View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by INRIA a CCSD electronic archive server

Espace intrinsèque d'un graphe et recherche de communautés

Alain Lelu, Martine Cadot

To cite this version:

Alain Lelu, Martine Cadot. Espace intrinsèque d'un graphe et recherche de communautés. Première conférence sur les Modèles et l'Analyse des Réseaux : Approches Mathématiques et Informatique - MARAMI 2010, Oct 2010, Toulouse, France. pp.1. hal-00516865

HAL Id: hal-00516865

https://hal.archives-ouvertes.fr/hal-00516865

Submitted on 24 Mar 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Espace intrinsèque d'un graphe et recherche de communautés.

Alain Lelu*,†,‡ — Martine Cadot*,††

* LORIA, Nancy - France
† Université de Franche-Comté/LASELDI, Besançon - France
Alain.Lelu@univ-fcomte.fr
†† Université de Nancy/Département Informatique, Nancy - France
Martine.Cadot@loria.fr
‡ Institut des Sciences de la Communication du CNRS, Paris - France

Résumé

La recherche de communautés dans un graphe se heurte à des problèmes épineux de représentation (formes convexes, recouvrantes, individus isolés…) dont l'abord optimal est réalisé par les méthodes spectrales, basées sur les dimensions propres du Laplacien de ce graphe. Déterminer le nombre de dimensions à prendre en considération est essentiel pour beaucoup d'applications. On s'attaque ici à ce problème dans le cadre de graphes non-orientés et non pondérés, qui inclut un type de graphe courant dans les applications de réseaux biologiques et sociaux, ceux munis d'une distribution des degrés de leurs nœuds en loi de puissance. Nous proposons à cet effet un test de randomisation, indépendant des lois de distribution. Après un petit exemple introductif, nous validons d'abord notre approche sur un graphe artificiel de ce type comportant deux communautés, puis sur deux graphes de test « Football League » et « Mexican Politician Network », où nous montrons à partir des résultats d'une méthode densitaire de clustering le caractère optimal du nombre de dimensions extraites.

***Mots-clés** : graphe, laplacien d'un graphe, réduction de dimensions, dimension intrinsèque, test de randomisation, clustering de graphe, méthode densitaire de clustering, graphe sans échelle, extraction de communautés, éboulis de Cattell.*

Abstract

Determining the number of relevant dimensions in the eigen-space of a graph Laplacian matrix is a central issue in many spectral graph-mining applications. We tackle here the problem of finding out the "right" dimensionality of Laplacian matrices, especially those often encountered in the domains of social or biological graphs: the ones underlying large, sparse, unoriented and unweighted graphs, often endowed with a power-law degree distribution. We present here the application of a randomization test to this problem. After a small introductive example, we validate our approach first on an artificial sparse and scale-free graph, with two intermingled clusters, then on two real-world social graphs ("Football-league", "Mexican Politician Network"), where the actual, intrinsic dimensions appear to be 10 and 2 respectively ; we illustrate the optimality of the transformed dataspaces both visually and numerically, by means of a density-based clustering technique and a decision tree.

***Keywords** : graph, graph Laplacian, dimensionality reduction, intrinsic dimension, randomization test, dominant eigen-subspace, graph clustering, density clustering method, scale-free graph, Cattell's scree.*

1. INTRODUCTION ET PROBLEMATIQUE : LA DIMENSION INTRINSEQUE D'UN GRAPHE.

Les méthodes spectrales sont de plus en plus utilisées pour extraire de la connaissance à partir de graphes. Ainsi le partitionnement spectral de graphe, qui consiste à grouper les sommets en communautés dans l'espace des premiers vecteurs propres de la matrice d'adjacence, ou de matrices qui en dérivent, est considéré comme une voie privilégiée d'amélioration de la qualité de ces partitions (Von Luxburg, 2007). Ou encore certaines caractéristiques spectrales sont considérées comme des indicateurs pertinents pour extraire des "motifs de graphes" dans des applications biologiques (Banerjee, 2008). Sans compter l'importance (y compris économique…) des indicateurs spectraux de centralité, comme *PageRank* (Brin et al., 1998), pour l'exploitation des réseaux sociaux et de connaissance. Deux questions se posent quand on utilise les méthodes spectrales :

- Quelle transformation de l'espace des données est la plus adéquate ? Bien que le débat ne soit pas clos, un consensus se dessine pour prendre en considération l'espace des principaux vecteurs propres de l'une ou l'autre des matrices « laplaciennes » qui se déduisent de la matrice d'adjacence et que nous détaillerons plus loin.

- Combien de dimensions principales choisir dans cet espace pour y observer au mieux des éléments caractéristiques ou y réaliser des traitements, comme l'extraction de communautés ? C'est à cette question que nous essaierons de répondre ici. Dans le cadre général de matrices de données rectangulaires beaucoup de réponses ont été proposées ; la plupart recherchent – visuellement ou à partir des différences secondes entre valeurs propres consécutives – un "coude" (ou décrochement, *gap*) séparant celles qui sont significatives de celles qui ne représentent que du bruit (Cattell, 1996). D'autres font l'hypothèse de distributions statistiques spécifiques (Bouveyron *et al.*, 2009), au risque de ne pas être adaptées au cas des grands graphes « sans échelle » (*scale-free*) présents dans nombre d'applications sociales ou biologiques.

Alors que les comparaisons statistiques avec les modèles "nuls" d'un graphe, c'est à dire avec des versions randomisées de ce graphe, attirent de plus en plus l'intérêt pour des tâches comme la découverte de motifs dans les graphes biologiques (Milo *et al.*, 2002), aucune proposition n'a été faite, à notre connaissance, pour délimiter au moyen d'une méthodologie statistique rigoureuse ce que nous nommerons désormais l'*espace intrinsèque* d'un graphe. Nous nous limiterons ici au cas des graphes non pondérés et non orientés. Ceci constitue notre contribution originale, et nous présenterons en section 2 quelques travaux proches, puis nous ferons en section 3 quelques rappels sur les méthodes spectrales et certaines approches pionnières. La section 4 décrira notre méthode de génération de matrices randomisées et test TourneBool dans le cas général, déjà publiée, et la spécifiera pour le cas de graphes non orientés et non pondérés, ce qui est nouveau. Après un exemple introductif sur un petit graphe à quatre communautés, nous présenterons en section 5 trois applications : l'une sera consacrée à un graphe sans échelle généré artificiellement et comportant deux communautés de sommets ; les autres utilisent respectivement les graphes sociaux réels "Football League" et "Mexican Politician Network". L'optimalité de cet espace sera évaluée quantitativement en y appliquant, pour plusieurs nombres de dimensions retenues, 1) un algorithme densitaire de clustering dont les résultats seront comparés à la "vraie structure" de classes au moyen d'une F-mesure, 2) un jeu de règles de discrimination évalué de même.

2. APPROCHES VOISINES

Les auteurs de (Milo *et al.*, 2002) comparent un graphe orienté à ses versions randomisées, ayant le même nombre de nœuds et une distribution identique des degrés entrants et sortants, dans le but de détecter des sous-graphes orientés significatifs qu'ils nomment « motifs de réseaux ». Comme leur objectif, éloigné du nôtre, se concentre sur la détection de modules élémentaires, constitutifs des réseaux biologiques, ils astreignent leurs modèles d'hypothèse nulle à d'autres contraintes, (par exemple s'agissant d'extraire les 4-motifs, ils imposent à leurs 3-motifs la même distribution que dans le réseau d'origine). La contribution (Banerjee, 2002) a d'autres objectifs bio-inspirés, comme la découverte de motifs fusionnés, ou celle de duplications. Elle montre que les spectres complets des laplaciens des graphes à comparer gardent des traces caractéristiques de ces événements. Elle n'explore pas les propriétés de versions randomisées des graphes, ni des parties dominantes de leurs spectres.

Dans (Lelu, Cadot, 2009) et (Lelu, Cadot, 2010) nous avons comparé des matrices binaires (textes × mots) à leurs versions randomisées, pour faire apparaître les liens (et anti-liens) statistiquement valides dans les deux graphes des relations entre mots et relations entre textes. Bien que très liées aux méthodes et outils exposés ici, ces approches n'ont pas abordé jusqu'à présent la recherche du « bon » sous-espace de représentation d'un graphe, comme c'est notre but ici. La contribution (Gionis et al., 2007) aborde comme nous l'avons fait dans d'autres travaux cités plus haut le problème du nombre de dimensions significatives d'une matrice binaire rectangulaire, mais de façon heuristique, en se basant sur une unique matrice randomisée.

3. ESPACE INTRINSEQUE D'UN GRAPHE

Nous appellerons *espace intrinsèque* d'un graphe l'espace de représentation réduit dans lequel se trouvent concentrées et mises en évidence ses caractéristiques structurelles « intéressantes », comme certains regroupements de nœuds en clusters ou en chaînes, caractéristiques qui ne se retrouvent pas dans ses variantes randomisées (i.e. à même répartition des degrés des nœuds). On se focalisera ici sur les critères d'intérêt les plus unanimement reconnus dans la littérature, à savoir les caractéristiques spectrales de l'une ou l'autre des matrices laplaciennes du graphe, i.e. ses valeurs propres. Les vecteurs propres correspondant aux valeurs propres significatives constituent alors son espace de représentation intrinsèque.

A notre connaissance, la première application aux graphes de l'analyse spectrale remonte à (Benzécri, 1973) (qui reprenait le cours polycopié de 1969 *Sur l'analyse de la correspondance définie par un graphe*), dans lequel l'Analyse Factorielle des Correspondances (AFC) était appliquée à la matrice d'adjacence d'un graphe. Rappelons que l'AFC (Greenacre, 2007 ; Lebart et al., 1984) repose sur la décomposition aux valeurs singulières d'une matrice **Q** issue du tableau de correspondance **X** :

$$\mathbf{Q} = \mathbf{D_r}^{-1/2} \mathbf{X} \mathbf{D_c}^{-1/2}$$

(à noter que pour un graphe non orienté et non pondéré, on applique une telle décomposition à un **X** symétrique et à valeurs binaires), où $\mathbf{D_r}$ et $\mathbf{D_c}$ sont les matrices diagonales des sommes en lignes et en colonnes. La décomposition aux valeurs singulières de **Q** s'écrit :

$$\mathbf{Q} = \mathbf{U} \mathbf{\Lambda} \mathbf{V'}$$

où $\mathbf{\Lambda}$ est la matrice diagonale des valeurs singulières (parmi lesquelles : $\lambda_1 ... \lambda_L = 1$, L étant le nombre de composantes connexes ; et $1 > \lambda_{L+1} > ... > \lambda_R > 0$, R étant le rang de **X**). Les matrices **U** et **V** rassemblent les vecteurs propres pour les lignes et les colonnes respectivement. Les facteurs d'AFC **F** et **G** s'en déduisent, par multiplication avec des matrices diagonales :

$$\mathbf{F} = x_{..}^{1/2} \mathbf{D_r}^{-1/2} \mathbf{U} \mathbf{\Lambda} \qquad \mathbf{G} = x_{..}^{1/2} \mathbf{D_c}^{-1/2} \mathbf{V} \mathbf{\Lambda}$$

où $x_{..}$ est le total de **X**. (Benzécri, 1973) a proposé des solutions analytiques pour des graphes simples comme les anneaux ou les grilles. Dans (Lebart, 1984) l'auteur a généralisé à l'analyse de la contiguïté, et illustré en montrant que le plan factoriel (F2, F3) de l'AFC de la matrice de contiguïté entre les départements français reconstituait l'allure de la carte de France.

Une lignée de recherche indépendante initiée par (Chung, 1997) a défini deux matrices "laplaciens normalisés de graphes", à savoir le laplacien symétrique (**I – Q**), où **I** est la matrice identité (on a $\lambda_1 ... \lambda_L = 0$, L étant le nombre de composantes connexes ; $0 < \lambda_{L+1} < ... < \lambda_R$, R étant le rang de **X**), et sa variante "marche aléatoire" $\mathbf{I} - \mathbf{D_r}^{-1} \mathbf{X}$. A noter que le vecteur propre dominant de $(\mathbf{D_r}^{-1}\mathbf{X})'$ (plus précisément, de $\alpha (\mathbf{D_r}^{-1} \mathbf{X})' + (1- \alpha) (1/N) \mathbf{11'}$, où **1** désigne le vecteur-colonne unitaire, pour "forcer" l'existence d'une seule composante connexe) n'est autre que l'indicateur *PageRank* de centralité (Brin *et al.*, 1998). A noter aussi que la somme algébrique des valeurs propres des laplaciens de graphes est égale à la somme des degrés pour le laplacien simple, non normalisé, $\mathbf{D_r} - \mathbf{X}$, au nombre de sommets pour les laplaciens normalisés **I – Q** et $\mathbf{I} - \mathbf{D_r}^{-1} \mathbf{X}$ (et donc nulle pour **Q** et $\mathbf{D_r}^{-1} \mathbf{X}$ qui nous serviront dans la suite de l'exposé).

La partition spectrale de graphe consiste à grouper les nœuds dans l'espace des K plus importants vecteurs propres – pour une revue cf. (Chung, 1997) – et constitue une voie de recherche de plus en plus active. Jusqu'à présent, à notre connaissance, la détermination du nombre K, quand la distribution des degrés sort des modèles classiques (loi binomiale, etc.), n'a pas reçu de réponse plus satisfaisante que la classique détermination visuelle ou par examen des différences secondes d'une discontinuité dans la séquence des valeurs propres (Cattel, 1966) – ce qui ne pose pas de problème pour les petits graphes, mais passe difficilement à l'échelle de centaines ou milliers de noeuds. Nous avons pu constater par exemple que pour un graphe des relations entre les 3700 mots les plus fréquents d'un corpus documentaire aucune évidence visuelle de cassure ne ressortait de l'examen de l'éboulis des mille premières valeurs propres, ni de celui du signal (très bruité) de leurs différences secondes, et aucune discontinuité n'était perceptible dans l'inflexion régulière de leur somme cumulée.

4. TEST DE RANDOMISATION POUR ETABLIR LA DIMENSION INTRINSEQUE D'UN GRAPHE

4.1 Le test TourneBool et son application au problème

Notre but n'est pas de simuler un graphe aléatoire avec une suite de degrés donnés, mais, pour un graphe donné, de générer de la façon la plus directe et rigoureuse possible une suite de graphes aléatoires indépendants de même suite de degrés. Les deux points de ce cahier des charges ne sont pas remplis par le « configuration model » (Molloy & Reed, 1995) (qui ne garantit pas l'adéquation rigoureuse à une distribution des degrés donnée) et ses dérivés, qui compliquent ce modèle, cf. par exemple (Viger & Latapy, 2005).

TourneBool est une méthode de génération de N versions aléatoires (“randomisées”, N souvent égal à 100 ou 200) d’un tableau de données binaires, à marges lignes et colonnes inchangées, et de test statistique de toute quantité construite sur ce tableau, par comparaison avec les N valeurs trouvées sur les tableaux randomisés. Il est à noter que les principes de génération de matrices aléatoires à marges fixes, en partant d’une matrice binaire donnée, semblent avoir été découverts indépendamment plusieurs fois dans plusieurs domaines d’application : écologie (Connor *et al.*, 1979 ; Cobb *et al.*, 2003), sociologie (Snijders, 2004), combinatoire (Ryser, 1964). En ce qui nous concerne, l’un de nous a présenté (Cadot, 2005) un algorithme de permutation basé sur des échanges rectangulaires (un échange rectangulaire à la croisée des lignes i1 et i2 et des colonnes j1 et j2 est possible sans modifier les marges si les cases (i1,j1) et (i2,j2) valent 1 alors que les cases (i1,j2) et (i2,j1) valent 0) ; il incorpore un contrôle de la convergence de l’algorithme pour éviter tout biais. Sa justification théorique, exposée dans (Cadot, 2006), est basée sur la notion, originale à notre connaissance, d’échange en cascade, opération qui transforme une matrice booléenne en une autre matrice de mêmes marges – et à l’inverse, il a été montré dans ce même mémoire que toute matrice booléenne pouvait être transformée ainsi en toute autre de mêmes sommes marginales en un nombre fini de telles cascades. Dans le domaine des graphes, nous avons appliqué cette approche pour créer des graphes de liens (et d’anti-liens) valides entre variables booléennes (les mots) à partir de corpus textuels (Lelu , Cadot, 2009). A notre connaissance, l’application de ce type de méthode de randomisation à la détermination du meilleur espace de représentation d’un graphe est originale.

L’algorithme de génération et de contrôle des matrices randomisées est détaillé dans (Lelu , Cadot, 2010). Il permet de faire abstraction de la “structure d’arrière-plan” commune à toutes les matrices de mêmes sommes marginales, donc de prendre en compte tout type de données binaires, à la fois 1) en acceptant tout type de loi de distribution marginale, et 2) sans nécessité de spécifier la loi de probabilité suivie par les marges. En termes de graphes, cet algorithme s’interprète comme une suite d’échanges croisés de liens entre paires d’arêtes, quand ces échanges sont possibles. D’autres applications de cette génération de matrices aléatoires ont été présentées plus haut en section 2, mais elles n’abordent pas le problème de l’espace intrinsèque d’un graphe.

A noter que les tests de permutation, dont dérivent les tests de randomisation, ont été démontrés comme les plus “puissants”, c’est à dire minimisant le risque bêta pour un risque alpha donné[1] (Droesbeke, Finne, 1996).

4.2. Application aux graphes non orientés et non pondérés

En l’état, TourneBool peut être appliqué aux matrices d’adjacence de graphes bipartis, non orientés et non pondérés, car leurs éléments non-nuls sont inclus dans deux matrices binaires rectangulaires, symétriques l’une de l’autre par rapport à la diagonale, et cette disposition est compatible avec le processus de génération décrit ci-dessus. Mais pour les versions randomisées de matrices d’adjacence de graphes non orientés et non pondérés, des contraintes supplémentaires doivent être ajoutées au moment de permettre (ou pas) un échange rectangulaire : la matrice doit rester symétrique (on n’autorise l’échange rectangulaire aux intersections des lignes (i1,i2) et des colonnes (j1,j2) que s’il est également possible aux intersections des lignes (j1,j2) et des colonnes (i1,i2)), et sa diagonale doit rester vide (pas d’échange quand un des éléments est sur la diagonale).

Quelle matrice dérivée de la matrice d’adjacence faut-il prendre en compte pour la détermination de l’espace propre pertinent ? Rappelons tout d’abord le résultat bien établi en analyse de données selon lequel l’information pertinente, non bruitée, réside dans les éléments propres dominants d’une matrice de données (Benzécri, 1973). Ensuite, dans le cas de la matrice **Q** décrite en section 3, (Benzécri, 1973), (Chung 1997) et bien d’autres travaux ont montré que sa plus grande valeur propre, de multiplicité L (L étant le nombre de composantes connexes du graphe) valait 1. Il en va de même de la matrice $\mathbf{D_r}^{-1}\ \mathbf{X}$ dont l’espace de représentation semble préféré par beaucoup d’auteurs.

Notre test permet de trouver quelles sont les valeurs propres de **Q** ou $\mathbf{D_r}^{-1}\ \mathbf{X}$ *significatives*, c’est-à-dire qui s’écartent de celles qui résultent du hasard, extraites à partir de versions randomisées de X. Notre test s’écrit ainsi :

- Extraire la séquence complète des valeurs propres de **Q** ou $\mathbf{D_r}^{-1}\ \mathbf{X}$, λj étant la valeur propre de rang j.
- Engendrer un échantillon suffisant (**X1**, **X2**, ... **Xp**) de versions randomisées de la matrice d’origine **X** (par ex. 200 matrices).
- Pour chaque version randomisée Xi de X (i varie de 1 à p), extraire la séquence complète des valeurs propres de **Qi** ou $\mathbf{D_r}^{-1}\ \mathbf{Xi}$, λij étant la valeur propre de rang j.
- Pour chaque rang j, positionner la valeur propre λj de la matrice d’origine par rapport à la séquence ordonnée des valeurs propres ($\lambda i_0 j$, $\lambda i_1 j$, …, $\lambda i_p j$) de même rang des matrices simulées : si la valeur propre λj est dans l’intervalle des valeurs randomisées situées au seuil de significativité (par ex. entre la 2ème et la 199ème pour un seuil de confiance de 99%), elle est attribuée au hasard, sous contrainte de même distribution des degrés, sinon elle est déclarée s’écarter significativement des valeurs obtenues par hasard.

[1] Risque alpha : risque de conclure à une différence qui n'existe pas (« faux positif », « bruit ») ; risque bêta : risque de ne pas mettre en évidence une différence qui existe réellement (« faux négatif », « silence »).

Seules les valeurs significatives sont conservées. Des remarques sont à faire, se déduisant de l'usage du test TourneBool :

1) Il est à noter que la première valeur propre est 1 pour la matrice d'origine comme les matrices simulées, ce qui permet de l'écarter. Dans le cas d'un graphe avec k composantes connexes, les k valeurs suivantes de la matrice d'origine sont égales à 1, mais il y a très peu de chances qu'elles atteignent 1 pour les matrices simulées, ce qui les rend significatives dans le cas le plus courant..

2) Lors du fonctionnement de ce algorithme adapté aux matrices de graphes, nous avons pu constater sur nos divers exemples que les valeurs propres significatives se succédaient de façon ininterrompue, la première de celle-ci étant la deuxième valeur propre, jusqu'au moment où elles basculaient dans la non-significativité, pour y rester jusqu'à la fin. Ce qui nous a permis de déterminer un espace intrinsèque cohérent avec les choix habituels (en utilisant des valeurs propres consécutives).

5. METHODOLOGIE : EXEMPLE D'UN PETIT GRAPHE ARTIFICIEL A QUATRE COMMUNAUTES

Pour commencer par un cas simple nous avons créé un graphe de quatre cliques, de tailles voisines, dans lequel nous avons ajouté et retranché aléatoirement une faible quantité d'arcs, sans boucles ni arcs multiples. Ce qui se traduit par la matrice d'adjacence **M0** de la figure 1, bruitée avec 17 % de « blancs » dans les zones « noires », et 12 % pour l'inverse.

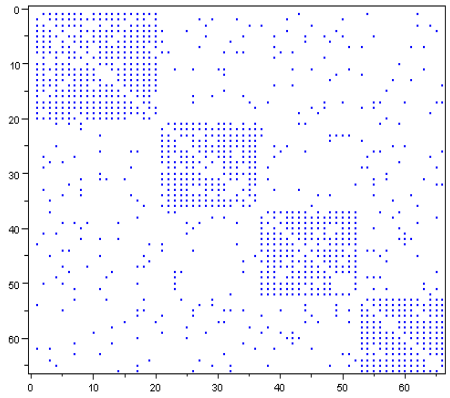

Figure 1. *La matrice d'adjacence **M0** avec 4 communautés*

On calcule alors la matrice $\mathbf{D^{-1}\,M0}$ (cf. section précédente) où **D** est la matrice diagonale des degrés de la matrice d'adjacence **M0**, puis ses 66 valeurs propres et vecteurs propres correspondants U1 à U66. Nous avons choisi de créer par TourneBool 200 matrices **M1** à **M200** randomisées à partir de **M0.**

Nous avons porté sur la figure 2 gauche : 1) en trait bleu la séquence des valeurs propres de $\mathbf{D^{-1}\,M0}$ (qui peuvent être positives ou négatives) dans l'ordre où les fournit l'algorithme QR, standard dans les librairies de calcul numérique, 2) on procède de même pour chacune des 200 simulations $\mathbf{D^{-1}\,Mxx}$, et on obtient ainsi, pour chaque valeur propre d'un rang donné, 200 valeurs qu'on ordonne. Les quantiles 2 et 199 de cette série sont les bornes à 99% des intervalles de confiance de la valeur propre. Et de même pour les quantiles 5 et 196 qui fournissent les bornes de l'intervalle de confiance à 95%. Ces quatre bornes figurent en rouge sur le graphique de la figure 2. L'algorithme fournit en premier les valeurs propres importantes, qu'elles soient positives ou négatives, les valeurs insignifiantes étant rejetées à la fin.

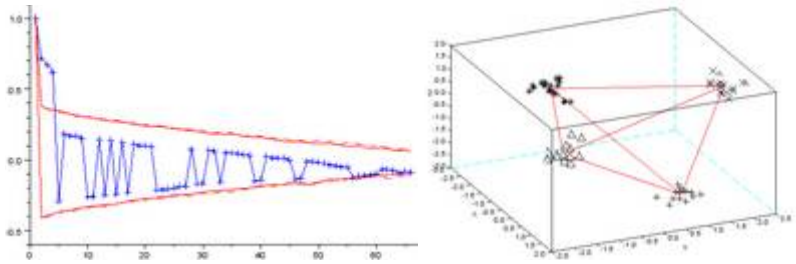

Figure 2. Gauche : *La séquence des 66 premières valeurs propres issues de* $\boldsymbol{D^{-1}}$ ***M0*** *(en bleu) comparées aux bornes des intervalles de variation de ses dérivées randomisées.* Droite : *les 4 communautés (en noir) dans l'espace des 3 vecteurs propres U2, U3, U4.*

On remarque que seules les trois premières valeurs propres non triviales de la matrice **M0** d'origine sont significatives, car en dehors des bornes entre lesquelles oscillent les autres : ceci est cohérent avec l'existence de quatre communautés faiblement liées entre elles, nécessitant donc un espace de dimension trois pour se représenter distinctement autour des quatre sommets d'un tétraèdre (cf. figure 2 droite en rouge). Les vecteurs propres U2 à U4 définissent donc l'espace intrinsèque du graphe considéré.

Nous avons préféré cette représentation, induite par notre algorithme, à celle plus habituelle en « éboulis » de valeurs propres triées par valeurs algébriques décroissantes. En effet la contrainte de somme algébrique constante des valeurs propres – ici égale à zéro – crée dans cette dernière un phénomène de compensation : si certaines de ces valeurs sont largement au-dessus de l'intervalle de variation des matrices aléatoires, d'autre peuvent passer au-dessous par la suite. La figure 4 illustre volontairement le phénomène pour l'exemple montré en section 6. Nous avons constaté aussi qu'utiliser les valeurs absolues triées améliore la situation, mais ne la résout pas.

6. VALIDATION SUR UN GRAPHE ARTIFICIEL « ZIPFIEN » A DEUX COMMUNAUTES : Y A-T-IL UNE SEULE DIMENSION INTRINSEQUE ?

Nous nous attachons ici à reproduire deux caractéristiques importantes d'un grand nombre de réseaux sociaux ou biologiques : 1) une distribution en loi de puissance de leurs degrés ; 2) une structure de communautés rarement en tout ou rien : on pourrait plutôt la décrire comme une appartenance progressive, nuancée, des individus à chaque communauté, et structurée autour d'amas plus denses. En d'autres termes, les communautés sont souvent imbriquées, recouvrantes, et rarement « orthogonales » au sens où le seraient des composantes connexes d'un même graphe.

6.1. Génération du graphe artificiel

Nous allons créer une telle structure dans un cas relativement simple, mais réaliste, celui de deux communautés recouvrantes de tailles inégales, et dont les degrés des nœuds suivent une loi de puissance. Ignorants au moment de nos travaux du générateur de graphes sans échelle, à nombre voulu de clusters, décrit dans (Lancichinetti & Fortunato, 2009), nous avons suivi un processus ad-hoc sans prétention à une rigueur absolue ni capacité de généralisation : partis d'une matrice bloc-diagonale à 2×2 blocs à valeurs binaires (les deux blocs diagonaux « noirs », les deux autres « gris »), nous en avons réalisé la morphose en deux étapes. Tout d'abord sa transformation en matrice de réels positifs dont les sommes en lignes et colonnes suivent une loi de puissance donnée. Ensuite, comme l'ont fait aussi les auteurs cités ci-dessus, ces nombres ont été transformés pour la matrice finale en probabilités de tirer des valeurs 1 plutôt que zéro, aboutissant à la matrice 732 × 732 montrée figure 3, sans évidence visuelle de sa structure en deux communautés.

A noter que le générateur de Lancichinetti et Fortunato peut aussi introduire des perturbations dans la distribution demandée des degrés, ou même ne pas aboutir au nombre désiré de communautés. Mais l'essentiel de l'usage de telles matrices engendrées est bien de simuler de façon contrôlée des graphes réels très bruités, sans plus.

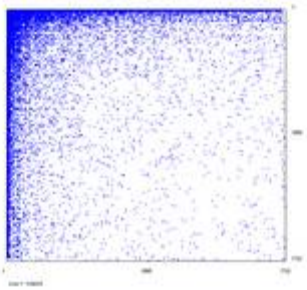

Figure 3. *La matrice d'adjacence **M0** avec 2 communautés, un degré minimum 4, et une distribution des degrés en loi de puissance*

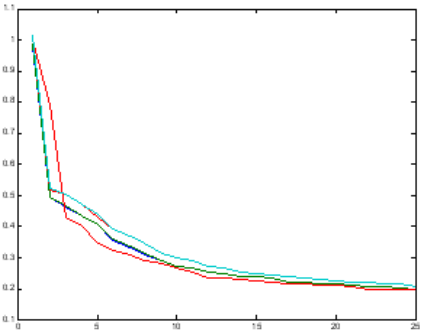

Figure 4. *L'"éboulis" classique des 50 premières valeurs propres issues de* $\boldsymbol{D^{-1}\,M0}$ *(en rouge) comparées aux intervalles de variation de ses dérivées randomisées.*

6.2. Traitements et résultats

On calcule alors la matrice $\mathbf{D^{-1}\,M0}$ (cf. section 4) où **D** est la matrice diagonale des degrés de la matrice d'adjacence **M0**, puis ses 50 plus grandes valeurs propres, et vecteurs propres correspondants U1 à U50. Nous avons choisi de créer par TourneBool 200 matrices **M1** à **M200** randomisées à partir de **M0** : bien qu'il soit recommandé, en matière de test de randomisation, de réaliser au moins mille simulations, la figure 4 (et 7, 9, pour les expériences suivantes) montre que notre choix initial de 200, plus économe en temps de calcul, était suffisant pour créer une forte continuité visuelle entre les intervalles de variation de chaque valeur propre d'ordre k (courbe verte supérieure), ainsi qu'une faible variabilité de ces valeurs, pour l'ensemble des matrices simulées – ce qui pourrait ne pas être le cas pour des statistiques plus locales que des valeurs propres, comme celles de co-occurrences utilisées dans (Lelu, Cadot, 2010). Par ailleurs nous avons opté pour un seuil de significativité à 1% de risques de se tromper car l'autre seuil courant, 5%, se traduisait par un intervalle de variation plus étroit, et aurait pu avoir comme effet de rendre significative une valeur propre de plus. En fait, notre choix montre qu'il n'en est rien, car cette valeur propre reste inférieure dans les deux cas à la valeur médiane, et changer de seuil de significativité ne peut de ce fait influer sur le nombre de valeurs propres significatives, pour nos quatre exemples du moins.

La figure 4 les confronte à la séquence des valeurs propres issues de la matrice d'origine : seule la « première » de ces valeurs propres (en fait la deuxième, puisque par construction la 1ère valeur propre de $\mathbf{D^{-1}\,M0}$ est « 1 »), domine les intervalles de confiance pris en séquence décroissante, mettant en évidence la structuration en deux pôles recouvrants imposée aux données. Ce que traduit visuellement la matrice **M0** réordonnée en lignes et en colonnes dans l'ordre des valeurs du vecteur U2 (fig. 5).

On remarquera que cette séquence traverse le « couloir de confiance » des matrices simulées, du fait de la contrainte de constance de la somme algébrique des valeurs propres : ce fait nous a amené à préférer à cette représentation en éboulis la représentation en « entonnoir » qui suit l'ordre d'extraction des éléments propres par l'algorithme QR, décrite en section précédente.

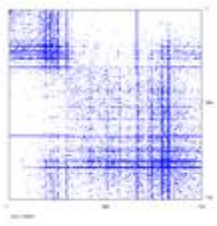

Figure 5. *La structure de **M0** en 2 communautés entremêlées mise en évidence en ordonnant lignes et colonnes selon U2*

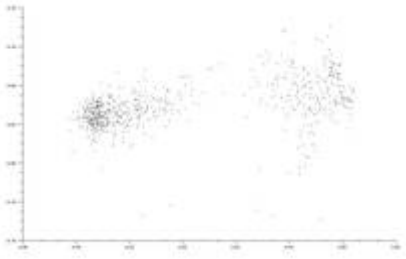

Figure 6. *La projection des 732 sommets dans le plan (U2,U3).*

La figure 6 présente la projection des sommets du graphe dans le plan (U2, U3) et confirme le peu d'information pertinente apportée par U3. L'axe U2 constitue donc la seule dimension intrinsèque du graphe **M0**, et met en évidence un continuum de sommets, indépendamment de leur degrés respectifs, entre les deux noyaux denses des communautés mises en évidence, comme le montre clairement la figure 5.

Une telle structure en communautés recouvrantes peut aussi être vue comme la superposition d'une partition stricte et d'un « bruit » : dans notre cas les deux quadrants anti-diagonaux de la figure 5 délimités par la valeur 0 des projections sur l'axe U2 comportent 1402 liens, soit 27% du total de 6462 liens du graphe. Ceci confirme que notre méthode, comme c'est le plus souvent le cas pour les méthodes spectrales, présente de bonnes capacités de filtrage et de résistance au bruit dans les données.

7. ESPACE INTRINSEQUE DE RESEAUX SOCIAUX REELS : Y A-T-IL DES COMMUNAUTES, ET SI OUI, COMBIEN ?

7.1 Football League

On trouve sur Internet les données (*http://www-personal.umich.edu/~mejn/netdata/*) du graphe social « Football League » des matchs joués entre les 115 clubs de la ligue de football américain pendant la saison 2000 (Girvan, Newman, 2002). Ce jeu de données présente l'intérêt d'inclure la structure sociale "théorique" de ces clubs répartis sur 12 groupements régionaux (dits *conferences*), à rapprocher de la structure réelle qui émane, de façon non supervisée, de l'ensemble des matchs joués. Cette caractéristique peu courante permet de quantifier au moyen d'indicateurs comme le F-score, issus de méthodes supervisées, la qualité des méthodes de détection non-supervisée de communautés dans les graphes.

La figure 7 découle du test TourneBool, aux seuils de confiance de 95% et 99%, sur 200 matrices d'adjacence randomisées (mêmes justifications que pour les exemple précédents) : les 10 "premières" valeurs propres (N°2 à N°11) de la matrice $\mathbf{D}^{-1}\,\mathbf{M0}$ (où **D** est la matrice diagonale des degrés de **M0**) dominent clairement le « couloir » de confiance issu des 200 matrices correspondantes **M1** à **M200** . Toutes les autres à partir de la 11^{ième} (N° 12) se trouvent dans ce couloir. Ce qui nous invite, d'un point de vue purement statistique, à penser que la dimension intrinsèque de ce graphe est 10, alors que du point de vue géométrique, nous pouvions attendre une dimension intrinsèque de 11=12-1.

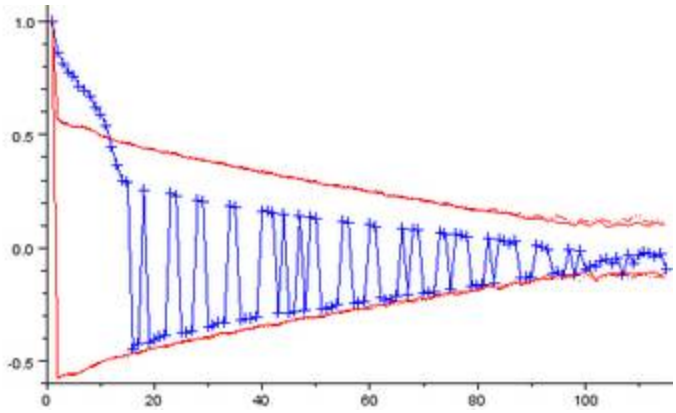

Figure 7. *L'"entonnoir" des valeurs propres issues du graphe social Football-League : en bleu celles de la matrice d'origine, en rouge les bornes des intervalles de variation de ses dérivées randomisées.*

A ces deux points de vue, nous pouvons en ajouter un troisième, qui s'appuie sur l'interprétation visuelle possible dans l'espace des p premiers vecteurs propres pour proposer une valeur adaptée de p.

La figure 8 montre en effet la projection des clubs dans les plans (U2, U3) et (U12, U13), où les matchs intra-conférences sont représentés en traits pleins et les match inter-conférences en traits pointillés. Faute de place, nous ne pouvons montrer les plans intermédiaires, mais il ressort de ces figures que bon nombre de « conférences » apparaissent comme des noyaux visuellement denses, alors que d'autres s'étalent ou se séparent en sous-blocs. Par contraste, les axes U13, U14, etc. ne semblent pas mettre en évidence de structure interprétable. L'espace des 10 ou 11 premiers vecteurs

propres (U2 à U12) “normalise” les phénomènes de groupe, quel que soit le nombre d’individus statistiques impliqués : les petits phénomènes sont mis en exergue autant que les grands, mais ils le sont par des axes de rangs éloignés.

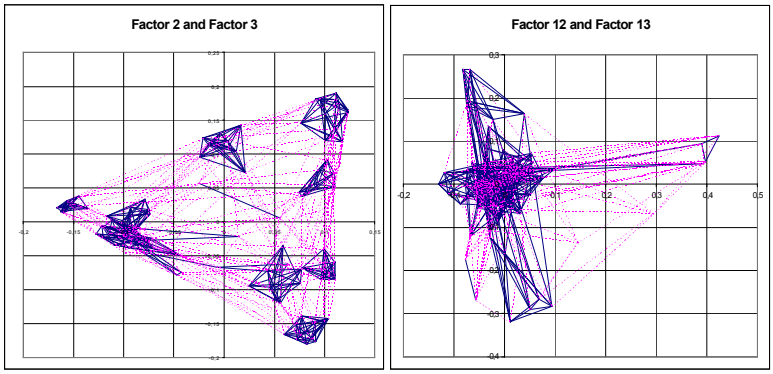

Figure 8. *Les plans (U2,U3) et (U12,U13).*

Nous avons testé quantitativement cette remarque en créant semi-automatiquement un jeu de règles de discrimination dans l’espace 11-dimensionnel U2 à U12 : le tableau 1 montre quelles « conférences » peuvent être précisément, voire parfaitement reconstituées à partir de cet espace, et lesquelles ne le peuvent pas, étant moins géographiquement ancrées, comme *Sun Belt* ou *Independents*. On peut remarquer que l’axe U12 intervient dans ces deux conférences mal reconstituées, ce qui est un argument supplémentaire pour éliminer cet axe et choisir la solution à 10 dimensions au lieu de 11, comme le suggérait l’utilisation du test Tournebool.

Pour terminer, nous avons adopté un quatrième point de vue, celui de la classification. Nous avons réalisé un partitionnement spectral dans l’espace propre du graphe pour évaluer l’accord entre partition non supervisée (les communautés extraites) et partition de référence (les « conférences ») quand on fait varier le nombre de dimensions propres prises en compte. D’où une exigence de stabilité dans les mesures qui nous a fait éliminer les méthodes de type K-means, dépendantes de l’initialisation, pour retenir une méthode densitaire (Lelu, 1994) dénommée Analyse en Composantes Locales (ACL) et utilisant un noyau « cosinus seuillé » : à toute valeur du seuil s correspond un paysage de densité et un seul, vers les sommets duquel montent en gradient un ensemble de vecteurs **u** initialisés par les vecteurs-données **x** (tous ces vecteurs sont normalisés) au moyen de la loi d’apprentissage :

$$\mathbf{u} := \mathbf{u} + \alpha\eta\mathbf{x} \; ; \quad \mathbf{u} := \mathbf{u}/||\mathbf{u}|| \quad \text{(normalisation)}$$

avec $\eta = [\cos(\mathbf{x},\mathbf{u}) - s]^+$ (troncature de la projection de **x** sur **u**), où α est une constante d’apprentissage petite. Par construction, le nombre de maxima diminue par paliers depuis N, nombre de vecteurs-données distincts, pour s= 1, jusqu’à se stabiliser quand on diminue s, à la valeur 1.

Conférences	**Règles**	**Nombre de**			
		T	***TP***	***FP***	***F-score***
0-Atlantic Coast	U5<-0.146	9	9	0	1
1-Big East	U10<-0.136	8	8	0	1
2-Big Ten	U6>0.1 & U3>0.05	11	11	0	1
3-Big Twelve	U4>0.115	12	12	0	1
4-Conference USA	U7<-0.1 & U10>0.1	10	9	0	0.95
5-Independents	U11<-0,01& U12>0.12	5	2	0	0.57
6-Mid-American	U2>0.05 & U3>0.1	13	13	0	1
7-Mountain West	U9>0.16	8	8	0	1
8-Pacific Ten	U2<-0.132	10	10	0	1
9-Southeastern	U3<-0.131	12	12	0	1
10-Sun Belt	3.U11+2.U12>0.7	7	7	3	0,82
11-Western Athletic	U6>0.06 & U7<-0,1	10	8	1	0,84
Accord global		**115**	**109**	**4**	**0,956**

TABLE I. REGLES DE DISCRIMINATION ISSUES DE L’ESPACE 11-DIMENSIONNEL PERTINENT POUR RECONSTITUER LES 12 “CONFERENCES” DE FOOTBALL AMERICAIN.

Pour passer des maxima du paysage de densité à une partition stricte, un deuxième paramètre, l'écart minimal ε entre maxima, doit être défini. Comme nous disposons de la « partition vraie » des données en « conférences », un indice F-score d'accord global entre cette structure et la partition obtenue peut être calculé (à partir du nombre de liens dont les deux nœuds-extrémités sont ou ne sont pas de la même classe), et faire l'objet d'une optimisation en grille sur les paramètres s et ε : ce F-score est maximum (.934) dans l'espace U2 à U12, choisi sur des considérations géométriques il diminue légèrement quand on réduit cet espace (.931 dans l'espace U2 à U11) désigné comme intrinsèque par TourneBool, et plus nettement quand on l'étend (.915 dans l'espace U2 à U13). L'axe U12 n'apporte pratiquement pas d'information utile, ce qui confirme le résultat du test TourneBool.

7.2 Mexican Politician Network

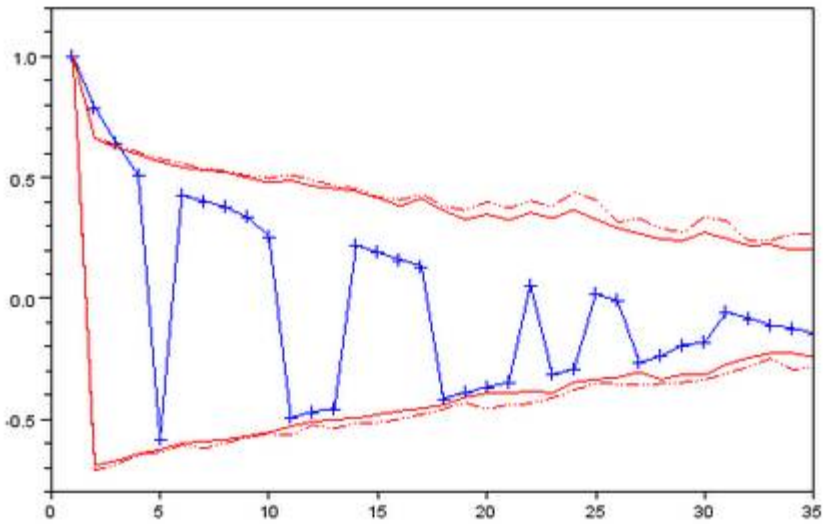

Figure 9. *L'"entonnoir" des valeurs propres issues du graphe social Mexican Politician : en bleu celles de la matrice d'origine, en rouge les bornes des intervalles de variation de ses dérivées randomisées (trait plein :95% ; tirets :99%) .*

On trouve sur Internet le graphe social « Mexican Politician Network » (*http://vlado.fmf.uni-lj.si/pub/networks/data/esna/Mexican. htm*) des liens entre 35 hommes politiques mexicains à la fin du 20e siècle, exploité par (de Nooy *et al.*, 2004) et mis en ligne. Une partie d'entre eux sont des militaires, et il est intéressant d'examiner si cette caractéristique structure ou non le réseau global du pouvoir politique dans ce pays. Les auteurs de (Chen *et al.* 2009) ont comparé sur cet exemple les deux méthodes non-supervisées HMaxMin, maximisant leur critère Max-Min Modularity, et l'algorithme hiérarchique N (Clauset *et al.* 2004) utilisant la notion de modularité simple. Leur critère de comparaison entre les partitions trouvées de façon non-supervisée et la partition de référence militaires/civils est l'*Adjusted Rand Index* (ARI) (Hubert, Arabie 1985).

La figure 9 découle du test TourneBool, aux seuils de confiance de 95% et 99%, sur 2000 matrices d'adjacence randomisées : les deux "premières" valeurs propres (N°2 à N°3, ce graphe ne comportant qu'une seule composante connexe) de la matrice $\mathbf{D^{-1}\,M0}$ dominent le « couloir » de confiance issu des 2000 matrices correspondantes.

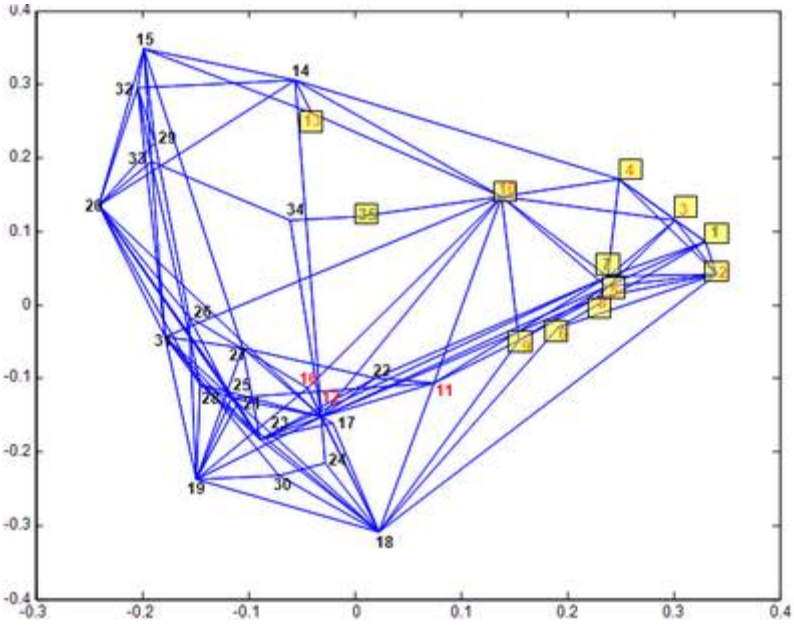

Figure 10. *L'espace intrinsèque (U2,U3) du graphe des liens entre hommes politiques mexicains. En rouge les militaires, en surligné jaune notre partition optimale.*

La figure 10 montre la projection des 35 hommes politiques dans le plan (U2, U3) auquel se limite l'espace intrinsèque de ce graphe. Les arêtes de ce graphe sont représentées en traits plein. Les militaires sont signalés par des numéros rouges : on constate visuellement une organisation en trois pôles, l'un dense et principalement militaire, à l'exception de deux civils qui y sont centraux (F. Madero, E. Portes Gil), et deux autres plus lâches incluant quelques militaires.

L'application de notre méthode ACL dégage ces trois pôles pour une large plage des seuils de cosinus, mais maximise l'indicateur ARI autour de la valeur -0.44 du seuil, pour laquelle deux communautés sont dégagées (celle à dominante militaire est surlignée en jaune sur la figure 10). Ce maximum est de 0.620, et surpasse de beaucoup les meilleures valeurs trouvées par les méthodes HMaxMin et N (0.359 et 0.255 respectivement).

Comme pour le graphe Football League, nous avons comparé les valeurs maximales de notre critère de qualité de classification d'une part en réduisant l'espace à la seule composante U2, puis en l'étendant aux 3 composantes U2 à U4 : les valeurs de l'indicateur ARI sont alors respectivement de 0.512 et 0.604, confirmant ainsi le caractère optimal de l'espace intrinsèque (U2,U3).

8. CONCLUSIONS, PERSPECTIVES

Notre contribution originale a été de montrer que la dimension de l'espace dans lequel réaliser des opérations spectrales sur un graphe, ici la recherche de communautés, pouvait être déterminée par la méthode de génération de matrices randomisées et test statistique TourneBool, adaptée ici au cas de matrices d'adjacence symétriques – ceci de façon plus rigoureuse et plus généralisable à des matrices de grandes dimensions que l'heuristique du « coude » dans l'éboulis des valeurs propres. Nous avons vérifié empiriquement l'optimalité de cet espace intrinsèque sur quatre graphes, les uns artificiels, les autres réels. Le premier, petit et comportant quatre cliques bruitées, nous a permis d'exposer la méthode et de la justifier par une nouvelle façon de représenter la séquence des valeurs propres, comparée à l'« entonnoir » des bornes de celles des matrices simulées. Pour le second, constitué de plus de 700 sommets et construit avec deux communautés recouvrantes, TourneBool a correctement caractérisé comme unidimensionnel l'espace intrinsèque de ce graphe sans échelle. Pour le troisième (« Football League ») notre test a identifié un espace intrinsèque à 10 dimensions dans lequel est confirmée visuellement l'existence des 12 "conférences" officielles, la plupart bien regroupées dans cet espace, d'autres plus transversales. La qualité des résultats a été mesurée et elle n'augmente pas réellement quand on passe de 10 dimensions intrinsèques à 11, et chute quand on passe à 12. Le même constat a pu être fait sur le quatrième graphe, celui des relations entre hommes politiques mexicains de la fin du 20ème siècle, de dimension intrinsèque 2 et dans lequel notre méthode de clustering densitaire ACL a pu trouver une partition en deux communautés beaucoup plus proche de la partition militaires/civils de référence que celles citées dans la littérature.

L'utilisation que nous avons faite de cet espace réduit et « sphéré » (même variance dans toutes les directions) suggère qu'il est possible de répondre à des questions de fond concernant la recherche de communautés dans les graphes, comme : certains nœuds forment-ils des clusters « naturels » (forte cohésion interne et fort éloignement des autres nœuds), et si oui lesquels ? D'autres nœuds ne sont-ils pas mieux décrits, plutôt que par un rattachement plus ou moins arbitraire au cluster le plus proche, par d'autres concepts : communauté mono-élément (= « outlier », sommet marginal), pluri-appartenance, ou représentation sur une échelle continue entre deux pôles (ou plus) ?

En particulier notre méthode densitaire de clustering ACL, peu adaptée aux espaces de données brutes qui recèlent d'importantes différences de densité relative, et de ce fait en jachère depuis une quinzaine d'années, trouve dans ce type d'espace normalisé un terrain idéal. Ce que confirment nos bons résultats d'accord entre la structure en clusters « théorique » et la structure réelle trouvée par ACL sur les graphes « Football League » et « Mexican Politician Network ». Sans compter la capacité de cette méthode à constituer des clusters à divers niveaux de granularité, en fonction des plages de notre paramètre de seuillage des projections.

Au passage, nous avons donc établi deux points moins importants, mais nécessaires à notre démonstration :

- une nouvelle représentation visuelle des séquences de valeurs propres d'une matrice laplacienne d'un graphe, pour les comparer à celles des bornes correspondantes de ses contreparties randomisées.

- notre méthode densitaire de clustering ACL, indépendante des conditions initiales, à la différence des K-means et de leurs dérivées, ne semble pas présenter dans l'espace intrinsèque les inconvénients constatés dans l'espace d'origine des données et dus à des différences considérables de densité des clusters.

Nous sommes conscients que le choix des deux paramètres du test TourneBool (pourcentage de confiance et nombre de matrices simulées) peut influer sur la stabilité et la reproductibilité des résultats, tout en dépendant de la taille et du pourcentage de remplissage du tableau des données. Ceci mérite une étude en soi, qui sort du cadre du présent article. Et conforter par la théorie la remarque 2 de l'algo.

Bien du travail reste à faire pour renforcer et décliner les points établis ci-dessus : par exemple passer d'une échelle d'un millier de noeuds à celle des "graphes de terrain" aujourd'hui disponibles – ce problème se ramène alors à celui de la recherche des éléments propres dominants dans des matrices clairsemées (*sparse*) et de très grandes dimensions, étudié depuis plusieurs décennies et à mettre ici en œuvre techniquement. La référence (Hernandez *et al.*, 2008) montre qu'une architecture parallèle à p processeurs Xeon de l'année 2008 peut traiter des matrices d'adjacence de l'ordre de 100 000 p nœuds. On peut encore tester la robustesse des dimensions intrinsèques obtenues, par injection de « bruit » dans la matrice d'adjacence. Quoiqu'il en soit, le sous-espace délimité par TourneBool semble constituer une base stable pour construire des représentations élaborées d'un graphe, qu'elles soient visuelles ou qu'elles permettent des traitements comme la recherche de communautés et leur évolution au cours du temps.

Remerciements : merci aux relecteurs anonymes pour leurs critiques avisées et leurs nombreuses suggestions de corrections.

9. BIBLIOGRAPHIE

Banerjee, J. On the spectrum of the normalized graph Laplacian, Linear Algebra and its. Applications, 428, 3015-3022, (2008)

Benzécri J.-P. L'analyse des données (3 tomes) Dunod (1973)

Bouveyron, C., Celeux G., Girard S., Intrinsic Dimension Estimation by Maximum Likelihood in Probabilistic PCA, Statistics and Computing 17(4) (2007)

Brin S., Page L., The PageRank hypertextual Web Search Engine, Computer Networks and ISDN Systems, vol. 30, pp. 107-117, (1998)

Cadot M., A simulation technique for extracting robust association rules. In: CSDA (2005)

Cadot M., "Extraire et valider les relations complexes en sciences humaines : statistiques, motifs et règles d'association". PhD thesis, Université de Franche-Comté, (2006)

Cattell R. B., The scree test for the number of factors. Multivariate Behavioral Research, 1(2), (1966). 245-276

Chen J., Zaiane O.R., Goebel R., Detecting Communities in Social Networks using Max-Min Modularity, SIAM International Conference on Data Mining (SDM'09), Sparks, Nevada, USA, April 30- May 2, (2009)

Chung F.R.K., Spectral Graph Theory, (CBMS Regional Conference Series in Mathematics, No. 92), American Mathematical Society, (1997)

Clauset A., Newman M. E. J., Moore C., Finding community structure in very large networks. Physical Review E 70, 066111 (2004)

Cobb G., Chen Y., An application of Markov chain Monte Carlo to community ecology. The American Mathematical Monthly (2003) pp 264–288

Connor E., Simberloff D., The assembly of species communities: Chance or competition? Ecology (1979) pp 1132–1140

de Nooy W., Mrvar A., Batagelj V., *Exploratory Social Network Analysis with Pajek.* Cambridge: Cambridge University Press, 2004), Chapter 12.

Droesbeke J., Finne J., Inférence non-paramétrique – Les statistiques de rangs. Editions de l'Université de Bruxelles (1996)

Fisher R., The use of multiple measurements in taxonomic problems. Annals of Eugenics (1936) pp 179–188

Gionis, A., Mannila, H., Mielikäinen, T., and Tsaparas, P., Assessing data mining results via swap randomization. ACM Trans. Knowl. Discov. Data (2007)

Girvan M. and Newman M. E. J., Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99, 7821–7826 (2002)

Greenacre, Correspondence Analysis In Practice (interdisciplinary Statistics) Chapman & Hall/crc Interdisciplinary Statistics Series, (2007)

Hernandez V., Roman J.E., and Tomàs A., A robust and efficient parallel SVD solver based on restarted Lanczos bidiagonalization. Electronic Transactions on Numerical Analysis,vol.31, pp.68-85, (2008).

Hubert L. and Arabie P., "Comparing partitions". *Journal of Classification* **2** (1): 193–218 (1985).

Lancichinetti A., Fortunato S., Benchmark for testing community detection algorithms on directed and weighted graphs with overlapping communities. Physical Review E 80, 016118 (2009)

Lebart L., Morineau A. et Warwick K., Multivariate Descriptive Statistical Analysis, John Wiley and sons, New-York, (1984)

Lebart L., Correspondence Analysis of Graph Structure - In. Comm. Meeting of the Psychometric Society , Bulletin Technique du CESIA, vol 2, (1984) 5-19

Lelu A., Cadot M., Statistically valid links and anti-links between words and between documents: applying TourneBool randomization test to a Reuters collection. Advances in Knowledge Discovery and Management (AKDM), Ritschard G. et Studer M. eds., pp. 327-344, Springer-Verlag (2010)

Lelu A., Slimming down a high-dimensional binary datatable: relevant eigen-subspace and substantial content. In COMPSTAT 2010 (2010)

Lelu A., Cadot M.. Graphes des liens et anti-liens statistiquement valides entre les mots d'un corpus textuel, Extraction et gestion de connaissance 2009 (EGC'09), (2009)

Lelu, A.: Clusters and factors: neural algorithms for a novel representation of huge and highly multidimensional data sets. In: Diday E., Lechevallier Y. & al. (eds): New Approaches in Classification and Data Analysis, 241-248 Springer-Verlag, Berlin (1994).

Manly B., Randomization, Bootstrap and Monte Carlo methods in Biology. Chapman and Hall/CRC (1997)

Milo, Shen-Orr, Itzkovitz, Kashtan, Chklovskii, and Alon, Network Motifs: Simple Building Blocks of. Complex Networks, Science vol. 298 (2002)

Molloy M., Reed B., A critical point for random graphs with a given degree sequence. Random structures and Algorithms, vol. 6 (2/3):161-180 (1995)

Ryser H., Recent Advances in Matrix Theory. Madison (1964)

Snijders T., Enumeration and simulation methods for 0-1 matrices with given marginals. Psychometrika (2004) pp 397–417

Viger F., Latapy M., Random generation of large connected simple graphs with prescribed degree distribution. LNCS, COCOON (11-th international conference Computing and Combinatorics), Kunming, Yunnan, China. (2005).

Von Luxburg L., A Tutorial on Spectral Clustering. Statistics and Computing 17(4) : 2007