Analyse statistique de la périphérie des graphes de réseaux sociaux

Rapport de stage - M2 Physique

Ecole Normale Supérieure de Lyon

Lionel Tabourier

Stage effectué sous la direction de Christophe Prieur.
Laboratoire d'Informatique et d'Algorithmique : Fondements et Applications.
Université de Paris 7.

Avril à Juillet 2006

Remerciements

Je souhaiterais remercier Christophe Prieur, pour le temps et l'attention qu'il m'a consacré tout au long de ce stage. Je remercie également Toufik Bennouas, Pascal Pons, Mohamed Bouklit, Fabien Baille et Houy Kuoy pour l'aide qu'ils m'ont fournie au cours de ce travail.

Table des matières

1 Introduction générale : motivations et objectifs

Le sujet de ce stage entre dans le domaine de la modélisation des systèmes complexes. L'ambition générale de cette discipline est la recherche d'une meilleure compréhension des systèmes contenant un grand nombre de constituants de différentes natures, ayant entre eux des interactions non-triviales.

Le choix de cette thématique de travail tient d'abord au très vaste champ d'investigation qu'ouvrent les travaux effectués récemment sur ce terrain. En particulier l'application de ces méthodes aux sciences de l'homme. C'est sur ce type de problématiques que je souhaite effectuer ma thèse, et j'avais pour objectif de faire de ce stage une entrée en matière pour mon travail de recherche à venir.

Comme il s'agit d'un domaine relativement neuf, et dont l'objet peut prêter à discussion, le risque de choisir une voie de recherche qui soit stérile ou scientifiquement infondée me semble plus élevé qu'ailleurs. Je comptais donc aussi utiliser ce stage pour alimenter une réflexion d'ensemble sur les questions que l'on peut chercher à résoudre et celles qui mènent à une impasse.

Par ailleurs, l'étude des systèmes complexes est interdisciplinaire, d'abord parce que les objets de recherche viennent de domaines très variés (e.g. de la génétique, de la biophysique, de l'informatique ou encore de la sociologie), et aussi parce que les méthodes employées font appel à des compétences issues de différentes disciplines.

Comme le travail de recherche utilise de grandes bases de données et nécessite souvent des algorithmes efficaces de traitement de ces données, une part importante de la recherche sur les systèmes complexes (et sur les graphes en particulier) est effectuée par des informaticiens. Il me paraissait donc important d'avoir un aperçu de leurs préoccupations et de leurs méthodes de travail sur les questions de réseaux sociaux. C'est pourquoi, notamment sur les conseils de Michel Morvan, j'ai choisi d'effectuer mon stage au LIAFA, sous la direction de Christophe Prieur, dont les sujets de recherche me semblaient correspondre particulièrement à ce sur quoi je souhaitais travailler.

Au cours de cette étude, on cherche à décrire des phénomènes collectifs à l'aide de réseaux, c'est-à-dire des ensembles d'objets ou de personnes connectés ou maintenus en liaison, et par extension, l'ensemble des liaisons ainsi établies. Ici, on s'intéresse particulièrement aux réseaux sociaux, dont une partie importante est constituée de réseaux extraits du Web.

On va chercher à réunir les membres du réseau par affinités ou caractéristiques communes et voir comment les différents groupes ainsi constitués communiquent les uns avec les autres.

Les axes de réflexion sur lesquels on se propose de travailler ici visent à donner des éléments de réponses aux questions suivantes : par quelle catégorie d'agents les informations transitent-elles d'un groupe à un autre ? Quelle est la dynamique d'évolution des groupes constitutifs du réseau décrit ?

Pour modéliser un réseau, on peut employer le vocabulaire mathématique issu de la théorie des graphes, dont on donne ci-après les éléments indispensables à ce travail.

2 Contexte et définitions fondamentales

2.1 Définitions sur la notion de graphe

Un graphe est un objet mathématique constitué d'un ensemble de n noeuds reliés entre eux par des liens qui peuvent être éventuellement orientés (on parle alors d'*arcs*) ou pondérés. Le nombre de liens établis par chacun des n sommets est son *degré* (ou sa *connectivité*).

Ainsi, le graphe peut être intégralement représenté à l'aide d'une matrice $(n \times n)$, dite *matrice d'adjacence* telle que :

$m_{ij} = \begin{cases} \text{poids (ou 1 pour un graphe non-pondere) s'il existe un lien de } i \text{ vers } j. \\ 0 \text{ sinon.} \end{cases}$

Et si le graphe n'est pas orienté, $m_{ij} = m_{ji}$. Des illustrations de cette définition sont données dans l'annexe A du rapport.

On dit qu'un ensemble de sommets du graphe est *connexe* s'il existe toujours un chemin pour lier deux sommets de l'ensemble.

2.2 Réseaux étudiés à l'aide de graphes

Un graphe ne contient pas directement d'information sur le contenu de ce qu'il représente : natures des agents, et des relations entre ceux-ci. Alors on peut en employer pour formaliser des environnements où une description complète des liens ferait intervenir un grand nombre de paramètres. Il est alors possible de décrire une grande diversité d'objets. On expose dans cette partie les graphes utilisés au cours de ce stage, et les données à partir desquelles on les a conçus. Par ailleurs, on a placé à la fin du rapport un résumé reprenant les caractéristiques essentielles et donnant une représentation de ces graphes.

2.2.1 Des graphes du Web

Le plus souvent, on les construit ainsi : un noeud représente une page web, un arc sera l'existence d'un lien hypertexte de la page i vers la page j (graphe orienté, non pondéré). Ce terrain d'étude suscite l'intérêt d'une communauté importante de chercheurs, notamment en raison des enjeux économiques qui lui sont associés (par exemple, le classement "Pagerank" effectué par le moteur de recherche Google utilise la matrice d'adjacence de ces graphes). On trouve une revue intéressante des moteurs de recherche pour Internet dans [1].

Dans cette catégorie, on utilise notamment des graphes de blogs, qui sont réalisés à partir d'une méthode d'exploration du Web dite *de parcours en largeur*. Le principe en est le suivant : on prend un sommet (la racine), et on répertorie tous ceux qui lui sont connectés, puis pour chacun des noeuds ainsi répertoriés, on effectue la même tâche et ainsi de suite. On oppose souvent cette méthode à une autre : le *parcours en profondeur*, qui consiste à prendre une origine, aller sur un des sommets adjacents, puis de ce sommet, aller à un autre etc ; ainsi, on s'éloigne rapidement de la racine, alors qu'avec un parcours en largeur, on répertorie tous les noeuds au voisinage de l'origine.

Dans ce cas particulier de réseau, on répertorie pour chaque blog les URL des blogs figurant sur son *blogroll*, c'est-à-dire la liste des blogs recommandés par l'auteur. Pour faciliter la lecture et la comparaison des différents résultats produits dans ce rapport, j'indiquerai régulièrement la base de données dont ils sont tirés. Pour les données de blogs, on dispose de trois sources distinctes : "Scorpione" et "Yacine" d'une part (tirés l'un et l'autre du réseau Skyblog), "Dupin" d'une autre (les noms sont attribués en fonction du blog racine du réseau).

On dispose également de données d'une autre nature recueillies sur des blogs. Dans la base "Full", on a enregistré les différents *posts* (les commentaires) laissés par des lecteurs ayant eux-mêmes réalisé un blog. On établira un lien entre deux blogs si les deux *bloggers* ont eu l'occasion de dialoguer par le biais de posts.

A partir de données collectées sur le Web, on a aussi construit des graphes décrivant l'encyclopédie libre en ligne Wikipedia. On dispose pour chaque page des noms (ou adresses IP) des rédacteurs qui ont participé à l'article. On travaille alors sur le graphe que l'on qualifie de *biparti*, c'est-à-dire qui contient deux familles de noeuds, ici les pages et les rédacteurs, que l'on relie si le rédacteur a participé à la page.

2.2.2 D'autres graphes de réseaux sociaux

On dispose de la liste des administrateurs des firmes cotées au SBF 250 (l'extension du CAC 40 à 250 entreprises) entre 1994 et 2004 (base de données "Admin" [1]). Là encore, on peut construire plusieurs types de graphes à l'aide de ces données, e.g. les noeuds sont des administrateurs et on lie deux administrateurs s'ils collaborent dans un même conseil d'administration. Autre possibilité : les noeuds sont des administrateurs ou des firmes (graphe biparti), et on connecte un administrateur à un groupe s'il siège au conseil d'administration. Le graphe que l'on étudiera à partir de ces données ne comprend que les noeuds (administrateurs ou firmes) présents au cours des onze années durant lesquelles on a recueilli des données. On isole ainsi une sorte de noyau dur au sein de ce réseau.

Les graphes construits sur des données d'acteurs (qui sont en libre accès sur le site D'A.L. Barabasi [2]) sont construits en connectant des acteurs qui participent à un même film. Comme pour les données de blogs, la collecte est réalisée en parcourant le réseau en largeur : on choisit un noeud racine et pour chacun de ses films, on recueille la liste des acteurs qui y ont également pris part (on se limite aux films et aux acteurs répertoriés dans *the internet movie database*) ; puis on procède de même pour chacun des acteurs de cette liste et ainsi de suite. On n'a utilisé ici qu'une petite partie de la totalité de cette base.

D'autres graphes utilisés sont réalisés à l'aide de la base de données "Tra". Celle-ci est constituée à partir des actes de mariage du XIX^{eme} siècle d'hommes français dont le nom commence par Tra. Les professions du père

[1] Le travail de collecte de données sur les blogs a été réalisé par Dominique Cardon, et celui sur les conseils d'administration par Nathalie Del Vecchio qui ont eu l'amabilité de les mettre à ma disposition. Pour une étude plus approfondie sur ce second sujet, on peut citer : [2].

[2] www.nd.edu/~networks/resources.htm

du marié et du marié figurent sur l'acte, on peut alors construire par exemple un graphe pondéré et orienté, où on trace un arc de la profession du père vers la profession du fils, le poids correspondant au nombre d'occurrences de ce lien[3]. Pour cette base de données, ce type de graphes serait sans doute le plus significatif ; cependant, le graphe que l'on manipulera ici est dépourvu d'orientation et de pondération, car certains algorithmes de traitement des données dont on fera usage ne prennent pas ces caractéristiques en compte.

Enfin, la base des Forums Sociaux Européens ("FSE") est constituée des associations qui ont participé aux séries de conférences axées sur l'altermondialisme à Londres, Florence et Paris, au cours de l'année 2004. On construira deux familles de graphes depuis ces données :

- Dans la première, on lie deux associations lorsqu'elles ont participé à une même conférence.
- Dans la seconde, les graphes sont bipartis : les sommets sont les associations et les conférences ; on les relie lorsque l'association était représentée à la conférence.

2.3 L'information contenue dans la topologie du graphe

L'analyse des graphes permet d'accéder à une "géométrie" des relations entre agents. Mais on peut penser que celle-ci a une signification qui va au-delà de cet aspect formel.

On peut donner un exemple simple d'une caractéristique des graphes de réseaux sociaux qui donne un aperçu du sens que l'on peut associer à sa topologie : la propriété de "petit monde" [5]. De tels graphes présentent les particularités suivantes :

- Le chemin le plus court pour aller d'un noeud quelconque à un autre est de longueur (en nombre de liens) bien inférieure, en moyenne, à celle d'un réseau régulier (comme une grille de maille carrée à deux dimensions). En cela, le petit monde se rapproche de réseaux où les noeuds seraient connectés de manière aléatoire. C'est ce qui justifie qu'en regardant toute la population française, en examinant par exemple le réseau de qui a serré la main à qui et en choisissant deux individus au hasard, la probabilité pour qu'il y ait au moins un chemin de longueur inférieure à cinq ou six poignées de mains entre ces deux individus est de l'ordre de 0.5.
- Mais le petit monde se distingue fortement du réseau aléatoire lorsque l'on considère son *clustering* : la probabilité pour que deux voisins d'un même sommet soient connectés entre eux. Pour un petit monde, et selon le principe "les amis de mes amis sont mes amis", cette valeur est relativement élevée, en comparaison du clustering d'un réseau aléatoire (qui tend vers 0 lorsque le nombre de noeuds augmente, à connectivité moyenne constante).

Dans de nombreux travaux récents, on cherche à identifier à l'aide de la topologie des groupes qui seraient des communautés constitutives du graphe. Chacun de ces groupes est lui-même un sous-graphe, auquel on peut appliquer les mêmes outils de description que pour le graphe complet.

2.4 Communautés de sommets

Lorsqu'on modélise un réseau à l'aide d'un graphe, une communauté est décrite comme une partie dense (en liens) d'un graphe globalement peu dense (il n'y a pas de seuil formel).

On espère que l'analyse du graphe concorde avec les informations que l'on a obtenu sur le réseau par d'autres moyens, en particulier par l'examen de son contenu. Ainsi les algorithmes réalisés par des chercheurs du Liafa sont testés actuellement sur des graphes de Wikipedia, pour voir si les communautés décelées correspondent effectivement à des thématiques voisines de contenu des pages.

On verra par la suite que notre travail sur les graphes est subordonné à la façon dont on détecte les communautés. On va donc présenter très brièvement les différentes méthodes couramment utilisées pour découper le graphe en communautés, puis on expliquera un peu plus en détails le principe des deux programmes employés ici.

2.4.1 Etat de l'art

La question de la détection des communautés d'un graphe a donné lieu, ces dernières années, à un nombre important de publications. Dans la plupart de ces travaux, le découpage en communautés est une partition : chaque noeud appartient à une et une seule communauté. Cela sera aussi le cas des algorithmes utilisés au cours de ce stage.

[3] Je remercie ici Maurizio Gribaudi et Jean-Pierre Pélissier par l'intermédiaire de qui j'ai pu travailler sur cette base, et qui m'ont apporté un point de vue sociologique sur ces données auxquelles ils ont consacré une partie de leurs recherches (cf [3] et [4]).

Un certain nombre de méthodes courantes utilisant des outils d'algèbre linéaire sur la matrice d'adjacence, comme par exemple la bissection spectrale [6], permettent de réaliser des partitions du graphe mais en connaissant au préalable le nombre et la taille des communautés à détecter.

A l'opposé, les approches agglomératives ne nécessitent pas de connaître les tailles communautaires. On y mesure un indice qui doit représenter la distance entre deux noeuds du graphe ; la façon dont on définit la distance dépend de l'algorithme. Initialement, les communautés contiennent un unique noeud, puis, à chaque étape, on regroupe deux communautés proches (au sens de cette distance). De plus, on doit définir un critère d'arrêt, sans lequel le procédé serait itéré jusqu'à ce que ce qu'il n'y ait plus qu'une communauté pour tout le graphe. On peut considérer que les programmes employés au cours de ce stage : Walktrap et Cosmoweb, entrent dans cette catégorie.

La démarche inverse est aussi possible : on part d'une unique communauté pour tout le graphe, puis on retire à chaque étape des liens. De cette manière, on remplace le graphe par des sous-graphes connexes et chaque élément connexe est assimilé à une communauté, puis on itère le procédé sur chacun de ces sous-graphes. On peut évoquer dans cette catégorie l'algorithme de Girvan et Newman [7] ; dans ce cas, les arêtes retirées sont celles qui ont la plus grande "centralité d'intermédiarité", c'est-à-dire le nombre de plus courts chemins d'un noeud à un autre passant par ce lien. En affirmant que ces liens sont des liens intercommunautaires, on utilise le fait qu'il y a peu de liens connectant une communauté à une autre, point sur lequel on va revenir plus en détails.

2.4.2 Walktrap

Cet algorithme a été réalisé par Pascal Pons (pour plus de détails sur le fonctionnement de Walktrap, on consultera d'ailleurs [8]). Il utilise le constat intuitif qu'une marche aléatoire tend à rester piégée dans une communauté. On effectue donc sur le graphe une marche aléatoire courte (de l'ordre de cinq pas), dont la règle est que pour un noeud lié à k autres sommets, la probabilité d'accéder à chacun des voisins est de 1/k (en tous cas pour des réseaux non-pondérés, auxquels on se restreint pour expliquer le fonctionnement de Walktrap).

De cette manière, on peut évaluer la probabilité de se trouver au sommet j en partant de i après avoir effectué p pas. Estimée pour chaque couple de sommets du graphe, cette probabilité permet de définir la distance dans le graphe (on n'entrera pas dans le détail de la définition de la distance entre noeuds ou entre communautés). Un premier découpage communautaire est réalisé de manière à ce que la distance entre des noeuds d'une communauté soit faible devant la distance entre des noeuds de communautés différentes.

Ensuite, Walktrap agglomère ces communautés : on définit un indice dont le rôle est d'évaluer la qualité de la partition. Ici il s'agit de la modularité introduite par Newman [9], notée Q et définie par :

$$Q = \sum_i (e_{ii} - a_i^2)$$

où $a_i = \sum_j e_{ij}$ et e_{ij} est la fraction des arêtes du graphe liant la communauté i à la communauté j, a_i est donc la fraction des liens ayant au moins une extrêmité dans i.

L'algorithme prend alors les deux communautés les moins éloignées et les fusionne. Puis il réévalue les distances entre communautés et réalise une nouvelle fusion et ainsi de suite, jusqu'à n'obtenir qu'une seule communauté recouvrant tout le graphe. Ensuite, pour chacun des découpages intermédiaires, on évalue la modularité Q, et le partage communautaire conservé est celui qui donne la valeur de Q la plus élevée.

2.4.3 Cosmoweb

Cosmoweb est un programme de partition communautaire et de visualisation des réseaux du web (d'ailleurs, il ne prend pas en compte d'éventuelles pondérations puisque les graphes du web en sont dépourvus). Il a été conçu par Toufik Bennouas et al., et on peut se référer à [10] pour des informations complémentaires sur sa réalisation.

Il utilise un modèle inspiré des modèles gravitationnels de l'univers pour regrouper les sommets en communautés. En effet, on peut voir le web (et les communautés sociales en général) comme un système s'organisant à plusieurs échelles : simple page, site ... , un peu comme un système stellaire fait partie d'une galaxie, qui elle-même constitue avec d'autres galaxies une structure d'amas etc. Et à chaque échelle, les pages se regroupent autour d'une page d'autorité, de la même manière que les corps peu massifs gravitent autour des corps qui le sont plus.

Les sommets sont initialement distribués aléatoirement dans un espace à 3D, et ils se voient attribuer une masse correspondant à leur Pagerank. Sans entrer dans les détails de l'évaluation du Pagerank, on peut dire que cette méthode introduite par Google consiste à estimer l'importance d'une page, en s'appuyant sur le critère suivant :

une page est d'autant plus importante qu'un grand nombre de pages pointent (en termes de liens hypertextes) vers cette page, et que ces pages sont elles-mêmes importantes. Pour plus de détails sur le Pagerank et Google, on peut se reporter à [11] ou à [12].

S'il existe un lien entre deux noeuds, ils sont attirés l'un par l'autre sous l'effet d'une force à laquelle on donne la même forme que la force newtonnienne de gravitation. Au bout d'un pas de temps, on obtient donc des noeuds répartis de manière non-homogène dans l'espace, ce qui permet de réaliser un premier découpage communautaire. On procède ensuite à une agglomération du même type que celle de Walktrap : on réunit les communautés de manière à augmenter la modularité.

Ensuite, on procède à un transfert de masse : suivant le Pagerank, que l'on modifie en fonction de la proximité des noeuds dans l'espace 3D, chaque noeud se voit attribuer une nouvelle masse, c'est-à-dire, dans le cas de pages web, une nouvelle valeur d'autorité. Puis, on laisse à nouveau évoluer le système gravitationnel durant un pas de temps. On itère le procédé jusqu'à ce qu'il ne soit plus possible d'augmenter la modularité du système au cours de la phase d'agglomération.

2.5 Notion de périphérie

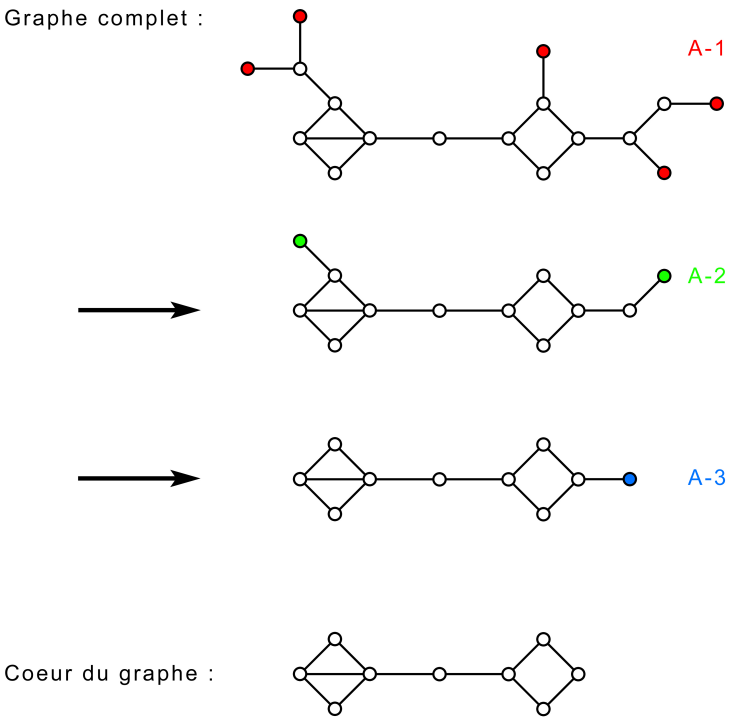
2.5.1 Définition

On partitionne les noeuds du graphe en deux sous-ensembles, que l'on appellera *coeur* et *périphérie*. Par extension, on utilisera aussi ces termes pour désigner les sous-graphes induits. On peut décrire la périphérie d'un ensemble connexe de noeuds comme la partie arborescente externe de cet ensemble. Pour en donner une définition plus formelle, on va utiliser le concept de *core*, introduit par Seidman en 1983 [13], qu'on trouve plus souvent dans la littérature.

Soit $G = (N, L)$ un graphe; où N est l'ensemble des noeuds, L est l'ensemble des liens. Le sous graphe $G_k = (N_k, L_k)$ est le k-core de G si pour tout $n \in N_k$ le degré de n dans G_k est supérieur ou égal à k, et G_k est le sous-graphe maximum ayant cette propriété.

Avec cette définition, le coeur de la communauté coïncide avec le 2-core du graphe, soit G_2 avec les notations ci-dessus.

Pour déterminer la périphérie, on procède de la façon représentée sur les figures suivantes :



On élimine d'abord les noeuds de degré un (on note ici l'ensemble des sommets supprimés alors A-1). Le nouvel ensemble de sommets et de liens comporte de nouveaux noeuds de degré 1, on itère le procédé : on supprime ces

noeuds de connectivité 1 (ensemble A-2) et ainsi de suite (A-3, A-4, ...) jusqu'à ce qu'il n'y ait plus de noeuds de degré 1 dans l'ensemble restant, i.e. le coeur du graphe.

2.5.2 Périphérie de communautés

De la même manière que l'on définit la périphérie d'un graphe, on peut définir la périphérie de tout sous-graphe, et en particulier de ses communautés. Par la suite, on va principalement concentrer notre attention sur l'arborescence des communautés.

Pour supprimer toute ambiguïté sur les périphéries dont il est question, on notera celle du graphe $A(G)$, et celles des communautés $A(C)$ (et si l'on doit préciser le numéro k de la communauté : $A(C_k)$).

2.5.3 Liaisons entre communautés

Par définition, les liens entre communautés sont plus rares que les liens internes, le nombre de liens pouvant servir de jonctions intercommunautaires est donc faible.

Pour un réseau dont on a fait une partition communautaire, les noeuds jouant une fonction d'intermédiaire vers d'autres communautés méritent d'être examinés avec attention. Par exemple, on peut penser que l'information qui transite par un tel noeud d'un ensemble vers un autre a une signification particulière. C'est une idée que de nombreux sociologues mettent en avant dans le cas des réseaux sociaux. Selon les auteurs, le phénomène est qualifié de différentes façons, certains parlent par exemple de "trous structuraux" [14].

Pour illustrer cette notion, on peut prendre le cas de la recherche d'expertise dans un réseau social. Ici, le graphe est constitué d'individus (les sommets), un lien représentant l'existence d'une communication par mail entre deux agents. La recherche d'une expertise est supposée se faire par le transit d'un mail d'un individu à un autre en ne passant que par les liens déjà existant (c'est-à-dire qu'on n'envoie la demande d'expertise qu'aux agents avec qui on a déjà eu l'occasion de communiquer par mail). Ce mail voyage alors d'un noeud à l'autre jusqu'à trouver quelqu'un qui soit susceptible de répondre à la demande.

Si le mail doit sortir de la communauté du demandeur, il va emprunter un des rares liens de la communauté du demandeur vers l'extérieur. Par conséquent même si le flux d'information circulant par ces liens n'est pas important en quantité, il est important pour le fonctionnement du réseau, ici il est nécessaire pour trouver l'expert (sur l'exemple précis des recherches d'expertise dans une entreprise : [15]).

Ce problème suscite particulièrement l'intérêt des sociologues. On peut comprendre pourquoi sur un point au moins : dans un réseau où l'intérêt individuel entre en ligne de compte, le noeud qui sert de pont vers l'extérieur est en position de force, puisqu'il a la possibilité d'influencer le flux d'information de sa communauté vers une autre (c.f. [16] ou [17]).

Ce sont ces observations qui ont amené à concentrer notre attention sur les périphéries de communautés. En effet, un noeud de l'arborescence de la communauté peut être considéré comme "en marge" de celle-ci. On peut penser que le fait de se situer en périphérie indiquerait que le noeud fait partie de la "zone frontière" entre deux communautés, et il serait alors possible qu'il existe une certaine corrélation entre le rôle d'intermédiaire entre communautés et l'appartenance ou non d'un noeud à l'ensemble $A(C_k)$.

La première partie de ce travail cherche à examiner cette hypothèse.

3 Etude statistique du graphe

3.1 Introduction

Le but de cette partie du projet est de réaliser un traitement systématique des données permettant
- de réaliser des découpages communautaires à l'aide des programmes de partition Walktrap et Cosmoweb, ou des découpages aléatoires du graphe.
- de déterminer la périphérie de chaque communauté ainsi détectée.
- puis d'en déduire des statistiques permettant de valider ou non l'hypothèse formulée précédemment.

On ne développera pas les problèmes de complexité des outils utilisés. Précisons simplement que le traitement des données est fait à l'aide de scripts bash, sed et awk, en raison de leur bonne vitesse d'exécution et de leur simplicité d'écriture. Par ailleurs, on doit éviter autant que possible d'insérer une boucle dans une autre, car si l'on manipule n noeuds, la complexité de ce type de structure est en $O(n^2)$; or les scripts ne sont avantageux en termes de temps de calcul par rapport aux langages de programmation traditionnels que pour des complexités inférieures (en $O(n)$ par exemple).

Avant de discuter de l'analyse statistique des graphes, on va évoquer différents éléments qui me semblent importants pour comprendre la démarche suivie.

3.2 Problème de la collecte des données

Pour que les statistiques aient une signification, on doit manipuler des bases des données suffisamment vastes. Leur collecte se fait alors souvent de manière automatique (notamment pour celles extraites du web), ou par des personnes distinctes. On va voir que ceci est à l'origine de nombreux biais, souvent difficiles à contrôler (que l'on regroupera sous le terme de *bruit*). Il faut essayer de l'éliminer ou au moins de l'avoir à l'esprit pour pouvoir interpréter correctement les résultats.

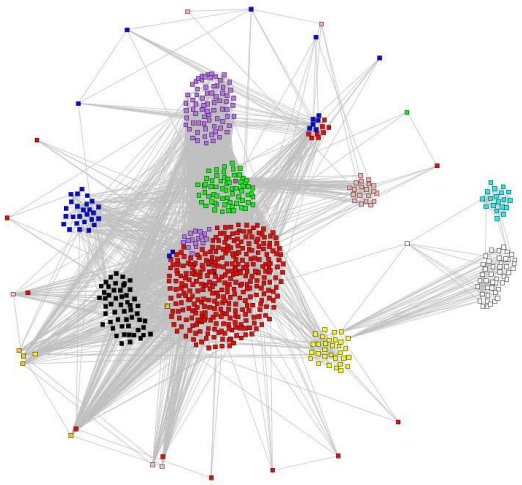
On peut citer, pour illustrer le type de problèmes rencontrés :

- Les noeuds non-significatifs : prenons les graphes résultant du réseau Wikipedia dont les pages (noeuds) sont reliées si elles ont un contributeur commun. Wikipedia est équipé de robots : ce sont des programmes se déplaçant sur le réseau, dont le rôle est d'effectuer de petites tâches systématiques (corriger l'orthographe etc).
  Lors de la collecte des données, le robot n'est pas distingué d'un utilisateur usuel, mais son comportement est très différent : par exemple, s'il modifie un grand nombre de pages rapidement, il diminue la longueur moyenne du chemin d'un noeud à un autre. Il doit donc être éliminé de la base, au moins si l'on cherche à analyser le comportement d'un rédacteur Wikipedia "conventionnel".
- Dans ce premier cas, le biais est clairement identifié, mais dans d'autres situations, il est bien plus difficile d'éliminer le bruit car celui-ci tient à la nature même de l'information. On peut mettre en évidence ce type de difficultés sur la base de données "Tra" évoquée précédemment. On souhaite en extraire des informations sur l'hérédité professionnelle au XIX^{eme} siècle.
  Mais il se pose des problèmes sur lesquels il est d'autant plus difficile de trancher qu'ils relèvent d'avantage des sciences de l'homme que de l'informatique. Par exemple, comment distinguer deux mots identiques mais qui peuvent être associés, selon l'environnement (milieu social, rural ou urbain...) à des réalités tout à fait différentes, comme la profession "domestique" ?

  On sera amené à considérer à nouveau ces problèmes auxquels on est continuellement confronté lorsque l'on travaille sur des données de réseaux sociaux réels.

3.3 Conditions intiales de Cosmoweb, définition du degré d'analogie

Comme Cosmoweb utilise des conditions initiales aléatoires, la partition obtenue change à chaque réalisation de l'algorithme. Le graphe ci-dessous est produit à l'aide du logiciel Guess [4](comme d'ailleurs tous les graphes de ce rapport).



[4] graphexploration.cond.org

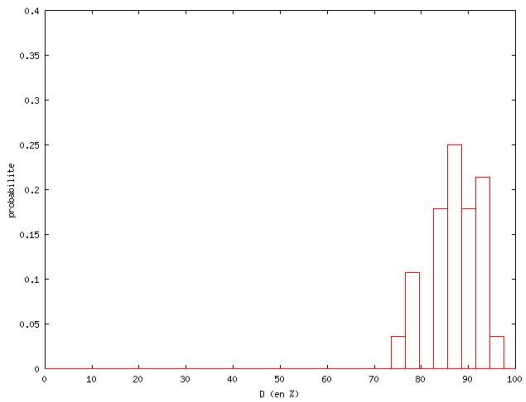
Il s'agit d'un graphe de blogs (base Dupin). On y a regroupé les sommets d'une même communauté obtenue à l'aide d'une première partition Cosmoweb, et les noeuds d'une même couleur appartiennent à une même communauté selon une seconde partition Cosmoweb.

Ce type de visualisation permet d'avoir une première impression de l'incidence du caractère aléatoire des conditions initiales sur le découpage communautaire. On constate que la différence se fait pour une large part sur des noeuds qui sont rattachés à des communautés importantes selon un découpage, et isolés en communautés singletons dans l'autre partition.

En conséquence, le nombre de communautés produit par l'un et l'autre découpages n'est pas nécessairement un bon indicateur de leur concordance. Sur cet exemple précis, on recense 12 communautés dans la partition associée aux couleurs, 29 pour celle associée à la position, et pourtant, il y a manifestement une correspondance entre l'une et l'autre.

On cherche donc un critère plus approprié pour estimer le degré de similitude entre deux découpages communautaires. On a retenu le suivant : on dénombre les liens pour lesquels les deux partitions communautaires s'accordent pour désigner le lien comme intercommunautaire ou intracommunautaire. Pour reprendre l'exemple ci-dessus, le graphe contient 3192 liens, les deux partitions s'accordent pour 2879 d'entre eux, on dira donc que les découpages sont analogues à (environ) 90.2% [5].

On voudrait évaluer la stabilité de ce degré d'analogie, que l'on notera D (dans l'exemple précédent, $D \simeq 0.902$) . On compare alors plusieurs couples de partitions Cosmoweb (toujours sur le même graphe), les résultats obtenus sont indiqués dans le diagramme ci-dessous : on trace en fonction de la valeur de D, le nombre normalisé d'occurrences obtenues.



Sur ce graphe, on a donc en moyenne une correspondance de l'ordre de 87 % entre deux partitions Cosmoweb. Cependant, cette caractéristique fluctue avec la topologie du graphe, comme on pourra le voir dans l'annexe B.

3.4 Comparaison des algorithmes de détection communautaire

On compare ici des résultats produits par Walktrap et Cosmoweb. Cela nous permettra de voir sur quels points la partition dépend de la méthode utilisée, et d'approfondir des critères de comparaison qui nous seront utiles par la suite.

3.4.1 Remarque préliminaire

Walktrap sélectionne systématiquement le plus grand ensemble connexe du graphe pour en faire la partition communautaire. On restreint donc notre étude à cette partie du graphe. Notons que pour les graphes du web (Wikipedia, blogs...), la collecte de données se faisant le plus souvent de proche en proche, le plus grand connexe s'identifie souvent à l'ensemble du graphe. En revanche, ce n'est pas nécessairement vrai pour tous (comme ceux construits sur Tra, cités précédemment).

3.4.2 Degré d'analogie

La représentation ci-dessous est réalisée selon le même principe que ce qui a été fait précédemment pour Cosmoweb : on a regroupé dans le plan les noeuds selon leur communauté Walktrap, on associe à chaque couleur

[5] Les résultats des calculs effectués sur les graphes seront toujours donnés à trois chiffres significatifs, car cela semble être la précision suffisante pour ne pas masquer les effets qu'on mettra en évidence par la suite.

une communauté Cosmoweb.

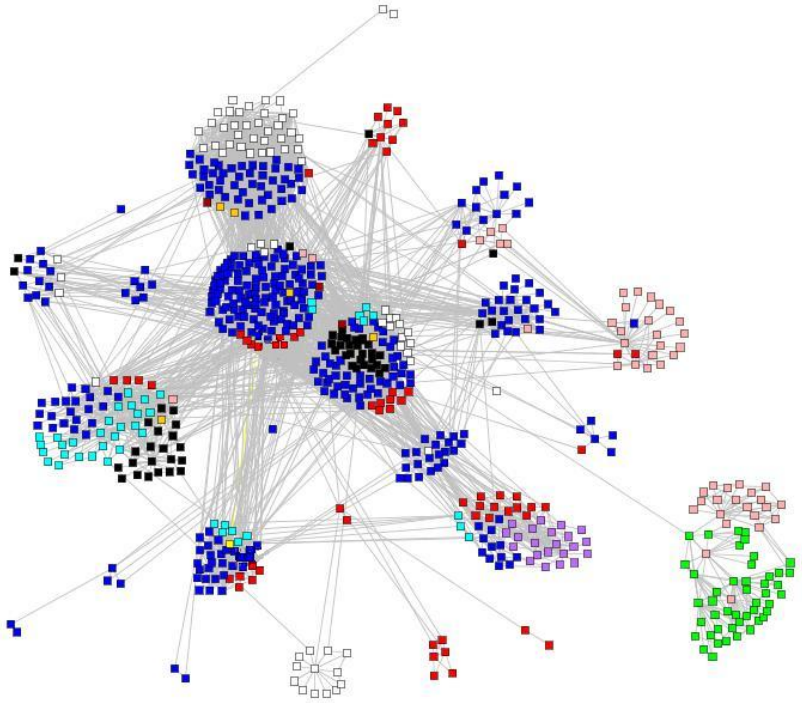
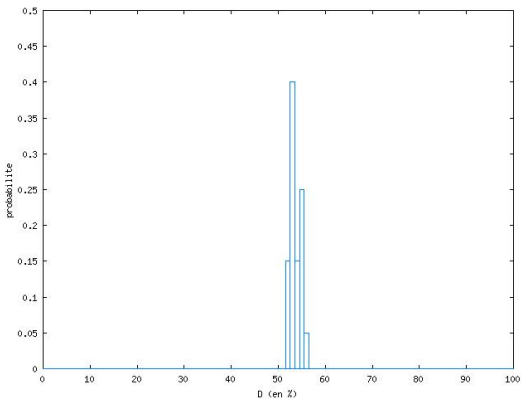


FIG. 1 – Comparaison des partitions Walktrap et Cosmoweb sur le graphe des blogs Dupin.

Le graphe représenté ici est construit depuis les données de blogs Dupin. L'impression visuelle laissée par ce graphe est que les partitions obtenues sont assez différentes.

Pour examiner plus objectivement le recouvrement des communautés, on utilise le degré D d'analogie que l'on a défini dans la partie 3.3. On dénombre sur ce graphe 3192 liens, Cosmoweb en donne 1717 intracommunautaires, Walktrap : 2411 ; 1714 sont qualifiés identiquement par les deux programmes soit 53.7% (ou D=0.537).

On peut constater que D est stable avec les conditions initiales de Cosmoweb. Pour illustrer ceci, on trace comme précédemment le nombre d'occurrences normalisé en fonction de D, mais cette fois, on mesure l'accord entre la partition Walktrap et diverses partitions Cosmoweb du graphe des blogs Dupin.



A quoi peut-on comparer la valeur de D pour mieux comprendre sa signification ?

On connait pour chacune des deux partitions le nombre de liens inter- et intracommunautaires. On les note respectivement $l_{\rightarrow}$ et $l_{\leftarrow}$; de plus, on emploie l'indice C pour la partition Cosmoweb, l'indice W pour la Walktrap.

Les partitions communautaires tendent à minimiser le nombre de liens intercommunautaires par rapport aux liens intracommunautaires. On sait alors que D ne peut pas être plus grand que $\frac{l-|l_{\rightarrow,W}-l_{\rightarrow,C}|}{l}$, et pas plus petit que $\frac{l-(l_{\rightarrow,W}+l_{\rightarrow,C})}{l}$, où l est le nombre total de liens du graphe. Sur notre exemple, $D \in [0.313; 0.802]$, mais cet encadrement est trop large pour donner une indication nette.

Avec les $l_{\rightarrow}$, $l_{\leftarrow}$ et l, on peut calculer la grandeur :

$$\overline{D} = \frac{l_{\rightarrow,W}.l_{\rightarrow,C}}{l^2} + \frac{l_{\leftarrow,W}.l_{\leftarrow,C}}{l^2}$$

que l'on qualifiera de *degré estimé d'analogie*. On peut l'interprêter comme la valeur de D si les méthodes de partition communautaire sont totalement décorrélées. En effet, si tel est le cas, chacune des deux méthodes cherche à minimiser les liens intercommunautaires, indépendemment l'une de l'autre. Alors, une fois le nombre de liens intercommunautaires de chaque partition déterminé, on s'attend à ce que la probabilité pour chaque lien d'être désigné identiquement par les deux partitions soit donnée par :

$$P(\rightarrow, W).P(\rightarrow, C) + P(\leftarrow, W).P(\leftarrow, C)$$

où $P(\rightarrow, W)$ désigne la probabilité pour le lien d'être intercommunautaire selon la partition Walktrap.

Sur le graphe évoqué précédemment, $\overline{D} = 0.519$. On constate systématiquement un faible écart tel que $D > \overline{D}$: on a reproduit ci-dessous les deux valeurs pour plusieurs conditions initiales différentes Cosmoweb. Pour rendre compte plus précisément de la différence entre D et $\overline{D}$, on utilise $(D\text{-}\overline{D})/D$, qui est donc ici systématiquement positif (quoique faible).

	D	$\overline{D}$	$\frac{D-\overline{D}}{D}$ (%)				
Essai 1	0.537	0.519	+ 3.4	Essai 5	0.555	0.527	+ 5.0
Essai 2	0.559	0.531	+ 5.0	Essai 6	0.568	0.535	+ 5.8
Essai 3	0.527	0.502	+ 4.7	Essai 7	0.555	0.530	+ 4.5
Essai 4	0.547	0.525	+ 3.8	Essai 8	0.536	0.507	+ 5.4
				Essai 9	0.537	0.514	+ 4.3

Pour vérifier la validité de ce qui précède, on utilise le programme Randcom, dont le principe est le suivant : on lui fournit la liste des tailles de communautés, et il attribue à chaque sommet du graphe une communauté de manière aléatoire, avec pour seule contrainte que le découpage réalisé soit constitué de communautés ayant les tailles données dans la liste. De cette manière, si le sens que l'on a associé à $\overline{D}$ est exact, on doit observer la même valeur pour D et $\overline{D}$ entre un découpage Randcom et un découpage Cosmoweb (ou entre Randcom et Walktrap). Donc, $(D\text{-}\overline{D})/D$ doit être proche de zéro (inférieur en module aux valeurs reportées dans le tableau précédent).

C'est effectivement ce que l'on constate [6] :

Essais Randcom/Walktrap	D	$\overline{D}$	$\frac{D-\overline{D}}{D}$ (%)	Essais Randcom/Cosmoweb	D	$\overline{D}$	$\frac{D-\overline{D}}{D}$ (%)
Essai 1	0.303	0.298	+ 1.7	Essai 1	0.481	0.491	- 2.1
Essai 2	0.293	0.297	- 1.4	Essai 2	0.483	0.485	- 0.4
Essai 3	0.290	0.292	- 0.7	Essai 3	0.466	0.484	- 3.9

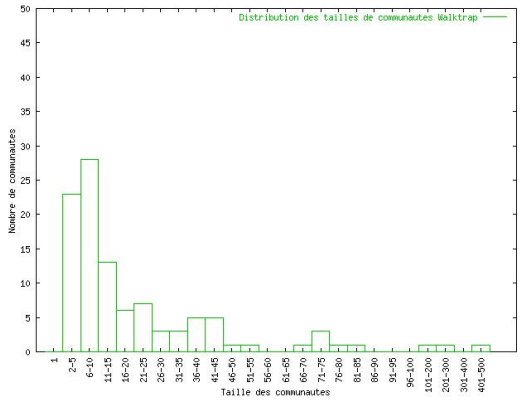
On peut toutefois se demander dans quelle mesure ces observations dépendent du type de graphe sur lequel a été réalisée l'étude. On donne dans l'annexe C des éléments montrant que l'on peut faire le même type d'observations sur d'autres graphes, mais l'écart entre D et $\overline{D}$ est plus ou moins significatif.

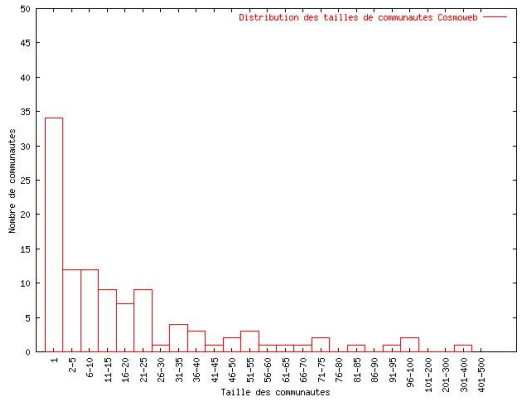
Cette observation indique qu'il existe une corrélation entre les découpages Walktrap et Cosmoweb : non seulement l'un et l'autre programmes cherchent à minimiser le nombre de liens externes, mais les deux méthodes ont tendance à désigner les mêmes liens du graphes comme externes. Néanmoins cette tendance reste faible et fluctue selon le graphe examiné.

3.4.3 Tailles des communautés

L'examen des tailles de communautés nous indique un point sur lequel les découpages sont clairement distincts : les distributions des tailles communautaires n'ont pas la même allure. On a pris ici l'exemple de cette distribution pour le graphe de blogs Yacine, en regroupant les communautés par catégories de taille.

[6] Sur les tests effectués, les tailles des "communautés" Randcom sont alignées sur celles des communautés détectées par le programme avec lequel on compare.





On met à part les communautés contenant un unique noeud, car celles-ci sont souvent nombreuses avec Cosmoweb (en fait leur nombre varie beaucoup selon les conditions initiales) et rares avec Walktrap (de l'ordre de 1% du nombre total de communautés, à l'exception du graphe Tra, qui présente un nombre de singletons très élevé). C'est d'ailleurs essentiellement sur les communautés de petites tailles que la différence est notable. Ceci n'est pas particulier au graphe Yacine mais très général, on a mis en annexe D un autre exemple de ce type.

3.4.4 Conclusion

En fait, cette analyse met surtout en évidence un problème qui est inhérent aux partitions de graphes : il n'y a pas de manière idéale de réaliser un découpage communautaire. L'usage de la modularité peut donner l'illusion qu'on a un critère pour hiérarchiser deux partitions produites par deux algorithmes différents. Cependant, bien que cela donne une indication, on ne peut pas trancher de manière définitive.

D'abord, on peut trouver des modularités comparables pour l'un et l'autre programme (on trouvera les valeurs de Q pour différents graphes dans le tableau de la partie 3.5.1), alors que les découpages communautaires sont nettement distincts.

Par ailleurs, l'efficacité du découpage est une question qui contient une part d'arbitraire. Pour reprendre le problème de la distribution des tailles de communautés, on a vu que Cosmoweb peut générer beaucoup de singletons, au contraire de Walktrap. L'une et l'autre approches ont leur justification : si un sommet est connecté à un grand nombre de communautés, il semble acceptable de le considérer comme un élément à part et une communauté à lui-seul, mais il est tout aussi recevable de le rapprocher de la communauté avec laquelle il serait le plus lié.

3.5 Corrélation entre la périphérie et le caractère intercommunautaire

3.5.1 Première approche

Pour savoir s'il existe effectivement une corrélation, on n'utilise que des graphes non-orientés et non-pondérés, puis on procède ainsi :

- On compte, pour la totalité de chaque graphe, le nombre de liens intercommunautaires $l_{\rightarrow}$; parallèlement, on recense aussi le nombre d'arêtes liant deux noeuds appartenant à des coeurs de communautés, distinctes ou non (on les notera liens C-C, et leur nombre l_{c-c}), le nombre de liens d'un noeud de coeur à un noeud arborescent (C-A), enfin le nombre de liens entre noeuds arborescents (A-A).
- On fait alors une estimation du nombre de liens C-C, C-A et A-A intercommunautaires que l'on penserait obtenir s'il n'y avait aucune corrélation entre le caractère intercommunautaires des liens et la propriété d'un noeud d'être dans l'arborescence d'une communauté. Ainsi, on attend un nombre de liens C-C intercommunautaires $\overline{l_{c-c,\rightarrow}}$ donné par :

$$\overline{l_{c-c,\rightarrow}} = l_{c-c}.\frac{l_{\rightarrow}}{l}$$

- On compare ces valeurs aux valeurs effectivement observées sur le graphe : $l_{c-c,\rightarrow}$, $l_{c-a,\rightarrow}$ et $l_{a-a,\rightarrow}$. On introduit alors un paramètre pour estimer la corrélation, défini par :

$$\alpha = \frac{\overline{l_{c-c,\rightarrow}} - l_{c-c,\rightarrow}}{\overline{l_{c-c,\rightarrow}}}$$

Ainsi, α devrait être nul s'il n'y a pas de corrélation. On pourrait choisir d'estimer la corrélation avec les liens C-A ou A-A, mais les liens C-C sont souvent les plus nombreux, ce qui permet de minimiser les fluctuations statistiques.

Avant de mesurer α pour des partitions communautaires, on teste la fiabilité de la méthode proposée à l'aide de Randcom : comme dans ce cas le caractère intercommunautaire et la périphérie sont totalement décorrélés, on doit mesurer un α proche de 0. Sur 20 essais menés sur le graphe Dupin, on observe effectivement des valeurs groupées autour de 0 (entre -0.022 et +0.010), on trouvera la distribution en annexe E.

D'autre part, il nous faut préciser la convention choisie pour les communautés singletons : le sommet est considéré comme un noeud de coeur. Cela n'a pas beaucoup d'importance pour Walktrap qui génère peu de communautés singletons. En revanche, pour Cosmoweb ce nombre peut varier assez largement d'un essai à un autre, et dans le cas où il est élevé, cela pèse statistiquement pour un α_C faible, car les liens connectant ces singletons au reste du graphe sont intercommunautaires et lient un sommet de type C à un sommet indéterminé. C'est aussi cette convention qui explique que le programme Randcom a une légère tendance à évaluer α négatif (c.f. annexe E).

On voudrait donc estimer les fluctuations de α_C causées par les conditions initiales de Cosmoweb, ce que l'on fait ici sur deux exemples : Dupin et FSE biparti. On représente ci-dessous la distribution obtenue pour le graphe Dupin ; en revanche cela n'est pas nécessaire pour FSE, où les résultats sont très groupés : sur 25 essais, tous sont compris dans l'intervalle [-0.235 ;-0.226].

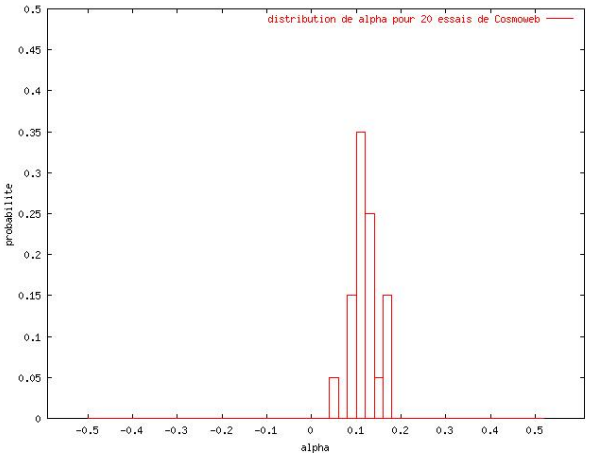


FIG. 2 – Mesure de α_C pour 20 essais de Cosmoweb sur le graphe Dupin

L'estimation de la largeur à mi-hauteur de ces courbes donne une idée de l'étalement caractéristique des valeurs de α_C : environ 0.080 pour Dupin, 0.005 pour FSE biparti, cette évaluation est très approximative étant donné l'allure des distributions. On voit que selon le graphe considéré, l'étalement varie d'un facteur qui peut être supérieur à 10.

On peut maintenant donner la valeur de α (selon les deux découpages communautaires) et quelques autres grandeurs caractéristiques (modularité Q etc) pour les différents graphes :

Graphe	Nombre de noeuds dans le graphe	Nombre de liens dans le graphe	Noeuds dans A(G) (pourcentage du total)	Q_C	Q_W	α_C	α_W
Dupin	724	3192	99 (13.7 %)	0.497	0.603	+ 0.137	+ 0.044
Scorpione	604	1410	160 (26.5 %)	0.561	0.742	+ 0.211	+ 0.000
Yacine	2702	5477	839 (31.1 %)	0.763	0.836	+ 0.201	+ 0.135
Full	2656	3327	1796 (67.6 %)	0.613	0.350	- 0.622	+ 0.091
Tra	1226	4821	587 (47.9 %)	0.230	0.128	+ 0.044	+ 0.042
FSE biparti	3131	5178	1769 (56.5 %)	0.555	0.596	- 0.226	- 0.226
FSE simple	2262	21630	38 (1.7 %)	0.401	0.542	+ 0.114	+ 0.004
Wikipedia	18067	43668	14000 (77.5 %)	0.179	0.307	- 0.127	- 0.231
Acteurs	1298	11068	24 (1.8 %)	0.755	0.737	+ 0.141	+ 0.005
Admin	691	1184	307 (44.9 %)	0.467	0.550	- 0.304	- 0.048

En ce qui concerne α, on ne peut pas déceler de tendance particulière, de plus, α_C et α_W sont souvent très différents. La dépendance de α avec les autres paramètres du graphe n'est pas simple; l'inverse serait d'ailleurs surprenant, car le coefficient α cherche à rendre compte d'une propriété complexe du graphe, et on ne s'attend donc pas à ce que toutes les dépendances de α soient contenues dans un unique paramètre. Cependant, on peut voir sur le tableau qu'en général, lorsque A(G) représente une grande proportion des noeuds du graphe (Full, Wikipedia ...), α est faible voire négatif.

Cela nous amène à penser que l'ensemble de noeuds A(G) joue un rôle tout à fait particulier. Comme $A(G) \subseteq \bigcup A(C_k)$, tous les liens de A(G) sont de type A-A par construction du graphe, et non en raison du partage communautaire. Par conséquent, ils vont peser statistiquement en faveur de liens A-A intracommunautaires, car les communautés qui sont partiellement ou totalement dans A(G), comme n'importe quelle communauté, ont plus de liens internes que de liens externes.

3.5.2 Statistique sur le coeur du graphe

On va donc réaliser la même étude sur le coeur du graphe : on supprime de la base de données les noeuds et les liens qui ne sont pas dans C(G), on effectue un nouveau découpage communautaire, puis on calcule la valeur de α. On donne également pour chacun des coeurs de graphe les valeurs de D et $\overline{D}$, (entre les deux partitions), ainsi que d'autres paramètres caractéristiques de C(G) :

Coeur de Graphe	Nombre de noeuds	Nombre de liens	Q_C	Q_W	α_C	α_W	D	$\overline{D}$	$\frac{D-\overline{D}}{D}$ (%)
Dupin	625	3093	0.480	0.592	+ 0.154	+ 0.069	0.512	0.503	+ 1.8 %
Scorpione	444	1250	0.563	0.716	+ 0.295	+ 0.108	0.760	0.718	+ 5.5 %
Yacine	1863	4638	0.734	0.839	+ 0.306	+ 0.323	0.807	0.739	+ 8.4 %
Full	860	1531	0.468	0.650	+ 0.240	+ 0.410	0.495	0.499	- 0.8 %
Tra	639	4234	0.095	0.151	+ 0.226	+ 0.130	0.538	0.560	- 4.1 %
FSE biparti	1362	3409	0.411	0.547	+ 0.305	+ 0.174	0.463	0.482	- 4.1 %
FSE simple	2224	21592	0.267	0.546	+ 0.064	+ 0.007	0.677	0.583	+ 13.3 %
Wikipedia	4067	29668	0.094	0.118	+ 0.108	+ 0.002	0.623	0.664	- 6.6 %
Acteurs	1274	11044	0.722	0.773	+ 0.044	+ 0.007	0.813	0.795	+ 2.2 %
Admin	384	877	0.317	0.500	+ 0.094	+ 0.175	0.423	0.477	- 12.8 %

La valeur de $(D\text{-}\overline{D})/D$ est moins élevée pour les coeurs de graphe ci-dessus que pour les graphes complets (se reporter à l'annexe C), et ceci est particulièrement visible pour les graphes ayant une périphérie importante (comme Full, FSE biparti ou encore Admin). On peut justifier cette observation en supposant que les partitions Walktrap et Cosmoweb s'accordent mieux pour A(G) que pour G complet : dans la périphérie du graphe, les deux programmes ont une probabilité plus élevée de donner à un lien la même qualification : intra- ou intercommunautaire.

Maintenant, en ce qui concerne α lui-même, tous les résultats indiquent la même tendance, que cela soit avec Cosmoweb ou avec Walktrap (même si la corrélation est presque toujours plus nette sur le découpage Cosmoweb) : $\alpha \geq 0$, et ce pour tous les graphes sur lesquels on a effectué ce travail. Cela n'aurait sans doute pas beaucoup de sens de comparer les valeurs de α d'un graphe à l'autre, tant les contenus des graphes examinés (et les topologies associées) sont différents. Aucun paramètre mesuré ne semble étroitement lié à la valeur de α.

Cette observation confirme qu'il existe effectivement une corrélation entre le caractère périphérique d'un noeud et la tendance à établir des liens vers les autres communautés. Considérant que la valeur maximum théorique de α

est 1, cette corrélation semble faible pour la plupart des graphes examinés ; toutefois on peut souvent majorer α de manière bien plus précise : prenant l'exemple du coeur du graphe Acteurs pour la partition Walktrap, on dénombre seulement 19 liens (sur 11044) qui ne soient pas de type C-C ; comme $l_{c-c,\rightarrow,W} = 1309$, $\alpha_W \leq \frac{19}{1309} \simeq 0.015$. Ce cas est particulièrement net (d'ailleurs, on peut se demander si le nombre de liens C-A et A-A n'est pas ici trop faible pour qu'on puisse accorder une signification au signe de α_W), on indique les bornes supérieures de α en annexe F pour les différents coeurs de graphe. Pour avoir une idée plus précise de l'amplitude de la corrélation, c'est à cette valeur maximum qu'il faut comparer α.

3.5.3 Résistance au bruit

Comme on l'a dit dans la partie 3.2, les bases sont bruitées en raison de problèmes tenant essentiellement à la collecte des données. On veut s'assurer que les grandeurs évaluées sur le graphe sont peu perturbées par les différents types de bruit.

Pour tester la stabilité des différents paramètres, on va utiliser deux protocoles correspondant à deux formes de bruit auxquelles on a été confronté sur les bases de données utilisées.

Duplication de noeuds Lors de la collecte des données des forums sociaux, les noms des organisations apparaissent sous plusieurs formes distinctes. Par exemple, on trouve "I.G.M." ou "IG Metall" pour désigner le même groupe. Autant que possible, on a cherché à homogénéiser les noms, éliminer les différences de typographie, les coquilles etc... Mais ce travail est long et fastidieux car on ne peut pas le faire simplement à l'aide de scripts.

On va chercher à reproduire l'effet que peut avoir ce type de bruit. On prend une fraction f des noeuds de la base (on la suppose dépourvue de bruit), et pour ces sommets, on propose deux dénominations différentes : l'ancienne (notée A), et la nouvelle (B). Ensuite, à chaque occurrence de A dans la liste non bruitée des liens, on a une de probabilité 0.5 de modifier A en B.

Les mesures ne sont effectuées qu'avec le programme Walktrap, pour éviter d'ajouter aux fluctuations dues au bruit les fluctuations associées aux conditions initiales de Cosmoweb. Comme on a observé ce type de bruit sur la base FSE, on effectue les tests sur le graphe du coeur de FSE (biparti). On mesure également le degré d'analogie entre le graphe bruité et la référence (le graphe non bruité) ; ainsi, lorsque D et $\overline{D}$ sont du même ordre, on peut considérer que les découpages communautaires sont décorrélés, c'est-à-dire que pour Walktrap, tout se passe comme si le graphe bruité et le graphe de référence n'avaient plus rien de commun.

Les résultats de ce protocole sont reportés dans le graphique suivant :

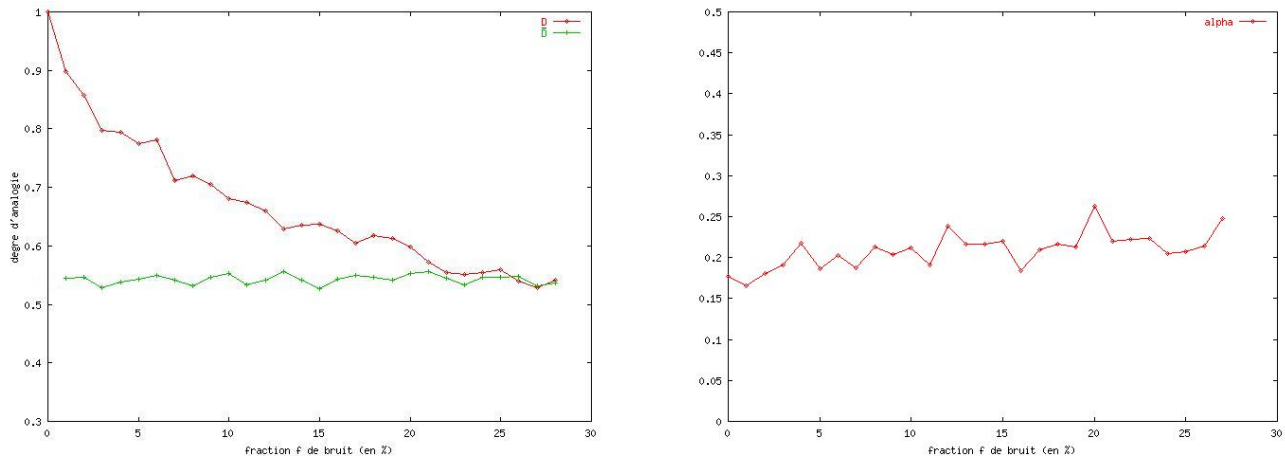


FIG. 3 – A gauche : D et $\overline{D}$ pour différentes valeurs de f ; à droite : α pour différentes valeurs de f.

Lorsque D et $\overline{D}$ deviennent comparables (pour $f \simeq 0.23$), on peut estimer que le découpage du graphe bruité et du graphe de références sont décorrélés. Autrement dit, on perd l'essentiel de l'information que produit le découpage communautaire Walktrap sur le graphe non-bruité.

On voit par ailleurs que la pente de la courbe D est d'abord élevée en valeur absolue et décroît jusqu'à être quasi-nulle, évoquant l'allure d'une exponentielle décroissante. Par conséquent, un faible bruit de ce type suffit à détériorer considérablement l'information du découpage Walktrap réalisé.

En ce qui concerne α, on peut constater qu'il fluctue autour d'une valeur moyenne qui croît lentement avec f. On peut observer que le nombre de communautés augmente lui aussi lentement avec f (résultats mis en annexe G) ; or, à nombre de liens égal, l'augmentation du nombre de commnautés entraîne une baisse du nombre moyen de liens intracommunautaires, et donc une augmentation du nombre de noeuds dans $\bigcup A(C_k)$ (car plus le rapport (nombre de noeuds)/(nombre de liens) d'une communauté est élevé, plus la périphérie devrait être importante), il n'est donc pas étonnant de constater une telle augmentation.

Noeud omniprésent Lors de la collecte des données de blogs (type Scorpione), certains liens ont été établis entre le blog visité et des pages qui ne correspondait pas à d'autres blogs, mais à divers sites, comme par exemple le portail de skyblog [7]. Dans la base de données, ces pages sont considérées comme des sommets usuels, mais leur degré est souvent très supérieur à celui des autres noeuds.

On va donc essayer de tester l'impact de noeuds non significatifs de forte connectivité. Pour ce faire, on introduit un noeud supplémentaire dans la base de données que l'on connecte à une fraction g des sommets du graphe, sélectionnés aléatoirement. On utilisera le programme Walktrap uniquement, sur le coeur du graphe Scorpione.

Comme le nombre de liens varie entre le graphe de référence et les graphes bruités, il faut ici préciser les valeurs de D et $\overline{D}$ mesurées. Les liens présents dans le graphe bruité qui n'existaient pas dans la référence seront considérés comme différents lorsque l'on évalue D. En notant l_b le nombre de liens associé au graphe bruité et l_{ref} celui associé au graphe de référence, la valeur estimée $\overline{D}$ sera alors donnée par :

$$\overline{D} = \frac{l_{ref}}{l_b}.(\frac{l_{b,\rightarrow}.l_{ref,\rightarrow}}{l_b.l_{ref}} + \frac{l_{b,\leftarrow}.l_{ref,\leftarrow}}{l_b.l_{ref}})$$

Chacune des grandeurs ci-dessus faisant référence au découpage communautaire Walktrap. On obtient les résultats suivants :

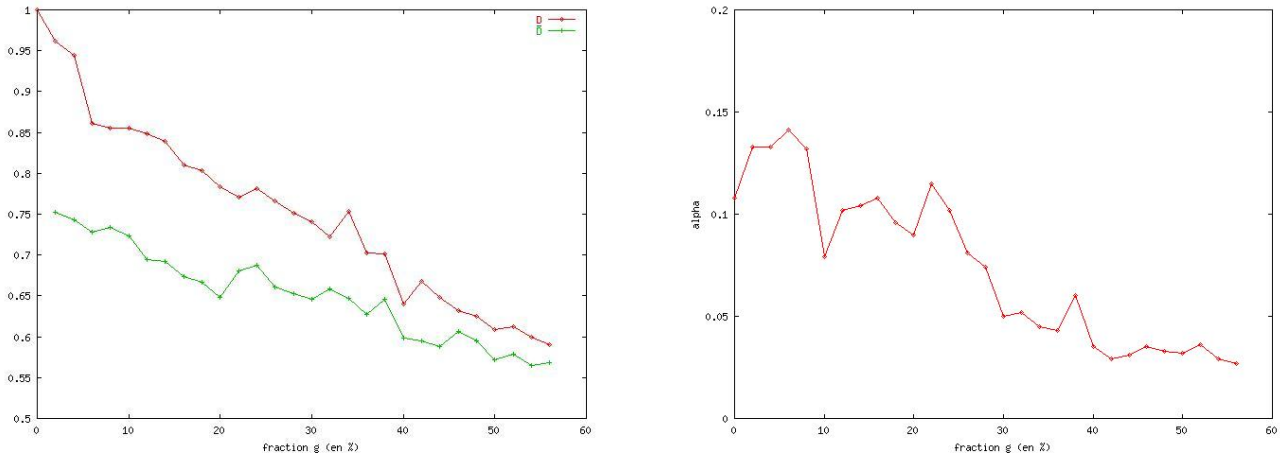


FIG. 4 – A gauche : D et $\overline{D}$ pour différentes valeurs de g ; à droite : α pour différentes valeurs de g.

L'évolution relative de D et $\overline{D}$ indique que le découpage communautaire Walktrap est assez peu sensible à cette forme de bruit : D converge lentement vers $\overline{D}$. Ce résultat n'est pas intuitif : on pourrait croire qu'un noeud fortement connecté tend à réduire assez nettement la taille des chemins entre noeuds, modifiant largement la structure communautaire de Walktrap, celle-ci étant basée sur des marches aléatoires. Cependant le coeur du graphe Scorpione est de petite taille (444 noeuds), les plus courts chemins entre deux sommets sont donc de faibles longueurs ; ce qui peut justifier la faible sensibilité à cette forme de bruit. L'incidence serait sans doute plus forte sur un graphe plus grand.

C'est aussi la petite taille du graphe qui explique les fortes fluctuations de α, notamment pour les faibles valeurs de g. La valeur de alpha diminue en moyenne avec g ; en fait, $\overline{l_{c-c,\rightarrow}}$ et $l_{c-c,\rightarrow}$ augmentent l'un et l'autre avec un écart à peu près constant. C'est le résultat qu'on attendrait si le noeud parasite modifie la structure communautaire seulement dans son environnement immédiat ; c'est ce que semble confirmer les visualisations du graphe avec Guess mises en annexe H.

[7] www.skyblog.com

3.5.4 Corrélation et catégories de communautés

Comme on l'a vu lorsque l'on a séparé A(G) de C(G) pour évaluer α, on peut chercher à séparer les sommets en sous-groupes pour trouver des comportements caractéristiques d'une catégorie de noeuds.

De la même façon, on peut penser que toutes les communautés décelées, selon leur taille, leur forme... ne jouent pas le même rôle dans le graphe. En particulier, on tente d'identifier un groupe de communautés dans lequel la corrélation périphérie/caractère intercommunautaire est plus nette que sur l'ensemble du coeur.

On ne va développer ici qu'un exemple simple du travail que l'on peut faire dans cet objectif : on partage les communautés en deux groupes, les communautés de grandes tailles (associées au sous-graphe O(G)) et celles de petites tailles (sous-graphe I(G)). Pour ce faire, on doit choisir un critère : on scindent les communautés de manière à ce qu'il y ait un nombre à peu près égal de noeuds dans I(G) et O(G).

On obtient alors sur le groupe O(G) de certains coeurs de graphes les résultats suivants (on n'effectue le travail qu'avec Walktrap) :

Graphe	Nombre de grandes communautés	Nombre de liens restant dans O(G)	rappel : α_W pour tout C(G)	α_W pour O(G)	borne sup de α_W pour O(G)
Yacine	9	2432	0.323	0.373	1
Full	6	873	0.410	0.410	1
FSE biparti	5	1842	0.174	0.166	0.623
Admin	10	470	0.175	0.123	0.540

Les autres résultats, jugés moins significatifs, sont consignés dans l'annexe I.

On constate pour le graphe Yacine une nette augmentation de α_W, mais pas pour les autres graphes, c'est même l'effet opposé pour le graphe Admin. Le critère choisi n'est donc pas adapté pour détecter des communautés où la correspondance entre la périphérie et les liens intercommunautaires serait particulièrement élevée ; ou du moins ce critère ne s'applique pas à tout type de graphe avec la même efficacité.

Pour le cas de graphes où la dépendance de α_W avec la taille communautaire est avérée (Admin, Yacine), on pourrait essayer d'affiner à nouveau la description en introduisant d'autres classes de taille ; cependant, plus on cherche à découper le graphe en sous-parties, moins l'analyse statistique est fiable : on peut calculer la valeur de α pour n'importe quel sous-graphe, mais elle perd sa signification (comme dans les cas reportés dans l'annexe H). Plus exactement, cela revient à adopter un point de vue "microscopique" sur le graphe : on examine non plus un comportement statistique global, mais une situation particulière, mettant en jeu un petit nombre de noeuds.

3.6 Conclusion

On a montré dans cette partie que, dans le coeur du graphe, il existe effectivement une correspondance statistique entre l'appartenance à la périphérie communautaire et la propension à établir des liens vers l'extérieur. Cela signifie que, dans une certaine mesure, on peut comprendre l'arborescence comme une zone frontière de la communauté.

Cependant, cette description convient plus ou moins selon la communauté considérée. On a cherché à identifier des communautés qui présentaient une corrélation plus élevée que la moyenne en utilisant le critère de la taille, ce qui ne s'est pas avéré très efficace. On peut envisager d'autres catégories de communautés et d'autres critères, mais cette approche conduit naturellement à sonder le graphe plus en détails.

4 Approche dynamique

On peut supposer que l'importance d'un lien dans un réseau peut se traduire par un comportement dynamique particulier. Reprenant l'analyse sociologique des trous structuraux évoquée en 2.5.3, si la position d'intermédiaire intercommunautaire est une position de force dans certains réseaux, un agent chercherait à la conserver au cours du temps. Comme précédemment, on va tenter de trouver une trace statistique de ce phénomène sur les graphes.

4.1 Le graphe étudié

Pour pouvoir observer des données dynamiquement, on utilise un graphe dans lequel les liens évoluent avec le temps. De plus, la notion de trou structurel étant associée aux situations où les agents du graphe tirent un intérêt

individuel de leur position, il semble que les données des conseils d'administration soient adaptés à l'étude que l'on veut mener.

Il y a toutefois un grand nombre de graphes réalisables depuis ces données. On a choisi de travailler sur le noyau dur défini dans la partie 2.2.2, en conservant le coeur et l'arborescence.

4.2 Durée de vie des liens

On se propose de chercher si la "durée de vie" des liens du graphe est liée à la position périphérique des noeuds qu'ils relient. On porte une attention particulière aux liens intercommunautaires, en se fondant ici encore sur l'idée qu'ils sont les voies de transit d'informations essentielles. On entend ici par durée de vie le nombre d'années pendant lesquelles un lien existe effectivement.

Le dénombrement donne les valeurs moyennes de la durée de vie pour différents ensembles de liens, selon le découpage Walktrap et plusieurs découpages Cosmoweb correspondant à différentes conditions initiales. On notera la durée de vie d'un lien : τ (exprimé en années), et les ensembles de liens seront notés selon le même principe que les grandeurs définies dans les parties 3.4 et 3.5, mais avec des majuscules : L pour l'ensemble des liens du graphe, $L_{\rightarrow}$ pour les liens intercommunautaires etc.

Essai	τ moyen sur L	τ moyen sur $L_{\rightarrow}$	τ moyen sur L_{c-c}	τ moyen sur L_{c-a}	τ moyen sur L_{a-a}	τ moyen sur $L_{c-c,\rightarrow}$	τ moyen sur $L_{c-a,\rightarrow}$	τ moyen sur $L_{a-a,\rightarrow}$
Walktrap	7.71	6.20	6.49	8.66	9.33	5.45	6.78	7.94
Cosmoweb 1	7.71	6.71	6.63	7.94	7.70	7.08	6.63	6.67
Cosmoweb 2	7.71	6.72	7.02	7.78	7.84	7.57	6.52	6.63
Cosmoweb 3	7.71	6.70	6.60	7.76	7.96	7.06	6.55	6.78
Cosmoweb 4	7.71	6.68	6.83	7.80	7.91	7.15	6.36	6.56
Cosmoweb 5	7.71	6.72	6.62	7.89	7.77	7.08	6.60	6.74

Avec le découpage Walktrap, les liens intercommunautaires mettant en jeu au moins un noeud de $\bigcup A(C_k)$ ont une durée de vie en moyenne supérieure à celle des liens C-C intercommunautaires. Cela pourrait être un indice montrant que les liens C-A ou A-A sont des voies de passage importantes que l'on cherche à maintenir dans le graphe, on les appellera *liens forts*. Cependant, on voit que ce n'est pas vrai pour le découpage Cosmoweb, où c'est même l'effet inverse que l'on observe.

Peut-être est-ce une propriété de l'algorithme Walktrap que d'associer les liens C-A ou A-A et les liens forts, mais il semble plus probable que cela soit la topologie particulière du graphe Admin qui en soit la cause, et donc que ces résultats ne puissent pas être extrapôler à d'autres graphes et d'autres contextes. Admin est construit de manière arbitraire : on n'a conservé que le noyau dur, de plus la collecte des données s'étend sur 11 ans, est-ce une période suffisante pour cibler des liens forts ? Et bien sûr l'estimation de la force d'un lien relève aussi d'une appréciation subjective. Il est donc difficile d'évaluer la qualité de la détection de liens forts sans faire l'analyse du contenu sociologique des données.

4.3 Des éléments pour l'étude microscopique

Le but d'une étude statistique du graphe est de donner une tendance ou un comportement moyen, et non d'isoler une situation particulière. Elle ne semble donc pas adaptée à l'objectif qui était le nôtre pour cette partie du travail : identifier des trous structuraux, ou en tous cas des configurations qui pourraient y être associées, et examiner leur comportement dynamique. Pour ce faire, il paraît plus approprié de regarder précisément deux communautés, voir comment évoluent leurs connections avec le temps etc. On donne dans cette partie une illustration du travail qui peut être fait par l'analyse "microscopique" du graphe, sans toutefois aller très loin dans cette direction, dont le thême et les méthodes s'éloignent de l'orientation que l'on souhaitait donner au stage.

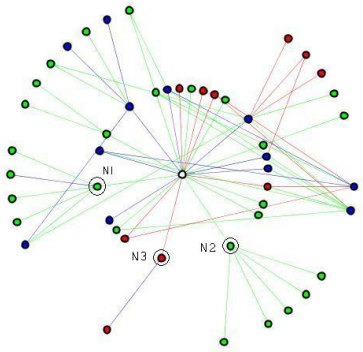
Une difficulté majeure à laquelle on est confronté est le manque de lisibilité d'une représentation sous forme de graphe. Par ailleurs, l'utilisation des découpages communautaires pour identifier d'éventuels trous structuraux s'avère délicate. En effet, on ne trouve pas facilement la situation idéale de deux communautés de tailles significatives qui ne seraient liées que par un ou deux ponts dont on pourrait examiner l'évolution.

On choisit donc un autre mode de représentation des données qui permet de contourner ces deux difficultés : on sélectionne un conseil d'administration et tous les administrateurs qui lui ont été connectés entre 1994 et 2004, puis tous les conseils liés sur cette période à ces administrateurs. En partant de l'hypothèse que deux firmes ayant un administrateur commun sont alliées (ce qui est sans doute très simplificateur), on pense pouvoir observer les mouvements des alliances concernant ce conseil.

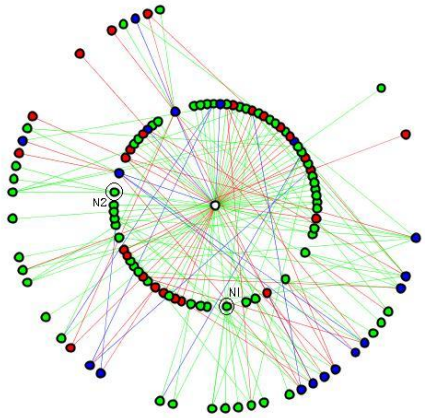
Les graphes suivants sont construits sur ce mode de représentation : au centre, on place la racine du graphe, le premier cercle autour de lui est constitué de ses administrateurs, le second des firmes "alliées".

D'autre part, on trace en rouge les liens qui existaient en 1994 mais pas en 2004, en vert ceux qui sont présents en 2004 mais pas en 1994 et en bleu ceux présents en 1994 et 2004. On en déduit si les conseils étaient alliés du sommet central en 1994 mais plus en 2004 (noeuds rouges), s'il l'étaient en 1994 et 2004 (noeuds bleus), ou seulement en 2004 (noeuds verts).

On peut observer sur ce premier graphe des administrateurs qui assurent seuls la liaison entre le noeud central et plusieurs autres conseils d'administration ; cette position pourrait correspondre aux situations recherchées.



On peut alors remarquer que le noeud central a créé des liens avec des administrateurs eux-mêmes fortement connectés (par exemple N1, N2), en revanche les liens qui disparaissent concernent surtout des noeuds peu connectés (N3). C'est un comportement qui semble assez fréquent, on le retrouve dans ce second graphe (N1, N2) :



D'autres exemples de graphes réalisés sur ce modèle sont en annexe J. On n'a pas eu l'occasion de constater le cas d'un administrateur fortement lié se déconnectant de la racine. On peut alors suggérer que les alliances se constituent "plus vite" qu'elles ne se désagrègent.

5 Conclusion générale

Les graphes constituent un moyen simple et polyvalent d'analyse des réseaux sociaux mais dépourvu de structure prédéfinie : toute l'information qu'ils contiennent se résume à la liste de leurs liens (et leurs éventuelles orientations et pondérations). Il est donc naturel de chercher une organisation dans la topologie du graphe, et les méthodes de détection communautaire sont des outils efficaces dans cet objectif.

Toutefois, la partition du graphe est une méthode dont on perçoit vite les limites : associer un sommet à un et un seul groupe donne nécessairement une vision très simplificatrice du réseau. On cherche donc à affiner cette description. Le travail produit au cours de ce stage montre qu'on peut interprêter plus précisément les découpages communautaires : une communauté peut être scindée en deux parties, son coeur et sa périphérie, et il semble que

la seconde soit une sorte de frontière vers l'extérieur ; cela permet de mieux estimer le degré d'implication d'un sommet dans sa communauté.

Mais la tendance mise en évidence est une moyenne effectuée sur la totalité du graphe, elle ne permet pas l'identification de situations précises. En particulier, on souhaiterait cibler sur l'ensemble du graphe les noeuds et les liens essentiels à son bon fonctionnement. On utilise alors des méthodes complémentaires avec le point de vue global, consistant en une analyse élémentaire du graphe et du contenu sociologique qui lui est associé.

Références

[1] K. Efe, V. Raghavan, C.H. Chu, A.L. Broadwater, L. Bolelli, and S. Ertekin. The shape of the Web and its implications for searching the Web, 2000.

[2] N. Del Vecchio. To what extent do small world structures of board interlocks follow locked strategies ? In *The 21st EGOS Colloquium 2005 "Unlocking Organizations"*, 2005.

[3] M. Gribaudi. *Les formes de l'expérience*, chapter Les discontinuités du social. Un modèle configurationnel, pages 187–225. Albin Michel, b. lepetit (ed.) edition, 1995.

[4] J.-P. Pélissier, D. Rébaudo, M.H.D. Van Leeuwen, and I. Maas. Migration and endogamy according to social class : France,1803–1986. *International Review of Social History*, 50 :219–246, 2005.

[5] D.J. Watts and S.H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393, 1998.

[6] A. Pothen, H.D. Simon, and K.P. Liou. Partitioning sparse matrices with eigenvectors of graphs. *SIAM J. Matrix Anal. Appl.*, 11(3) :430–452, 1990.

[7] M. Girvan and M.E.J. Newman. Community structure in social and biological networks. *PNAS*, 99(12) :7821–7826, 2002.

[8] P. Pons. Rapport de dea : Algorithmes des grands réseaux d'interactions : détection de structure de communautés. Accessible sur la page personnelle de Pascal Pons, sur le site du Liafa : www.liafa.jussieu.fr/~pons.

[9] M.E.J. Newman. Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, 69(066133), 2004.

[10] T. Bennouas, M. Bouklit, and F. de Montgolfier. Un modèle gravitationnel du web. Accessible sur la page personnelle de Fabien de Montgolfier, sur le site du Liafa : www.liafa.jussieu.fr/~fm.

[11] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7) :107–117, 1998.

[12] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking : Bringing order to the web. Technical report, 1998.

[13] S.B. Seidman. Network structure and minimum degree. *Social Networks*, 5, 1983.

[14] A. Degenne and M. Forsé. *Les réseaux sociaux*. Armand Colin, Coll. U, Paris, 2004.

[15] J. Zhang and M.S. Ackerman. Searching for expertise in social networks : A simulation of potential strategies. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, pages 71–80, 2005.

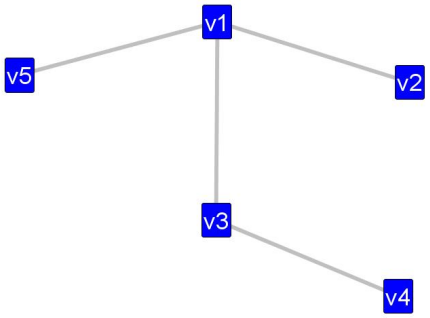
[16] R. Burt. Structural holes and good ideas. *American Journal of Sociology*, 2004.

[17] M. Granovetter. The strength of weak ties. *American Journal of Sociology*, pages 1360–1380, 1978.

Annexes

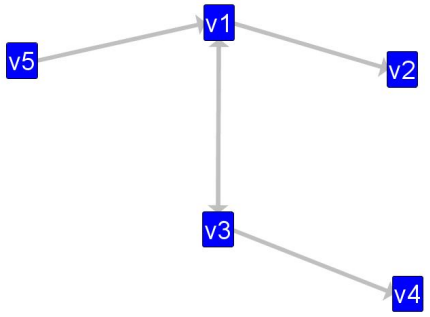
Annexe A : Matrices d'adjacence, graphe associé.

Un graphe non-orienté (non-pondéré) et sa matrice d'adjacence associée :



$$\begin{pmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Un graphe orienté (non-pondéré) et sa matrice d'adjacence associée :



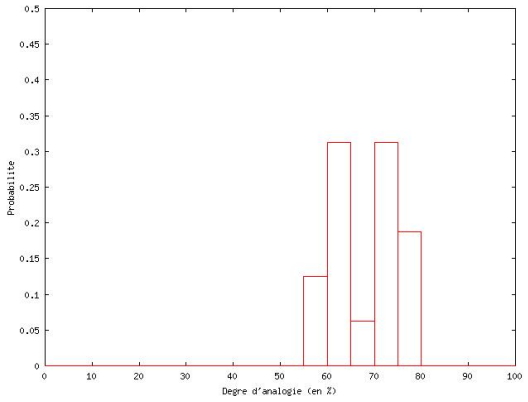
$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Dans ces deux situations, on a pris des termes diagonaux nuls, ce qui correspond à des graphes sans boucle (pas de lien d'un noeud sur lui-même).

Annexe B : Stabilité du découpage Cosmoweb sur FSE biparti.

On fait ici sur le graphe FSE biparti le même type d'étude que ce qui a été fait sur Dupin dans la partie 3.3.

On trace la distribution du degré d'analogie pour plusieurs (16) essais de Cosmoweb :



On constate que sur ce graphe, le degré d'analogie moyen entre deux distributions Cosmoweb est inférieur à 70 %.

Annexe C : Degré d'analogie des partitions Walktrap et Cosmoweb sur différents graphes.

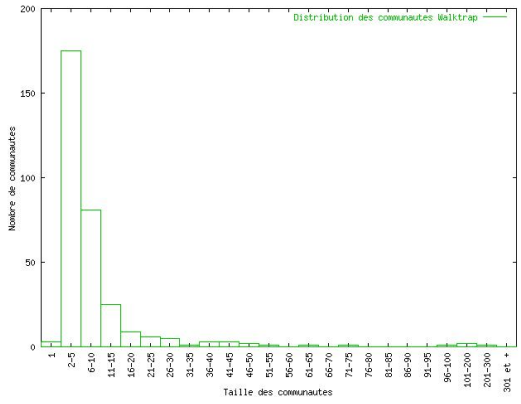
On effectue ici sur le graphe FSE biparti le même travail que ce qui a été fait dans la partie 3.4.2 sur Dupin. Les résultats sont du même type, mais la différence entre D et $\overline{D}$ est plus significative, et les valeurs de D et $\overline{D}$ fluctuent beaucoup moins avec les conditions initiales de Cosmoweb :

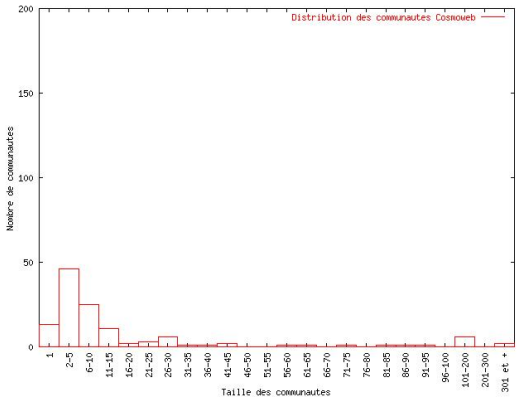
	D	$\overline{D}$	$\frac{D-\overline{D}}{D}$ (%)				
Essai 1	0.650	0.533	+ 18.0 %	Essai 5	0.649	0.533	+ 17.8 %
Essai 2	0.649	0.533	+ 17.8 %	Essai 6	0.650	0.530	+ 18.5 %
Essai 3	0.649	0.533	+ 17.8 %	Essai 7	0.649	0.533	+ 17.8 %
Essai 4	0.650	0.532	+ 18.2 %	Essai 8	0.649	0.533	+ 17.8 %
				Essai 9	0.649	0.531	+ 18.2 %

Enfin, on fait la comparaison pour tous les autres graphes à notre disposition, mais en réalisant un unique essai (i.e. un seul jeu de conditions initiales pour Cosmoweb). Les résultats obtenus sont reportés dans le tableau suivant :

Graphe	D	$\overline{D}$	$\frac{D-\overline{D}}{D}$ (%)	Graphe	D	$\overline{D}$	$\frac{D-\overline{D}}{D}$ (%)
Scorpione	0.755	0.683	+ 9.5 %	Wikipedia	0.506	0.491	+ 3.0 %
Yacine	0.828	0.748	+ 9.7 %	FSE simple	0.647	0.553	+ 14.5 %
Full	0.631	0.586	+ 7.1 %	Acteurs	0.873	0.785	+ 10.1 %
Tra	0.463	0.452	+ 2.8 %	Admin	0.580	0.498	+ 14.1 %

Annexe D : Distribution comparée des tailles communautaire sur le graphe FSE biparti.





Annexe E : Valeurs de α obtenues sur les découpages aléatoires Randcom.

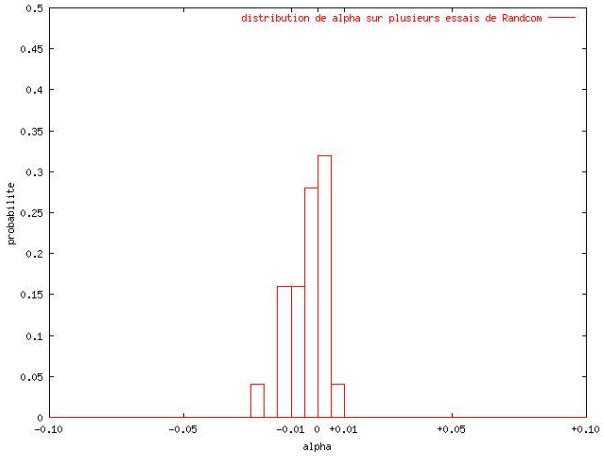


FIG. 5 – On effectue 20 essais de Randcom sur le graphe Dupin, la valeur de α obtenue est toujours très proche de 0, sachant que théoriquement, α peut aller jusqu'à + 1.

Annexe F : Borne supérieure de α sur différents coeurs de graphe.

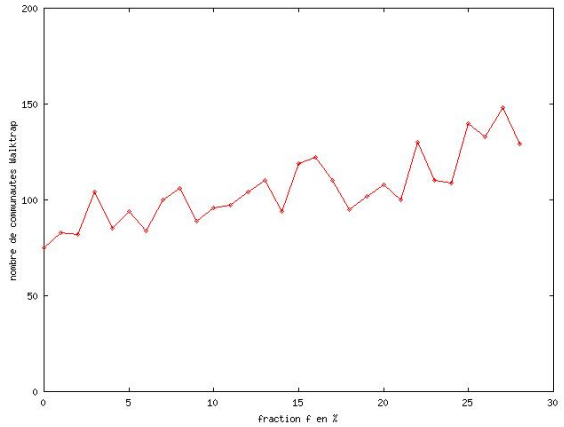
On n'utilise ici qu'une seule partition Cosmoweb, et on compare pour les partitions Walktrap et Cosmoweb les valeurs mesurées aux bornes supérieures de α, calculées selon :

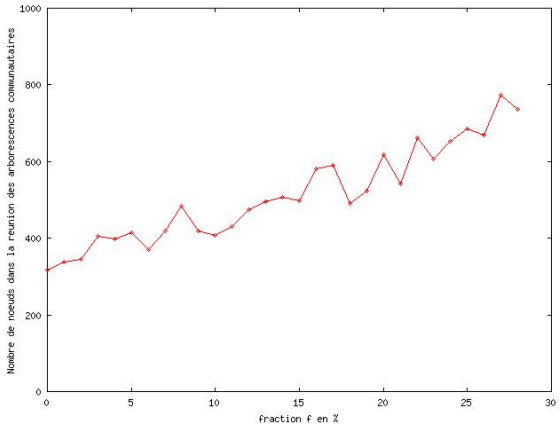
$$\frac{l_{c-a} + l_{a-a}}{\overline{l_{c-c,\rightarrow}}}$$

Coeur de graphe	α_C mesuré	borne sup. pour α_C	α_W mesuré	borne sup. pour α_W
Dupin	0.154	0.471	0.069	0.165
Scorpione	0.295	0.692	0.108	0.299
Yacine	0.306	0.712	0.323	0.692
Full	0.240	1	0.410	1
Tra	0.226	0.346	0.130	0.358
FSE biparti	0.305	1	0.174	1
FSE simple	0.064	0.099	0.007	0.020
Wikipedia	0.108	0.781	0.002	1
Acteurs	0.044	0.058	0.007	0.015
Admin	0.094	1	0.175	1

Annexe G : Résistance à la duplication des noeuds.

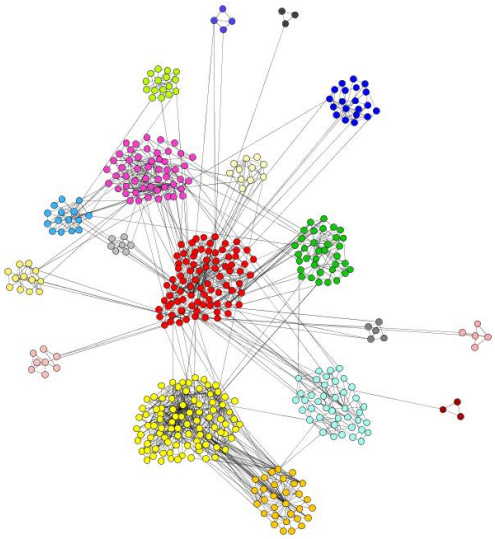
A gauche : évolution du nombre de communautés avec f selon le découpage Walktrap.
A droite : évolution du nombre de noeuds dans la réunion des arborescences de communautés avec f. On rappelle ici que le coeur du graphe FSE biparti comporte 1362 noeuds.

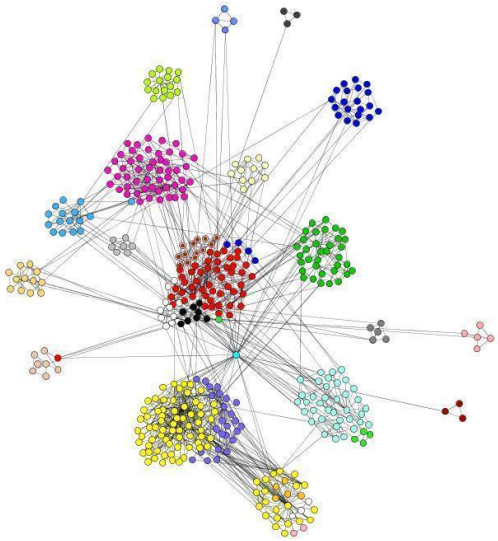




Annexe H : Graphe bruité par le noeud omniprésent.

On utilise Guess pour représenter l'évolution des communautés avec le bruit (type noeud omniprésent). On trace d'abord le graphe de référence avec ses communautés Walktrap, puis le graphe bruité ($g = 10\%$), on associe à chaque couleur une communauté du nouveau découpage pour voir comment celui-ci se superpose à l'ancien, et on constate que le noeud parasite (couleur cyan) perturbe le graphe localement.





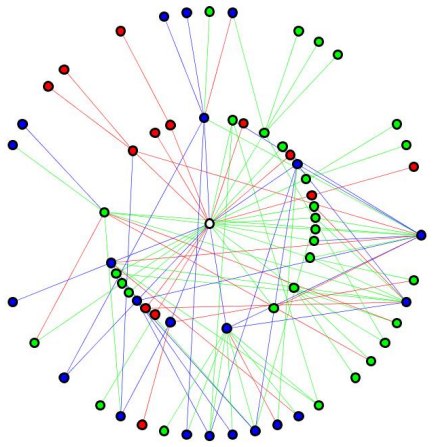
Annexe I : Compléments corrélation et grandes communautés.

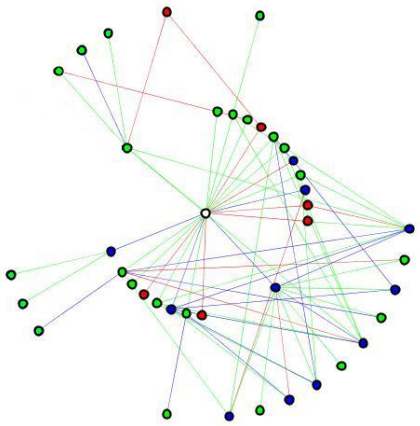
On n'a pas fait figurer ces résultat dans le corps du rapport car le très faible nombre de liens de types C-A ou A-A restant conduit à penser que le calcul de α n'est pas significatif. De faibles valeurs de l_{a-a} et l_{c-a} se traduisent par une borne supérieure de α_W également faible (sauf pour Scorpione où le nombre de liens C-C intercommunautaires attendus est lui aussi faible). On rappelle que la borne supérieure de α_W est donnée par $\frac{l_{a-a}+l_{c-a}}{\overline{l_{c-c,\rightarrow}}}$.

Graphe	Nombre de grandes communautés	Nombre de liens restant dans O(G)	α_W pour tout C(G)	α_W pour O(G)	borne sup de α_W pour O(G)
Acteurs	5	5219	0.007	0.005	0.008
Dupin	4	2159	0.069	0.037	0.091
FSE simple	5	15122	0.007	0.003	0.009
Scorpione	3	687	0.108	0.111	0.722
Tra	2	2432	0.130	0.029	0.095

Annexe J : Graphes centrés sur un Conseil d'Administration

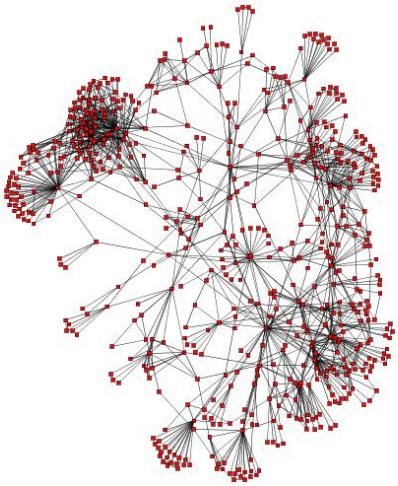
On donne ici d'autres exemples de graphes centrés sur un conseil d'administration, avec l'évolution de ses alliances entre 1994 et 2004.





Les graphes étudiés

On a mis dans cette annexe les graphes cités dans ce rapport, leurs caractéristiques principales, et une représentation simple obtenue avec Guess (option de visualisation GEM).

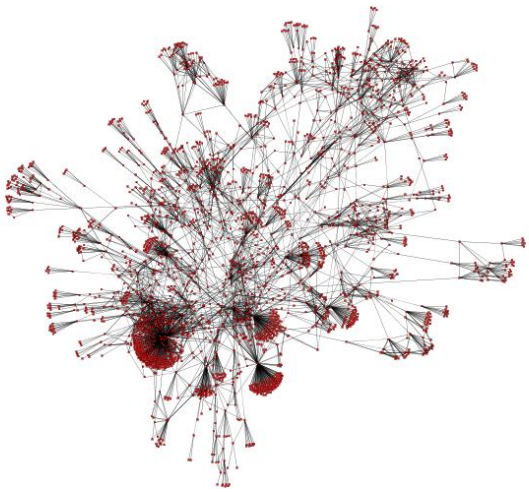


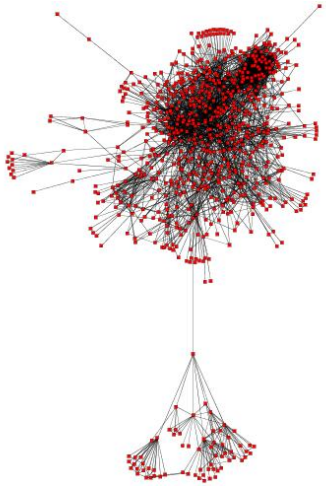
← Scorpione

- Graphe de blogs, deux blogs sont liés si au moins un figure sur le blogroll de l'autre.
- nombre de noeuds : 604
- nombre de liens : 1410
- nombre de noeuds/de liens dans C(G) : 444/1250
- nombre de noeuds dans A(G) : 160

Yacine →

- Graphe de blogs, deux blogs sont liés si au moins un figure sur le blogroll de l'autre.
- nombre de noeuds : 2702
- nombre de liens : 5477
- nombre de noeuds/de liens dans C(G) : 1863/4638
- nombre de noeuds dans A(G) : 839



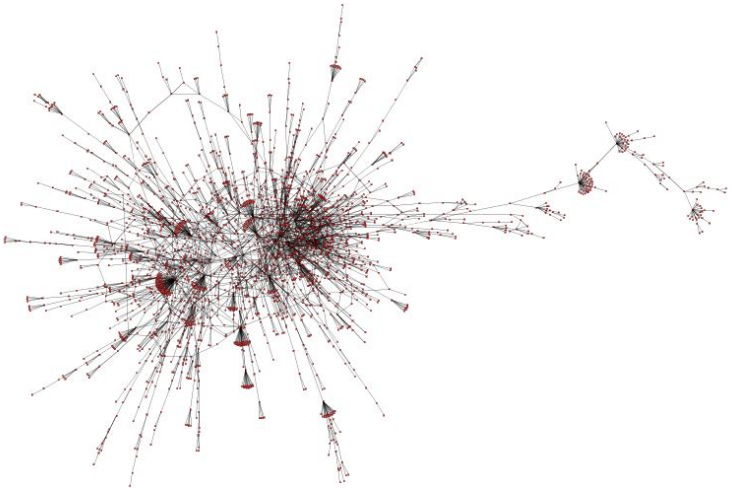


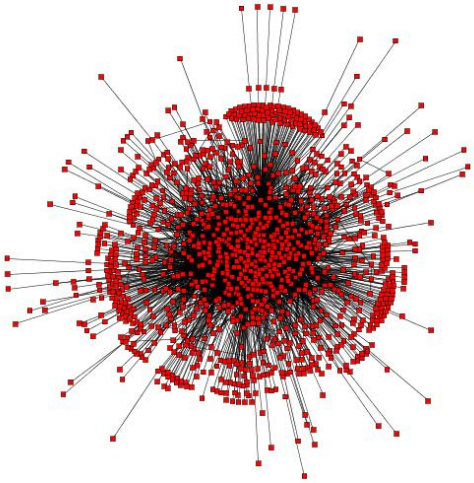
← Dupin

- Graphe de blogs, deux blogs sont liés si au moins un figure sur le blogroll de l'autre.
- nombre de noeuds : 724
- nombre de liens : 3192
- nombre de noeuds/de liens dans C(G) : 625/3093
- nombre de noeuds dans A(G) : 99

Full →

- Graphe de blogs, deux blogs sont liés si les deux bloggers ont communiqué par posts.
- nombre de noeuds : 2656
- nombre de liens : 3327
- nombre de noeuds/de liens dans C(G) : 860/1531
- nombre de noeuds dans A(G) : 1796



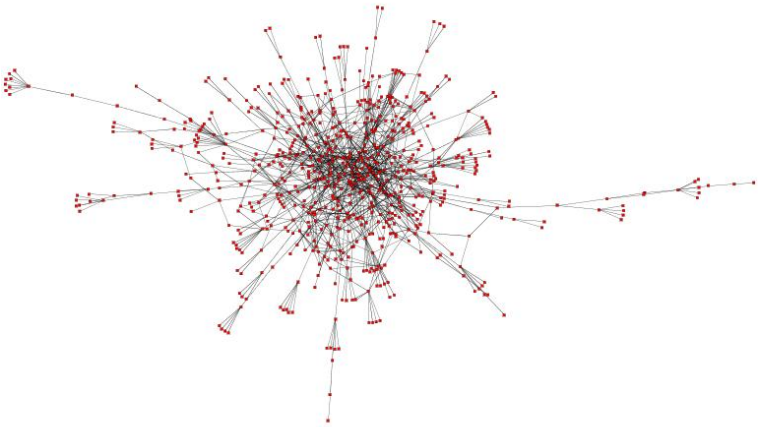


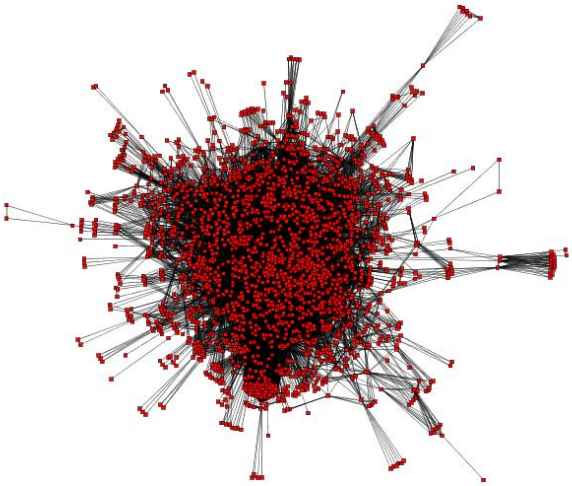
← Tra

- Graphe de généalogie du travail, les sommets sont des professions, deux sommets sont liés si le père a exercé une profession et son fils l'autre.
- nombre de noeuds : 1226
- nombre de liens : 4821
- nombre de noeuds/de liens dans C(G) : 639/4234
- nombre de noeuds dans A(G) : 587

Admin →

- Graphe des conseils d'administration biparti, un noeud-administrateur est lié à un noeud-conseil s'il y siège.
- nombre de noeuds : 691
- nombre de liens : 1184
- nombre de noeuds/de liens dans C(G) : 384/877
- nombre de noeuds dans A(G) : 307



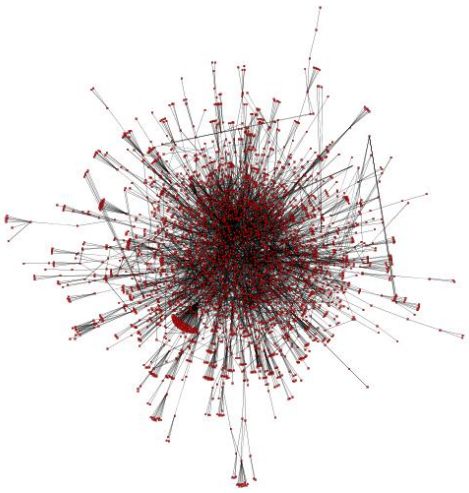


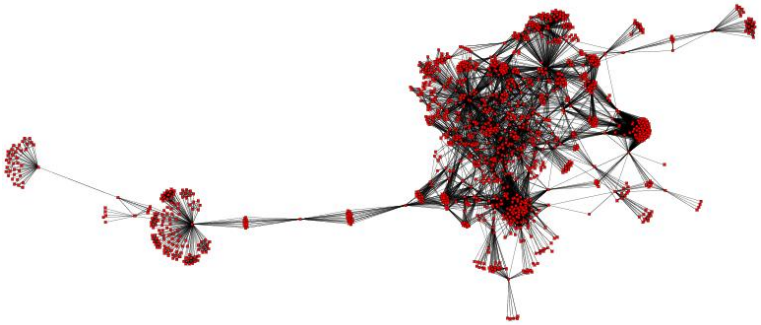
← FSE simple

- Graphe des Forums Sociaux, les sommets sont des associations, deux sommets sont liés si les deux associations ont participé à la même conférence.
- nombre de noeuds : 2262
- nombre de liens : 21630
- nombre de noeuds/de liens dans C(G) : 2224/21592
- nombre de noeuds dans A(G) : 38

FSE biparti →

- Graphe des Forums Sociaux biparti, un noeud-association est lié à un noeud-conférence si elle y a participé.
- nombre de noeuds : 3131
- nombre de liens : 5178
- nombre de noeuds/de liens dans C(G) : 1362/3409
- nombre de noeuds dans A(G) : 1769





← Acteurs

- Graphe des Acteurs, deux sommets sont liés s'ils ont participé à un même film.
- nombre de noeuds : 1298
- nombre de liens : 11068
- nombre de noeuds/de liens dans C(G) : 1274/11044
- nombre de noeuds dans A(G) : 24

Wikipedia (Le nombre de liens élevé contenus dans Wikipedia ne permet pas de le visualiser avec Guess.)

- Graphe de Wikipedia biparti, un noeud-page est lié à un noeud-rédacteur s'il a contribué à la page.
- nombre de noeuds : 18067
- nombre de liens : 43668
- nombre de noeuds/de liens dans C(G) : 4067/29668
- nombre de noeuds dans A(G) : 14000