
Extraction de structures macroscopiques dans des grands graphes par une approche spectrale

B. Jouve* - P. Kuntz - F. Velin****

* GRIMM- Département de Mathématiques et Informatique
Université Toulouse Le Mirail
5, allées A. Machado
31058 Toulouse cedex
jouve@univ-tlse2.fr

** IRIN - Ecole Polytechnique de l'Université de Nantes
La Chantrerie - BP 60609
44306 Nantes cedex 3
{Pascale.Kuntz, François.Velin}@polytech.univ-nantes.fr

RÉSUMÉ. Dans de nombreux domaines dont le Web est un exemple paradigmatique, la croissance continue de la taille des graphes de relations mis en jeu nécessite, préalablement à l'application d'algorithmes de fouille ou de visualisation spécifiques, la décomposition des graphes en leurs principales composantes "macroscopiques". Les méthodes spectrales consistent à plonger le graphe dans un espace euclidien de sorte que les sommets fortement reliés soient représentés dans une même partie de l'espace et les sommets sans ou avec peu de connections soient représentés à distance. Nous nous focalisons ici sur une méthode factorielle et présentons dans un cadre unifié les différentes versions d'une distance bien adaptée pour les cas des graphes non orientés, orientés et pondérés.

ABSTRACT. In numerous fields whose WWW is a paradigmatic example, the size of the considered relationship graphs is continually growing and, often requires a step of decomposition of the graphs into their main macroscopic components preliminary to the application of specific complex mining algorithms. Spectral methods consist in embedding the graph in a Euclidean space such that strongly connected vertices are represented by close points in the space whereas those with few or no connections are distant. We here focus on a factor analysis approach and we present the different versions of a well-adapted distance for oriented, non-oriented and weighted graphs in a unified approach

MOTS-CLÉS : graphe, dissimilarité, plongement euclidien, analyse factorielle, Web mining

KEY WORDS: graph, dissimilarity, Euclidean embedding, factor analysis, Web mining

1. Introduction

Un grand nombre de méthodes d'extraction de connaissances (ECD) ont été développées initialement pour des n-uplets de base de données relationnelles décrivant les caractéristiques des objets de la base par des attributs. Si les problèmes liés à cette classe de données dominent encore en partie la production scientifique du domaine, l'analyse des données "structurées" de grande taille devient un enjeu majeur. Parmi celles-ci, on trouve en particulier les réseaux de relations modélisés par des graphes, un des principaux domaines d'application étant le Web (ex. [CHA00], [BRO00]). Les tailles des réseaux considérés peuvent défier les calculs et interdire toutes représentations imagées intelligibles de par leurs seules dimensions. Pour pouvoir les appréhender et appliquer des algorithmes spécifiques de complexité parfois élevée, il est souvent nécessaire d'en extraire préalablement les principales composantes "macroscopiques". Il s'agit alors de décomposer le graphe initial en sous-graphes de tailles praticables telles que la cohésion locale de chacun des sous-graphes (abondance des relations intra-graphe, ...) et la qualité de séparation des sous-graphes (rareté des liaisons inter-graphes ...) soient assurées.

Les approches spectrales ont connu ces dernières années un regain d'intérêt pour ce type de problème. Etant donné un graphe $G = (V, E)$ avec un ensemble de sommets V et un ensemble d'arêtes E –ou d'arcs selon que l'on considère ou non une orientation–, l'idée consiste à plonger G dans un espace géométrique X , souvent euclidien, de sorte que les sommets fortement reliés soient représentés dans une même partie de l'espace et les sommets sans ou avec peu de connections soient représentés à distance. Les coordonnées des sommets définissant le plongement dans X sont déduites de la décomposition spectrale d'une matrice de produit scalaire. Selon les champs d'applications, on peut schématiquement distinguer deux grandes approches : celles basées sur la décomposition du Laplacien discret du graphe et les méthodes factorielles.

1.1. Les approches spectrales

Les premières ont été initialement développées dans le domaine de la Conception Assistée par Ordinateur pour le partitionnement et le placement de circuits VLSI (ex. [ALP95]). Des analogies entre les problèmes posés par ces technologies et ceux rencontrés actuellement pour l'analyse structurelle de liens entre sites ont conduit récemment à leur utilisation en ECD ([DIN 01], [VEL01]). Rappelons que le Laplacien discret d'un graphe G décrit par sa matrice d'adjacence A est défini par $Q = De - A$ où De est la matrice diagonale des degrés. Pour les définitions sur les graphes nous renvoyons à [BER 83]. Les coordonnées des sommets de V dans l'espace X de dimension k sont données par les vecteurs propres $\mu_0, \mu_1, \dots, \mu_k$ associées aux plus petites valeurs propres $0 \leq \lambda_0 \leq \lambda_1 \leq \dots \leq \lambda_k$ de Q . Différentes propriétés du Laplacien permettent de justifier cette utilisation dans le cadre défini plus haut ([HAL70], [CHU97]). En particulier, on peut montrer que pour un vecteur colonne arbitraire x de composantes x_1, \dots, x_n alors

$$\lambda_k = \underset{x \perp P_{k-1}}{\text{Min}} \frac{\sum_{(i,j) \in E} (x_i - x_j)^2}{\sum_i x_i^2} \quad [1]$$

où P_{k-1} est le sous-espace engendré par les vecteurs propres μ_i pour $i \leq k-1$ [CHU97]. Ainsi, les composantes des premiers vecteurs propres sont proches selon la métrique euclidienne usuelle pour les sommets adjacents et éloignées pour des sommets sans liens.

Les approches “ factorielles ” consistent à définir préalablement sur l’ensemble des sommets une mesure de dissimilarité d –ou de similarité– qui rende compte de la densité des relations entre les sommets et à plonger le graphe dans X en optimisant un critère d’adéquation entre la dissimilarité initiale d et celle associée à l’espace métrique X . Bien que les applications de ce type d’approche pour l’analyse de graphes ne soient pas récentes (ex. [LEB 84], [BEN 73], [TIN 71]), leur intérêt a été relancé pour l’extraction de structures dans les grands réseaux par des recherches menées ces dernières années dans des domaines variés (en neurosciences [JOU98b], [SIM94], [YOU92], en CAO [KUN00], en réseaux sociaux [RIC 97], en linguistique [PLO 98]).

1.2. Contribution

Dans cette communication nous nous restreignons à cette dernière approche en nous focalisant sur les plongements euclidiens. Dans une première partie, nous rappelons brièvement les résultats classiques d’algèbre linéaire qui sous-tendent ces représentations. Dans une seconde partie, nous discutons du choix d’une dissimilarité d pour le problème de l’extraction de composantes dans des graphes et présentons dans un cadre unifié les différentes versions d’une distance bien adaptée pour les cas non orienté, orienté et pondéré. Dans une troisième partie, nous synthétisons et complétons différents résultats théoriques et expérimentaux épars qui permettent de caractériser les plongements de graphes obtenus dans les différents cas.

2. Espace de représentation euclidien - Rappels

Dans la suite, nous considérons qu’une dissimilarité d est définie sur l’ensemble V des n sommets du graphe G ; pour chaque couple de sommets $(i, j) \in V \times V$, $d(i, j)$ est positive, symétrique ($d(i, j) = d(j, i)$) et telle que $d(i, i) = 0$. Nous rappelons que d est une *distance* si elle est définie c-à-d. si $\forall (i, j) \in V \times V, i \neq j \Rightarrow d(i, j) \neq 0$ et vérifie l’inégalité triangulaire $\forall (i, j, k) \in V^3, d(i, j) \leq d(i, k) + d(k, j)$. Si la première condition n’est pas vérifiée d est une semi-distance.

Nous nous restreignons ici au cas où l'espace X de plongement du graphe est un espace euclidien. Dans la pratique, d'autres espaces de représentations peuvent être considérés, comme l'espace l_1 où la métrique définie sur l'espace est une distance de Manhattan mais les problèmes de plongement et d'approximation sont généralement plus complexes [DEZ 97].

Définition 1. Une dissimilarité d sur V est une *distance Euclidienne* s'il existe un ensemble $\{P_i\}_{i=1,n}$ de points dans un espace Euclidien (X, δ) tel que la distance $\delta(P_i, P_j)$ entre n'importe quel couple de points soit égale à $d(i, j)$ dans cet espace. La dimension de d est la dimension de l'espace vectoriel défini par les points P_i . Il est évident que cette dimension est majorée par $n - 1$.

Si $n = 3$ une représentation euclidienne (ou plongement isométrique) de (V, d) est possible lorsque d est une distance. En revanche, pour $n > 3$, cette condition n'est plus suffisante. Différentes conditions ont été formulées (ex. [GOW 82]); nous retenons ici pour des facilités de calculs et d'interprétation celle fournie par Torgerson [TOR 58]. Celle-ci repose sur la matrice W définie par

$$W_{ij} = \frac{1}{2} (d^2(i, \cdot) + d^2(j, \cdot) - d^2(\cdot, \cdot) - d^2(i, j))$$

où

$$d^2(i, \cdot) = \frac{1}{n} \sum_{j=1,n} d^2(i, j) \quad \text{et} \quad d^2(\cdot, \cdot) = \frac{1}{n} \sum_{i=1,n} d^2(i, \cdot) \quad [2]$$

Il est bien connu en Analyse Factorielle que d est Euclidienne si et seulement si W est semi-définie positive. Dans ce cas, ses valeurs propres sont positives et, si on note $\mu_1, \mu_2, \dots, \mu_p$ une base orthonormée de vecteurs propres de W associés aux p valeurs propres strictement positives $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ ($p \leq n - 1$) alors

$$d^2(i, j) = \sum_{k=1,p} \lambda_k (\mu_{i,k} - \mu_{j,k})^2 \quad [3]$$

3. Choix d'une distance sur le graphe

D'après [2], les caractéristiques du plongement du graphe sont intrinsèquement liées au choix de d . Une des distances parmi celles les plus utilisées sur les graphes est celle du plus court chemin. Cette distance est bien adaptée pour représenter certains graphes ayant de nombreuses symétries comme les graphes de permutation par exemple. En revanche, elle ne permet pas de mettre en évidence distinctement des classes ; sa valeur est définie par une seule "liaison" entre les sommets et non par une "densité" de liens. De plus, on peut montrer facilement que la distance du plus court chemin est euclidienne uniquement pour des chaînes élémentaires ou des graphes complets. Pour ces raisons nous nous focalisons ici sur les distances qui

tiennent compte des relations locales de chacun des sommets. Partant d'une distance sur un graphe simple, nous discutons ensuite d'une extension au cas orienté puis au cas pondéré.

3.1. Cas des graphes simples

Un graphe simple $G=(V,E)$ pouvant être décrit par une matrice d'adjacence binaire A , les indices dits de présence-absence de la littérature taxonomique peuvent *a priori* s'appliquer. Soit a_i le vecteur binaire associé à la i -ème ligne de A , la dissimilarité entre i et j est une fonction des paramètres a , b , c et d suivants :

$$a = \langle a_i, a_j \rangle, b = \langle a_i, \mathbf{1} - a_j \rangle, c = \langle \mathbf{1} - a_i, a_j \rangle, \text{ et } d = \langle \mathbf{1} - a_i, \mathbf{1} - a_j \rangle \quad [4]$$

où \langle, \rangle est un produit scalaire sur \mathbb{R}^n et $\mathbf{1}$ le vecteur unité. Les entiers a et d dénombrent la présence et l'absence simultanée de 1 dans les deux vecteurs a_i et a_j , alors que b et c dénombrent les occurrences où la présence de 1 dans un des 2 vecteurs est associée à une absence dans l'autre. Hubalek [HUB82] a dénombré plus de quarante indices de ce type dans des domaines d'application très divers. Un grand nombre d'entre eux (Jaccard, Ochai, Russel-Rao, *etc.*) peuvent être regroupés dans deux familles

$$d_\alpha(i, j) = 1 - \frac{a}{m_\alpha(a+b, a+c)} \quad \text{et} \quad d_\theta(i, j) = \frac{b+c}{\theta a + b+c} \quad [5]$$

où

$$m_\alpha(a+b, a+c) = \left(\frac{(a+b)^\alpha + (a+c)^\alpha}{2} \right)^{1/\alpha}$$

est la moyenne de Cauchy entre les quantités $a+b$ et $a+c$, représentant respectivement les nombres de 1 dans a_i et a_j .

Les dissimilarités d_α et d_θ ne sont pas euclidiennes – et ne sont même pas toujours des distances- mais leurs racines carrées sont des semi-distances euclidiennes pour $\alpha \geq 0$ et $\theta \leq 2$ ([CAI 96], [GOW 86]).

Outre ses propriétés géométriques, une analyse expérimentale des distances associées aux valeurs entières de α et θ sur des graphes nous a conduit à privilégier la racine carrée de l'indice d_α avec $\alpha = 1$ – appelé indice de Czekanowski-Dice – associé à la moyenne arithmétique m_1 .

Cet indice peut s'interpréter sur un graphe en fonction de la relation de voisinage.

Soit $V(i)$ l'ensemble des voisins de i sur G : $V(i) = \{j \in V ; (i,j) \in E\}$. Alors,

$$d_1(i, j) = 1 - \frac{a}{((a+b) + (a+c))/2} = \frac{b+c}{(a+b) + (a+c)} = \frac{|V(i)\Delta V(j)|}{|V(i)| + |V(j)|} \quad [6]$$

où Δ est l'opérateur de la différence symétrique.

Afin que les sommets d'un graphe complet soient à distance nulle, nous supposons qu'il existe une boucle en chaque sommet : $\forall i \in V, (i, i) \in E$. Cette hypothèse permet aussi de pouvoir adapter les définitions des distances d_α et d_θ aux graphes orientés avec la seule condition qu'ils soient connexes, sans risque de division par 0 (voir 3.2). L'indice d_l rend bien compte de la densité locale : deux sommets sont proches si et seulement s'ils ont de nombreux voisins communs et peu de différents.

3.2. Cas des graphes simples orientés

En présence de graphes orientés, $V^+(i)$ et $V^-(i)$ désignent respectivement l'ensemble $\{j \in V, (i, j) \in E\}$ des successeurs et $\{j \in V, (j, i) \in E\}$ des prédécesseurs d'un sommet i d'un graphe orienté $G=(V, E)$.

De la même manière que pour les graphes non orientés, un graphe simple orienté est décrit par une matrice binaire A mais non nécessairement symétrique. Il est possible d'adapter les dissimilarités précédentes en considérant simultanément A comme une table de successeurs et de prédécesseurs des sommets, ce qui consiste à utiliser les résultats du 3.1. sur A et A' où A' est la transposée de A . Comme pour A , on définit pour A' les nombres a', b', c' et d' par les définitions [4] et les familles d'indices de dissimilarité d_α et d_θ dont peut être muni V s'écrivent alors

$$d_\alpha = \frac{1}{2}(d_\alpha^+ + d_\alpha^-) \text{ et } d_\theta = \frac{1}{2}(d_\theta^+ + d_\theta^-) \quad [7]$$

avec d'une part,

$$d_\alpha^+(i, j) = 1 - \frac{a}{m_\alpha(a+b, a+c)} \text{ et } d_\alpha^-(i, j) = 1 - \frac{a'}{m_\alpha(a'+b', a'+c')}$$

et d'autre part,

$$d_\theta^+(i, j) = \frac{b+c}{\theta a+b+c} \text{ et } d_\theta^-(i, j) = \frac{b'+c'}{\theta a'+b'+c'}$$

On notera que si le graphe orienté est symétrique alors $d^+ = d^-$ et on retrouve les mêmes expressions que dans le cas non orienté.

Si W^+ et W^- sont les matrices de Torgerson associées à $\sqrt{d_\alpha^+}$ et $\sqrt{d_\alpha^-}$ (resp. $\sqrt{d_\theta^+}$ et $\sqrt{d_\theta^-}$), $W = \frac{1}{2} \cdot (W^+ + W^-)$ est la matrice de Torgerson associée à $\sqrt{d_\alpha}$ (resp. $\sqrt{d_\theta}$). La semi-positivité éventuelle des deux matrices W^+ et W^- entraîne donc celle de W . Ceci permet notamment de conclure au caractère Euclidien de $\sqrt{d_1}$ qui est ici définie par

$$d_1^+(i, j) = 1 - \frac{a}{((a+b)+(a+c))/2} = \frac{|V^+(i)\Delta V^+(j)|}{|V^+(i)|+|V^+(j)|} \text{ et } d_1^-(i, j) = \frac{|V^-(i)\Delta V^-(j)|}{|V^-(i)|+|V^-(j)|} \quad [8]$$

Deux sommets sont proches suivant d_1 si et seulement s'ils ont de nombreux successeurs communs et peu de différents, ainsi que de nombreux prédécesseurs communs et peu de différents.

3.3. Cas des graphes simples pondérés

Une extension assez naturelle de d_1 peut être proposée dans le cas de graphes simples pondérés. Reprenons la formule 5 pour $\alpha = 1$ avec les définitions d'origine de a , b et c :

$$d_1(i, j) = 1 - \frac{2 \cdot \langle a_i, a_j \rangle}{\|a_i\|^2 + \|a_j\|^2} \quad [9]$$

où $\langle a_i, a_j \rangle = \sum a_{ik} a_{jk}$ et $\|a_i\|^2 = \sum a_i^2$.

Dans le cas pondéré les coefficients a_{ij} de la matrice d'adjacence valent 0 si $(i, j) \notin E$ ou le poids de (i, j) sinon. Si les nombres a_i sont binaires on retrouve l'expression de d_1 donnée dans le cas des graphes simples. Définir l'extension de d_1 de cette façon permet de conserver certaines propriétés géométriques, notamment le caractère Euclidien de $\sqrt{d_1}$. En effet, en introduisant la similarité $s_1 = 1 - d_1$ on peut utiliser les techniques de démonstration présentées par exemple dans [GOW 86] pour prouver que s_1 est semi-définie positive, et donc que $\sqrt{d_1}$ est Euclidienne.

4. Quelques caractéristiques

Différentes caractéristiques ont été mises en évidence sur des classes de graphes définies sur la figure 1. Dans la synthèse présentée par cette figure, nous introduisons, outre l'orientation et la pondération, un troisième paramètre qui est la densité d'arêtes (ou d'arcs) et qui joue un rôle majeur dans l'analyse spectrale des graphes.

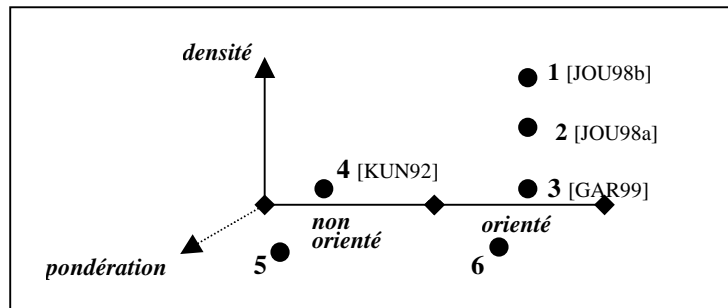


Figure 1. Récapitulatif des différentes classes de graphes analysées

La densité est définie par le rapport du nombre d'arêtes (ou d'arcs) sur le nombre de sommets. Dans l'étude des graphes peu denses, la notion de composante dense est centrale.

Définition 2. Une composante dense d'un graphe est un sous-graphe connexe dont la densité d'arêtes (ou d'arcs) est supérieure à celle du graphe et qui est maximal pour cette propriété.

Qu'il s'agisse du cas avec ou sans orientation, la dissimilarité d_i est non graduée : les valeurs extrémales valent toutes 1. En effet, $d_i(i, j)$ est uniquement fonction des voisinages de i et j et vaut 1 dès lors que i et j ne sont pas adjacents et n'ont pas de voisins communs. Ainsi, les composantes denses du graphe sont *grosso modo* distribuées sur une hypersphère de \mathbb{R}^p . On observe que la dimension p est de l'ordre du nombre de composantes moins 1. A titre illustratif, la figure 2 montre le plongement obtenu dans \mathbb{R}^3 pour un graphe d'une classe de graphes pseudo-aléatoires $G_{GAR}(k, nv, p_{int}, p_{ext})$ où k est le nombre attendu de composantes de nv sommets et p_{int} (resp. p_{ext}) la probabilité d'une arête $\{i, j\}$ si i et j appartiennent à la même composante (resp. si i et j appartiennent à des composantes différentes), toutes les arêtes étant choisies indépendamment [GAR 90].

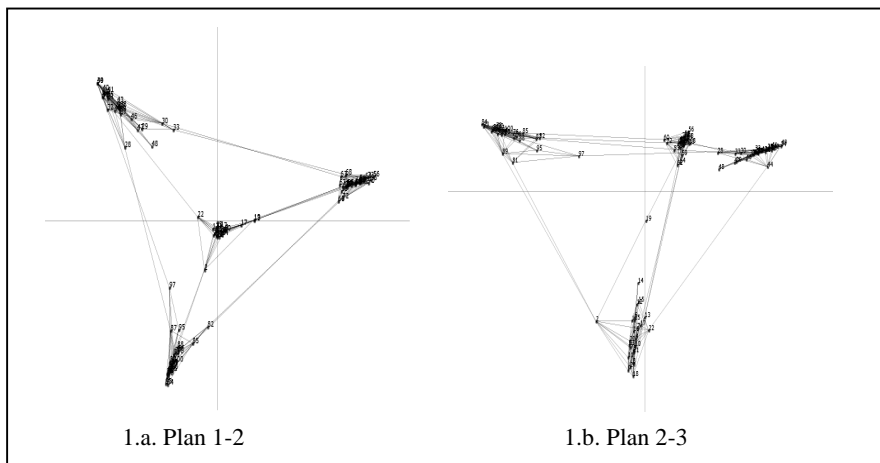


Figure 2. Plongement d'un graphe $G=G_{GAR}(4, 10, 0.35, 0.004)$ dans \mathbb{R}^3 avec $\sqrt{d_i}$. Cette représentation est la restriction sur l'espace des 3 premiers axes principaux d'inertie d'un plongement isométrique de G sur \mathbb{R}^{39} muni de la distance euclidienne usuelle.

De la même manière, les plongements proposés ici sont adaptés à l'étude des "graphes de petits mondes", qui sont des graphes de faible diamètre, peu denses, et constitués de plusieurs composantes denses. Cette classe de graphes, dont on commence à fournir de bons modèles [WAT 99], semble être la base de nombreux réseaux relationnels réels comme le World Wide Web [ADA 99].

4.1. Graphes orientés

Dans certains cas de graphes orientés, la forme du plongement peut donner des indications sur la présence éventuelle de propriétés "macroscopiques" portées par l'orientation en particulier les sources et les puits. Rappelons qu'un sous-graphe est une source (resp. un puits) si aucun arc n'y entre (resp. n'en sort). Ferré et Jouve [FER01] ont montré récemment, pour des graphes orientés dont toutes les composantes denses sont des ensembles d'articulation, comme par exemple sur le graphe de la figure 3.b., que la densité d'une composante dense diminue si sa configuration se rapproche de celle d'une source ou d'un puits. Dans le plongement cette information semble principalement portée par les dimensions supérieures à p si le nombre de composantes est égal à $p+1$.

Cette situation est illustrée sur la figure 3. Comme pour les graphes simples on construit des graphes pseudo-aléatoires $G_{GAR}(k, nv, p_{in}, p_{out})$ orientés. Pour illustrer la mise en évidence d'un ensemble puits, on va, successivement dans (a) et (b), imposer quelques contraintes aux graphes G_{GAR} utilisés :

a - Cas d'un graphe dense orienté : on considère le plongement dans \mathbb{R}^2 de $\tilde{G}_{GAR}(4, 10, 0.6, 0.3)$ construit à partir de $G_{GAR}(4, 10, 0.6, 0.3)$ et tel que la composante constituée des sommets $\{31, \dots, 40\}$ soit un puits. Cet ensemble puits est révélé dans le plongement par le premier axe (figure 3.a.).

b - Cas d'un graphe non-dense orienté : à partir de $G_{GAR}(4, 10, 0.6, 0.2)$, on construit une chenille de composantes denses comme indiqué en haut de la figure 3.b. La première contrainte impose que la composante dense numéro 7 soit un puits, les autres contraintes étant les absences de connections entre certaines composantes. Notons que ces dernières contraintes transforment $G_{GAR}(4, 10, 0.6, 0.2)$ en un graphe non dense. Dans un processus d'agrégation des sommets à l'aide d'une classification ascendante hiérarchique avec le critère du saut minimum sur le plongement global, les puits et les sources se distinguent en s'agrégeant avant les autres composantes.

4.2. Vers une extension aux graphes pondérés

Nous étudions actuellement l'extension des résultats obtenus sur les graphes binaires à une classe de graphes "faiblement" pondérés. Plus précisément, il s'agit de graphes $G(\varepsilon)$ dont la pondération peut être vue comme une petite perturbation de graphes binaires G .

Dans ce cas, si $A(\varepsilon)$ et A sont respectivement les matrices d'adjacence de $G(\varepsilon)$ et G , on peut écrire $\tilde{A} = A + \varepsilon U$ où

$$\varepsilon = \frac{\|A(\varepsilon) - A\|}{\|A(\varepsilon) - A\| + \|A\|} \in [0;1] \quad [10]$$

est la perturbation et U une matrice dépendant de ε . Cette perturbation affecte aussi la matrice des dissimilarités entre les sommets du graphe. Mais pour une petite perturbation, c'est-à-dire pour ε petit, et dans les conditions d'application de la théorie matricielle des perturbations (ex. [KAT 66]), les valeurs et vecteurs propres de la matrice de Torgerson sont holomorphes et les résultats sur les graphes binaires peuvent s'étendre alors par continuité.

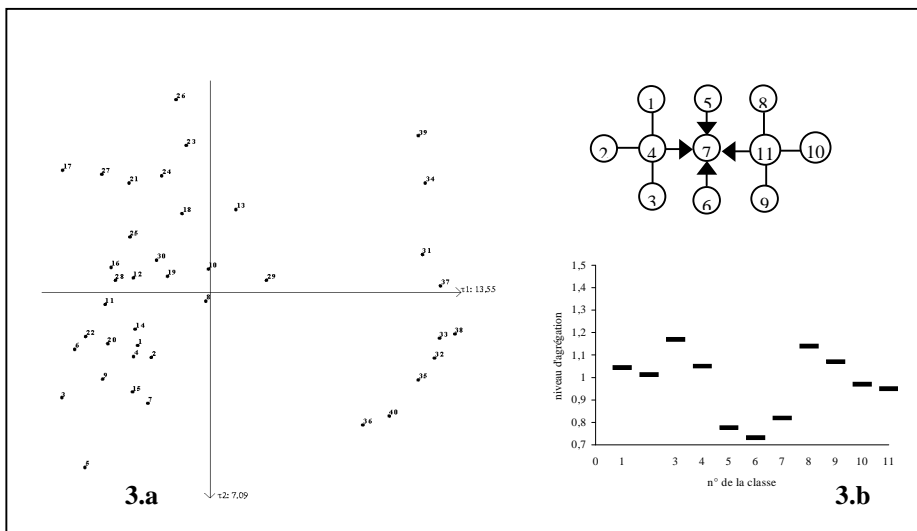


Figure 3. Plongements de graphes orientés et mise en évidence des puits

3.a. Plongement d'un graphe dense $G=G_{GAR}(4, 10, 0.6, 0.2)$ avec une composante puits $\{31, \dots, 40\}$. Celle-ci est révélée sur le premier axe (à droite).

3.b. Plongement d'un graphe dont les composantes sont organisées par une structure en chenille avec un puits. Le graphique représente le niveau d'agrégation des classes par une CAH à saut minimum dans l'espace global de plongement. On retrouve les composantes denses, et les puits et les sources se distinguent en s'agrégeant avant les autres composantes.

Bibliographie

- [ADA 99] ADAMIC L.A., “The small world web”, <http://www.parc.xerox.com/istl/groups/iea/www/smallworld.html>.
- [ALP 95] ALPERT C., KAHNG A. “Recent directions in netlist partitioning : a survey”, *Integration*, vol. 19, 1995, p. 1-81.
- [BEN 73] BENZECRI J.P., *L'analyse des données – Tome 2*, Dunod, Paris, 1973.
- [BER 83] BERGE C., *Graphes*, Gauthier-Villars, Paris, 1983.
- [BRO 00] BROUDER C., KUMAR A. R., *et al.*, “Graph structure in the Web”, 9th International World Wide Web Conference, Foretec Seminar, 2000.
- [CAI 96] CAILLIEZ F., KUNTZ P., “Contribution to the study of the metric and Euclidean structures of dissimilarity”, *Psychometrika*, n° 61, 1996, p. 241-253.
- [CHA 00] CHAKRABARTI S., “Data mining for hypertext : a tutorial survey”, *SIGKDD Explorations*, vol. 1, n° 2, 2000, p. 1-11.
- [CHU 97] CHUNG F.R.K., *Spectral graph theory*, Regional Conf. Series in Mathematics, N°92, American Mathematical Society, 1997.
- [DEZ 97] DEZA M., LAURENT M., *Geometry of cuts and metrics*, Springer, 1997.
- [DIN 01] DING C., HE X. “A spectral method to separate disconnected and nearly-disconnected Web graph components”, *Proc. of the 7th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2001, p. 275-280.
- [FER 01] FERRE L., JOUVE B., “An example of vertex partitioning of a digraph for searching a pseudo sink set”, *Rapport Interne*.
- [GAR 90] GARBERS J., PROMEL H., STEGER A., “Finding clusters in VLSI circuits”, *IEEE Int. Conf. on Computer Aided Design*, 1990, p. 520-523.
- [GAR 99] GARGALLO Y., JOUVE B., “Le Web vu comme un réseau orienté”, *Colloque Comprendre les usages d'Internet*, 1999, ENS, Paris.
- [GOW 66] GOWER J.C., “Some distance properties of latent root and vector methods used in multivariate data analysis”, *Biometrika*, vol. 53, 1966, p. 315-328.
- [GOW 82] GOWER J., “Euclidean distance geometry”, *Math. Scientist*, vol. 7, 1982, p. 1-14.
- [GOW 86] GOWER J., LEGENDRE P., “Metric and Euclidean structures of dissimilarity coefficients”, *J. of Classification*, vol. 3, 1984, p. 5-48.
- [HAL 70] HALL K.M., “An r-dimensional quadratic placement algorithm”, *Management Science*, vol. 17, n° 3, 1970, p. 219-229.
- [HUB 82] HUBALEK Z., “Coefficients of association and similarity based on (presence, absence) : an evaluation”, *Biological Rev.*, vol. 57, 1982, p. 669-689.

- [JOU 98a] JOUVE B., "A new partitioning of large tournaments", *Rapport du CAMS*, Ecole des Hautes Etudes en Sciences Sociales, n°157, 1998.
- [JOU 98b] JOUVE B., ROSENSTIEHL P., IMBERT M., "A mathematical approach to the connectivity between the cortical areas of the macaque monkey", *Cerebral Cortex*, vol. 8, 1998, p. 28-39.
- [KUN 92] KUNTZ P., "Représentation euclidienne d'un graphe abstrait en vue de sa segmentation", *Thèse*, 1992, Ecole des Hautes Etudes en Sciences Sociales, Paris.
- [KUN 00] KUNTZ P., HENNAUX F., "Numerical comparisons of two spectral decompositions for vertex clustering", *Data Analysis, Classification and Related Methods- Proc. of IFCS'2000*, 2000, Springer Verlag, p. 581-586.
- [LEB 84] LEBART L., "Correspondence analysis of graph structures", *Bulletin technique du CESIA*, vol. 2, n° 1-2, 1984, p. 5-19.
- [PLO 98] PLOUX S., VICTORRI B., "Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes", *Traitement Automatique des Langues*, vol. 39, n°1, 1998, p. 161-182.
- [RIC 97] RICHARDS W.D., SEARY A.J., "Convergence analysis of communication networks", *Working paper*, 1997, S. Fraser University, Burnaby, Canada.
- [SIM 94] SIMMEN M., GOODHILL G., WILISHAW D., "Scaling and brain connectivity", *Nature*, vol. 369, 1994, p. 448-450.
- [TIN 71] TINKLER K.J., "The physical interpretation of eigenfunctions of dichotomous matrices", *Inst. Br. Geog. Trans.*, vol. 369, 1971, p. 17-46.
- [TOR 58] TORGERSON W.S., *Theory and methods of scaling*, Wiley, 1958.
- [VEL 01] VELIN F., KUNTZ P., BRIAND H., "Web cartography for online site promotion : an algorithm for clustering Web resources", *Proc. of the IEEE Int. Conf. on Data Mining*, accepté.
- [WAT 99] WATTS D.J., *Small Worlds : the dynamics of networks between order and randomness*, Princeton : Princeton University Press, 1999.
- [WEL 75] WELLES J., WILLIAMS H., *Embeddings and extensions in analysis*, Berlin, Springer-Verlag, 1975
- [YOU 92] YOUNG M.P., "Objective analysis of the topological organization of the primate cortical visual system", *Nature*, vol. 358, 1992, p. 152-155.
-