Modèles de génération des graphes de collaborations

Ghislain Romaric Meleu, Paulin Melatagia Yonta

Modèles de génération des graphes de collaborations

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'InformatiqueUniversité 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*

RÉSUMÉ. *Nous proposons deux modèles de croissance de graphe basé sur la formation des cliques. Le premier modèle est itératif et 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. Dans le second modèle, on suppose que les sommets du réseau généré par le modèle précédent sont affiliés à des organisations, chaque organisation pouvant-être affiliée à une organisation de niveau supérieur. Nous déduisons les réseaux hiérarchiques relatifs à ces affiliations. Nous montrons que les réseaux générés par ces deux modèles sont petit-monde et sans-échelles. La distribution des degrés des réseaux du modèle 1 suit une Loi de Puissance de paramètre $1+1/\lambda$ et ces réseaux ont un coefficient de clustering élevé et une faible densité.*

ABSTRACT. *We propose two models of growing network based on the formation of cliques. The first 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 means of vertices in a clique and λ is the average proportion of old vertices per clique. In the second model, we assume that the vertices of the network generated by the previous model are affiliated with organizations; each organization may be affiliated with a higher level organization. We deduce the hierarchical networks for these affiliations. We show that the networks generated by these two models are small-world and scale-free. The degree distribution of networks of model 1 following Power Law with parameter 1 + $1/\lambda$ and thoses network have high clustering coefficient and weak density.*

MOTS-CLÉS : *Réseaux de collaboration, Clique, Affiliation, Attachement préférentiel, Petits-mondes.*

KEYWORDS: *Collaborative networks, Clique, Affiliation, Preferential attachment, Small-worlds.*

2e soumission à *MARAMI-2015*, le 10/09/2015

1. Introduction

Dans de nombreux contextes applicatifs (Newman, 2001 ; i Cancho et Solé, 2001), on rencontre de grands graphes appelés graphes de terrain. La plupart de ces grands graphes ont des propriétés statistiques communes. Notamment, ils ont une densité très faible, une distance moyenne faible, une distribution de degrés en Loi de Puissance, et un fort coefficient de clustering (Newman, 2003 ; Guillaume et Latapy, 2004). Lorsque pour un réseau on a une distance moyenne l tel que $l \approx \log(n)$, on dit que le réseau est petit-monde (Watts et Strogatz, 1998). Le coefficient de clustering est la moyenne sur tous les nœuds, du rapport du nombre de liens entre les voisins d'un nœud avec le nombre maximal possible (Watts et Strogatz, 1998) alors que la transitivité correspond à la proportion de triangles (clique de trois nœuds) parmi toutes les triades connectées (Barrat et Weigt, 2000). Malgré de nombreuses tentatives, générer des graphes ayant toutes ces propriétés reste un problème ouvert. Plusieurs modèles de génération de graphes ont été proposés dans la littérature.

Leskovec et al.(Leskovec *et al.*, 2005) ont observé deux phénomènes surprenants sur les graphes de terrains. Tout d'abord, la plupart de ces graphes se densifient au cours du temps, avec le nombre d'arêtes croissant de façon super-linéaire par rapport au nombre de nœuds. Deuxièmement, la distance moyenne entre les nœuds se rétrécit au fil du temps, contrairement aux attentes qui voudraient que la distance moyenne croisse lentement en fonction du nombre de nœuds (comme $O(\log n)$ ou $O(\log(\log n))$. Cela pose de nouveaux défis pour la modélisation mathématique de réseaux de terrains. Leskovec et al.(Leskovec *et al.*, 2005) proposent deux modèles permettant de reproduire ces caractéristiques.

Les modèles itératifs permettent d'itérer un processus de construction qui produira un graphe ayant les propriétés souhaitées. Le processus de croissance le plus célèbre est sans doute l'attachement préférentiel (Price, 1976 ; Barabási et Albert, 1999). Les nœuds arrivent un à un, ils sélectionnent les anciens sommets avec une probabilité proportionnelle à leurs degré et créent des liens. Le mérite de ce processus est qu'il reproduit la distribution des degrés en Loi de Puissance et la faible densité. Il ne permet malheureusement pas d'obtenir le coefficient de clustering rencontré en pratique (Newman, 2003).

Jean-Loup Guillaume et Matthieu Latapy proposent une approche basée sur les graphes bipartites (Guillaume et Latapy, 2004 ; Guillaume et Latapy, 2006) pour générer les graphes de terrains. Ces modèles ont le mérite de reproduire des graphes avec des distributions de degré en Loi de Puissance, une faible distance moyenne et un coefficient de clustering moyen élevé mais, les réseaux générés tendent à avoir une très forte densité par rapport aux graphes d'origine.

Silvio Lattanzi et al. (Lattanzi et Sivakumar, 2009) proposent un modèle mathématique donc l'idée sous-jacente est que dans les réseaux sociaux, il y a deux types d'entités : les acteurs et les sociétés ; qui sont liés par affiliation du premier au second. Ces relations peuvent être naturellement considérées comme des graphes bipartites appelé réseaux d'affiliation. Le réseau social entre les acteurs résultant du graphe bi-

parti est obtenu par projection à savoir, remplacer chaque nœud de la société dans le réseau d'affiliation par une clique dans le graphe projeté. En ce sens, ce modèle est un modèle de génération de graphe à partir des graphes bipartites.

En parcourant les entêtes des articles scientifiques, nous avons construit et analysé trois réseaux à savoir : les réseaux des auteurs, des laboratoires et des institutions. Nous entendons par entête de l'article, la description du titre de l'article, des noms des auteurs et de leurs affiliations. Le terme institution est utilisé pour se référer à une université, un centre de recherche ou une institution de recherche. Les réseaux sont corrélés à cause des relations d'affiliations qui existent entre les acteurs des trois réseaux. En effet, un auteur est affilié à au moins un laboratoire et un laboratoire est affilié à une institution. Nous disons alors que ces réseaux sont hiérarchiques et sont déduits du réseau de collaborations des auteurs.

Nous généralisons ce système à une structure hiérarchique où un acteur est affilié à une organisation qui peut elle aussi être affiliée à une organisation de niveau supérieur et ainsi de suite. Les relations entre entités à un même niveau sont déduites de celles existantes entre entités de niveau inférieur. Contrairement aux réseaux d'affiliation présentés dans (Lattanzi et Sivakumar, 2009), l'affiliation n'est pas l'élément qui crée la relation entre les acteurs ou les organisations ; mais les interactions des acteurs. Par exemple, deux auteurs ne sont pas en relation parce qu'ils sont membre du même laboratoire, mais parce qu'ils ont collaboré pour la publication d'un article. Les réseaux des laboratoires et des institutions sont déduit de cette collaboration. Expliquer ces systèmes suppose qu'on puisse maitriser à la fois comment les acteurs interagissent, mais aussi les règles d'affiliations aux organisations afin de mieux expliquer les réseaux des organisations qui sont déduits des interactions de ces acteurs. A notre connaissance, il n'existe pas encore de modèle permettant de reproduire des réseaux ainsi présenté. En effet, les modèles de génération de graphes actuels reproduisent un seul réseau à la fois. Il faudrait donc produire chaque réseau indépendamment. Procéder de la sorte engendrerait une perte d'information en termes de corrélation entre les réseaux. En effet dans la définition des réseaux déduits des co-publications, on peut remarquer que les acteurs interagissant dans les collaborations sont les auteurs. Les réseaux des laboratoires et des institutions se déduisent de ces interactions par des relations d'affiliations. Nous imaginons donc qu'il doit exister des relations entre les propriétés observés dans ces graphes(laboratoires et institutions) et celui des auteurs.

Dans cet article nous proposons deux modèles pour capturer l'évolution de ce type de réseaux. Le premier modèle est un modèle itératif qui permet d'expliquer la formation des réseaux de co-publications entre auteurs. Il peut-être utilisé pour reproduire tout réseau de collaboration. Dans le second modèle, on suppose que les sommets du réseau généré par le modèle précédant sont affiliés à des organisations, chaque organisation pouvant-être affiliée à une organisation de niveau supérieur. Nous déduisons les réseaux hiérarchiques induits de ces affiliations.

2. Analyse de quelques réseaux de collaborations

Nous présentons l'analyse de trois réseaux de collaboration corrélés issue du CARI[1]. Il s'agit :

– du réseau des auteurs (CA_AUT) : un sommet est un auteur qui a publié au moins un article et une arête existe entre deux auteurs s'ils ont collaboré ensemble à la publication d'au moins un article.

– du réseau des laboratoires de recherche (CA_LAB) : un sommet est un laboratoire dont au moins un auteur a publié un article et une arête relie deux laboratoires s'il existe au moins un article co-publié par des auteurs des deux laboratoires.

– du réseau des institutions (CA_INST) : un sommet est une institution qui possèdent des auteurs ayant publié au moins un article et une arête relie deux institutions s'il existe au moins un article co-publié par des auteurs de deux laboratoires rattachés chacun à une des institutions.

Si un article contient n_s auteurs (respectivement laboratoires ou institutions), on crée une clique de taille n_s dans le réseau correspondant. L'ensemble de données utilisées est constitué de 1070 auteurs, 590 laboratoires, 293 institutions de recherche pour 646 articles.

Nous analysons aussi le graphe de co-publication dans arXiv section physique des hautes énergies (High energy physics theory ou HepTh). Pour ce jeu de données nous avons construit uniquement le graphe des auteurs, les métadonnées ne contenant pas la description des affiliations de ces auteurs. Les données ont été obtenu du site du projet Stanford Network Analysis[2]. Les données couvrent les articles publiés dans la période de Janvier 1992 à Avril 2003. Cette source de données contient 29554 articles et nous avons extrait 11913 auteurs.

Pour les deux jeux de données utilisés, nous avons une proportion moyenne de 0.3 et 0.7 d'anciens auteurs par article respectivement pour CARI et HepTh. Nous avons remarqué que, pour le réseau du CARI, le nombre de nouvelles arêtes et le nombre de nouveaux sommets suivent la même dynamique. Ceci nous laisse comprendre que les nouvelles arêtes sont principalement engendrées par l'arrivée des nouveaux sommets qui créent des relations d'une part avec les anciens sommets, mais aussi avec d'autres nouveaux sommets impliqués dans les mêmes articles. Par contre, la variation du nombre des nouveaux auteurs et du nombre des nouvelles arêtes ont des dynamiques contraires dans le réseau Hepth. Ceci sous-entend que les nouvelles arêtes sont formées majoritairement entre les anciens sommets et leur nombre n'est donc pas fortement lié à l'arrivée de nouveaux sommets. Les proportions moyennes d'anciens auteurs par article, l'un faible (CARI) et l'autre élevé peuvent expliquer ce comportement. Cette analyse nous a permis d'émettre l'hypothèse que la proportion moyenne d'anciens auteurs par article joue un rôle important pour les propriétés des réseaux.

1. www.cari-info.org
2. http ://snap.stanford.edu/data/cit-HepTh-abstracts.tar.gz

	CA_AUT	CA_LAB	CA_INST	HepTh
Nombre de sommets	1070	592	293	11913
Nombre d'arêtes	1349	514	375	15509
Degré moyen	2.52	1.73	2.56	2.6
Distance moyenne	5.27	5.47	3.86	7.47
Densité	0.0022	0.003	0.008	0.0002
Coefficient de Clustering	0.86	0.67	0.56	0.59
Transitivité	0.54	0.32	0.16	0.23

Tableau 1 – Propriétés structurelles des différents réseaux

Nous avons observé dans les deux réseaux des auteurs que les composantes de taille 3 sont généralement des graphes complets, ceux de taille 4 et 5 et la majorité des composantes de taille 6 sont formés par un noyau complet de 2 à 4 nœuds avec des feuilles. Elles résultent de la fusion de singletons avec des petits groupes complets. Le graphe de HepTh lui est dominé par une composante géante qui présente une distribution des degrés en Loi de Puissance et est petit monde. Il est plus proche des graphes des auteurs analysé dans la littérature (Newman, 2001 ; Barabási *et al.*, 2002). Le tableau 1 présente la synthèse des propriétés des structurelles des différents réseaux.

D'après les observations ci-dessus, nous pouvons postuler que la dynamique structurelle du réseau des auteurs est basée sur deux processus de création des liens : la fusion de petites composantes au cours des éditions et l'arrivée de nouveaux auteurs.

3. Les modèles proposés

3.1. *Modèle 1 : réseaux de collaboration*

Ce modèle simule à chaque itération une collaboration entre acteurs et en déduit les relations dans le réseau. C'est donc un modèle de croissance basé sur la formation des cliques. 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 créer une clique avec en moyenne $\lambda\eta$ anciens sommets et $(1-\lambda)\eta$ nouveaux sommets et l'insère au graphe ; η est le nombre moyen de sommets dans la clique et λ la proportion d'anciens sommets dans la clique. Les paramètres du modèle sont :

– $P(x = i)$ la distribution du nombre d'acteurs par collaboration ($\eta = \sum iP(x = i)$)

– λ la proportion d'ancien acteurs par collaboration

L'algorithme de génération des collaborations est le suivant :

Algorithme 1 Génération des graphes de collaboration(GGC)

1) Pour $\mathbf{t = 1}$ à $\mathbf{N_c}$ faire {N_c est le nombre de collaboration à générer.}

2) $\mathbf{n \leftarrow nb\_Acteurs(P)}$

3) Pour $\mathbf{i = 1}$ à $\mathbf{n}$ Faire

4) Sélectionner un ancien sommet avec une probabilité λ suivant l'attachement préférentiel ou créer un nouveau sommet avec une probabilité $\mathbf{1 - \lambda}$

5) Créer une clique avec les sommets créés et/ou sélectionnés.

3.2. *Modèle 2 : Réseaux hiérarchiques*

Supposons maintenant que chaque individu du modèle précédent est affilié à une organisation qui peut elle aussi être affiliée à une organisation de niveau supérieur et ainsi de suite. Dans ce contexte un sommet $x = (id, aff)$, à un niveau i , est représenté par son identifiant id et celle de son affiliation aff au niveau $i + 1$. Nous proposons donc les définitions suivantes pour les réseaux hiérarchiques :

Définition 1 : Une collaboration hiérarchique de hauteur h, $h \geq 1$ est une collaboration où on peut-extraire à partir des affiliations h réseaux de collaborations. Nous noterons $\mathbf{C}^h$, l'ensemble des collaborations hiérarchiques de hauteur h.

Définition 2 : On définit le niveau d'une collaboration hiérarchique par :

1) les acteurs qui interagissent sont au niveau 0

2) le niveau $i \geq 1$ est constitué des affiliations des sommets de niveau $i - 1$.

Définition 3 : Soit $c \in \mathbf{C}^h$, l'ensemble $\mathbf{G}_c$ des graphes hiérarchiques de collaborations de c est $\mathbf{G}_c = \{G, G_1, \ldots, G_{h-1}\}$ où G_i est le graphe des collaborations des sommets du niveau i et G est le graphe au niveau 0.

Définition 4 : Le vecteur d'affiliation est l'ensemble $\Lambda_c = \{\lambda, \lambda_1, \ldots, \lambda_{h-1}\}$ où $\lambda_i, i > 1$ est la probabilité qu'un nouveau sommet du niveau $i - 1$ soit affilié à une ancienne organisation du niveau i et λ est la probabilité de sélection d'un ancien sommet pour une collaboration au niveau 0

Pour tout nouveau sommet affilié à une nouvelle organisation, il faut affilier cette nouvelle organisation au niveau supérieur à l'aide du vecteur d'affiliation. On peut adopter d'utiliser la même politique d'affiliation sur tous les niveaux ou alors définir une politique d'affiliation différente à chaque niveau. Nous proposons dans l'algorithme 2 un processus qui affilie les nouveaux sommets aux organisations. En première approximation, nous supposons que les sommets s'affilient aux organisations suivant l'attachement préférentiel (AP) à tous les niveaux.

Pour construire le graphe hiérarchique du niveau $i \geq 1$, pour chaque collaboration, on crée une clique avec les affiliations des sommets du niveau $i - 1$. Nous désignerons ce modèle par modèle de génération des graphes de collaborations hiérar-

Algorithme 2 Procédure de création des sommets et leurs affiliations

Entrée $x = (id, aff)$, $\Lambda = \{\lambda, \lambda_1, \ldots, \lambda_H\}$,$niv$

Sortie $x = (id, aff)$

```
1)  SI (niv = H) Alors
2)        x.aff = 0
3)        Stop
4)  FinSi
5)  p ← random()
6)  Si (p ≤ λ_{niv+1}) Alors
7)     Sélectionner suivant AP une organisation y au niveau niv + 1
8)     x.aff ← y.id
9)     Stop
10) Sinon
11)    créer une nouvelle organisation y = (id, 0) au niveau niv + 1
12)    x.aff ← y.id
13)    Affiliation(y, Λ, niv + 1)
14) Finsi
```

chiques(GGCH). Les paramètres du modèle de construction de réseaux hiérarchiques sont :

– $P(x = i)$: La distribution du nombre d'acteurs par collaboration au niveau 0

– $\Lambda_c = \{\lambda, \lambda_1, \ldots, \lambda_{h-1}\}$: le vecteur d'affiliation

Un exemple de collaboration hiérarchique est la collaboration pour les co-publications scientifiques où ont a : au niveau 0 le graphe des auteurs, au niveau 1 le graphe des laboratoires déduit des affiliation des auteurs aux laboratoires et au niveau 2 le graphe des institutions déduit des affiliations des laboratoires aux institutions.

4. Propriétés des modèles

4.1. *Modèle 1 : Réseaux de collaborations*

Proposition 1

Soit t le nombre de collaborations générés par le modèle **GGC**. Pour $t >> 1$, l'espérance du nombre de sommets du réseau est :

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

Preuve

A chaque itération on sélectionne η sommets avec une proportion moyenne de $1-\lambda$ nouveaux auteurs. Ainsi à chaque itération on a $(1-\lambda)\eta$ nouveaux sommets. On a donc

$$n_t = n_0 + t(1-\lambda)\eta \approx t(1-\lambda)\eta \ \texttt{pour} \ t >> 1$$

n_o représente le nombre de sommets de la phase d'initiation.

Proposition 2

Soit t le nombre de collaborations générés par le modèle **GGC**. Pour $t >> 1$, l'espérance du nombre d'arêtes dans le réseau ; en supposant négligeable pour des besoins de simplification l'existence des arêtes multiples ; est donnée par :

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

Preuve

On génère un graphe complet entre η sommets par collaboration. Nous négligeons le fait deux anciens sommets sélectionnés soient connecté. Par conséquent chaque collaboration crée $\frac{1}{2}(1-\lambda)\eta[(1-\lambda)\eta-1]$ arêtes entre les sommets de la collaboration. Pour t collaborations, nous avons :

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

Proposition 3

Des deux propositions précédentes on déduit la densité δ_t pour $t >> 1$:

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

Théoreme 1

Le degré moyen du réseau généré par le modèle **GGC**, si nous négligeons le fait que deux anciens collaborateurs sélectionnés à chaque collaboration peuvent avoir d'anciennes relations, est :

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

Preuve

Une variante de la définition de la densité est $\delta = \frac{\bar{d}}{n_t-1} \Rightarrow \bar{d} = \frac{(\eta-1)}{(1-\lambda)}$ en se basant sur la proposition résumé par 3.

Théoreme 2

La distribution de degré d'un réseau généré par le modèle **GGC**, suit une Loi de Puissance et est donné (pour les grandes valeurs de degré k) par :

$$p_k \approx k^{-\left(1+\frac{1}{\lambda}\right)} \qquad [5]$$

Preuve

La probabilité qu'un ancien sommet soit sélectionné pour une collaboration est :

$$\frac{k}{\sum jp_j} p_k \quad \texttt{où} \quad \sum jp_j = \bar{d} \qquad [6]$$

La fraction de sommet degré k sélectionné quand une collaboration est crée est :

$$\lambda\eta \frac{k}{\sum jp_j} p_k$$

Soit $p_{k,t}$ la valeur p_k au temps t. La variation de la fraction $n_t p_k$ de sommets de degré k par l'ajout de $(1-\lambda)\eta$ sommets est :

$$(n_t + (1-\lambda)\eta)p_{k,t+1} - n_t p_{k,t} = \frac{\lambda\eta}{\bar{d}} \left[(k-\eta+1)p_{k-\eta+1,t} - kp_{k,t}\right]$$

L'équation à l'état stationnaire c'est-à-dire quand $p_{k,t+1} = p_{k,t} = p_k$, lorsqu'on sait que tout sommet sélectionné gagne $\eta - 1$ relations et en remplaçant le degré moyen par son expression de l'équation 4 , est :

$$\texttt{A:}\ p_{\eta-1} = \frac{1}{1+\lambda} \quad \texttt{et B:}\ p_k = \frac{k-\eta+1}{k+\frac{1}{\lambda}(\eta-1)} p_{k-\eta+1}$$

$$\begin{aligned} & & p_k &= \frac{\left(\frac{k}{\eta-1}-1\right)\ldots 1}{\left(\frac{k}{\eta-1}+\frac{1}{\lambda}\right)\ldots\left(2+\frac{1}{\lambda}\right)} \cdot \frac{1}{1+\lambda} \\ \texttt{A et B} \Rightarrow & & &= \frac{\Gamma\left(\frac{k}{\eta-1}\right)\Gamma\left(1+\frac{1}{\lambda}\right)}{\Gamma\left(\frac{k}{\eta-1}+1+\frac{1}{\lambda}\right)} \\ & & &= B\left(\frac{k}{\eta-1}, 1+\frac{1}{\lambda}\right) \end{aligned} \qquad [7]$$

$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ est la beta-function de Legendre, qui est asymptotique à a^{-b} pour les grande valeurs de a et b constant, et par conséquent :

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

Corrolaire 1

D'après le théorème 2, la distribution de degré suit une Loi de Puissance de paramètre $\left(1+\frac{1}{\lambda}\right) > 2$. L'on peut donc conclure, en se basant sur les travaux de (Cohen et Havlin, 2003) que la distance moyenne des réseaux générés par le modèle **GGC** sont des réseaux petit-monde dont la distance moyenne l est donné par :

$$l \approx \log(n_t) \quad [8]$$

Particulièrement, pour $\lambda \geq 0.5$, les réseaux sont ultra-petit-mondes c'est à dire :

$$l \approx \log\log(n_t) \quad [9]$$

Théoreme 3

Pour un sommet de degré $k > \eta$ et pour $\eta > 2$, le coefficient de clustering généré par **GGC** est borné par :

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

Preuve

Soit $k \geq \eta$ le degré d'un sommet. On cherche à déterminer C_k le coefficient de clustering d'un sommet de degré k. On suppose qu'on créé en des collaborations avec en moyenne η acteurs. On considère deux situation : soit les voisins d'un sommet collaborent souvent entre eux, soit ils ne collaborent pas du tout. Dans le cas où les voisins ne collaborent jamais, la structure du graphe qui se résume aux collaboration de ce sommet a la forme d'une étoile (Fig.1). Dans le cas où ils collaborent, tous les voisins peuvent avoir collaboré entre eux dans ce cas $C_k = 1$.

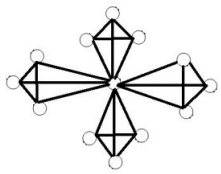


Figure 1 – Sommet ayant participé à des collaborations isolées avec 4 sommets par collaboration

Soit E_k le nombre d'arêtes entre les voisins d'un sommet de degré k. On rappelle que

$$C_k = 2\frac{E_k}{k(k-1)}$$

Dans le cas où le graphe associé aux collaborations du sommet est une étoile on a :

$$E_k = \frac{1}{2}\frac{k}{\eta - 1}(\eta - 1)(\eta - 2) = \frac{1}{2}k(\eta - 2)$$

D'où

$$\frac{\eta - 2}{k - 1} \leq C_k \leq 1$$

4.2. *Modèle 2 : Réseaux hiérarchiques*

Soit $c \in \mathbf{C}^n$ une collaboration hiérarchique et η le nombre moyen d'acteurs par collaboration au niveau 0, $\Lambda_c = \{\lambda, \ldots, \lambda_n\}$ le vecteur d'affiliation et $\mathbf{G}_c = \{G, \ldots, G_{n-1}\}$ l'ensemble des graphes hiérarchiques $G_i = (V_i, E_i)$.

Remarques : Les réseaux du niveau 0 ont les mêmes propriétés que les graphes du modèle 1. En effet, les affiliations n'influencent pas dans le processus de création des collaborations au niveau 0 dans le modèle que nous avons proposé.

Proposition 4

Le nombre de sommets $|V_i|$, à chaque niveau i après la création de t collaborations est :

$$|V_i|_t = t(1 - \lambda)(1 - \lambda_1) \ldots (1 - \lambda_i)\eta \qquad [11]$$

Preuve

La preuve est faite par récurrence.

– au niveaux 0 ont a $t(1 - \lambda)\eta$ après t collaborations soit $(1 - \lambda)\eta$ nouveaux sommets par collaboration.

– supposons qu'au niveau i on a $t(1 - \lambda)(1 - \lambda_1) \ldots (1 - \lambda_i)\eta$ sommets après t itérations, soit $(1 - \lambda)(1 - \lambda_1) \ldots (1 - \lambda_i)\eta$ nouveaux sommets par collaboration et montrons qu'on aura $t(1 - \lambda)(1 - \lambda_1) \ldots (1 - \lambda_{i+1})\eta$ sommets au niveau $i + 1$ après t itérations.

Pour $(1 - \lambda)(1 - \lambda_1) \ldots (1 - \lambda_i)\eta$ nouveaux sommets par collaboration au niveau i, les nouveaux sommets au niveau $i + 1$ sont générés par la fraction de ces sommets qui s'affilient au nouvelles organisation au niveau $i + 1$. Or la probabilité de s'affilier à une nouvelle organisation au niveau $i + 1$ est $(1 - \lambda_{i+1})$. On aura donc $(1 - \lambda)(1 - \lambda_1) \ldots (1 - \lambda_{i+1})\eta$ nouveaux sommets.

Théorème 4

Soit $\bar{d}_i$ le degré moyen du graphe au niveau i,

$$\bar{d}_{i+1} \leq \frac{\bar{d}_i}{1 - \lambda_{i+1}} \qquad [12]$$

Preuve

La somme des degrés au niveau i est $M_i = \bar{d}_i * n_i$. Si on suppose le processus où il y très peu de collaborations entre sommets d'une même organisation du niveau $i+1$ alors on a $M_{i+1} \approx M_i$. En effet, dans ces conditions toute arête du niveau i engendre une arête au niveau $i+1$. Le degré moyen au niveau $i+1$ sera donc :

$$\frac{M_i}{V_{i+1}} = \bar{d}_i \frac{(1-\lambda)(1-\lambda_1)\dots(1-\lambda_i)\eta}{(1-\lambda)(1-\lambda_1)\dots(1-\lambda_i(1-\lambda_{i+1})\eta} = \frac{\bar{d}_i}{1-\lambda_{i+1}}$$

Dans tous les autres cas, le nombre d'arêtes résultant du passage d'un graphe du niveau i au niveau $i+1$ engendre une réduction de la somme des degrés causé par la collaboration entre sommets affiliés à la même organisation au niveau i.

5. Expérimentation

5.1. *Modèle 1 : Réseaux de collaborations*

Après avoir extrait les paramètres des jeux de données en notre procession, nous avons comparé les principales caractéristiques des réseaux obtenus aux réseaux réels. Les paramètres des deux réseaux se différencient par : la proportion moyenne des anciens auteurs par articles qui est faible pour le CARI (0.3) et forte pour HepTh (0.7) ; la distribution du nombre de sommets par article qui est monotone et décroissante pour HepTh alors que pour le CARI elle est croissante entre les points 1 et 2 et décroissante entre 2 et 7 et enfin par le nombre d'article à générer qui est faible pour le CARI (646) et élevé pour HepTh (29554). Pour comparer la dynamique, nous produisons pour chaque année, un nombre aléatoire d'article identique au nombre réel observé. Nous avons constaté que le modèle **GGC** reproduit très bien les caractéristiques principales observées (Voir Fig. 2. Particulièrement, il reproduit dans le même ordre de grandeur la distribution des degrés, le coefficient de clustering, la transitivité, la densité et la distance moyenne. Les réseaux simulés, comme les réseaux réels sont constitués par de nombreuses composantes connexes.

5.2. *Modèle 2 : Réseaux hiérarchiques*

Nous avons utilisé les co-publications extraites du CARI pour construire le vecteur d'affiliations de cette collaboration hiérarchique et simulé notre modèle. Nous donnons ici quelques illustrations comparative entre les réseaux générés et les réseaux réels pour les laboratoires et les institutions (Fig 3). On peut constater que le modèle **GGCH** reproduit très bien la distribution des degrés et le fort coefficient de clustering des graphes hiérarchique. Ces graphes ont aussi de faible distance moyenne. A partir des informations sur les affiliations et les processus qui gouvernent la croissance du réseau au niveau 0, nous sommes parvenus à reproduire un ensemble de graphes corrélés.

6. Conclusion

Dans cet article, nous proposons deux modèles de générations de graphes. **GGC** permet de générer les réseaux de collaborations. Dans **GGCH**, nous supposons que les sommets de **GGC** sont affiliés aux organisations et déduisons les réseaux de ces organisations à partir des affiliations. Nous qualifions ces réseaux de hiérarchiques. Les deux modèles permettent de générer les réseaux petit-mondes. Les expérimentations effectuées montrent que les modèles permettent de reproduire les réseaux très proches des réseaux de terrains observés.

Une propriété importante, elle aussi observée dans de nombreux graphes de terrains est la présence d'une structure de communautés. Il existe peu de modèles réalistes permettant de produire des réseaux possédant une structure de communautés. La question de savoir si les modèles présenté ici peuvent s'étendre de manière à intégrer cette propriété, voire la contrôler, est donc très pertinente et sera abordé dans la suite de ce travail. En plus il serait souhaitable de faire une étude comparative des réseaux générés par **GGC** et d'autres modèles. Pour **GGCH** la définition de la notion de réseau hiérarchique ainsi défini ouvre un vaste champ pour les démonstrations formelles des propriétés déduit à partir des réseaux du niveau 0.

7. Bibliographie

Barabási A.-L., Albert R., « Emergence of scaling in random networks », *science*, vol. 286, n° 5439, p. 509-512, 1999.

Barabási A.-L., Jeong H., Néda Z., Ravasz E., Schubert A., Vicsek T., « Evolution of the social network of scientific collaborations », *Physica A : Statistical mechanics and its applications*, vol. 311, n° 3, p. 590-614, 2002.

Barrat A., Weigt M., « On the properties of small-world network models », *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 13, n° 3, p. 547-560, 2000.

Cohen R., Havlin S., « Scale-free networks are ultrasmall », *Physical review letters*, vol. 90, n° 5, p. 058701, 2003.

Guillaume J.-L., Latapy M., « Bipartite structure of all complex networks », *Information processing letters*, vol. 90, n° 5, p. 215-221, 2004.

Guillaume J.-L., Latapy M., « Bipartite graphs as models of complex networks », *Physica A : Statistical Mechanics and its Applications*, vol. 371, n° 2, p. 795-813, 2006.

i Cancho R. F., Solé R. V., « The small world of human language », *Proceedings of the Royal Society of London B : Biological Sciences*, vol. 268, n° 1482, p. 2261-2265, 2001.

Lattanzi S., Sivakumar D., « Affiliation networks », *Proceedings of the forty-first annual ACM symposium on Theory of computing*, ACM, p. 427-434, 2009.

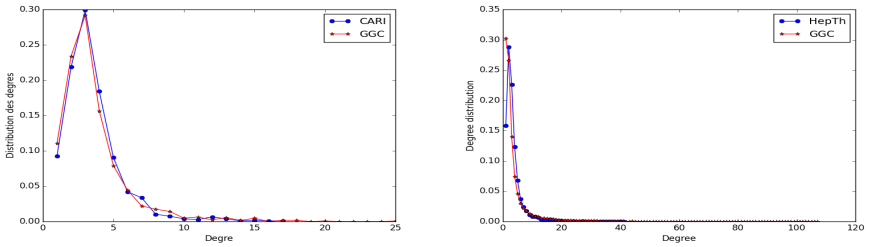
Leskovec J., Kleinberg J., Faloutsos C., « Graphs over time : densification laws, shrinking diameters and possible explanations », *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, ACM, p. 177-187, 2005.

Newman M. E., « The structure of scientific collaboration networks », *Proceedings of the National Academy of Sciences*, vol. 98, n° 2, p. 404-409, 2001.

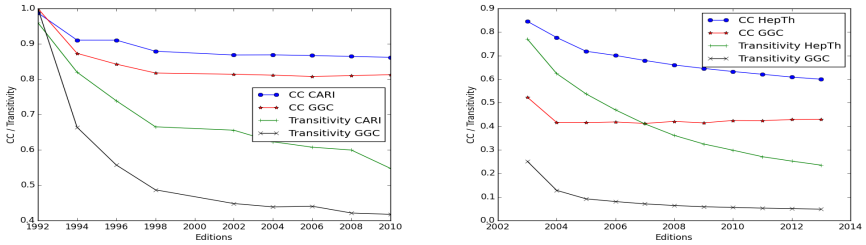
Newman M. E., « The structure and function of complex networks », *SIAM review*, vol. 45, nº 2, p. 167-256, 2003.

Price D. J. d. S., « A general theory of bibliometric and other cumulative advantage processes », *Journal of the American society for Information science*p. 293, 1976.

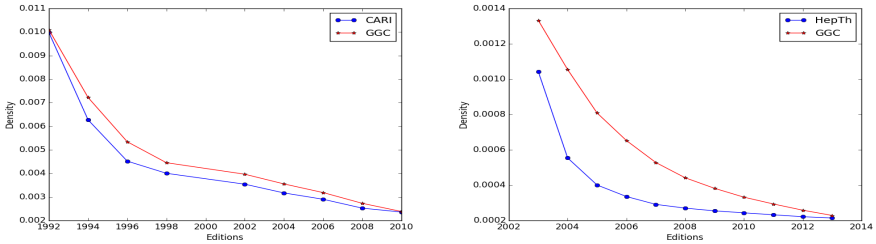
Watts D. J., Strogatz S. H., « Collective dynamics of 'small-world'networks », *nature*, vol. 393, nº 6684, p. 440-442, 1998.



(a) Distribution des degrés

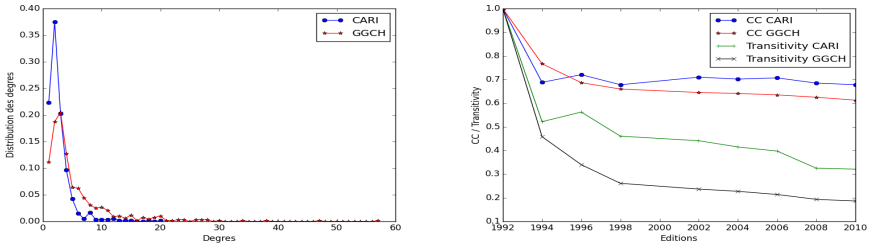


(b) Coefficient de clustering

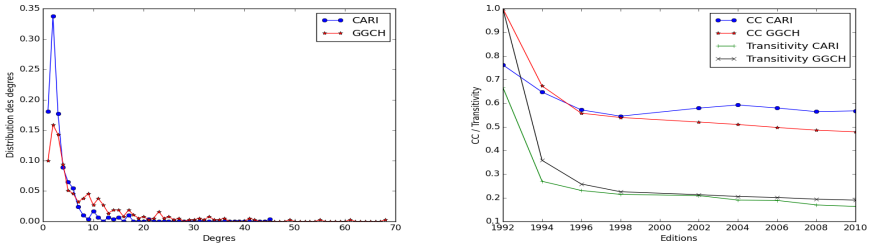


(c) Densite

Figure 2 – Propriétés comparées du réseau des auteurs



(a) Laboratoires



(b) Institutions

Figure 3 – Propriétés comparées du modèle GGCH et des réseaux des institutions et des laboratoires :