

Modèles de génération des graphes de collaboration multi-niveaux

SEMINAIRE AU LIP6

MELEU Ghislain Romaric et Paulin MELATAGIA

UMI 209 UMMISCO

12 octobre 2015



Tables des matières

- 1 Généralités
- 2 Quelques Approches de modélisation
- 3 Les graphes de co-publications
- 4 Modèles proposés
- 5 Conclusion et perspectives
- 6 Bibliographie

Tables des matières

- 1 Généralités
- 2 Quelques Approches de modélisation
- 3 Les graphes de co-publications
- 4 Modèles proposés
- 5 Conclusion et perspectives
- 6 Bibliographie

Graphes de terrains

- Les grands graphes sont utilisés pour modéliser de nombreux phénomènes
- On les regroupe en quatre catégories
 - les réseaux sociaux,
 - les réseaux biologiques,
 - les réseaux technologiques,
 - les réseaux d'informations.

On les désigne généralement par l'expression **graphes de terrains**

Propriétés des graphes de terrains

- de nombreuses caractéristiques communes.
 - Distribution des degrés en Loi de Puissance ;
 - fort coefficient de clustering ;
 - faible densité ;
 - faible distance moyenne ;
 - présence des communautés ...

Dynamique des systèmes

Les systèmes caractérisés par de nombreuses interactions entre les sommets.

- Ajout/suppression de nouveaux sommets.
- Création/suppression des liens entre sommets.

Problèmes

- Comprendre et expliquer la dynamique

Solution proposés

- Utilisation des modèles mathématiques et informatiques

Tables des matières

- 1 Généralités
- 2 Quelques Approches de modélisation
- 3 Les graphes de co-publications
- 4 Modèles proposés
- 5 Conclusion et perspectives
- 6 Bibliographie

Caractéristiques des modèles

- ❶ types de graphes utilisés
 - graphe bipartite
 - graphe simple
 - graphe multiplexe
- ❷ création des sommets
 - statique
 - dynamique
- ❸ et de création des liens
 - aléatoire uniforme
 - rattachement au k-plus proche voisins
 - l'attachement préférentiel

Attachement préférentiel

Les modèles de Price [1] et Barabási[2] sont basé sur deux Principes

- **La Croissance** : A chaque itération un nouveau sommet est ajouté dans le réseau
- **L'attachement préférentiel** : Les nouveaux sommets s'attachent aux anciens avec des probabilités proportionnelles à leurs degrés.

Modèle de Barabasi

$$P(k_i) = \frac{k_i}{\sum k_j}$$

Modèle de Price

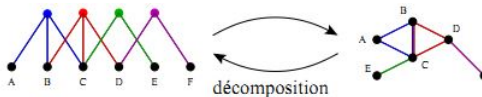
$$P(k_i) = \frac{k_i+1}{\sum (k_j+1)}$$



Modèle de J.L Guillaume et M. Latapy [3, 4]

Une proposition basée sur trois résultats :

- la plupart des grands réseaux d'interactions peuvent être décomposés en graphes bipartis non triviaux,
- on peut engendrer un graphe biparti aléatoire dont la distribution des degrés est donné à l'avance
- la structure bipartie peut être vue comme responsable des propriétés observées sur les réseaux d'interactions



Génération d'un graphe à partir de sa version bipartite et décomposition en cliques pour obtenir un graphe biparti.

Modèle de S. Lattanzi et D. Sivakumar [5]

Une proposition basée sur les observations suivantes :

- les réseaux sociaux présentent deux types d'entités : les acteurs et les sociétés,
- ces entités sont liés par affiliation des acteurs au sociétés
- ces relations peuvent être considérées comme des graphes bipartites (réseaux d'affiliations)
- le réseau social entre les acteurs est obtenu par projection du graphe bipartite

Tables des matières

- 1 Généralités
- 2 Quelques Approches de modélisation
- 3 Les graphes de co-publications**
- 4 Modèles proposés
- 5 Conclusion et perspectives
- 6 Bibliographie

Les différents graphes

A partir des entêtes des articles, nous pouvons construire :

- ❶ Le réseau des auteurs : un sommet est un auteur qui a publié au moins un article et une arête existe entre deux auteurs s'ils ont publié ensemble.
- ❷ Le réseau des laboratoires de recherche : 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 un(des) auteur(s) des deux laboratoires.
- ❸ Le réseau des institutions : un sommet est une institution qui possède un(des) auteur(s) ayant publié au moins un article et une arête relie deux institutions s'il existe au moins un article publié par un(des) auteur(s) de deux laboratoires rattachés chacun à une des institutions.

Observations

- 1 Une publication met en œuvre plusieurs entités organisés de façon hiérarchique.
- 2 Les réseaux sont corrélés à cause des relations d'affiliations.
- 3 Les auteurs sont les entités qui permettent la construction des relations.
- 4 Chaque réseau est un réseau de collaboration.

Illustrations

Community Analysis and Link Prediction in Dynamic Social Networks

Blaise Ngonmang, Emmanuel Viennet, Maurice Tchuente, and Vanessa Kamga

B. Ngonmang (✉) • V. Kamga

Université de Paris 13, Sorbonne Paris Cité, L2TI, 93430 Villetaneuse, France

Université de Yaoundé 1, IRD UMI 209 UMMISCO-LIRIMA, Equipe IDASCO, BP 812

Yaoundé, Cameroon

e-mail: blaise.ngonmang@univ-paris13.fr; vansylvania.kamga@lirima.org

E. Viennet

Université de Paris 13, Sorbonne Paris Cité, L2TI, 93430 Villetaneuse, France

e-mail: emmanuel.viennet@univ-paris13.fr

M. Tchuente

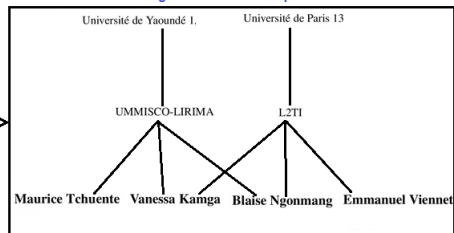
Université de Yaoundé 1, IRD UMI 209 UMMISCO-LIRIMA,

Equipe IDASCO, BP 812 Yaoundé, Cameroon

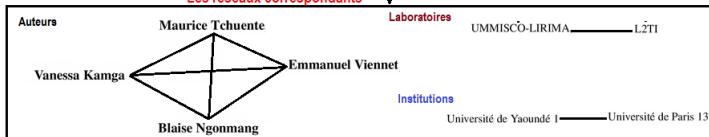
e-mail: maurice.tchuente@lirima.org



Organisation hiérarchique des entités



Les réseaux correspondants



Propriétés des réseau du CARI

L'ensemble des données utilisées est constituée de 646 articles.

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

TABLE: Propriétés structurelles des différents réseaux. CA_AUT : réseau des auteurs, CA_LAB : réseau des laboratoires, CA_INST : réseau des institutions

Généralisation

- Nous généralisons ce système à une structure hiérarchique où :
 - un acteur est affilié à une organisation
 - l'organisation pouvant être affiliée à une organisation de niveau supérieur . . .
 - les acteurs collaborent (publication, films,...)
 - les relations entre acteurs sont déduits des collaborations
 - les relations entre entités à un même niveau d'organisation sont déduites de celles du niveau inférieur.

Nous disons alors que ces réseaux sont hiérarchiques

Problémantique

- Comprendre ces systèmes demande de :
 - ① comprendre les interactions entre les acteurs
 - ② comprendre les règles d'affiliations aux organisations
- afin de mieux expliquer les réseaux des organisations qui sont déduits des interactions de ces acteurs.

Tables des matières

- 1 Généralités
- 2 Quelques Approches de modélisation
- 3 Les graphes de co-publications
- 4 Modèles proposés**
- 5 Conclusion et perspectives
- 6 Bibliographie

Idée de base

Une collaboration est formé par :

- Deux acteurs ou plus.
- Un acteur dans une collaboration est soit
 - déjà présent dans le réseau (ancien)
 - un nouvel arrivant (nouveau)
- Une collaboration génère une clique entre tous les participants

Les paramètres du modèle sont :

- $P(x = i)$ la distribution du nombre d'acteurs par collaboration
($\eta = \sum i P(x = i)$)
- λ la proportion d'anciens acteurs par collaboration

Réseaux de collaboration

Algorithme 1 Génération des graphes de collaboration(GGC)

- ❶ Pour $t = 1$ à N_c faire
 - ❷ $n \leftarrow \text{nb_Acteurs}(P)$
 - ❸ Pour $i = 1$ à n Faire
 - ❹ Sélectionner un ancien sommet avec une probabilité λ
 suivant l'attachement préférentiel ou créer un
 nouveau sommet avec une probabilité $1 - \lambda$
 - ❺ Créer une clique avec les sommets créés et/ou sélectionnés.
-

Quelques concepts

Définition : 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 $i - 1$.

Quelques concepts

Définition 2 : Le vecteur d'affiliation est $\Lambda_c = \{\lambda, \lambda_1, \dots, \lambda_{h-1}\}$
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
- λ est la probabilité de sélection d'un ancien sommet pour une collaboration au niveau 0

Réseaux hiérarchiques

- Un sommet x est représenté par un couple (id, aff) où :
 - id est son identifiant
 - aff celui de son affiliation au niveau $i + 1$
- Les paramètres du modèle sont :
 - $P(x = i)$: La distribution du nombre d'acteurs par collaboration au niveau 0
 - $\Lambda_c = \{\lambda, \lambda_1, \dots, \lambda_{h-1}\}$: le vecteur d'affiliation
- Pour construire le graphe hiérarchique du niveau $i \geq 1$
 - pour chaque collaboration et pour tout niveau $i > 0$
 - créer une clique avec les affiliations des sommets du niveau $i - 1$

Réseaux hiérarchiques

Algorithme 2 Procédure de création des affiliations

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

Sortie $x = (id, aff)$

- **SI** ($niv = H$) **Alors**
- $x.aff = 0$
- **Stop**
- **FinSi**
- $p \leftarrow \text{random}()$

Réseaux hiérarchiques

Algorithme 3 Procédure de création des affiliations(suite)

- **Si** ($p \leq \lambda_{niv+1}$) **Alors**
- Sélectionner suivant AP une organisation y au niveau $niv + 1$
- $x.aff \leftarrow y.id$
- **Stop**
- **Sinon**
- créer une nouvelle organisation $y = (id, 0)$ au niveau $niv + 1$
- $x.aff \leftarrow y.id$
- **Affiliation**($y, \Lambda, niv + 1$)
- **Fin**

Propriété du réseau des acteurs (niveau 0)

Nombre de sommets

$$n_t = t(1 - \lambda)\eta$$

Nombre d'arêtes

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

Degré moyen

$$\bar{d} = \frac{\eta - 1}{1 - \lambda}$$

Propriété du réseau des acteurs (niveau 0)

Distribution des degrés

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

- La probabilité qu'un ancien sommet soit sélectionné est :
 $\frac{k}{\sum_j j p_j} p_k$ où $\sum_j j p_j = \bar{d}$
- La variation des sommets de degré k est :

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

Propriété du réseau des acteurs (niveau 0)

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

$$p_k = \frac{(\frac{k}{\eta-1}-1)\dots 1}{(\frac{k}{\eta-1}+\frac{1}{\lambda})\dots(2+\frac{1}{\lambda})} \cdot \frac{1}{1+\lambda}$$

$$\begin{aligned} A \text{ et } B \Rightarrow &= \frac{\Gamma\left(\frac{k}{\eta-1}\right)\Gamma\left(1+\frac{1}{\lambda}\right)}{\Gamma\left(\frac{k}{\eta-1}+1+\frac{1}{\lambda}\right)} \\ &= B\left(\frac{k}{\eta-1}, 1+\frac{1}{\lambda}\right) \end{aligned} \quad (1)$$

Propriétés des réseaux de niveau > 0

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

Nombre de sommets

$$|V_i|_t = t(1 - \lambda)(1 - \lambda_1) \dots (1 - \lambda_i)\eta$$

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

Degré moyen

$$\bar{d}_{i+1} \leq \frac{\bar{d}_i}{1 - \lambda_{i+1}}$$

Paramètres

	CARI	HEPTH
Nombre de collaboration	646	29554
λ	0.3	0.7
Λ_c	0.3, 0.56, 0.35	—

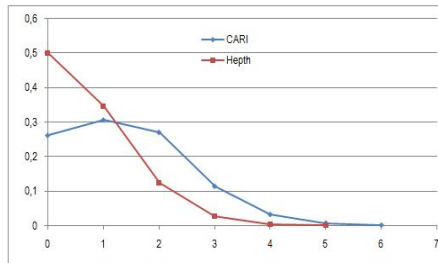
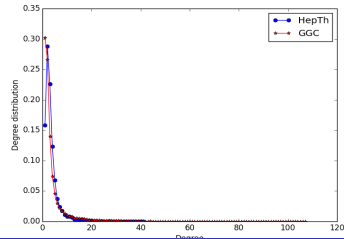
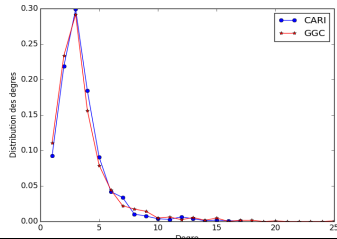
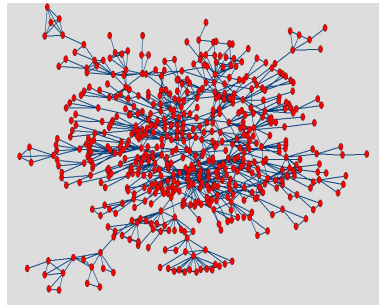
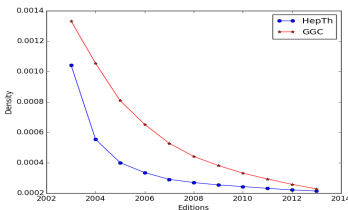
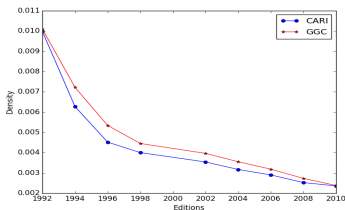
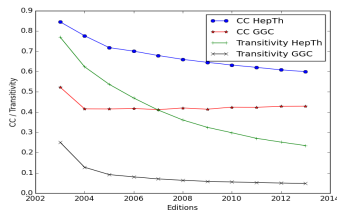
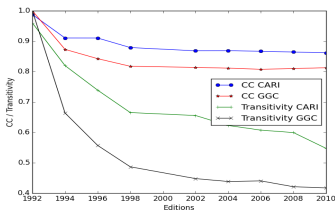


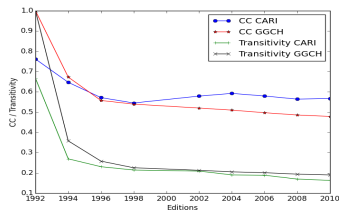
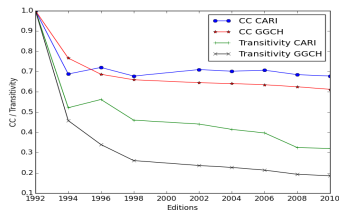
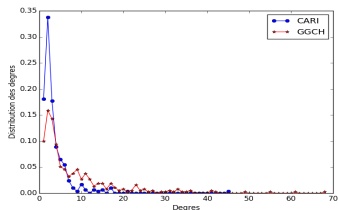
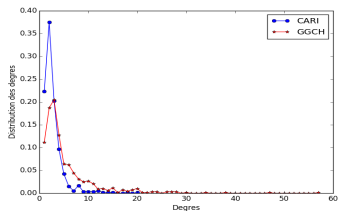
FIGURE: Distribution du nombre de sommets par collaboration



Propriétés comparées des réseaux des auteurs



Résultats des réseaux hiérarchiques



Tables des matières

- 1 Généralités
- 2 Quelques Approches de modélisation
- 3 Les graphes de co-publications
- 4 Modèles proposés
- 5 Conclusion et perspectives**
- 6 Bibliographie

Conclusion

Nous avons présenté :

- un modèle de génération de graphe multi-couches
- basé sur la création de clique (ancien et nouveaux sommets)
- associé à un processus d'affiliation à des organisations.
- La sélection des sommets étant faite suivant l'attachement préférentiel.
- Les simulations effectuées montrent que le modèle reproduit les propriétés des deux graphes de publications étudiés.

Perspectives

Les orientations futures de ce travail sont :

- exprimer les propriétés des réseaux des organisations à partir du réseau des acteurs ;
- valider les modèle avec les jeux de données autres que les réseaux de co-publications ;
- Identifier des fonctions d'affiliations des acteurs aux organisations (et des organisations à d'autres organisations de niveau supérieur) plus à même de modéliser la réalité

Tables des matières

- 1 Généralités
- 2 Quelques Approches de modélisation
- 3 Les graphes de co-publications
- 4 Modèles proposés
- 5 Conclusion et perspectives
- 6 Bibliographie**

Références

-  de Solla Price, D. (1976). A General Theory of Bibliometric and Other Cumulative Advantage Processes. Journal of the American Society for Information Science, 293.
-  Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. science, 286(5439), 509-512.
-  J. L. Guillaume And M. Latapy(2006). Bipartite Graphs As Models Of Complex Networks. Physica A : Statistical Mechanics and its Applications, 371(2), 795-813.
-  J. L. Guillaume And M. Latapy (2004). Bipartite structure of all complex networks. Information processing letters, 90(5), 215-221.
-  Lattanzi, S., & Sivakumar, D. (2009, May). Affiliation networks. In Proceedings of the forty-first annual ACM symposium on Theory of computing (pp. 427-434). ACM.