53	0.002
EGC	2741	3723	7.7	0.86	0.37	2.71	0.001
CM model of EGC	2866	4411	5.65	0.74	0.25	3.07	$0,002$
HepTh	11913	15509	7.50	0.59	0.23	2.6	0.00021
CM model of HepTh	14868	25275	4.41	0.43	0.05	3.4	0.00022
Producers	69241	278446	5.7	0.76	0.31	8.04	0.00011
CM model of Producers	193684	1098364	5.1	0.62	0.07	11.34	0.000585

Table 3. *Comparaison 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 δ.*



(a) CARI



(b) Producers

Figure 10. *Distribution and proportion of vertices in the connected components.*

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 the proportion of old vertices by collaboration.

We conducted a theorical analysis of the model and the result of simulations were compared with four reals datasets. The parameters for the simulations were extracted from those datasets. It appears that the generated networks have the distributions that follow Power Law, have a low average distance, a high clustering coefficient, high transitivity and low density. Therefore, we can say that the proposed model reproduces random networks with characteristics similar to some 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 high transitivity in these graphs suggest the existence of many communities.

8. References

[1] A. Rapoport. Spread of information through a population with sociostructural bias: I. Assumption of transitivity. Bulletin of Mathematical Biophysics , 15:523-533, 1953.

[2] A. L. Barabási , R. Albert : Emergence of scaling in a random networks, Science 286:509–512, (1999).

[3] A. Rapoport. Nets with distance bias. Bulletin of Mathematical Biophysics, 13:85-91, 1951.

[4] A. Rapoport. Contribution to the theory of random and biased nets. Bulletin of Mathematical Biophysics, 19:257-277, 1957

[5] A. L. Barabási, H. Jeong, Néda, E. Ravasz, A. Schubert, T. Vicsek: Evolution of the social network of scientific collaborations, Physica A 311(3-4), 590-614 (2002).

[6] D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. Nature, 393(6684):440-442, 1998.

[7] D. J. de S. Price : A general theory of bibliometric and other cumulative advantage processes, J. Amer. Soc. In-form. Sci. 27, 292–306,(1976)

[8] D. J. Watts. Small worlds : the dynamics of networks between order and randomness. Princeton University Press, 1999.

[9] G. Bianconi and A. L. Barabási. Competition and multiscaling in evolving networks. EPL (Europhysics Letters), 54(4), 436 (2001).

[10] G. Ergun and G. J. Rodgers. Growing random networks with fitness. Physica A: Statistical Mechanics and its Applications, 303(1), 261-272(2002).

[11] J. L. Guillaume And M. Latapy (2004). Bipartite structure of all complex networks. Information processing letters, 90(5), 215-221.

[12] J. L. Guillaume And M. Latapy. Bipartite Graphs As Models Of Complex Networks. Physica A: Statistical Mechanics and its Applications, 371(2), 795-813(2006)

[13] J. Leskovec, J. Kleinberg and C. Faloutsos. Graph Evolution: Densification and Shrinking Diameters. ACM Transactions on Knowledge Discovery from Data (ACM TKDD), 1(1), 2007.

[14] M. E. J. Newman : Scientific collaboration networks. I. Network construction and fundamental results. Physical Review E 64(1), 016131 (2001).

[15] M. E. J. Newman : Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Physical Review E 64(1) 016132 (2001).

[16] M. E. J. Newman : The structure of scientific collaboration networks, Proc. Natl. Acad. Sci. USA 98, 404–409, (2000).

[17] M. E. J. Newman: The structure and function of complex networks, SIAM Review 45(2), 167-256 (2003).

[18] P. Erdos and A. Rényi. On random graphs. Publicationes Mathematicae, 6:290-297, 1959.

[19] P. Erdos and A. Rényi. On the evolution of random graphs. Publ. Math. Inst. Hungar. Acad. Sci, 5:17-61, 1960.

[20] P. L. Krapivsky, G. J. Rodgers and S. Redner. Degree distributions of growing networks. Physical Review Letters, 86(23), 5401, 2001.

[21] P. L.Krapivsky and S. Redner. Organization of growing random networks. Physical Review E, 63(6), 066123, 2001.

[22] 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(2006). Model and empirical study on some collaboration networks. Physica A: Statistical Mechanics and its applications, 360(2), 599-616.

[23] R. Cohen and S. Havlin : Scale-free networks are ultrasmall. Physical review letters, 90(5), 058701(2003).

[24] R. Ferrer and R. V. Solé (2001). The small world of human language. Proceedings of the Royal Society of London. Series B: Biological Sciences, 268(1482), 2261-2265.

[25] R. Kumar,P. Raghavan,S. Rajagopalan,D. Sivakumar,A. Tomkins and E. Upfal. Stochastic models for the web graph. In Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on (pp. 57-65). IEEE. 2000

[26] R. Meleu and P.Y.Melatagia : Analyse et modélisation du CARI : Croissance de la communauté des chercheurs du CARI. In proceding of CRI 2013, Yaoundé, Cameroun 1,84-88(2013).

[27] S. Lattanzi and D. Sivakumar : Affiliation networks, Proceedings of the forty-first annual ACM ; symposium on Theory of computing, ACM, p. 427-434, 2009.

[28] S. Milgran. The small world problem. Psychology Today, 1(1):60-67, 1967.

[29] S. N. Dorogovtsev, A. V. Goltsev and J. F. F. Mendes. Pseudofractal scale-free web. Physical Review E, 65(6), 066122. 2002

[30] S. Wasserman, K. Faust: Social networks analysis : methods and applications. Cambridge University Press, Cambridge, UK, 1994.