

# Étude d'un modèle aléatoire conservant les bicliques maximales dans les graphes bipartis

Fabien Tarissan & Lionel Tabourier

stages@complexnetworks.fr  
http://www.complexnetworks.fr

LIP6 – CNRS – SU

**Contexte.** Dans de très nombreux contextes applicatifs, on rencontre de grands graphes n'ayant aucune structure simple apparente, que nous appellerons graphes de terrain (par opposition aux graphes explicitement construits par un modèle ou une théorie). Citons par exemple la topologie de l'internet (routeurs et câbles entre eux), les graphes du web (pages web et liens hypertextes entre elles), les échanges divers (pair-à-pair, e-mail, etc), mais aussi les réseaux sociaux, biologiques ou linguistiques.

Il est apparu récemment que la plupart de ces grands graphes ont des propriétés statistiques en commun. Notamment, ils ont une densité très faible, une distance moyenne faible, une distribution de degrés hétérogène, et une densité locale forte. Depuis lors, de nombreux travaux ont été menés visant à capturer ces propriétés dans des modèles, nécessaires tant pour effectuer des simulations que pour étudier formellement ces objets, et bien sûr pour en comprendre la nature.

Ainsi on peut citer les graphes aléatoires générés par le modèle d'Erdős-Rényi qui capturent la densité faible mais qui échouent à présenter une distribution des degrés hétérogène. D'autres modèles, comme celui de Barabási-Albert, réussissent à reproduire la distribution hétérogène des degrés mais avec une densité local faible, tandis que d'autres encore, comme le modèle de Watts and Strogatz, génèrent des graphes ayant les propriétés inverses. Enfin, le Configuration Model génère des graphes aléatoires respectant une séquence de degré fixée mais avec une densité local faible.

Le but de ce stage est d'explorer une nouvelle manière de générer des graphes aléatoires en s'appuyant sur la version bipartite du Configuration Model.

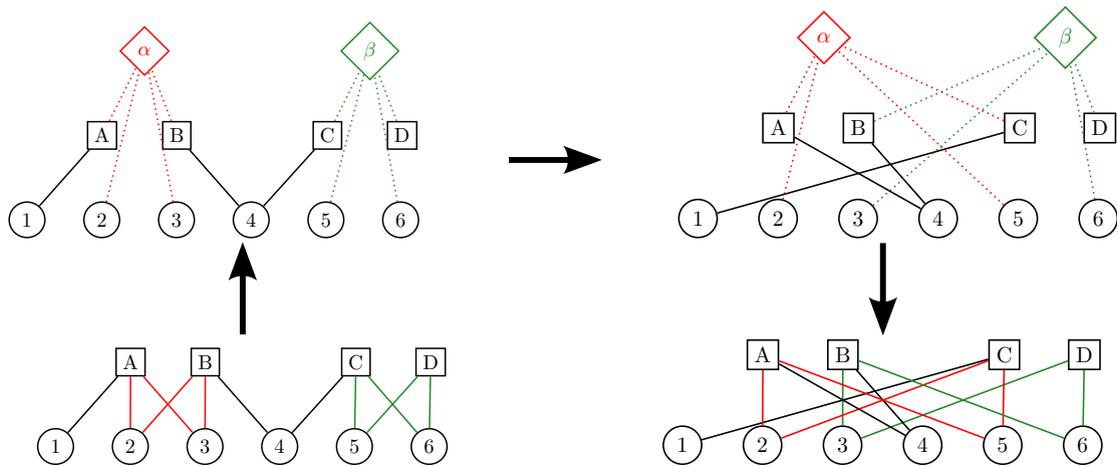
**Graphes bipartis.** En effet, bien que pertinente, la représentation des réseaux sous forme de graphes ne permet pas de rendre compte de la complexité inhérente à la structure hiérarchique de la plupart des réseaux. Si on considère par exemple les réseaux d'acteurs qui relient les acteurs d'un même film ou les réseaux de co-publications qui relient les auteurs publiant ensemble, il est bien plus réaliste de relier les acteurs aux films dans lesquels ils jouent, et les auteurs à leurs articles. Cette remarque a amené la communauté scientifique à s'intéresser aux graphes bipartis dans lesquels l'ensemble des nœuds est scindé en deux ensembles disjoints, et  $\top$  et  $\perp$  (par exemple les films et les acteurs), de façon à ce que les liens relient des nœuds du premier ensemble à des nœuds du second. Cet objet fondamental s'est révélé pertinent à la fois pour l'analyse et la génération aléatoire des graphes de terrains.

Or, si l'extension naturelle du Configuration Model pour les graphes bipartis permet bien de préserver les séquences de degrés des nœud  $\top$  et  $\perp$  (voir schéma),



il ne permet pas en revanche de conserver des structures plus fortes, comme les bicliques. Une simple modification de deux liens fait perdre les deux recouvrements dénotés par les deux bicliques (en rouge et en vert), qui ont totalement disparus après la phase aléatoire.

**Objectifs du stage.** Pour tenter d'améliorer la génération des graphes bipartis, et en particulier la conservation des structures de recouvrement présentes dans les réseaux réels, une proposition a été faite qui consiste à détecter dans un premier temps les structures recouvrantes dans le graphe biparti puis les représenter à l'aide d'un troisième niveau, définissant de fait un *graphe triparti* encodant la structure bipartie. Dès lors il devient possible de générer aléatoirement un graphe triparti puis de projeter le graphe généré afin de récupérer un graphe biparti, comme le montre la figure suivante :



Le but de ce stage est d'étendre cette proposition et d'étudier sa pertinence en tant que support pour générer des graphes réalistes. Plusieurs pistes peuvent être envisagées, parmi lesquelles on peut citer :

**Algorithme :** la phase de détection des structures recouvrantes passe par une énumération des bicliques maximales dans le graphe biparti. Cette phase est coûteuse si l'on fait une énumération exhaustive mais il est possible d'améliorer son efficacité en pratique si on s'appuie sur les propriétés attendues du graphe.

**Codage triparti :** une fois les bicliques maximales détectées, plusieurs stratégies peuvent être adoptées pour les représenter dans une structures tripartie. Il est donc nécessaire de comprendre quel impact ces stratégies vont avoir sur la phase de génération aléatoire

**Génération aléatoire :** une fois les bicliques maximales représentées par un graphe triparti, plusieurs générations aléatoires peuvent être envisagées. Il s'agit donc là de comprendre quelles sont les plus efficaces pour générer des graphes tripartis qui, une fois projetés, seront proches du graphe biparti original.