

Matrices de représentations généralisées, mesures spectrales et distances statistiques pour l'analyse et la classification de graphes et de signaux

Tristan Averty

▶ **To cite this version:**

Tristan Averty. Matrices de représentations généralisées, mesures spectrales et distances statistiques pour l'analyse et la classification de graphes et de signaux. Mathématique discrète [cs.DM]. Université de Brest, 2025. Français. ⟨NNT : 2025BRUN0017⟩. ⟨tel-05462154⟩

HAL Id: tel-05462154

https://theses.hal.science/tel-05462154v1

Submitted on 16 Jan 2026

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.



HAL Authorization

COLLEGE SCIENCES
DOCTORAL DE LA MER
BRETAGNE ET DU LITTORAL

Université de Bretagne Occidentale

Thèse de doctorat de

l'Université de Bretagne Occidentale

École Doctorale N° 598
Sciences de la Mer et du Littoral
Spécialité : *Observation de l'Environnement marin et Traitement de l'Information*

Par
Tristan AVERTY

Matrices de représentation généralisées, mesures spectrales et distances statistiques pour l'analyse et la classification de graphes et de signaux

Thèse présentée et soutenue à l'École navale (Lanvéoc), le 28 avril 2025
Unité de recherche : Institut de Recherche et d'Études Navales (IRENav), UR 3634

Rapporteurs avant soutenance :

Pierre BORGNAT	DR CNRS, ENS Lyon, Laboratoire de Physique (UMR 5672)
Nicolas TREMBLAY	CR CNRS, UGA, GIPSA-LAB (UMR 5216)

Composition du jury :

Président :	Vincent GRIPON	Professeur, IMT Atlantique, Lab-STICC (UMR 6285)
Examinateurs :	Sophie ACHARD	DR CNRS, UGA, Laboratoire Jean Kuntzmann (UMR 5224)
	Marc BARTHELEMY	DR CEA, CEA, Université Paris-Saclay, IPhT (UMR 3681)
	Thierry CHONAVEL	Professeur, IMT Atlantique, Lab-STICC (UMR 6285)
	Nicolas KERIVEN	CR CNRS, IRISA (UMR 6074)
Dir. de thèse :	Abdel-Ouahab BOUDRAA	PU, École Navale / Arts & Métiers ParisTech, IRENav (UR 3634)
Encad. de thèse :	Delphine DARÉ-EMZIVAT	MCF, École Navale / Arts & Métiers ParisTech, IRENav (UR 3634)

Invités :

Michel BERTHIER	PU, La Rochelle Université, Laboratoire MIA (UR 3165)

Remerciements

« Je ne connais pas la moitié d'entre vous autant que je le voudrais, et j'aime moins de la moitié d'entre vous à moitié moins que vous ne le méritez ! »

Le Seigneur des Anneaux, J.R.R. Tolkien

Les premières personnes que je souhaite remercier sont mes parents Joachim et Véronique : né d'un père militaire et d'une mère professeur de mathématiques, c'est grâce à leur éducation et leurs encouragements que je ne pouvais me retrouver ailleurs qu'à l'École navale pour y valider ce prestigieux diplôme universitaire. Ma fratrie ensuite : Kathleen, Margaux, Candice, Gabryel et Maxine qui, par l'hétérogénéité de leurs expériences personnelles, professionnelles et/ou scolaires, m'ont donné un feuilleton quotidien palpitant à suivre à des centaines de kilomètres. Je remercie mes grands-mères, Evelyne et Jacqueline, dont la pression mise jusqu'au dernier jour cachait en réalité une immense fierté. Aussi, je porte une attention singulière à ma partenaire de toujours, Mélanie, qui m'a supporté durant ces quatre années de thèse, et plus encore. Son altruisme, son respect, sa confiance et sa gentillesse m'ont grandement facilité la tâche. Nous avons construit ensemble une belle famille et, en temps voulu, je montrerai ce manuscrit à nos enfants, dont notre fils Morgan qui n'en comprendra pas un mot mais qui sera fier de son père comme je le suis de lui. J'adresse enfin une pensée émue à Patrick, mon grand-père et à Antoine, mon beau-père qui étaient admiratifs de ma jeune carrière de scientifique.

Pour mener à bien une thèse, il est important de s'entourer des bonnes personnes, de bons collègues, de bons amis : c'est grâce à la richesse de nos échanges que l'on avance et/ou que l'on se détend. Alors, je remercie Abdel, mon directeur de thèse, impressionnant par ses conseils toujours avisés et par sa pugnacité dans le monde de la recherche. Il débarque tous les matins du transrade avec de nouvelles idées qu'une thèse en 10 ans n'aurait su traiter. J'exprime également ma gratitude envers Delphine, mon encadrante, qui, grâce à une exigence constante saupoudrée d'une bienveillance maternelle, m'a toujours prodigué de bons conseils et a su me guider durant ces 4 années. Je suis également reconnaissant envers l'École navale, l'IRENav et leurs directions respectives pour leur confiance initiale et leur soutien tout au long de ma thèse, qu'il ait été humain et/ou financier. Je remercie aussi chaleureusement les membres de mon CSI, Thierry Chonavel et Michel Berthier, ainsi que Rozenn, Valérie, Éric, Marisnel, Joseph, Timothée, et tous mes collègues de l'IRENav. Aussi, je veux qu'Élodie et que mes fidèles amis, Adèle, Camille, Eulalie, Gabriel, Juliette, Manon, Nizar et Robin sachent que leur compagnie m'a toujours apporté joie, motivation et réconfort lorsque cela était nécessaire.

Enfin, je tiens à exprimer ma profonde gratitude aux membres du jury qui ont bien voulu évaluer ce travail. Je remercie tout particulièrement Pierre Borgnat et Nicolas Tremblay, pour avoir accepté d'être rapporteurs de cette thèse, pour le temps consacré à sa lecture attentive et pour la richesse de leurs remarques et suggestions. Je remercie également Vincent Gripon, Sophie Achard, Marc Barthélémy, Thierry Chonavel et Nicolas Keriven pour leur présence, leur bienveillance et la qualité de leurs échanges lors de la soutenance. Leurs observations m'ont permis de prendre du recul sur mes travaux et d'enrichir ma réflexion scientifique.

Avant-propos

Inscrit dans le cadre de l'obtention du diplôme de doctorat en mathématiques appliqués délivré par l'Université de Bretagne Occidentale, cette thèse, débutée le 15 octobre 2020, s'est déroulée au sein de l'Institut de Recherche de l'École navale (IRENav) sous la direction du Professeur des Universités Abdel-Ouahab Boudraa et encadrée par la Maître de Conférences Delphine Daré-Emzivat.

Les objets d'études de ce document sont les graphes, qu'ils proviennent de structures existantes ou qu'ils soient construits à partir de signaux. Une étude de leurs représentations matricielles classiques ainsi que l'introduction d'un cadre de généralisation sont proposés. Associés à des outils de théorie de l'information, ces méthodes permettent de classifier efficacement des graphes, de caractériser de manière singulière des séries temporelles ou encore d'identifier des vulnérabilités au sein de structures.

Cette thèse de doctorat est en réalité une extension de celle d'Hadj-Ahmed Bay-Ahmed intitulée « Classification des signaux et des graphes par approches spectrales algébriques ». Initiée en 2014, elle était dans l'air du temps car c'est au début des années 2010 que l'engouement autour des graphes et de leurs utilisations pour du traitement de signal est subitement réapparu. Aujourd'hui, à l'heure des données massives, il est facile de se rendre compte l'utilisation des graphes est véritablement essentielle.

Les travaux exposés dans ce manuscrit se veulent être des contributions devant permettre d'apporter des outils et des méthodes quant à la représentation et à la caractérisation de graphes, alimentant alors une communauté scientifique grandissante. Aussi, il ne sera pas rare dans ce travail de rencontrer des éléments provenant aussi bien de la théorie des graphes, du traitement du signal ou encore de la théorie de l'information pour répondre à ces problématiques.

Ce sujet de thèse étant exploratoire aussi bien dans la théorie que dans les applications, la difficulté majeure à l'écriture de ce document était de trouver le fil rouge qui permette une lecture et une compréhension claire du travail réalisé. En espérant que cette recherche fut une réussite.

Enfin, il est à noter que dans une démarche de science ouverte, les codes Matlab et Python ainsi que les données utilisés tout au long de cette thèse seront disponibles à la demande. Il en est de même pour le code source du manuscrit et les figures présentes dans ce dernier.

Bonne lecture.

Liste des publications

Actes de conférences nationales à comité de lecture

- ✍ **T. Averty**, D. Daré-Emzivat et A.-O. Boudraa. Détection d'épilepsie dans les signaux EEG par graphe de visibilité et un noyau de SVM adapté. *GRETSI*, pages 1–4, 2022
- ✍ **T. Averty**, D. Daré-Emzivat, A.-O. Boudraa et Y. Préaux. Approximation de l'entropie de von Neumann de graphes pour une analyse de vulnérabilité. *GRETSI*, pages 1–4, 2022
- ✍ **T. Averty**, D. Daré-Emzivat et A.-O. Boudraa. Sur la similarité spectrale des graphes par mesure de corrélation. *GRETSI*, pages 1–4, 2023

Articles de journaux internationaux

- ✍ **T. Averty**, A.-O. Boudraa et D. Daré-Emzivat. A New Family of Graph Representation Matrices : Application to Graph and Signal Classification. *IEEE Signal Processing Letters*, 31 :2935–2939, 2024
- ✍ **T. Averty**, A.-O. Boudraa et D. Daré-Emzivat. Hurst exponent estimation using natural visibility graph embedding in Fisher–Shannon plane. *Signal Processing*, 230 :109884, 2025
- ✍ **T. Averty**, Hadj-Ahmed Bay-Ahmed, D. Daré-Emzivat, A.-O. Boudraa et C. Richard. Identifying vulnerable links in large networks using von Neumann graph entropy. *IEEE Transactions on Signal and Information Processing over Networks*, 2025 (rédigé)

Table des matières

Remerciements **3**

Avant-propos **5**

Liste des publications **7**

Glossaire mathématique **13**

Introduction générale **15**

1 Généralités sur les graphes **25**

1.1 Rappels sur la théorie des graphes 25

1.1.1 Origine de la théorie des graphes 25

1.1.2 Graphes : adjacence et structure 26

1.1.3 Degré et distribution des degrés 28

1.1.4 Chemins, cycles, connexité et distance 32

1.1.5 Graphes pondérés 34

1.2 Structures particulières et graphes aléatoires 36

1.2.1 Quelques structures particulières 36

1.2.2 Modèles de graphes aléatoires d'Erdös-Renyi 36

1.3 Matrices de représentation et théorie spectrale de graphes 38

1.3.1 Matrice d'adjacence et propriétés spectrales 39

1.3.2 Matrice des degrés 42

1.3.3 Matrice(s) Laplacienne(s) : variantes et propriétés spectrales 43

1.3.4 Étude du problème de cospectralité 50

1.4 Transformation d'un signal en un graphe de visibilité 54

1.4.1 Graphe de visibilité naturelle 55

1.4.2 Graphe de visibilité horizontale 56

1.5 Conclusion 59

2 Classification et caractérisation de signaux par graphes de visibilité 61
2.1 Introduction . . . 61
2.2 Rappel sur les séparateurs à vaste marge (SVM) . . . 63
2.2.1 Principe . . . 63
2.2.2 Métriques de performance . . . 67
2.2.3 Validation croisée . . . 68
2.3 Classification de séries temporelles par graphe de visibilité . . . 70
2.3.1 Motivations et méthodes existantes . . . 70
2.3.2 Détection d'épilepsie dans des signaux EEG . . . 73
2.3.3 Détection d'anomalies magnétiques . . . 79
2.4 Caractérisation de processus aléatoires dans un plan informationnel . . . 81
2.4.1 Bruits colorés, mouvements Browniens et bruits Gaussiens fractionnaires . 82
2.4.2 Distributions extraites des graphes de visibilité . . . 85
2.4.3 Construction d'un squelette dans un plan informationnel . . . 88
2.4.4 Estimation du coefficient de Hurst . . . 95
2.5 Conclusion . . . 102

3 Vulnérabilité informationnelle d'un graphe 105
3.1 Introduction . . . 105
3.1.1 Mise en contexte . . . 105
3.1.2 Travaux sur la vulnérabilité de graphes . . . 106
3.1.3 Motivations . . . 107
3.2 Graphe en tant que système : matrice de densité et entropie . . . 109
3.2.1 Représentation d'un graphe dans le domaine quantique . . . 109
3.2.2 Matrice de densité . . . 109
3.2.3 Entropie de von Neumann d'un graphe . . . 110
3.3 Vulnérabilité informationnelle d'une arête . . . 111
3.3.1 Saillance d'une arête . . . 112
3.3.2 Algorithme EIVP (*Edge Informational Vulnerability to Perturbation*) . . . 117
3.3.3 Corrélation avec d'autres attributs . . . 120
3.4 Approximations de l'entropie . . . 124
3.4.1 Approche « approximations quadratiques » . . . 126
3.4.2 Approche « théorie de perturbation matricielle » . . . 132
3.4.3 Erreurs et temps de calcul des différentes approximations . . . 136
3.5 Conclusion . . . 140

4 Généralisation des représentations conventionnelles de graphes **143**
4.1 Introduction 143
4.2 Matrices d'adjacence généralisées 145
4.2.1 État de l'art 145
4.2.2 Matrice $\mathbf{T}_\alpha$ 147
4.2.3 Plan de représentation $\mathbf{P}_{\alpha,k}$ 151
4.3 Mesure de similarité par corrélation spectrale 155
4.4 Classification spectrale de graphes et de signaux 156
4.4.1 Bases de données 156
4.4.2 Méthodes & résultats 157
4.5 Conclusion 162

Conclusions & perspectives **165**

A Attributs d'un signal discret $\mathbf{s} = (s_i)_{1 \leq i \leq n}$ **171**

B Algorithme de Davies et Harte **174**

Bibliographie **177**

Glossaire mathématique

Symbole	Signification
G	Graphe
$\mathcal{V}$	Ensemble de sommets
n	Nombre de sommets (aussi appelé ordre)
$\mathcal{E}$	Ensemble d'arêtes
m	Nombre d'arêtes
$\{i, j\}$	Arête entre les sommets i et j
$G \cup G'$	Union des graphes G et G'
$\mathcal{N}(i)$	Voisinage du sommet i
$\deg(i)$	Degré du sommet i
$\bar{d}(G)$	Degré moyen du graphe G
$\delta(G)$	Degré minimal du graphe G
$\Delta(G)$	Degré maximal du graphe G
$s(i)$	Force du sommet i
$\bar{s}(G)$	Force moyenne du graphe G
$\mathrm{dev}(G)$	Déviation des degrés
$\mathrm{var}(G)$	Variance des degrés
$Z_1(G)$	Indice de Zagreb du graphe G
$\mathbf{A}$	Matrice d'adjacence
$\mathbf{W}$	Matrice des poids
$\mathbf{D}$	Matrice des degrés
$\mathbf{B}$	Matrice d'incidence
$\mathbf{L}$	Matrice Laplacienne
$\mathbf{Q}$	Matrice Laplacienne sans-signe
$\mathbf{A}_\alpha$	Matrice d'α-adjacence
$\mathbf{L}_\alpha$	Matrice α-Laplacienne
$\mathbf{T}_\alpha$	Nouvelle matrice de représentation introduite dans cette thèse
$\mathbf{P}_{\alpha,k}$	Nouveau plan de représentation introduit dans cette thèse
$\mathbf{G}$	Matrice d'adjacence généralisée
$\mathbf{U}$	Matrice d'adjacence universelle
$\mathbf{I}$	Matrice identité
$\mathbf{J}$	Matrice remplie de 1
$\mathcal{L}$	Matrice Laplacienne normalisée
$\boldsymbol{\rho}$	Matrice de densité

suite à la page suivante

Symbole	Signification
$(\lambda_\ell)_{1\leq\ell\leq n}$	Spectre de la matrice d'adjacence
λ_n	Rayon spectral
$(\mu_\ell)_{1\leq\ell\leq n}$	Spectre de la matrice Laplacienne
μ_2	Valeur de Fiedler
$(\mu_\ell^+)_{1\leq\ell\leq n}$	Spectre de la matrice Laplacienne sans-signe
$(\chi_\ell)_{1\leq\ell\leq n}$	Spectre de la matrice Laplacienne normalisée
$(\nu_\ell)_{1\leq\ell\leq n}$	Spectre de la matrice de densité
$E_{\mathbf{M}}$	Énergie définie grâce à la matrice de représentation $\mathbf{M}$
$\mathrm{Tr}(\cdot)$	Opérateur trace
$\mathbf{M}^{\mathsf{H}}$	Opérateur hermitien (matrice $\mathbf{M}$ conjuguée et transposée)
$\mathcal{P}_n$	Graphe chaîne
$\mathcal{S}_n$	Graphe étoile
$\mathcal{C}_n$	Graphe cycle
$\mathcal{W}_n$	Graphe roue
$\mathcal{K}_n$	Graphe complet
$\mathcal{C}_{n_1,n_2}$	Graphe comète
$\mathcal{B}_n$	Graphe *Barbell*
$\mathcal{G}_{n,p}$	Graphe aléatoire d'Erdös-Rényi
$\mathcal{H}$	Hyperplan
H	Coefficient de Hurst
$B_H(t)$	Mouvement Brownien fractionnaire de coefficient de Hurst H
$G_H(t)$	Bruit Gaussien fractionnaire de coefficient de Hurst H
$S(\cdot)$	Entropie de von Neumann (ou entropie de Shannon)
$R_\alpha(\cdot)$	Entropie de Rényi de paramètre α
$F(\cdot)$	Information de Fisher
$\mathrm{KL}(\cdot\|\cdot)$	Divergence de Kullback-Leibler
$\mathrm{JSD}(\cdot\|\cdot)$	Distance de Jensen-Shannon
VG/FS	Méthode d'estimation du coefficient de Hurst par projection des graphes de visibilité dans le plan informationnel Fisher-Shannon
$\widehat{\mathbf{x}}, \widehat{\mathbf{X}}$	Version normalisée du vecteur $\mathbf{x}$, de la matrice $\mathbf{X}$, ...
$\widetilde{\mathbf{x}}, \widetilde{\mathbf{X}}$	Version perturbée du vecteur $\mathbf{x}$, de la matrice $\mathbf{X}$, ...
$\mathrm{SSC}(G_1, G_2)$	Similarité spectrale conjointe entre deux graphes G_1 et G_2
$\mathrm{CS}(G_1, G_2)$	Covariance spectrale entre deux graphes G_1 et G_2
$\mathrm{CorS}(G_1, G_2)$	Corrélation spectrale entre deux graphes G_1 et G_2

Par défaut et sauf cas particuliers, les lettres minuscules en gras (exemple : $\mathbf{x}$) désignent des vecteurs tandis que les lettres majuscules en gras (exemple : $\mathbf{M}$) désignent des matrices.

Introduction générale

« Le meilleur moyen d'avoir une bonne idée est d'en avoir beaucoup. »

Linus Pauling

Contexte et enjeux

Les graphes sont des objets mathématiques, composés de sommets et d'arêtes, particulièrement adaptés pour analyser l'information véhiculée par des données structurées, bien souvent collectées de manière massive (le terme *big data* est utilisé dans ce contexte) et provenant éventuellement de divers capteurs. Pour illustrer cette quantité astronomique de données, 64 Zo de données ont été créées en 2020 (ce qui représente environ 22 Go de données par habitant et par jour) pour une capacité de stockage mondiale d'environ 6.7 Zo [1]. Les réseaux sociaux (virtuels ou non) constituent des exemples naturels de données denses pouvant être analysées. C'est pour cette raison qu'une carte partielle d'Internet, réseau informatique le plus connu, est proposée en figure 1. Les sommets sont des adresses IP dont les connexions Internet entre elles sont représentées par des arêtes. L'utilisation des graphes pour représenter des données multicapteurs offre une stratégie intéressante pour modéliser les relations complexes entre différentes sources d'information. Chaque capteur peut être vu comme un sommet du graphe, tandis que les arêtes illustrent les interactions, les corrélations ou les flux d'information entre ces capteurs. Cette représentation facilite l'intégration et l'analyse de données hétérogènes, en capturant non seulement les valeurs individuelles, mais aussi les structures relationnelles sous-jacentes. Par exemple, dans les réseaux de capteurs dédiés à la surveillance environnementale, les graphes permettent de suivre la propagation de phénomènes comme la pollution ou les incendies en se référant aux capteurs de ces réseaux. Des algorithmes d'analyse de graphes, tels que la détection de

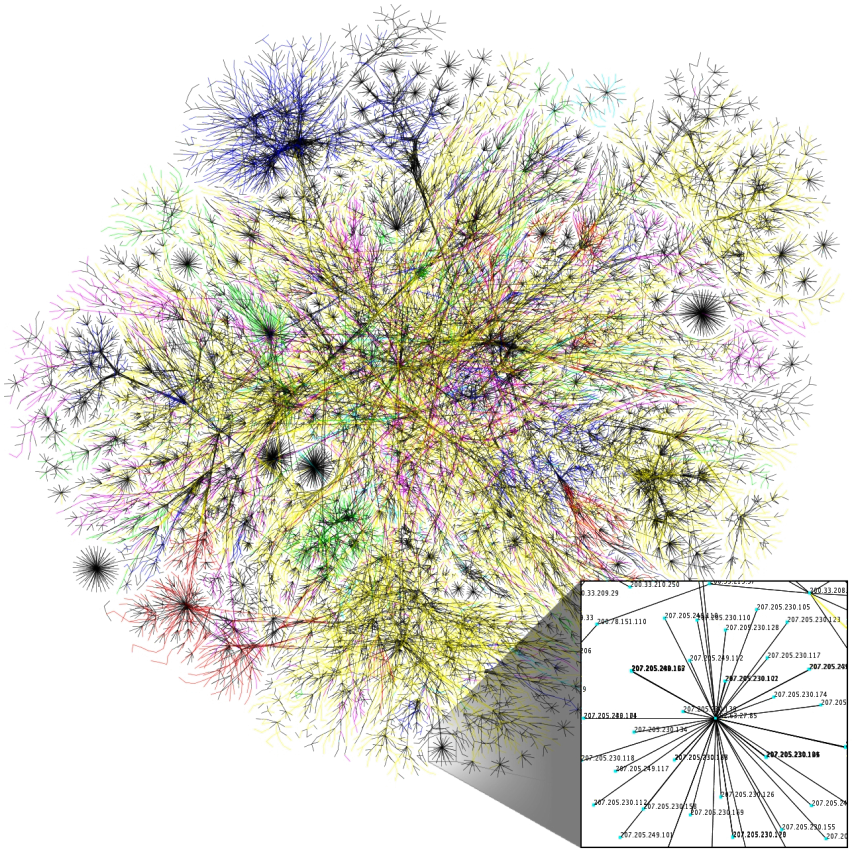


Figure 1 – Carte partielle d'Internet, basée sur les données du 15 juin 2005 situées à opte.org. Chaque arête relie 2 sommets, représentant 2 adresses IP. La longueur de chaque arête correspond au délai de connexion entre deux adresses.

communautés, peuvent également être appliqués pour extraire des motifs pertinents afin de permettre une bonne fusion de données et d'optimiser l'éventuelle prise de décision en temps réel. Un exemple de graphe modélisant un réseau multicapteurs de mesures de température est présenté en figure 2. Les sommets de ce graphe sont des capteurs de température reliés par des arêtes s'ils sont géographiquement proches.

Les données organisées sous forme de réseaux sont modélisées naturellement sous la forme de graphe. C'est le cas des réseaux sociaux, des réseaux informatiques, des réseaux d'approvisionnement en énergie ou encore des réseaux de transport (routiers, aériens, maritimes, ferroviaires, etc.) [3–7]. Dans le but d'étudier, de comprendre, d'analyser ou de classer ces réseaux, les graphes sont des outils adéquats, notamment car ces derniers peuvent être caractérisés par des grandeurs scalaires, appelées attributs structurels, et sont également traduits sous la forme d'une multitude de matrices de représentation, induisant alors l'utilisation d'éléments d'algèbre linéaire, en particulier les valeurs et vecteurs propres de ces matrices. Parmi ces dernières, citons par exemple la matrice d'adjacence $\mathbf{A}$,

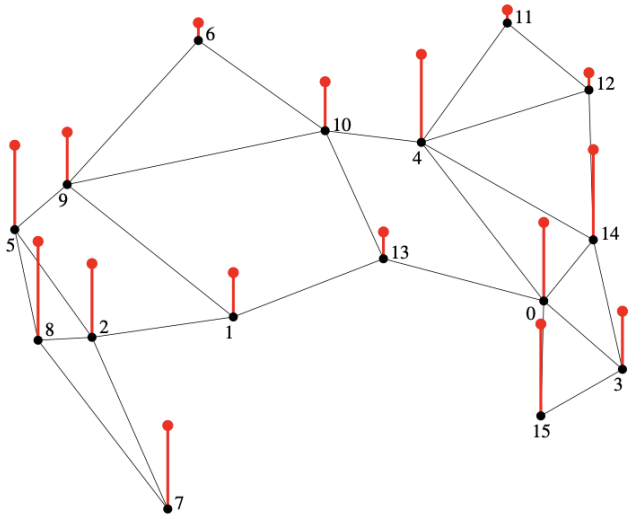


Figure 2 – Mesures de température multicapteurs représentées sous la forme d'un graphe où les sommets (capteurs) sont reliés par des arêtes s'ils sont géographiquement proches [2].

la matrice des degrés $\mathbf{D}$, la matrice Laplacienne $\mathbf{L}$ [8] ou encore la matrice Laplacienne sans-signe $\mathbf{Q}$ [9]. Comme les méthodes permettant l'analyse et l'étude du fonctionnement des réseaux (et donc des graphes les représentant) sont majoritairement basés sur ces matrices, la pertinence des informations qu'elles contiennent est alors capitale. C'est pour cela que la question relative à la « meilleure » matrice de représentation est toujours ouverte et est source de nombreux débats [10–12].

Parmi toutes les informations essentielles à extraire d'un réseau (et donc du graphe le modélisant), la recherche d'une éventuelle vulnérabilité de ses éléments suscite de l'intérêt. En effet, au regard de la situation géopolitique actuelle et des conflits qui règnent autour du globe mais aussi des conséquences liées aux changements climatiques[1], il est aisé de constater que les tensions se cristallisent autour d'infrastructures qui subissent un certain nombre de défaillances, volontaires ou non, et qui sont pourtant essentielles au bon fonctionnement de nations ou d'organisations. À titre d'exemples, environ 50 millions d'américains sont touchés par une importante panne de courant en août 2003 [3, 4] tandis que c'est près de la moitié de la population indienne qui en est victime en juillet 2012 [13]. En mars 2021, le canal de Suez se retrouve obstrué par le porte-conteneurs *Ever Given*, immobilisant alors plus de 400 navires sur cette route maritime, responsable à elle seule de plus de 12% du commerce mondial. En septembre 2022, c'est le réseau d'approvisionnement en gaz européen, *via* les pipelines NordStream 1 et NordStream 2 reliant l'Allemagne et la Russie, qui est victime d'un sabotage [14]. Enfin, ces dernières

1. Parmi ces conséquences, citons l'apparition de grandes routes terrestres et/ou maritimes empruntées par un grand nombre de réfugiés climatiques, le changement de réseau d'approvisionnement en électricité afin de bénéficier d'énergie verte, etc.

années, ce sont les câbles sous-marins, dont une illustration du réseau mondial est fournie en figure 3, qui sont au cœur des débats car de nombreux navires en détériorent involontairement à cause de leurs ancres [15], impactant de fait les réseaux sous-jacents. Ces derniers ne sont pas exempts de sabotage : le 17 et 18 novembre 2024, un câble de fibre optique est endommagé en mer Baltique tandis que le jour de Noël, le 25 décembre 2024, c'est le câble électrique Estlink 2 reliant la Finlande et l'Estonie qui est saboté [16]. Ces réseaux suscitent tous un intérêt particulier pour toute organisation qui souhaiterait paralyser le monde. Trouver des méthodes permettant d'identifier les vulnérabilités de ces réseaux est alors une tâche essentielle car cela permettrait de les rendre plus résilients.

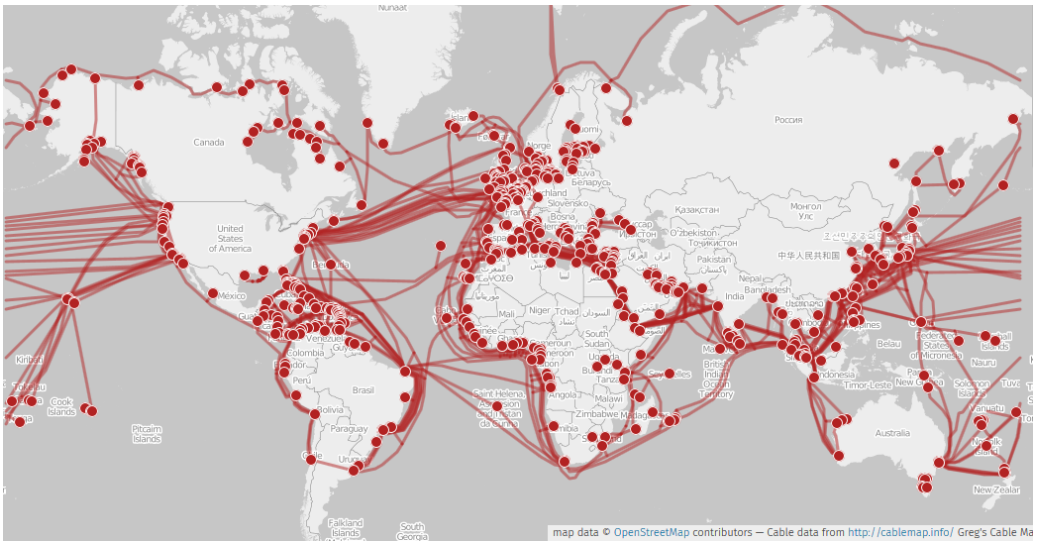


Figure 3 – Carte mondiale, en date du 21/07/2015, des câbles sous-marins de communication.

Alors que les réseaux physiques ont, pour la plupart, une représentation intuitive sous forme de graphes, ce n'est pas le cas des séries temporelles, objets largement étudiés dans ce manuscrit de thèse. Une série temporelle est une succession de valeurs scalaires traduisant l'évolution d'une variable au cours du temps. À première vue, aucun lien avec les graphes présentés ci-dessus. Toutefois, il est possible de définir un signal sur les sommets d'un graphe. Par exemple, le signal de la figure 2, composé des 16 mesures de température, est un signal défini sur les sommets du graphe construit grâce à la proximité géographique des capteurs. C'est ainsi que la thématique du traitement de signal sur graphes est apparue il y a quelques années. Dans ce domaine, les outils classiques du traitement de signal sont généralisés aux signaux définis sur les sommets d'un graphe. Parmi ces derniers, citons la convolution, la translation, la modulation [10–12, 17] ou encore la transformée de Fourier sur graphe [18, 19] et la transformée en ondelettes sur graphe [20]. Ces outils permettent de comprendre le graphe sur lequel est défini le signal comme un outil rendant capable une meilleure mise en valeur de l'information contenue dans ce signal. Au-delà de cette nouvelle thématique, d'autres passerelles existent entre signal

et graphe. En effet, il est possible de voir une série temporelle en un graphe. L'algorithme le plus utilisé est celui du graphe de visibilité, traduisant les échantillons comme des sommets reliés par des arêtes si ces derniers se « voient » géométriquement [21–23]. Les travaux de Lacasa *et al.* [21–23] et ceux de Gonçalves *et al.* [24, 25] ont montré l'apport des graphes de visibilité pour la caractérisation de séries temporelles.

Les éléments présentés précédemment permettent alors de traiter et d'analyser les signaux comme des graphes, avec des outils aussi bien issus de la théorie des graphes que de l'algèbre linéaire. Ainsi, dans ce contexte, classer des signaux revient à classer des graphes. Or, la classification de graphes, constituant une activité fondamentale en théorie des graphes, repose sur la comparaison de certaines de leurs propriétés structurelles et/ou spectrales au moyen de mesures de similarité intégrées dans des noyaux sur graphes[2] [28, 29]. Des travaux de la littérature proposent par exemple de comparer les plus courts chemins [30], d'autres de compter les sous-graphes types [31] quand certains proposent de comparer les spectres des matrices d'adjacence ou Laplacienne [32]. En effet, la théorie spectrale des graphes est un domaine qui permet d'établir certaines propriétés d'un graphe à travers les valeurs et vecteurs propres de ses matrices de représentation [8], conduisant à étudier sa connectivité, sa complexité ou sa régularité afin de comprendre sa structure ou encore à quantifier son contenu en information. Le verrou scientifique est alors de trouver quelles propriétés comparer et comment y parvenir. En guise d'exemples de tâches de classifications de graphes, il peut être envisagé de classer des structures moléculaires selon leur cancérogénicité ou non, de classer les flux intra-départements, de classer des circuits électroniques, et bien d'autres applications [28]. La classification de signaux (considérés comme des graphes) demeure d'actualité lors d'applications visant la détection d'épilepsie dans des signaux EEG ou encore la détection d'anomalies magnétiques. Ainsi, trouver des méthodes de classification de graphes et, par conséquent, des mesures de similarité adaptées, c'est contribuer à la conception d'outils nécessaires pour résoudre bon nombre de défis.

Motivations et verrous scientifiques

Après le rappel sur les enjeux de l'analyse des données de masse, aussi bien sous forme de graphes et/ou de séries temporelles, nous identifions quelques verrous scientifiques relatifs aux outils mathématiques ou algorithmiques associés. En effet, un certain nombre de méthodes de la littérature sont orientées vers l'extraction d'une quantité considérable d'attributs [33–35], quitte à en réduire *a posteriori* le nombre *via* une analyse en composantes principales, perdant de fait le sens physique des

2. Pour la classification de graphes, ces noyaux remplacent souvent ceux d'un SVM [26, 27] traditionnellement usités (notamment le noyau RBF).

attributs extraits *ab initio*. Certaines méthodes extraient ces nombreux attributs de manière automatique : c'est le cas notamment des réseaux de neurones [36]. Bien que les résultats obtenus montrent bien souvent une performance élevée de ces algorithmes, les attributs extraits n'ont pas encore trouvé de réelles interprétations physiques. Motivés par l'intérêt grandissant de la communauté scientifique pour la théorie des graphes, Lacasa *et al.* ont développé le graphe de visibilité [21], permettant de construire un graphe à partir d'une série temporelle à l'aide d'un simple critère géométrique. L'idée d'analyser des signaux grâce à leurs graphes de visibilité a déjà été explorée [37–39] et a permis d'obtenir de bons résultats en termes de caractérisation et de classification. Cela signifie que ces graphes véhiculent l'information nécessaire à l'analyse de signaux. À l'image d'un spectrogramme (transformée de Fourier à court terme) ou d'un scalogramme (transformée en ondelettes), l'information contenue dans la série temporelle serait alors mise en valeur différemment. Par ailleurs, s'il est question de classer des signaux, vus sous forme de graphe, une question naturelle se pose : **comment comparer les graphes de visibilité de manière efficiente ?** Cela passe par une mesure de similarité comme il en existe tant d'autres : comparaison des plus courts chemins [30], décompte des sous-graphes types [31], test d'isomorphisme [40]. La notion d'efficience est ajoutée car ces « noyaux sur graphes » relèvent bien souvent d'une complexité trop importante. C'est pour cela que nous avons construit un nouveau noyau comparant l'attribut structurel parmi les plus simples à extraire, à savoir la distribution des degrés.

Les graphes de visibilité possèdent de très bonnes propriétés, particulièrement si les signaux dont ils sont issus sont des processus stochastiques tels que les mouvements Browniens fractionnaires (fBm pour *fractional Brownian motion*) ou les bruits Gaussiens fractionnaires, paramétrés tous les deux par un coefficient appelé coefficient de Hurst [41], noté H et caractérisant des propriétés d'auto-similarité et de dépendance à plus ou moins long terme. En effet, nous montrons qu'il est possible, à partir de leurs simples distributions des degrés suivant une loi en puissance d'exposant γ, d'estimer finement le coefficient H à partir de γ [21]. Cependant, estimer ainsi H nécessite, *a priori*, l'estimation de γ, ce qui rend la méthode peu robuste. Les travaux de Gonçalves *et al.* [24,25] extraient quant à eux deux quantifieurs de la théorie de l'information des distributions de degrés des graphes de visibilité horizontale issus de fBm synthétiques, à savoir l'entropie de Shannon et l'information de Fisher. Ces deux données sont placées dans le plan Fisher-Shannon, développé en 2003 par Vignat et Bercher [42]. Lorsque l'expérience est renouvelée pour plusieurs fBm synthétiques disposant de coefficients de Hurst H variés, des nuages de points se créent en fonction des H, générant un squelette de référence, ouvrant la voie à une potentielle estimation de H. D'où la question suivante : **est-il possible d'estimer, à l'aide de graphes de visibilité et de manière robuste, le coefficient de Hurst d'un processus stochastique ?**

La question de la matrice de représentation d'un graphe est centrale. En effet, le calcul de l'énergie d'un graphe [43, 44], de son entropie de von Neumann [45] ou encore de la vulnérabilité de ses arêtes, fait appel au spectre d'une matrice de représentation, que ce soit la matrice d'adjacence $\mathbf{A}$, la matrice des degrés $\mathbf{D}$, la matrice Laplacienne $\mathbf{L}$, la matrice Laplacienne sans-signe $\mathbf{Q}$ [9] ou encore la matrice de densité $\boldsymbol{\rho}$ [46]. Ces matrices ont bien entendu été introduites pour diverses raisons : la matrice d'adjacence pour sa traduction immédiate du graphe en matrice, la matrice Laplacienne pour ses propriétés physico-mathématiques qui permettent de retrouver une forme quadratique, la matrice de densité pour son lien avec les états quantiques se présentant sous la forme d'un mélange d'états purs [46], etc. Les spectres donnent des informations essentielles sur la structure du graphe [8] : sa connectivité, sa capacité à être partitionnée ou encore le nombre possible d'arbres couvrants. Mais au-delà de ces attributs qui peuvent être extraits de graphes, les spectres permettent également de les classer [47] : c'est la classification spectrale de graphes. Trouver une mesure de similarité spectrale devient essentiel pour effectuer ce type de tâches. Toutefois, il est nécessaire de tenir compte des graphes $\mathbf{M}-$cospectraux [47–50], c'est-à-dire des graphes dont les structures sont différentes mais qui partagent le même spectre selon la matrice de représentation $\mathbf{M}$. Ainsi, à titre d'exemple, si deux graphes distincts G_1 et G_2 sont $\mathbf{L}-$cospectraux et qu'une simple distance euclidienne est calculée entre les spectres de leurs matrices Laplaciennes, alors cette distance serait nulle. Beaucoup de travaux ont étudié ce problème [32, 47, 49–52] et, parmi ces derniers, Bay-Ahmed *et al.* ont proposé une mesure de similarité conjointe qui prend en compte les spectres d'adjacence et les spectres Laplacien [53]. En effet, il a été montré qu'il existe beaucoup moins de graphes qui soient à la fois $\mathbf{A}-$ et $\mathbf{L}-$cospectraux. Il est également connu que la matrice Laplacienne sans-signe $\mathbf{Q}$ possède peu de graphes cospectraux [9]. Peut-être n'est-il pas nécessaire de calculer deux spectres pour effectuer une classification spectrale pertinente. De plus, il serait intéressant de construire une matrice de représentation dont le spectre conduit à une classification de graphes performante. Il est ainsi légitime de se poser la question suivante : **quelle matrice représente le mieux un graphe ?** Cette question est volontairement formulée en omettant l'adjectif « meilleure » car trouver la meilleure matrice de représentation dépend de la finalité souhaitée et il semble impossible de déterminer une matrice optimale en toute circonstance. Des travaux de la littérature ont proposé des matrices de représentation dites généralisées comme la matrice d'adjacence généralisée [49], la matrice d'adjacence universelle [50], la matrice Laplacienne déformée [54], la matrice d'$\alpha-$adjacence de Nikiforov [55], la matrice $\alpha-$Laplacienne [56] et bien d'autres. L'intérêt de ces matrices est qu'elles permettent, à l'aide de paramètres, d'évaluer graduellement l'importance des matrices de représentation classiques, notamment à travers leurs contenus spectraux. Certaines d'entre elles possèdent trop de paramètres, d'autres ne couvrent pas par toutes les matrices de représentation traditionnelles. De plus, une question naturelle à laquelle nous souhaitions

répondre était : **existe-t-il une mesure de similarité spectrale, ne requérant qu'un seul spectre et permettant une bonne classification de graphes ?**

Enfin, comme rappelé au début de cette introduction, les réseaux sont modélisés sous la forme de graphes et, depuis une trentaine d'années, il est question d'étudier la vulnérabilité de ces derniers en perturbant des arêtes, en supprimant certaines d'entre elles, en retirant un sommet mais aussi en ajoutant des arêtes ou encore des sommets pour simuler une éventuelle intrusion [57–61]. Le concept de vulnérabilité n'est toujours pas objectivement défini, si ce n'est d'être « un antonyme naturel de la robustesse » [62]. D'où la problématique de **comment quantifier la vulnérabilité d'une arête ?** Aussi, basé sur les outils utilisés jusqu'alors, **est-il possible de le faire en utilisant des métriques de la théorie de l'information ?** En effet, il existe un certain nombre de travaux qui ont apporté des outils permettant de quantifier la vulnérabilité d'arêtes ou, de manière relativement équivalente, de sommets, mais aucun n'utilise réellement d'outils de la théorie de l'information pour répondre à cette problématique [62]. L'article complet de Freitas *et al.* en cite des exemples [62]. En considérant le graphe comme un système physique établi dans le monde quantique, il est possible d'utiliser les travaux de Braunstein *et al.* dans lesquels est définie la matrice de densité $\boldsymbol{\rho}$ d'un graphe à partir de sa matrice Laplacienne $\mathbf{L}$ [46]. Ainsi, l'entropie de von Neumann peut s'appliquer à un graphe vu comme un système physique et peut se calculer à l'aide des valeurs propres de cette matrice de densité [45]. Cette entropie peut être vue comme une mesure de régularité ou encore de complexité, mais aussi, en utilisant un prisme « théorie de l'information », comme une manière de quantifier le contenu informationnel du système physique sous-jacent [45].

Plan de thèse

Le chapitre 1 introduit les rappels relatifs à la théorie des graphes, ses définitions usuelles, les différentes matrices de représentation, les propriétés connues de la théorie spectrale mais aussi l'intérêt des graphes de visibilité. Ces notions sont nécessaires à la compréhension du travail réalisé.

Dans le chapitre 2, nous abordons l'importance des graphes de visibilité à classer et caractériser des séries temporelles. Nous introduisons de nouvelles techniques de classification de signaux basées sur la comparaison, à l'aide de distances statistiques, des distributions de degrés des graphes de visibilité. Dans la seconde partie de ce chapitre 2, nous poursuivons cette approche mais en recourant cette fois-ci à l'utilisation de quantificateurs issus de la théorie de l'information, à savoir l'information de Fisher et l'entropie de Shannon pour caractériser des processus stochastiques. En particulier, nous développons une méthode d'estimation du coefficient de Hurst H définissant ces processus toujours à l'aide de leurs représentations sous la forme de graphes de visibilité.

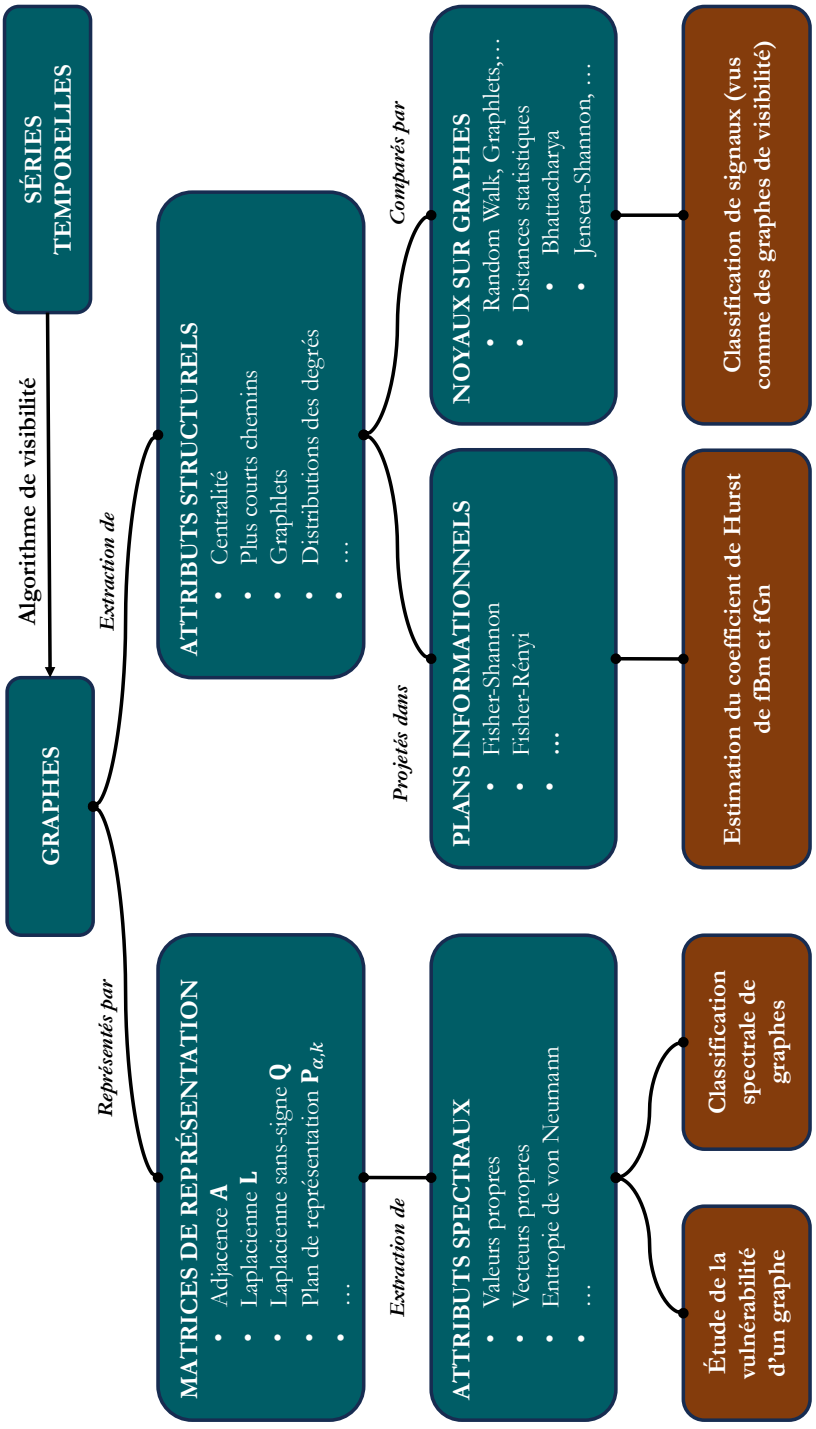
Le chapitre 3 traite du problème de la vulnérabilité des réseaux. Nous étudions l'intérêt de l'entropie de von Neumann, et en particulier son évolution lors de la perturbation des arêtes, conduisant à l'exploitation d'une mesure de vulnérabilité locale d'un graphe. Nous échangeons sur l'interprétation de cette entropie comme mesure du contenu informationnel d'un réseau considéré comme un système physique dont une légère perturbation modifierait de fait ce contenu. Nous développons alors un algorithme permettant l'obtention d'une carte des vulnérabilités éventuelles. L'entropie de von Neumann relevant d'une complexité cubique selon le nombre de sommets, nous introduisons deux approximations basées sur de l'analyse fonctionnelle et sur la théorie de perturbations matricielles, autorisant ainsi un calcul beaucoup plus rapide de l'algorithme précédemment mentionné.

Enfin, le chapitre 4 est consacré aux matrices de représentations généralisées des graphes. Ces dernières étant associées aux outils d'algèbre linéaire, la question se pose de trouver la meilleure au sens du contenu spectral. Des éléments de réponses sont apportés avec l'introduction d'une nouvelle matrice, notée $\mathbf{P}_{\alpha,k}$, généralisant celles traditionnellement utilisées dans la littérature et une nouvelle méthode de classification spectrale est développée, basée sur une corrélation entre valeurs propres.

Ce manuscrit s'achève par une conclusion générale qui propose une synthèse de tous les éléments clés de cette thèse. De plus, elle précise les différentes perspectives émanant de ce travail de recherche, allant de la validation de certains points, qu'ils soient théoriques ou pratiques, à l'exploration de nouvelles pistes de recherches, et ce, toujours en gardant à l'esprit que les relations entre graphes et signaux sont plutôt étroites et que les graphes sont présents dans presque tous les domaines scientifiques et techniques appuyant de fait toute avancée dans cette thématique.

Organigramme de la thèse

Nous proposons en page suivante une illustration permettant de mieux saisir le contexte scientifique dans lequel s'inscrit le travail réalisé. Le point de départ est de savoir comment est représentée l'information : sous la forme d'un graphe car elle provient d'un réseau naturel ou d'une série temporelle, pouvant être vue comme un graphe à l'aide de l'algorithme de visibilité. Puis la question de l'utilisation ou non de matrices de représentation se pose. Si des matrices de représentation sont considérées, elles le sont essentiellement pour leurs spectres et sinon, nous nous intéressons à l'utilisation d'attributs structurels pouvant être extraits des graphes.


SÉRIES TEMPORELLES
Algorithme de visibilité
GRAPHES
Représentés par
Extraction de
MATRICES DE REPRÉSENTATION
• Adjacence **A**
• Laplacienne **L**
• Laplacienne sans-signe **Q**
• Plan de représentation $\mathbf{P}_{\alpha,k}$
• …
Extraction de
ATTRIBUTS SPECTRAUX
• Valeurs propres
• Vecteurs propres
• Entropie de von Neumann
• …
Classification spectrale de graphes
Étude de la vulnérabilité d'un graphe
ATTRIBUTS STRUCTURELS
• Centralité
• Plus courts chemins
• Graphlets
• Distributions des degrés
• …
Projetés dans
Comparés par
PLANS INFORMATIONNELS
• Fisher-Shannon
• Fisher-Rényi
• …
NOYAUX SUR GRAPHES
• Random Walk, Graphlets,…
• Distances statistiques
• Bhattacharya
• Jensen-Shannon, …
Estimation du coefficient de Hurst de fBm et fGn
Classification de signaux (vus comme des graphes de visibilité)

Chapitre 1

Généralités sur les graphes

« Comprendre est le commencement d'approuver. »

Baruch Spinoza

Le principal objectif de ce chapitre est de présenter les outils essentiels aux méthodes développées au cours de ce travail de thèse. Un rappel sur l'origine de la théorie des graphes est proposé, suivi des définitions et notations relatives aux graphes. Une attention particulière est portée sur les matrices de représentation ainsi que sur les attributs structurels et spectraux déterminants pour toute tâche de classification. Les différentes méthodes permettant la transformation d'une série temporelle en graphe sont également décrites au cours de ce chapitre, en particulier l'algorithme de visibilité.

1.1 Rappels sur la théorie des graphes

1.1.1 Origine de la théorie des graphes

De nombreux réseaux naturels comme les réseaux électriques, les réseaux informatiques, les réseaux sociaux ou encore les protéines en chimie moléculaire peuvent être représentés sous la forme de graphes. Il est bien souvent admis que l'origine de la notion de graphe remonte au début du XVIII^{e} siècle lorsque Leonhard Euler, mathématicien suisse résidant à cette époque à Saint-Pétersbourg, rédige en 1735 un article intitulé « *Solutio problematis ad geometriam situs pertinentis* » publié 6 ans plus tard dans lequel il apporte une démonstration quant à la non-existence de solution au problème des 7 ponts de Königsberg [63]. Ce problème consistait à déterminer la possibilité, en partant de n'importe quel quartier de Königsberg visible en figure 1.1, de déambuler dans cette ville en passant une et une

seule fois par chacun des 7 ponts[1] tout en revenant à son point de départ [64]. Dans son article, Euler schématise la ville de Königsberg sous une forme assimilée à un objet appelé « graphe ». Il utilise des lettres majuscules (A, B, C et D) pour désigner les quartiers et des lettres minuscules (a, b, c, d, e, f et g) pour les ponts. Ces éléments sont présentés en figure 1.2.

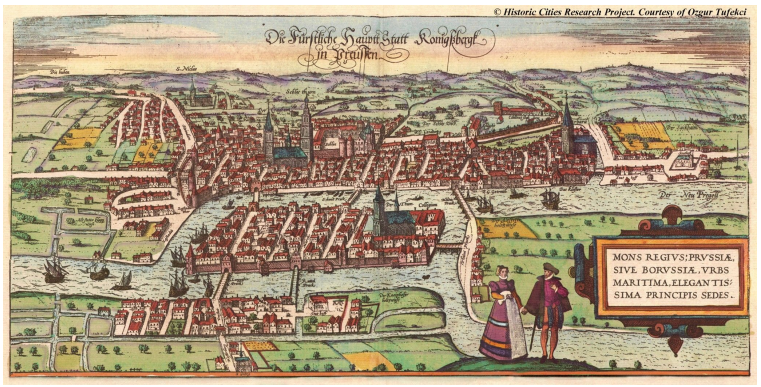


Figure 1.1 – Illustration de la ville prusse de Königsberg au XVI^{e} siècle issue du « *Civitates orbis terrarum* » de G. Braun.

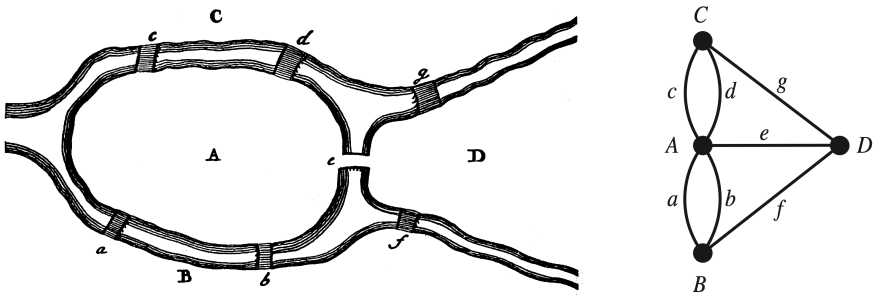


Figure 1.2 – À gauche, dessin d'Euler représentant la ville de Königsberg pouvant être schématisé en réalité par un multigraphe à 4 sommets (4 quartiers) et 7 arêtes (7 ponts), à droite.

1.1.2 Graphes : adjacence et structure

Bien que le cadre théorique permettant de confirmer la viabilité de la démonstration d'Euler n'ait été posé qu'en 1873 par un mathématicien allemand nommé Carl Hierholzer à titre posthume [65], un graphe $G = (\mathcal{V}, \mathcal{E})$ est un couple constitué d'un ensemble non-ordonné $\mathcal{V} = \{v_1, \ldots, v_n\}$ de n sommets, où $|\mathcal{V}| = n$ représente[2] l'ordre du graphe G, et d'un ensemble $\mathcal{E} = \{e_{ij}\} = \{\{v_i, v_j\}, 1 \leq i, j \leq n\} \subseteq \mathcal{V} \times \mathcal{V}$ de m arêtes[3]. Considérons un réseau informatique constitué de 5 ordinateurs, nœuds du réseau, et de 7 connexions Internet (liens) entre ces ordinateurs. Un exemple de

1. À bien analyser cette représentation, il semble que le 7^{e} pont ait été construit entre le XVI^{e} et le XVIII^{e} siècle.
2. Si $\mathcal{V}$ est un ensemble, $|\mathcal{V}|$ représente son cardinal, c'est-à-dire le nombre d'éléments de $\mathcal{V}$.
3. Dans la suite, on s'attachera à noter n le nombre de sommets (ordre) et m le nombre d'arêtes du graphe étudié.

graphe permettant de modéliser un tel réseau est représenté en figure 1.3. Ce graphe a 5 sommets $\mathcal{V} = \{1, 2, 3, 4, 5\}$ et 7 arêtes[4] $\mathcal{E} = \{\{1, 2\}, \{1, 4\}, \{1, 5\}, \{2, 3\}, \{2, 5\}, \{3, 4\}, \{4, 5\}\}$. Bien entendu, si les liens de ce réseau informatique représente les échanges de courriels, les arêtes du graphe peuvent être dirigés d'un sommet à un autre. Dans ce cas, on dit que le graphe est orienté. Cependant, l'introduction de ce type d'arêtes complexifie les définitions. Ainsi, sauf cas particuliers, tous les graphes de cette thèse sont non-orientés. Par ailleurs, on dit que le graphe de la figure 1.3 est simple car il ne possède ni boucles, c'est-à-dire d'arêtes $\{i, i\}$ reliant un sommet i à lui-même, ni arêtes multiples, c'est-à-dire plusieurs arêtes qui relient les mêmes sommets. Le graphe représentant la ville de Königsberg en figure 1.2 présente des arêtes multiples (les arêtes a et b relient tous les deux les sommets A et B tout comme les arêtes c et d qui relient tous deux les sommets A et C).

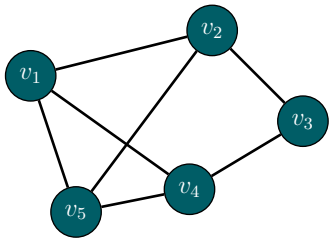


Figure 1.3 – Exemple de graphe simple ($n = 5$ sommets / $m = 7$ arêtes).

Un graphe $G' = (\mathcal{V}', \mathcal{E}')$ est un sous-graphe de G, noté $G' \subseteq G$, si les inclusions des ensembles respectifs $\mathcal{V}' \subseteq \mathcal{V}$ et $\mathcal{E}' \subseteq \mathcal{E}$ sont vérifiées. L'union des deux graphes G et G' peut également être définie comme le graphe $G \cup G' := (\mathcal{V} \cup \mathcal{V}', \mathcal{E} \cup \mathcal{E}')$. Enfin, les graphes G et G' sont dits isomorphes, propriété notée $G \simeq G'$, s'il existe une bijection $f : \mathcal{V} \longrightarrow \mathcal{V}'$ entre les ensembles de sommets de G et G' telle que $\{i, j\}$ soit une arête de G si et seulement si $\{f(i), f(j)\}$ est elle-même une arête de G' [66, 67].

Reprenons le graphe d'origine G : deux sommets i et j sont dits adjacents (ou voisins) s'ils sont connectés par une arête $\{i, j\}$[5]. Ainsi, le voisinage du sommet i, noté $\mathcal{N}(i)$, constitue l'ensemble des sommets lui étant adjacents, c'est-à-dire $\mathcal{N}(i) = \{j \in \mathcal{V} : \{i, j\} \in \mathcal{E}\}$. Pour définir certains graphes aux structures particulières, nous avons besoin de la notion de bipartition définie comme suit : le graphe G est dit biparti si son ensemble de sommets $\mathcal{V}$ est partagé en deux sous-ensembles $\mathcal{V}_1$ et $\mathcal{V}_2$ tels que deux sommets adjacents soient toujours dans deux sous-ensembles différents. La figure 1.4 montre un exemple de graphe biparti G et un sous-graphe de G.

4. Pour un graphe, on parle de sommets/arêtes tandis que pour le réseau sous-jacent, on utilise plutôt nœuds/liens.
5. Les abus de notation classiques seront utilisés : i pour le sommet v_i et $\{i, j\}$ pour définir une arête entre v_i et v_j.



(a) Graphe biparti G ($n = 7$ / $m = 6$). **(b)** Sous-graphe de G ($n = 4$ / $m = 3$).

Figure 1.4 – Exemple de graphe biparti et un sous-graphe. Les nuances de vert révèlent la bipartition.

1.1.3 Degré et distribution des degrés

Une notion très importante en théorie des graphes est celle de degré. En effet, c'est un attribut essentiel d'un sommet qui peut être extrait directement du graphe étudié. Il est également possible, dans le but de caractériser le graphe, de construire sa distribution des degrés, outil primordial au développement des méthodes présentées dans ce manuscrit de thèse. Soit un graphe $G = (\mathcal{V}, \mathcal{E})$ d'ordre n ayant m arêtes. Le degré d'un sommet i, noté $\deg(i)$, est le nombre d'arêtes reliant ce sommet. Un sommet ayant un degré nul est qualifié d'isolé. La valeur $\delta(G) := \min_{i\in\mathcal{V}} \deg(i)$ est appelée degré minimal du graphe et la valeur $\Delta(G) := \max_{i\in\mathcal{V}} \deg(i)$ est appelée degré maximal. Comme chaque arête a un sommet de départ et un sommet d'arrivée, les arêtes sont comptées deux fois si les degrés sont sommés, la formule de la somme des degrés s'exprimant alors par :

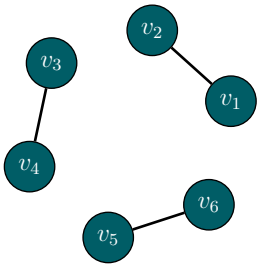
$$\sum_{i\in\mathcal{V}} \deg(i) = 2m. \tag{1.1}$$

Cette valeur est parfois appelée volume du graphe G [8]. À partir de la formule (1.1), il est également possible de calculer le degré moyen, noté $\bar{d}(G)$:

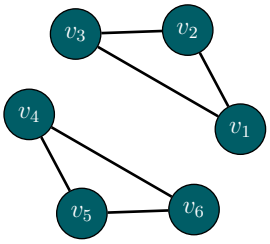
$$\bar{d}(G) = \frac{1}{n} \sum_{i\in\mathcal{V}} \deg(i) = \frac{2m}{n}. \tag{1.2}$$

Il est clair que l'inégalité $\delta(G) \leq \bar{d}(G) \leq \Delta(G)$ est vérifiée[6] avec égalité si et seulement si les n sommets du graphe ont le même degré k : le graphe est alors dit k–régulier. Ainsi, un graphe k–régulier d'ordre n admet $nk/2$ arêtes[7]. Un cas particulier de graphe régulier est le graphe complet à n sommets, noté $\mathcal{K}_n$, c'est-à-dire un graphe pour lequel tous ses sommets sont deux à deux adjacents. Ce graphe est ainsi $(n-1)$–régulier : il a alors $n(n-1)/2$ sommets. La figure 1.5 montre des exemples de graphes k–réguliers dont le graphe complet $\mathcal{K}_6$ d'ordre 6.

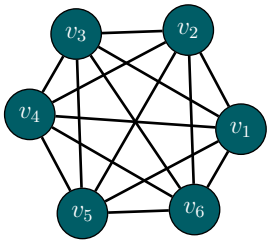
6. S'il y n'a pas d'ambiguïté sur le graphe étudié, on se contentera des notations δ, Δ et $\bar{d}$.
7. Notons que cela oblige nk à être pair et donne donc une condition pour générer des graphes réguliers.



(a) Graphe 1–régulier.



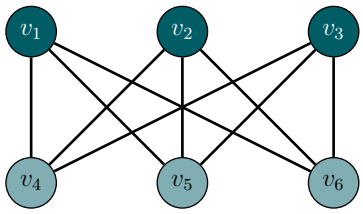
(b) Graphe 2–régulier.



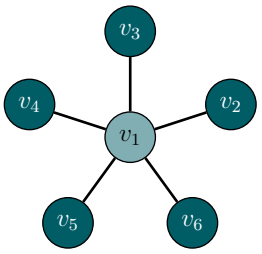
(c) Graphe complet $\mathcal{K}_6$ (graphe 5–régulier).

Figure 1.5 – Exemples de graphes k–régulier pour $k \in \{1, 2, 5\}$.

Il est possible de combiner plusieurs définitions rencontrées jusqu'alors. En effet, le graphe G est dit biparti complet s'il existe $\mathcal{V}_1, \mathcal{V}_2 \subseteq \mathcal{V}$ vérifiant la condition de bipartition et si chaque sommet de $\mathcal{V}_1$ est relié à tous ceux de $\mathcal{V}_2$ ou inversement. Un graphe biparti complet est souvent noté $\mathcal{K}_{n_1,n_2}$ où $n_1 = |\mathcal{V}_1|$ (resp. $n_2 = |\mathcal{V}_2|$). En particulier, le graphe biparti complet $\mathcal{K}_{1,n}$, c'est-à-dire le graphe biparti complet ne possédant qu'un seul sommet dans une des deux partitions est appelé graphe étoile et est noté $\mathcal{S}_n$. La figure 1.6 montre deux exemples de graphes bipartis complets dont un qui n'est autre que le graphe étoile $\mathcal{S}_5$.



(a) Graphe $\mathcal{K}_{3,3}$ ($n = 6$ / $m = 9$)



(b) Graphe $\mathcal{K}_{1,5}$ équivalent au graphe étoile $\mathcal{S}_5$ ($n = 6$ / $m = 5$)

Figure 1.6 – Exemples de graphes bipartis complets. Les nuances de vert révèlent la bipartition.

Pour caractériser la régularité[8] d'un graphe, il est adéquat d'introduire un coefficient $r(G)$ de régularité d'un graphe G défini par

$$r(G) \coloneqq \frac{\delta(G)}{\Delta(G)}. \tag{1.3}$$

Ce coefficient est toujours compris entre 0 et 1. Si $r(G) = 0$, cela signifie que le graphe G est très loin d'être régulier car il possède un sommet isolé. Si, à l'inverse, $r(G)$ vaut 1, cela signifie que le graphe est régulier car $\delta(G) = \Delta(G)$. Cependant, c'est le graphe étoile $\mathcal{S}_n$ qui a conduit à être prudent avec le coefficient de régularité (1.3) car il ne peut comparer deux graphes d'ordres différents. En effet, si deux

8. Savoir si un graphe est régulier est facile car la définition de régularité existe mais trouver une valeur qui caractérise cette régularité, au même titre que la robustesse ou la vulnérabilité d'un réseau que l'on abordera dans un chapitre suivant, relève souvent de la subjectivité.

graphes étoiles $\mathcal{S}_{n_1}$ et $\mathcal{S}_{n_2}$ sont considérés avec $n_1 \leq n_2$, alors $r(\mathcal{S}_{n_2}) \leq r(\mathcal{S}_{n_1})$ bien que ce soient deux graphes qui possèdent la « même » régularité. Pour pallier ce problème, deux autres mesures d'irrégularité[9] d'un graphe G ont été introduites : l'écart de degrés [68] (*degree deviation* en anglais) $\mathrm{dev}(G)$ et la variance de degrés [69] (*degree variance*) $\mathrm{var}(G)$, définies par

$$\mathrm{dev}(G) \coloneqq \sum_{i \in \mathcal{V}} \left| \deg(i) - \frac{2m}{n} \right|, \tag{1.4}$$

$$\mathrm{var}(G) \coloneqq \frac{1}{n} \sum_{i \in \mathcal{V}} \left(\deg(i) - \frac{2m}{n} \right)^2 \tag{1.5}$$

et vérifiant l'inégalité suivante [68] :

$$\frac{\mathrm{dev}^2(G)}{n^2} \leq \mathrm{var}(G) \leq \mathrm{dev}(G). \tag{1.6}$$

Il est clair que ces deux quantités sont égales à 0 si et seulement si le graphe est régulier (au sens théorique) et mesurent alors à quel point le graphe étudié « s'éloigne » de la régularité. Par exemple, le graphe $\mathcal{K}_{3,3}$ à gauche de la figure 1.6 vérifie $\mathrm{dev}(\mathcal{K}_{3,3}) = \mathrm{var}(\mathcal{K}_{3,3}) = 0$ car il est 3-régulier tandis que le graphe $\mathcal{S}_5$ à droite de cette figure vérifie $\mathrm{dev}(\mathcal{S}_5) = 20/3$ et $\mathrm{var}(\mathcal{S}_5) = 24/9$, deux valeurs bien supérieures à 0, ce qui est normal car ce graphe est loin d'être régulier.

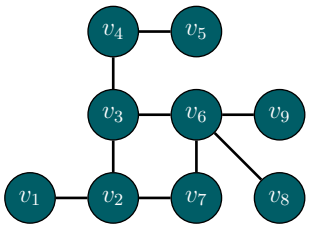
Bien que ce soient les informations les plus importantes concernant les sommets de G, les degrés sont bien souvent exploités sous la forme d'une distribution des degrés $\mathbf{p} = (p_k)_{1 \leq k \leq n}$ définie, dans sa version normalisée, par

$$p_k = \frac{|\{i \in \mathcal{V} : \deg(i) = k\}|}{n}. \tag{1.7}$$

Le fait que la distribution des degrés soit assimilée à une densité de probabilité représente son principal avantage. En effet, $p_k \geq 0$ pour tout $k \in [\![1, n]\!]$ et $\sum_{k=1}^{n} p_k = 1$. Grâce à cela, il est possible de retrouver le degré minimal δ et le degré maximal Δ mais pas seulement car le fait d'utiliser une distribution ouvre la possibilité de bénéficier des outils classiques de statistiques tels que la moyenne (et donc le degré moyen $\bar{d}(G)$ défini par l'équation (1.2)), la variance, le moment d'ordre 3 appelé coefficient d'asymétrie (*skewness* en anglais) ou encore le moment d'ordre 4 appelé coefficient d'aplatissement (*kurtosis*). Toutes ces grandeurs statistiques, aidant à une caractérisation et une classification efficace de graphes, seront les objets d'étude du prochain chapitre. Toutefois, il est à noter que la distribution des degrés est indifférente à une permutation des sommets au même titre que l'histogramme d'une image 2D est insensible à des rotations éventuelles. De plus, l'identification d'un graphe est impos-

9. Pour des raisons évidentes, l'irrégularité est souvent nommée hétérogénéité.

sible à partir de la seule connaissance de sa distribution des degrés. En effet, plusieurs graphes peuvent avoir la même distribution des degrés. La figure 1.7 affiche un graphe d'ordre 9, les degrés de ses sommets et sa distribution des degrés calculée grâce à l'équation (1.7). Il est clair que si les sommets 1 et 6 se retrouvent permutés, la distribution des degrés reste inchangée comme le montre la figure 1.8 et il est également assez simple de trouver un autre graphe que celui affiché ayant la même distribution des degrés. La figure 1.9 en montre un exemple : le graphe présenté sur cette figure est bien différent du graphe de la figure 1.7 et pourtant ils ont la même distribution des degrés.



Sommet i	**Degré** $\deg(i)$
1	1
2	3
3	3
4	2
5	1
6	4
7	2
8	1
9	1

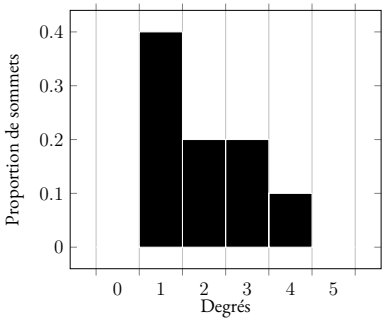
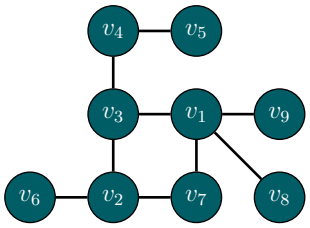


Figure 1.7 – Exemple d'un graphe d'ordre 9, degrés de ses sommets et sa distribution des degrés normalisée.



Sommet i	**Degré** $\deg(i)$
1	4
2	3
3	3
4	2
5	1
6	1
7	2
8	1
9	1

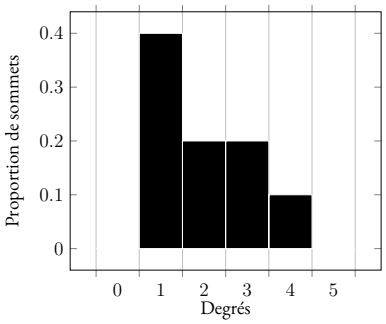
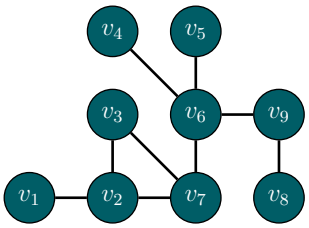


Figure 1.8 – Graphe de la figure 1.7 où les sommets 1 et 6 sont permutés : la distribution des degrés reste inchangée.

La théorie des graphes trouve des applications en chimie moléculaire en ce sens que les molécules peuvent souvent se représenter sous la forme de graphes dont les sommets désignent les atomes et les arêtes les liens entre ces derniers. Une manière de caractériser ces molécules est alors de calculer des quantités à partir des graphes les modélisant. Ces grandeurs sont quelquefois appelées descripteurs de graphes ou invariants de graphes[10] car elles doivent être capables de comparer deux graphes ayant des structures différentes. Une majorité de ces invariants peuvent s'exprimer de manière simple en fonction

10. Un invariant de graphe est une valeur qui reste inchangée par isomorphisme de graphes.



Sommet i	Degré $\deg(i)$
1	1
2	3
3	2
4	1
5	1
6	4
7	3
8	1
9	2

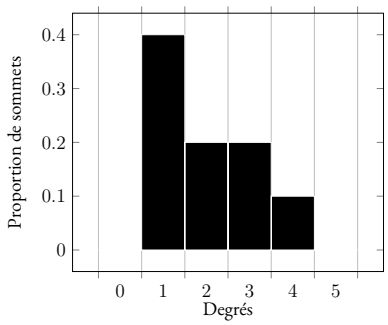


Figure 1.9 – Autre exemple d'un graphe d'ordre 9 possédant la même distribution des degrés que celui de la figure 1.7.

des degrés du graphe. Les indices de Zagreb $Z_1(G)$ et $Z_2(G)$ [70], l'indice hyper-Zagreb $Z_h(G)$ [71], l'indice de Randić $R(G)$ (quelquefois appelé indice de connectivité) [72], l'indice harmonique $H(G)$ [73] et le *inverse sum indeg index* $\mathrm{ISI}(G)$ [74] en sont des exemples :

$$Z_1(G) \coloneqq \sum_{i\in\mathcal{V}} \deg(i)^2 \tag{1.8}$$

$$Z_2(G) \coloneqq \sum_{\{i,j\}\in\mathcal{E}} \deg(i)\deg(j) \tag{1.9}$$

$$Z_h(G) \coloneqq \sum_{\{i,j\}\in\mathcal{E}} (\deg(i)+\deg(j))^2 \tag{1.10}$$

$$R(G) \coloneqq \sum_{\{i,j\}\in\mathcal{E}} \frac{1}{\sqrt{\deg(i)\deg(j)}} \tag{1.11}$$

$$H(G) \coloneqq \sum_{\{i,j\}\in\mathcal{E}} \frac{1}{\deg(i)+\deg(j)} \tag{1.12}$$

$$\mathrm{ISI}(G) \coloneqq \sum_{\{i,j\}\in\mathcal{E}} \frac{1}{\frac{1}{\deg(i)}+\frac{1}{\deg(j)}} = \sum_{\{i,j\}\in\mathcal{E}} \frac{\deg(i)\deg(j)}{\deg(i)+\deg(j)} \tag{1.13}$$

Cette liste, non exhaustive, dont les éléments proviennent majoritairement de la chimie moléculaire, montre d'ores et déjà que le nombre d'attributs qui peuvent être extraits d'un graphe se révèle très important.

1.1.4 Chemins, cycles, connexité et distance

Après s'être intéressé à la structure générale d'un graphe ou encore à la construction de sa distribution des degrés, il est maintenant question de décrire les éléments structurants éventuels que l'on peut trouver dans ce dernier tels que les marches, les cycles et les caractéristiques qui en découlent comme la

distance entre deux sommets, la notion de connectivité ou bien encore celle du diamètre d'un graphe. Comme cela a été précisé lors du rappel historique précédent, c'est vraisemblablement à travers ces éléments qu'est née la théorie des graphes telle qu'on la connait aujourd'hui.

Une marche entre des sommets i et j est une suite d'arêtes consécutives reliant i à j où un sommet peut être visité plus d'une fois et la longueur d'une marche est le nombre d'arêtes qu'elle contient. Une chaîne entre deux sommets i et j est, quant à elle, une suite d'arêtes consécutives reliant i à j, où les sommets visités sont tous distincts et la longueur d'une chaîne est le nombre d'arêtes qu'elle contient. La longueur de la plus courte chaîne entre deux sommets i et j est appelée distance entre i et j et est souvent notée $\mathrm{dist}(i, j)$. S'il n'existe pas de chaîne reliant i et j, il est commun de fixer $\mathrm{dist}(i, j) = \infty$ ou $\mathrm{dist}(i, j) = 0$. Des exemples de graphes et de chaînes sont présentés en figure 1.10.



Figure 1.10 – (À gauche) Graphe de la figure 1.3. (À droite) Deux chaînes reliant les sommets 1 et 2 : la chaîne **bleue** est de longueur 2 et la chaîne **rouge** est de longueur 3.

À présent, la notion de connexité d'un graphe peut être évoquée. En effet, un graphe est dit connecté (ou connexe) si et seulement si une chaîne entre i et j existe pour tout $i, j \in \mathcal{V}$. Une composante connexe (ou composante connectée) d'un graphe est un sous-graphe connexe de ce graphe. Un exemple de graphe connecté est donné en figure 1.11. Un arbre étant défini comme étant un graphe connecté acyclique, c'est-à-dire un graphe connecté qui ne contient aucun cycle, le graphe de la figure 1.11 est en réalité un arbre. Notons qu'un arbre est toujours biparti et qu'en guise de cas particulier, un graphe étoile est également un arbre. Enfin, un arbre couvrant d'un graphe désigne un arbre qui relie tous les sommets de ce graphe.

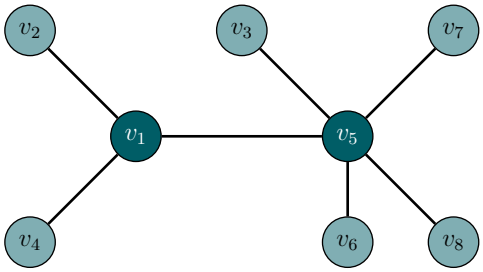


Figure 1.11 – Exemple d'arbre ($n = 8$ / $m = 7$). Les nuances de vert révèlent la bipartition.

Pour caractériser un graphe, en particulier une molécule comme cela a été évoqué plus tôt dans ce manuscrit, il est important d'extraire des quantités de ces graphes. Un invariant de graphe calculé à partir de la distance des plus courtes chaînes est l'indice de Wiener [75] :

$$W(G) = \frac{1}{2} \sum_{\substack{1 \leq i,j \leq n \\ i \neq j}} \mathrm{dist}(i,j) \tag{1.14}$$

Une autre information importante à extraire d'un graphe (qui n'est pas un invariant cette fois) est la longueur moyenne de la chaîne la plus courte, appelée aussi longueur de chaîne caractéristique [76] :

$$L(G) := \frac{1}{n(n-1)} \sum_{\substack{1 \leq i,j \leq n \\ i \neq j}} \mathrm{dist}(i,j). \tag{1.15}$$

Plus cette valeur est faible, plus les dynamiques sur le réseau[11] sont importantes car l'information doit parcourir, en moyenne, une distance plus courte. Les réseaux sociaux sont de bons exemples de graphes ayant des longueurs de chaîne caractéristique $L(G)$ assez faibles. En effet, dès le début du XX$^{\mathrm{e}}$ siècle, l'écrivain hongrois Frigyes Karinthy développe la théorie des « six degrés de séparation » dans laquelle il déclare que deux citoyens américains (pris au hasard) sont reliées en moyenne par une chaîne de six relations. Bien qu'il soit plutôt connu pour ses expériences sur l'autorité et la soumission, c'est le psychologue Stanley Milgram qui reprend cette théorie en tentant de la prouver expérimentalement dans les années 1960 [77], sans grand succès. Il lui revient le terme de petit-monde pour désigner ce genre de réseau et que c'est un domaine toujours actif de recherche, notamment grâce à la montée en puissance du numérique qui provoque un « rétrécissement du monde » : en 2016, la longueur de chaîne caractéristique sur le réseau social numérique Facebook était de 3.57 [78]. Au niveau mathématique, un réseau dit de petit-monde est caractérisé par un graphe ayant une longueur de chaîne caractéristique $L(G)$ proportionnelle au logarithme du nombre de sommets n : $L(G) \propto \log(n)$.

1.1.5 Graphes pondérés

Considérer et quantifier l'importance du lien unissant deux sommets conduit à l'obtention de graphes pondérés . Ainsi, des poids w_{ij} peuvent être attribués aux arêtes $\{i,j\}$. En général, les poids sont des nombres réels positifs ou nuls où un poids w_{ij} égal à 0 signifie que les sommets i et j ne sont pas adjacents. Il est clair qu'un graphe non-pondéré est en réalité un graphe pondéré où les poids

11. Les dynamiques de réseaux peuvent être de nature temporel lorsque la structure du réseau évolue au cours du temps (ce n'est pas le cas ici), ou de nature informationnel lorsque la structure du réseau n'évolue pas mais que des informations, des virus, des idées se propagent entre ces constituants.

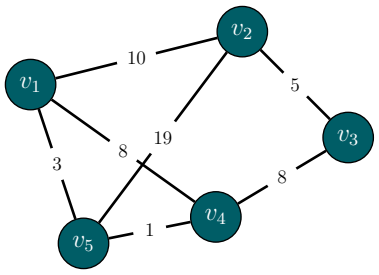
valent tous 1. Il va de soi qu'avoir un tel graphe pondéré permet de représenter plus d'informations quant au réseau sous-jacent. Supposons par exemple que le réseau électrique français soit modélisé sous la forme d'un graphe. Il semble très intéressant, pour caractériser ce réseau, d'ajouter les tensions des lignes électriques, qui pondèrent alors les arêtes du graphe [79]. On peut également citer le réseau de routes aériennes mondiales en les pondérant par le nombre d'avions passant par ces dernières ou encore un réseau de collaboration scientifique où les pondérations seraient le nombre d'articles écrits en commun [80]. Dans le cas d'un graphe pondéré, il convient d'introduire la force (ou degré pondéré) $s(i)$ d'un sommet i, définie comme la somme de toutes les pondérations des arêtes lui étant incidentes [81] :

$$s(i) = \sum_{j \in \mathcal{N}(i)} w_{ij}. \tag{1.16}$$

Il est alors possible de définir la force moyenne comme suit

$$\bar{s} = \frac{1}{n} \sum_{i=1}^{n} s(i). \tag{1.17}$$

Un exemple de graphe pondéré ainsi que l'affichage de la force de ses sommets est donné en figure 1.12. Si ce graphe représentait une carte de débits Internet entre ordinateurs, le débit entre les ordinateurs 2 et 5 serait beaucoup plus important que celui entre les ordinateurs 4 et 5 car $w_{2,5} = 19$ et $w_{4,5} = 1$.



Sommet i	Force $s(i)$
1	21
2	34
3	13
4	17
5	23

Figure 1.12 – Exemple d'un graphe pondéré ($n = 5$ / $m = 7$) et force de ses sommets.

Pour des graphes pondérés, la notion de distance entre sommets est, elle aussi, généralisée en ce sens où l'on suppose que le poids d'une arête représente sa longueur (ou le coût de l'interaction entre deux sommets) : on parle de distance pondérée de la plus courte chaîne entre les sommets i et j, notée $\mathrm{dist}_w(i, j)$ et définie comme la somme minimale des poids de toutes les chaînes reliant i et j. Ainsi, il possible de généraliser l'équation (1.15) en définissant la longueur moyenne pondérée de la plus courte chaîne :

$$L_w(G) := \frac{1}{n(n-1)} \sum_{\substack{1 \leq i,j \leq n \\ i \neq j}} \mathrm{dist}_w(i, j). \tag{1.18}$$

1.2 Structures particulières et graphes aléatoires

Pour étudier, analyser, établir des conjectures ou encore appliquer des méthodes sur des graphes, encore faut-il en disposer. Certains graphes existent de manière native car ils représentent des structures naturelles (réseaux informatiques, sociaux, électriques, etc.). Toutefois, ils sont difficilement accessibles en ligne, en libre accès et bien souvent dans un format non interprétable. Pour tester certaines méthodes, il est possible d'utiliser des structures particulières de graphes définies ci-après ou encore d'utiliser des modèles de générations de graphes aléatoires comme celui d'Erdös-Rényi, Watts-Strogratz ou encore Barabasi-Albert, qui ont l'avantage de créer des graphes possédant de bonnes propriétés vis-à-vis des réseaux qu'ils modélisent. La présentation de ces différentes structures et de ces modèles de génération, associés une description de leurs propriétés, fait l'objet de cette section.

1.2.1 Quelques structures particulières

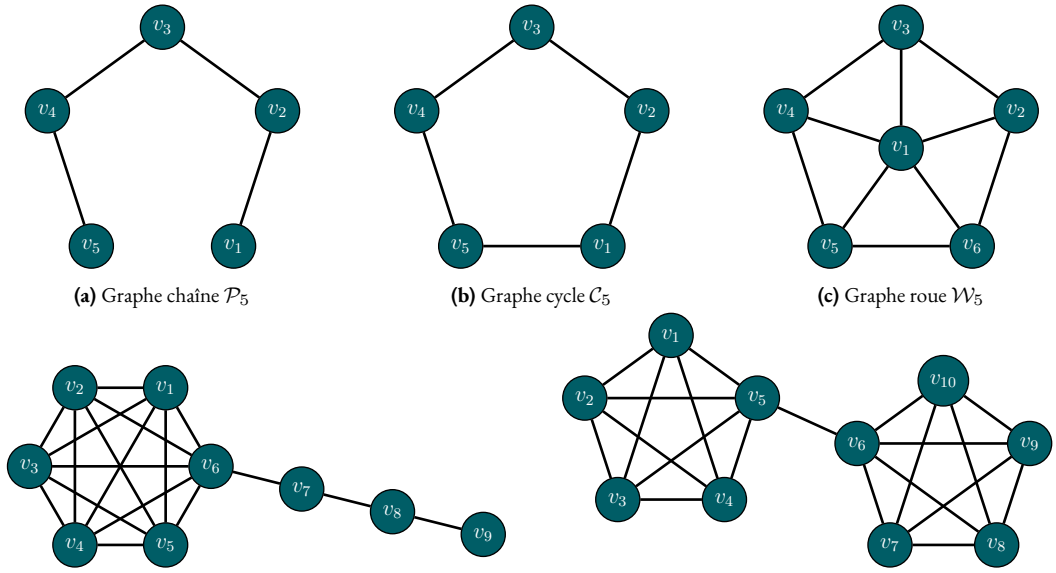
Outre les graphes complet $\mathcal{K}_n$ et étoile $\mathcal{S}_n$ qui ont été rencontrés dans la section précédente, il existe bien d'autres structures particulières qui vont être étudiées dans la suite de ce manuscrit de thèse. Parmi ces dernières, le graphe chaîne $\mathcal{P}_n$ d'ordre n est le graphe constitué des sommets $\{v_i\}_{1\leq i\leq n}$ et des arêtes $\{v_i, v_{i+1}\}_{1\leq i\leq n-1}$. Un graphe cycle $\mathcal{C}_n$ d'ordre n est le graphe constitué des sommets $\{v_i\}_{1\leq i\leq n}$ et des arêtes $\{v_1, v_n\} \cup \{v_i, v_{i+1}\}_{1\leq i\leq n-1}$. Un graphe cycle est toujours un graphe 2–régulier. Un graphe roue $\mathcal{W}_n$ est un graphe cycle $\mathcal{C}_n$ auquel un sommet relié à tous les autres est ajouté. Le graphe comète $\mathcal{C}_{n_1,n_2}$ est défini comme l'union d'un graphe complet $\mathcal{K}_{n_1}$ et d'un graphe chaine $\mathcal{P}_{n_2}$, tous deux reliés par une arête supplémentaire. Enfin, le graphe *Barbell*[12] $\mathcal{B}_n$ est défini comme l'union de deux graphes complets $\mathcal{K}_n$ reliés par une arête. Des exemples de ces cinq graphes particuliers font l'objet de la figure 1.13.

1.2.2 Modèles de graphes aléatoires d'Erdös-Renyi

Les graphes aléatoires d'Erdös-Rényi sont vraisemblablement les plus classiques et les plus simples à construire [83, 84]. En effet, un graphe aléatoire d'Erdös-Rényi $\mathcal{G}_{n,p} = (\mathcal{V}, \mathcal{E})$ est un graphe simple et non-orienté d'ordre $|\mathcal{V}| = n$ et pour lequel les variables aléatoires

$$X_{ij} = \begin{cases} 1, & \text{si } \{i,j\} \in \mathcal{E} \\ 0, & \text{sinon} \end{cases}, \quad 1 \leq i,j \leq n \tag{1.19}$$

12. À notre connaissance, à l'exception des travaux de thèse de H. A. Bay-Ahmed [82], cette appellation n'est pas retrouvée dans la littérature.



(a) Graphe chaîne $\mathcal{P}_5$ **(b)** Graphe cycle $\mathcal{C}_5$ **(c)** Graphe roue $\mathcal{W}_5$

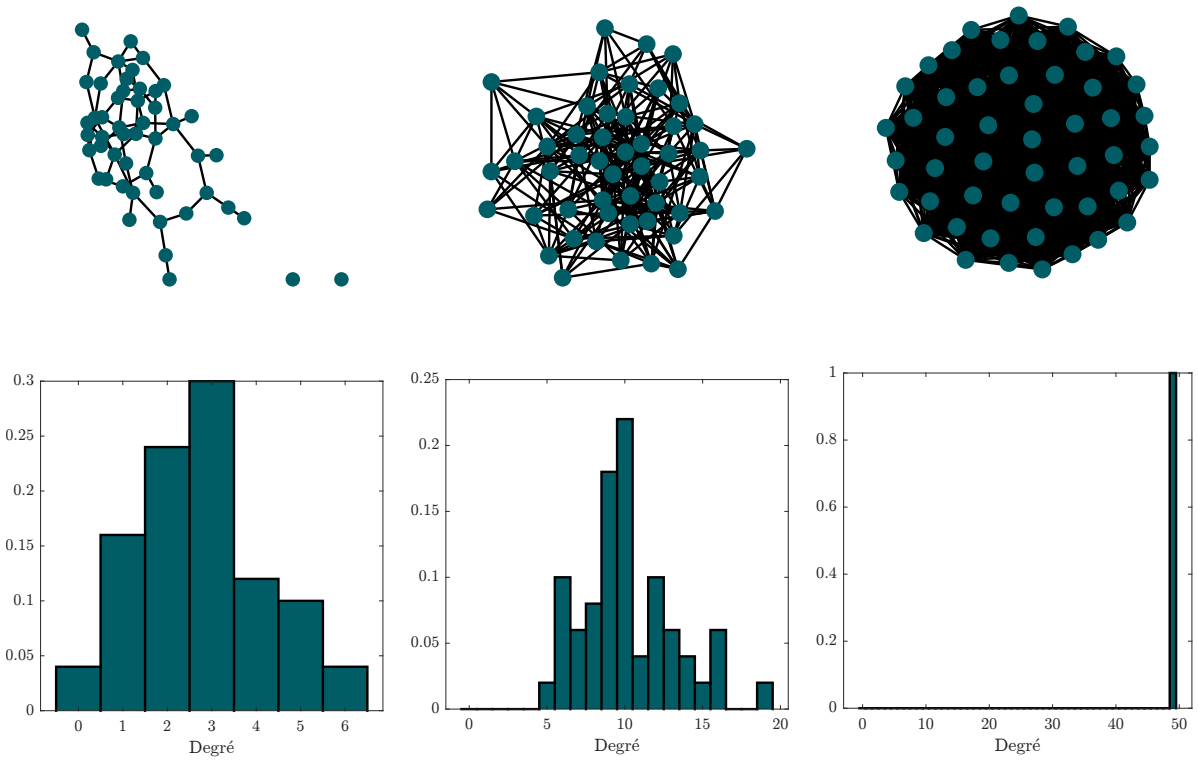
(d) Graphe comète $\mathcal{C}_{6,3}$ **(e)** Graphe *Barbell* $\mathcal{B}_5$

Figure 1.13 – Exemples de graphes particuliers.

sont des variables de Bernoulli indépendantes de paramètre p : $\mathbb{P}(X_{ij} = 1) = 1 - \mathbb{P}(X_{ij} = 0) = p$. En d'autres termes, chaque arête possible entre les sommets de $\mathcal{G}_{n,p}$ apparait avec une probabilité p, indépendamment des autres arêtes. Ainsi, le nombre d'arêtes m_p de $\mathcal{G}_{n,p}$ suit la loi binomiale de paramètres $n(n-1)/2$ et p car il y a, au maximum, $n(n-1)/2$ arêtes possibles. Des exemples de réalisations de graphes aléatoires d'Erdös-Rényi $\mathcal{G}_{n,p}$ avec un nombre de sommets égal à 50 et trois probabilités p de création des arêtes respectivement de 0.05, 0.2 et 1 sont présentés en figure 1.14. Sur cette figure, il peut être constaté que, plus la probabilité p d'apparition d'une arête augmente, plus le graphe devient dense, jusqu'à devenir le graphe complet $\mathcal{K}_{50}$ dans le cas où la probabilité p est égale à 1. Lorsque la probabilité p est trop faible, comme c'est le cas pour le graphe de gauche, alors un certain nombre de sommets isolés apparaissent. En reprenant les notations de l'équation (1.7), la distribution des degrés d'un graphe aléatoire d'Erdös-Rényi $\mathcal{G}_{n,p}$ est binomiale :

$$p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k} \tag{1.20}$$

Le nombre de sommets n'étant pas assez élevé, il ne peut être constaté sur la figure 1.14 le résultat pour lequel la distribution des degrés suit une loi de Poisson pour un nombre de sommets n grand et une valeur np constante [85]. De plus, cette version avec une probabilité d'apparition p d'une arête n'est pas l'originale introduite par Erdös et Rényi. En effet, cette dernière se note $\mathcal{G}_{n,m}$ et sa construction se fait en choisissant uniformément un sous-ensemble de m arêtes parmi les $n(n-1)/2$ possibles [83].



(a) $n = 50 \,/\, p = 0.05$ **(b)** $n = 50 \,/\, p = 0.2$ **(c)** $n = 50 \,/\, p = 1$

Figure 1.14 – Exemples de graphes aléatoires d'Erdös-Rényi $\mathcal{G}_{n,p}$ et leurs distributions de degrés normalisées.

Les graphes aléatoires construits grâce au modèle d'Erdös-Rényi ont deux limitations : leurs distributions de degrés ne convergent pas vers une loi de puissance[13] qui est observée dans bon nombre de réseaux réels (surtout des réseaux sociaux), et ce modèle ne produit pas de graphes ayant la propriété de petit-monde. C'est pourquoi Duncan Watts et Steven Strogratz ont introduit un modèle permettant de générer des graphes vérifiant la propriété de petit-monde [86]. Ayant d'ores et déjà présenté les limitations du modèle d'Erdös-Renyi, il en reste une que le modèle de Watts-Strogratz ne pallie pas : les graphes générés ne sont pas « sans-échelle », c'est-à-dire que leurs distributions de degrés ne suit pas une loi de puissance. C'est pourquoi Albert-László Barabási et Réka Albert ont créé un modèle permettant de générer des graphes possédant les propriétés voulues [76].

1.3 Matrices de représentation et théorie spectrale de graphes

Les graphes, bien que présents dans la quasi-totalité des domaines scientifiques et sociétaux, sont peu utilisés et analysés dans leurs formes structurelles. Les matrices de représentation qui leur sont

13. Une distribution suit une loi de puissance si elle peut s'exprimer grâce à un paramètre γ comme suit : $p_k \sim k^{-\gamma}$.

associées permettent alors de bénéficier de tous les outils d'algèbre linéaire à notre disposition tels que la trace, le déterminant, les exponentielles de matrice, les polynômes caractéristiques, les valeurs et vecteurs propres, les décompositions en valeurs singulières, etc. Il est commun de constater que deux « écoles de pensée » se retrouvent souvent confrontées dans la littérature, notamment dans le domaine du traitement de signal sur graphe : celle traitant de la matrice d'adjacence $\mathbf{A}$ [11] et celle traitant de la matrice Laplacienne $\mathbf{L}$ [10], et la question quant à la pertinence du spectre de chacune des matrices pour l'extraction d'informations liées aux graphes étudiés fait débat. En effet, les valeurs et vecteurs propres sont intéressants pour comprendre l'action d'un opérateur ou une forme quadratique, ce que la matrice d'adjacence $\mathbf{A}$ ne fournit pas contrairement à $\mathbf{L}$, bien qu'elle soit une manière beaucoup plus naturelle de représenter un graphe. Les informations spectrales issues de la matrice d'adjacence $\mathbf{A}$ ne sont pourtant pas dénuées d'intérêt : définition de l'énergie d'un graphe [87], association des fréquences [11], etc.

1.3.1 Matrice d'adjacence et propriétés spectrales

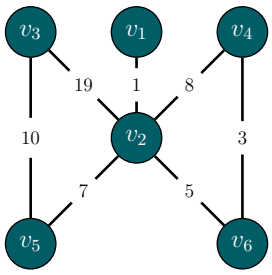
Soit un graphe $G = (\mathcal{V}, \mathcal{E})$ avec $|\mathcal{V}| = n$ sommets et $|\mathcal{E}| = m$ arêtes. La matrice de représentation la plus simple qui traduit directement le graphe G sous forme matricielle est vraisemblablement la matrice d'adjacence $\mathbf{A}(G)$ de taille $n \times n$ et constituée uniquement de 0 et de 1 :

$$\mathbf{A}(G) = [a_{ij}]_{1 \leq i,j \leq n} = \begin{cases} 1, & \text{si } \{i,j\} \in \mathcal{E} \\ 0, & \text{sinon.} \end{cases} \tag{1.21}$$

Il est clair que, dans le cas d'un graphe pondéré, les 1 peuvent être remplacés par les poids w_{ij} des arêtes pour donner la matrice d'adjacence pondérée (ou la matrice de poids) $\mathbf{W} = [w_{ij}]_{1 \leq i,j \leq n}$. Par ailleurs, dans le cas d'un graphe simple (pour rappel, c'est-à-dire sans boucle ni arête multiple), la diagonale est constituée de 0 et est symétrique car les graphes considérés sont non-orientés. Un exemple de graphe pondéré et sa matrice d'adjacence (pondérée ou non) est donné en figure 1.15.

La matrice d'adjacence $\mathbf{A}(G)$ d'un graphe simple et non-orienté étant symétrique et réelle, le théorème spectral nous dit que les valeurs propres $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ de cette matrice contenant les éventuelles multiplicités[14], formant le spectre d'adjacence de G, sont toutes réelles. La valeur propre maximale λ_n est appelée rayon spectral du graphe G. Ce spectre possède de nombreuses propriétés liées à la structure du graphe sous-jacent, en particulier le rayon spectral qui a fait l'objet de nombreux

14. Si la valeur propre λ d'une matrice $\mathbf{M}$ apparaît p fois dans le spectre, on dit qu'elle est de multiplicité p.



(a) Graphe G simple et pondéré

$$\begin{array}{c} \begin{matrix} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 \end{matrix} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \end{pmatrix} \begin{matrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{matrix} \end{array}$$

(b) Matrice d'adjacence $\mathbf{A}(G)$

$$\begin{array}{c} \begin{matrix} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 \end{matrix} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 19 & 8 & 7 & 5 \\ 0 & 19 & 0 & 0 & 10 & 0 \\ 0 & 8 & 0 & 0 & 0 & 3 \\ 0 & 7 & 10 & 0 & 0 & 0 \\ 0 & 5 & 0 & 3 & 0 & 0 \end{pmatrix} \begin{matrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{matrix} \end{array}$$

(c) Matrice des poids $\mathbf{W}(G)$

Figure 1.15 – Représentation d'un graphe grâce à sa matrice d'adjacence (pondérée ou non).

théorèmes : en 1957, Collatz et Sinogowitz ont par exemple prouvé que

$$\lambda_n - \frac{2m}{n} \geq 0 \tag{1.22}$$

avec égalité si et seulement si le graphe est régulier, faisant de cette valeur une bonne mesure de régularité [88]. En remarquant que la somme de la i^{e} ligne de la matrice d'adjacence $\mathbf{A}(G)$ d'un graphe G est égale au degré $\deg(i)$ du sommet i[15], et en notant respectivement δ et Δ le degré minimal et maximal de G, l'inégalité précédente peut être complétée grâce au théorème de Perron-Frobenius [89] :

$$\delta \leq \frac{2m}{n} \leq \lambda_n \leq \Delta. \tag{1.23}$$

Théorème de Perron-Frobenius appliqué aux graphes

Soit un graphe connecté G d'ordre supérieur ou égal à 2 et de matrice d'adjacence $\mathbf{A}(G)$ possédant un spectre $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. Les propriétés suivantes sont vérifiées :

1. La valeur propre maximale λ_n est positive ;
2. La valeur propre maximale λ_n admet une multiplicité égale à 1 associée à un vecteur propre appelé vecteur de Perron dont tous les coefficients sont positifs ;
3. Si θ et Θ sont respectivement le minimum et le maximum des sommes des éléments de chaque ligne de $\mathbf{A}$, alors $\theta \leq \lambda_n \leq \Theta$;
4. La suppression d'une arête fait décroître le rayon spectral λ_n ;
5. Toutes les valeurs propres vérifient $-\lambda_n \leq \lambda_\ell \leq \lambda_n$, pour tout $1 \leq \ell \leq n$.

15. Dans le cas d'un graphe pondéré, cette propriété est valable aussi pour la matrice des poids $\mathbf{W}$ où l'on retrouve non pas le degré $\deg(i)$ du sommet i mais sa force $s(i)$.

Une grandeur importante pour caractériser un graphe est son écart spectral $\lambda_n - \lambda_{n-1}$. Le théorème de Perron-Frobenius nous dit que, dans le cas d'un graphe connecté, la valeur propre maximale λ_n est de multiplicité 1, ce qui signifie que cet écart spectral est strictement positif. Il est alors possible de considérer l'écart spectral comme une mesure de connectivité en établissant que, plus cet écart est proche de 0, plus le graphe est proche de la déconnexion [90]. Par ailleurs, une autre propriété intéressante quant à la régularité de G émerge d'un vecteur propre de sa matrice d'adjacence. En effet, le graphe G est régulier si et seulement si le vecteur unitaire $(1, 1, \ldots, 1)$ est un vecteur propre de $\mathbf{A}(G)$ et lorsque c'est un vecteur propre, la valeur propre associée est égale au degré du graphe régulier [91].

Si G est un graphe non-orienté connecté à $n \geq 2$ sommets, alors sa matrice d'adjacence n'est pas semi-définie positive. En effet, une condition pour avoir la semi-définie positivité d'une matrice est que toutes ses valeurs propres soient positives. Or, ce n'est pas le cas car le théorème de Perron-Frobenius nous dit que $\lambda_n \geq 0$ donc il y a nécessairement une valeur propre négative car la trace[16] de cette matrice, qui n'est autre que la somme des valeurs propres, est nulle (car tous les coefficients diagonaux le sont) :

$$\mathrm{Tr}(\mathbf{A}) = \sum_{\ell=1}^{n} \lambda_\ell = 0. \tag{1.24}$$

Les puissances de la matrice d'adjacence $\mathbf{A}(G)$ ont une interprétation intéressante quant au graphe G [89, 91, 92]. En effet, le nombre de chaînes de longueur K entre les sommets i et j est égal au coefficient (i, j) de la matrice $\mathbf{A}^K$. Ainsi, le nombre de cycles de longueur 2 peut se calculer simplement grâce à la formule suivante :

$$\mathrm{Tr}\big(\mathbf{A}^2\big) = \sum_{\ell=1}^{n} \lambda_\ell^2 = 2m. \tag{1.25}$$

Enfin, l'une des applications de la théorie des graphes en chimie moléculaire est la correspondance étroite, *via* la méthode de Hückel, entre les valeurs propres de la matrice d'adjacence d'un graphe représentant un système d'hydrocarbures conjugués et les niveaux d'énergie E_π des orbitales moléculaires des électrons π dans ces systèmes. En effet, l'énergie E_π n'est autre que la somme des valeurs absolues des valeurs propres correspondant au graphe G représentant la molécule [87]. Inspirés par ce résultat, les travaux de Gutman *et al.* dans les années 1980 ont permis d'étendre le concept d'énergie à tous les graphes simples [70, 93–97]. Ainsi, un graphe G d'ordre n et possédant m arêtes de spectre d'adjacence $(\lambda_\ell)_{1 \leq \ell \leq n}$, l'énergie $E_{\mathbf{A}}(G)$ du graphe G est définie par :

$$E_{\mathbf{A}}(G) \coloneqq \sum_{\ell=1}^{n} |\lambda_\ell|. \tag{1.26}$$

16. La trace $\mathrm{Tr}(\mathbf{M})$ d'une matrice carrée $\mathbf{M} = [m_{ij}]_{1 \leq i,j \leq n}$ est égale à $\mathrm{Tr}(\mathbf{M}) = \sum_{i=1}^{n} m_{ii}$.

Une version $E_{\mathbf{W}}(G)$ dans le cas où le graphe est pondéré de matrice de poids $\mathbf{W}$ existe : les $(\lambda_\ell)_{1\leq\ell\leq n}$ dans l'équation (1.26) sont alors les valeurs propres de $\mathbf{W}$ [98]. L'énergie $E_{\mathbf{A}}(G)$ possède un certain nombre de propriétés, notamment les trois citées ci-après à savoir que l'énergie $E_{\mathbf{A}}(G)$ est positive ou nulle avec égalité si et seulement si le graphe G est vide (c'est-à-dire si $m = 0$). Si G est l'union de deux composantes déconnectées G_1 et G_2, alors $E_{\mathbf{A}}(G) = E_{\mathbf{A}}(G_1) + E_{\mathbf{A}}(G_2)$. Enfin, si G contient une composante connectée G_1 et que tous les autres sont des sommets isolés, alors $E_{\mathbf{A}}(G) = E_{\mathbf{A}}(G_1)$. L'énergie d'un graphe a fait l'objet de nombreuses études notamment pour l'approcher [87], lui trouver des bornes ou encore trouver les graphes d'énergies maximales [99–103].

En réalité, l'énergie $E_{\mathbf{A}}(G)$ d'un graphe G n'est qu'une variante de l'énergie $E_{\mathbf{M}}$ d'une matrice $\mathbf{M} \in \mathcal{M}_n(\mathbb{C})$, ayant pour spectre $(\nu_\ell)_{1\leq\ell\leq n}$, introduite par Bravo *et al.* en 2017 [104], en ce sens que, pour un graphe, la matrice $\mathbf{M}$ n'est autre que la matrice d'adjacence $\mathbf{A}$ qui a une trace nulle :

$$E_{\mathbf{M}} \coloneqq \sum_{\ell=1}^{n} \left| \nu_\ell - \frac{\operatorname{Tr}(\mathbf{M})}{n} \right| . \tag{1.27}$$

Toujours dans le domaine de la chimie moléculaire, une grandeur appelée indice d'Estrada, permettant de caractériser le repliement d'une protéine, peut être extraite de la matrice d'adjacence du graphe modélisant cette dernière. Soit G un graphe de spectre d'adjacence $(\lambda_\ell)_{1\leq\ell\leq n}$, l'indice d'Estrada $\mathrm{EE}(G)$ du graphe G est défini par [105, 106]

$$\mathrm{EE}(G) \coloneqq \operatorname{Tr}\left(\mathrm{e}^{\mathbf{A}}\right) = \sum_{\ell=1}^{n} \mathrm{e}^{\lambda_\ell} . \tag{1.28}$$

Il est également possible de définir la connectivité naturelle du graphe G à partir de son indice d'Estrada en calculant ce qui s'apparente à une valeur propre moyenne [107] :

$$\bar{\lambda}(G) \coloneqq \ln\left(\frac{1}{n}\,\mathrm{EE}(G)\right) = \ln\left(\frac{1}{n}\sum_{\ell=1}^{n} \mathrm{e}^{\lambda_\ell}\right) . \tag{1.29}$$

1.3.2 Matrice des degrés

Soit un graphe $G = (\mathcal{V}, \mathcal{E})$ avec $|\mathcal{V}| = n$ sommets et $|\mathcal{E}| = m$ arêtes. La matrice des degrés $\mathbf{D}(G)$ du graphe G est une matrice diagonale dont le i^{e} élément est égal au degré du sommet i :

$$[\mathbf{D}(G)]_{1\leq i,j\leq n} = \begin{cases} \deg(i), & \text{si } i = j \\ 0, & \text{sinon.} \end{cases} \tag{1.30}$$

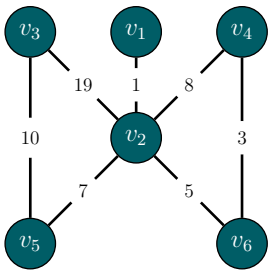
Dans le cas où le graphe G est pondéré, il est possible de calculer sa matrice des forces $\mathbf{S}(G)$ où l'on trouverait sur la diagonale les forces $s(i)$ des sommets i définies par l'équation (1.16) :

$$[\mathbf{S}(G)]_{1\leq i,j\leq n} = \begin{cases} s(i), & \text{si } i = j \\ 0, & \text{sinon.} \end{cases} \tag{1.31}$$

Un exemple de graphe pondéré représenté par sa matrice des degrés ou des forces est donné en figure 1.16. Le calcul de la trace de la matrice des degrés $\mathbf{D}$ nous permet de retrouver la formule de la somme des degrés (1.1) :

$$\mathrm{Tr}(\mathbf{D}) = \sum_{i=1}^{n} d_{ii} = \sum_{i=1}^{n} \deg(i) = 2m. \tag{1.32}$$

Enfin, il est à noter que, étant donné que la matrice des degrés $\mathbf{D}$ est diagonale, son spectre est constitué de ses coefficients diagonaux, et donc des degrés du graphe G. Faire une étude spectrale sur la matrice des degrés revient alors à faire une étude structurelle sur le graphe sous-jacent, *via* ses degrés.



(a) Graphe G simple et pondéré

$$\begin{array}{c} \begin{array}{cccccc} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{pmatrix} \end{array}$$

(b) Matrice des degrés $\mathbf{D}(G)$

$$\begin{array}{c} \begin{array}{cccccc} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 40 & 0 & 0 & 0 & 0 \\ 0 & 0 & 29 & 0 & 0 & 0 \\ 0 & 0 & 0 & 11 & 0 & 0 \\ 0 & 0 & 0 & 0 & 17 & 0 \\ 0 & 0 & 0 & 0 & 0 & 8 \end{pmatrix} \end{array}$$

(c) Matrice des forces $\mathbf{S}(G)$

Figure 1.16 – Représentation d'un graphe à partir de sa matrice des degrés.

1.3.3 Matrice(s) Laplacienne(s) : variantes et propriétés spectrales

Matrice Laplacienne L

La « concurrente » directe de la matrice d'adjacence $\mathbf{A}$ d'un graphe G est sans aucun doute la matrice Laplacienne $\mathbf{L}$ qui, alors qu'elle est définie à partir de $\mathbf{A}$, possède une interprétation physique grâce à une forme quadratique présentée ci-après, dont cette dernière est dépourvue. Bien qu'elle ait été utilisée dans les années 1970 dans le domaine de la chimie moléculaire [43,108], le cadre théorique quant à son analyse spectrale n'est fourni que dans les années 1980 et 1990 [8,109–112].

La matrice Laplacienne (quelquefois qualifiée de discrète) $\mathbf{L}(G)$ du graphe G, carrée de taille $n \times n$,

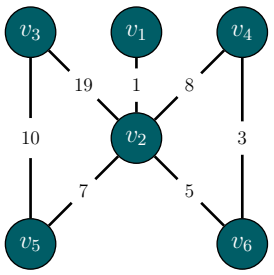
est définie comme la différence entre sa matrice des degrés et sa matrice d'adjacence :

$$\mathbf{L}(G) \coloneqq \mathbf{D}(G) - \mathbf{A}(G), \qquad [\mathbf{L}(G)]_{1\leq i,j\leq n} = \begin{cases} -1, & \text{si } i \neq j \text{ et } \{i,j\} \in \mathcal{E} \\ \deg(i), & \text{si } i = j \\ 0, & \text{sinon.} \end{cases} \tag{1.33}$$

Bien entendu, dans le cas où le graphe est pondéré, il est possible de lui définir sa matrice Laplacienne pondérée comme étant la différence entre sa matrice des forces et la matrice des poids :

$$\mathbf{L}(G) \coloneqq \mathbf{S}(G) - \mathbf{W}(G), \qquad [\mathbf{L}(G)]_{1\leq i,j\leq n} = \begin{cases} -w_{ij}, & \text{si } i \neq j \text{ et } \{i,j\} \in \mathcal{E} \\ s(i), & \text{si } i = j \\ 0, & \text{sinon.} \end{cases} \tag{1.34}$$

Une illustration de ces deux matrices permettant de représenter un graphe pondéré est donnée en figure 1.17.



(a) Graphe G simple et pondéré

$$\begin{array}{c} \begin{matrix} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 \end{matrix} \\ \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 5 & -1 & -1 & -1 & -1 \\ 0 & -1 & 2 & 0 & -1 & 0 \\ 0 & -1 & 0 & 2 & 0 & -1 \\ 0 & -1 & -1 & 0 & 2 & 0 \\ 0 & -1 & 0 & -1 & 0 & 2 \end{pmatrix} \end{array}$$

(b) Matrice Laplacienne $\mathbf{L} = \mathbf{D} - \mathbf{A}$

$$\begin{array}{c} \begin{matrix} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 \end{matrix} \\ \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 40 & -19 & -8 & -7 & -5 \\ 0 & -19 & 29 & 0 & -10 & 0 \\ 0 & -8 & 0 & 11 & 0 & -3 \\ 0 & -7 & -10 & 0 & 17 & 0 \\ 0 & -5 & 0 & -3 & 0 & 8 \end{pmatrix} \end{array} \begin{matrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{matrix}$$

(c) Matrice Laplacienne $\mathbf{L} = \mathbf{S} - \mathbf{W}$

Figure 1.17 – Représentation d'un graphe grâce à sa matrice Laplacienne (issue de la matrice de poids ou non).

Dans le cas d'un graphe simple et non-orienté, la matrice Laplacienne est symétrique et à coefficients réels : elle possède alors un spectre réel $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_n$, appelé spectre Laplacien où, par analogie avec la matrice d'adjacence, la valeur propre maximale μ_n est appelée rayon spectral Laplacien.

Toutefois, à la différence de la matrice d'adjacence, une des forces de cette matrice Laplacienne réside dans le fait qu'elle soit semi-définie positive. En effet, la matrice $\mathbf{L}$ est symétrique et à diagonale dominante[17]. De plus, la définition de la semi-définie positivité montre que, pour tout $\mathbf{x} \in \mathbb{R}^n_*$, le

17. Une matrice carrée à coefficients réels $\mathbf{M} \in \mathcal{M}_n(\mathbb{R})$ est dite à diagonale dominante lorsque la valeur absolue de chaque terme diagonal est supérieure ou égale à la somme des valeurs absolues des autres termes de sa ligne.

scalaire $\mathbf{x}\mathbf{L}\mathbf{x}^\top$ est positif ou nul. En effet,

$$\mathbf{x}\mathbf{L}\mathbf{x}^\top = \frac{1}{2}\sum_{\{i,j\}\in\mathcal{E}}(x_i - x_j)^2 \geq 0 \tag{1.35}$$

Dans le cas pondéré, l'expression prend la forme

$$\mathbf{x}\mathbf{L}\mathbf{x}^\top = \frac{1}{2}\sum_{\{i,j\}\in\mathcal{E}} w_{ij}(x_i - x_j)^2 \geq 0 \tag{1.36}$$

donnant par ailleurs la forme quadratique associée à la matrice Laplacienne $\mathbf{L}$ qui mesure la « régularité » du vecteur $\mathbf{x}$ sur les sommets du graphe G, permettant alors de définir la semi-norme[18] $\|\mathbf{x}\|_\mathbf{L} = \sqrt{\mathbf{x}\mathbf{L}\mathbf{x}^\top}$ [10, 113]. Grâce à la semi-définie positivité de la matrice Laplacienne d'un graphe, il est donc admis que son spectre Laplacien est composé de valeurs propres positives ou nulles. En l'occurrence, la première valeur propre μ_1 est toujours égale à 0 et associée au vecteur propre $\mathbf{1}_n \coloneqq (1, 1, \ldots, 1)$. Pour s'en convaincre, il suffit de remarquer que la somme de chaque ligne de cette matrice est nulle. Ainsi, on a bien $\mathbf{L}\mathbf{1}_n = 0\mathbf{1}_n$ ce qui prouve la propriété sus-citée.

Soit $\mathbf{L}$ la matrice Laplacienne d'un graphe G. Fiedler prouva que la multiplicité de la valeur propre $\mu_1 = 0$ de $\mathbf{L}$ correspond au nombre de composantes connectées de G [114, 115]. Il va de soi qu'une valeur propre particulièrement importante est alors μ_2 (la deuxième plus petite du spectre Laplacien en incluant les multiplicités) appelée valeur de Fiedler (ou connectivité algébrique). En effet, si celle-ci est égale à 0, cela signifie que le graphe est déconnecté et elle est strictement positive si le graphe est connecté. Une intuition est alors de dire que, plus la valeur de Fiedler est importante, plus le graphe est connecté avec une limite égale à $\mu_2 = n$, atteinte dans le cas où G est le graphe complet $\mathcal{K}_n$ (graphe le plus connecté possible). Le vecteur propre associé à la valeur de Fiedler est appelée vecteur de Fiedler et possède un intérêt particulier lorsqu'un partitionnement spectral est opéré sur le graphe G [116].

Comme pour la matrice d'adjacence $\mathbf{A}$, d'autres propriétés spectrales sont vérifiées par la matrice Laplacienne $\mathbf{L}$. Par exemple, un calcul de traces permet de retrouver des attributs structurels :

$$\mathrm{Tr}(\mathbf{L}) = \sum_{i\in\mathcal{V}}\deg(i) = \sum_{\ell=1}^{n}\mu_\ell = 2m \tag{1.37}$$

$$\mathrm{Tr}\big(\mathbf{L}^2\big) = \sum_{\ell=1}^{n}\mu_\ell^2 = 2m + \sum_{i\in\mathcal{V}}\deg(i)^2 = 2m + Z_1(G) \tag{1.38}$$

où $Z_1(G)$ est le premier indice de Zagreb du graphe G défini par l'équation (1.8). Par ailleurs, si le

18. Une semi-norme N est une « norme » ne vérifiant pas la propriété de séparation : $N(\mathbf{x}) = 0$ n'implique pas $\mathbf{x} = \mathbf{0}$.

graphe G est k–régulier, alors sa matrice des degrés est égale à $k\mathbf{I}_n$. Sa matrice Laplacienne est donc égale à $\mathbf{L} = k\mathbf{I}_n - \mathbf{A}$. Par conséquent, si $(\lambda_\ell)_{1\leq\ell\leq n}$ représentent les valeurs propres de $\mathbf{A}$, les valeurs propres $(\mu_\ell)_{1\leq\ell\leq n}$ de $\mathbf{L}$ vérifient

$$\mu_\ell = k - \lambda_\ell, \qquad \forall \ell \in [\![1, n]\!]. \tag{1.39}$$

La conjecture suivante souvent formulée repose sur cette propriété (1.39) [113, 117] : le spectre d'adjacence doit être étudié dans le sens inverse de celui du spectre Laplacien, c'est-à-dire que la première valeur propre de $\mathbf{A}$ correspond à celle de $\mathbf{L}$ et inversement. Par ailleurs, dans le cas particulier où G est le graphe complet $\mathcal{K}_n$, alors son spectre Laplacien n'est composé que de deux valeurs propres distinctes : 0 de multiplicité 1 et n de multiplicité $n-1$. En réalité, c'est une équivalence dont le sens réciproque est facile à vérifier. En effet, la somme des valeurs propres est égale à $n(n-1)$ mais doit aussi être égale à $2m$ d'après l'équation (1.37) : le seul graphe à n sommets possédant $m = n(n-1)/2$ arêtes est le graphe complet $\mathcal{K}_n$. En se servant de l'équation (1.39) car le graphe complet $\mathcal{K}_n$ est un graphe $(n-1)$–régulier, son spectre d'adjacence n'est alors composé que de deux valeurs propres distinctes : -1 de multiplicité $n-1$ et $n-1$ de multiplicité 1.

Un autre point intéressant concerne la suppression d'une arête et son impact sur les spectres. Pour la matrice d'adjacence $\mathbf{A}$, la suppression d'une arête du graphe dont elle est issue provoque une décroissance du rayon spectral (la valeur propre maximale de $\mathbf{A}$). Pour la matrice Laplacienne, une propriété plus importante, dite d'entrelacement des spectres, se démontre : soit un graphe G et $G' := G - e$ le graphe G auquel l'arête e a été retirée, de spectres Laplaciens respectifs $(\mu_\ell)_{1\leq\ell\leq n}$ et $(\mu'_\ell)_{1\leq\ell\leq n}$, alors [118]

$$0 = \mu'_1 = \mu_1 \leq \mu'_2 \leq \mu_2 \leq \cdots \leq \mu'_{n-1} \leq \mu_{n-1} \leq \mu'_n \leq \mu_n. \tag{1.40}$$

La matrice $\mathbf{L}$ est parfois appelée matrice Laplacienne combinatoire, matrice Laplacienne discrète, matrice de Kirchhoff ou matrice d'admittance. Pourquoi ces deux dernières dénominations ? Cela est vraisemblablement dû à deux constats : le premier est que la matrice Laplacienne $\mathbf{L}$ joue un rôle dans le *matrix-tree theorem* appelé aussi théorème de Kirchhoff caractérisant le nombre d'arbres couvrants dans un graphe. En effet, soit un graphe G de spectre Laplacien $(\mu_\ell)_{1\leq\ell\leq n}$, alors son nombre d'arbres couvrants T vérifie [119]

$$T = \frac{1}{n}\prod_{\ell=2}^{n} \mu_\ell. \tag{1.41}$$

Un autre élément provient du fait qu'un graphe à n sommets et m arêtes peut être vu comme un circuit électrique où les arêtes représentent des résistances unitaires (égales à $1\ \Omega$), rendant possible la

définition de la distance de résistance Ω_{ij} entre deux sommets i et j grâce à la formule suivante

$$\Omega_{ij} \coloneqq [\mathbf{L}^\dagger]_{ii} + [\mathbf{L}^\dagger]_{jj} - 2[\mathbf{L}^\dagger]_{ij} \tag{1.42}$$

avec $\mathbf{L}^\dagger$ l'inverse généralisée au sens de Moore-Penrose[19] de la matrice Laplacienne $\mathbf{L}$ [120, 121]. Par suite, une grandeur permettant de caractériser le graphe (et donc le réseau électrique sous-jacent), connues sous le nom d'indice de Kirchhoff[20], apparait comme une variante de l'indice de Wiener (1.14) définie par [123]

$$K(G) \coloneqq \frac{1}{2} \sum_{\substack{1 \leq i,j \leq n \\ i \neq j}} \Omega_{ij}. \tag{1.43}$$

En réalité, si le graphe G est connecté, son indice de Kirchhoff (1.43) et son spectre Laplacien $(\mu_\ell)_{1 \leq \ell \leq n}$ sont intimement liés par la relation suivante [120, 122] :

$$K(G) = n \operatorname{Tr}\left(\mathbf{L}^\dagger\right) = n \sum_{\ell=2}^{n} \frac{1}{\mu_\ell}. \tag{1.44}$$

À l'aide de l'énergie généralisée d'une matrice (1.27), une autre manière de caractériser un graphe G est d'étudier l'énergie Laplacienne $E_{\mathbf{L}}(G)$ de ce dernier [44] :

$$E_{\mathbf{L}}(G) \coloneqq \sum_{\ell=1}^{n} \left| \mu_\ell - \frac{2m}{n} \right|. \tag{1.45}$$

Bien entendu, son introduction a été motivée par celle de l'énergie $E_{\mathbf{A}}(G)$ d'un graphe (équation (1.26)) définie par Gutman *et al.*. Cette énergie Laplacienne peut être vue comme une mesure de complexité du graphe étudié [124]. Une recherche de graphes maximisant cette énergie a alors été envisagée [125].

L'énergie Laplacienne $E_{\mathbf{L}}(G)$ d'un graphe est définie à partir des valeurs propres de la matrice Laplacienne de ce même graphe. Le problème est qu'elle n'a pas les mêmes propriétés que l'énergie $E_{\mathbf{A}}(G)$ listée dans la section correspondante, d'où l'introduction de la Laplacian Energy Like $\mathrm{LEL}(G)$ d'un graphe G admettant $(\mu_\ell)_{1 \leq \ell \leq n}$ pour spectre Laplacien [126] :

$$\mathrm{LEL}(G) \coloneqq \sum_{\ell=1}^{n} \sqrt{\mu_\ell} \tag{1.46}$$

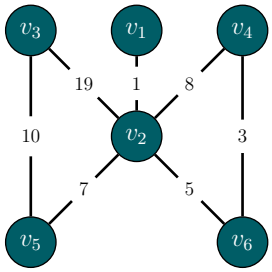
19. La notion d'inverse généralisée est requise car, 0 étant valeur propre de $\mathbf{L}$, cette dernière ne peut pas être inversible.
20. Trouver le nom de Kirchhoff ici n'est pas étonnant : en effet, c'est ce dernier qui a donné son nom aux lois éponymes caractérisant les résistances dans un réseau électrique [122].

Cette énergie possède de nombreuses propriétés, notamment celle caractérisant la complexité du graphe étudié et décrit bien les propriétés qui sont à l'origine de la majorité des descripteurs moléculaires [127, 128]. Toutefois, son nom est trompeur car, même si cette énergie est basée sur le spectre Laplacien, elle est plus proche de l'énergie $E_{\mathbf{A}}(G)$ que de l'énergie Laplacienne $E_{\mathbf{L}}(G)$ [128].

Matrice Laplacienne sans-signe Q

Bien entendu, d'autres « variantes » de la matrice Laplacienne existent comme la matrice Laplacienne sans-signe d'un graphe G, définie quant à elle comme la somme de sa matrice des degrés et de sa matrice d'adjacence et dont un exemple est donné en figure 1.18 :

$$\mathbf{Q}(G) \coloneqq \mathbf{D}(G) + \mathbf{A}(G). \tag{1.47}$$



(a) Graphe G simple et pondéré

$$\begin{array}{c} \begin{array}{cccccc} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 \end{array} \\ \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 5 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 0 & 1 & 0 \\ 0 & 1 & 0 & 2 & 0 & 1 \\ 0 & 1 & 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 1 & 0 & 2 \end{pmatrix} \end{array}$$

(b) Matrice Laplacienne sans-signe $\mathbf{Q} = \mathbf{D} + \mathbf{A}$

$$\begin{array}{c} \begin{array}{cccccc} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 \end{array} \\ \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 40 & 19 & 8 & 7 & 5 \\ 0 & 19 & 29 & 0 & 10 & 0 \\ 0 & 8 & 0 & 11 & 0 & 3 \\ 0 & 7 & 10 & 0 & 17 & 0 \\ 0 & 5 & 0 & 3 & 0 & 8 \end{pmatrix} \begin{array}{c} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{array} \end{array}$$

(c) Matrice Laplacienne sans-signe $\mathbf{Q} = \mathbf{S} + \mathbf{W}$

Figure 1.18 – Représentation d'un graphe *via* sa matrice Laplacienne sans-signe (issue de la matrice de poids ou non).

Cette matrice semble avoir été découverte en même temps que la matrice Laplacienne $\mathbf{L}$, soit dans les années 1970, mais il faudra attendre les années 2000 pour qu'elle soit réintroduite et étudiée [9, 48, 118]. Une fois de plus, la version pondérée existe si le graphe l'est également : $\mathbf{Q}(G) \coloneqq \mathbf{S}(G) + \mathbf{W}(G)$. En suivant la même stratégie pour la matrice Laplacienne discrète $\mathbf{L}$, il est possible de montrer que cette matrice est aussi semi-définie positive avec une valeur propre minimale égale à 0 toutefois, contrairement à la matrice Laplacienne $\mathbf{L}$, la multiplicité de cette valeur propre est cette fois égale au nombre de composantes biparties du graphe étudié [9]. L'énergie Laplacienne sans-signe $E_{\mathbf{Q}}(G)$ établie à partir de la matrice Laplacienne sans-signe $\mathbf{Q}$ de valeurs propres $(\mu_\ell^+)_{1\leq\ell\leq n}$ présente une similitude avec l'énergie Laplacienne[21] [129] :

$$E_{\mathbf{Q}}(G) \coloneqq \sum_{\ell=1}^{n} \left| \mu_\ell^+ - \frac{2m}{n} \right|. \tag{1.48}$$

21. Ce qui est tout à fait logique car l'énergie d'une matrice dépend de sa trace. Or, $\mathrm{Tr}(\mathbf{Q}) = \mathrm{Tr}(\mathbf{L}) = 2m$.

Matrice Laplacienne normalisée $\boldsymbol{\mathcal{L}}$

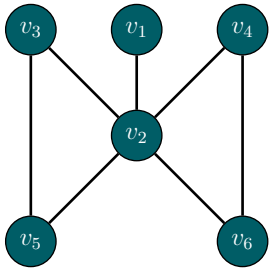
À bien des égards, notamment pour leurs comparaisons, il est souhaitable que les spectres des graphes étudiés soient bornés, peu importe la structure de ces derniers. C'est à cet objectif que répond la matrice Laplacienne normalisée $\boldsymbol{\mathcal{L}}(G)$, pièce maîtresse de l'analyse spectrale de Chung [8] et définie par

$$\boldsymbol{\mathcal{L}}(G) \coloneqq \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2} = \mathbf{I}_n - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2} \tag{1.49}$$

de coefficients

$$[\boldsymbol{\mathcal{L}}(G)]_{1\leq i,j\leq n} = \begin{cases} 1, & \text{si } i = j \text{ et } \deg(i) \neq 0 \\ -\frac{1}{\sqrt{\deg(i)\deg(j)}}, & \text{si } \{i,j\} \in \mathcal{E} \\ 0, & \text{sinon.} \end{cases} \tag{1.50}$$

La matrice $\boldsymbol{\mathcal{L}}$ est qualifiée de « normalisée » car seuls des 1 (ou des 0 s'il y a des sommets isolés) sont présents sur la diagonale. Un exemple d'une telle matrice de représentation de graphe est donné en figure 1.19.



(a) Graphe G simple et non-orienté

$$\begin{array}{c} \begin{array}{cccccc} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 \end{array} \\ \begin{pmatrix} 1 & -\frac{1}{\sqrt{5}} & 0 & 0 & 0 & 0 \\ -\frac{1}{\sqrt{5}} & 1 & -\frac{1}{\sqrt{10}} & -\frac{1}{\sqrt{10}} & -\frac{1}{\sqrt{10}} & -\frac{1}{\sqrt{10}} \\ 0 & -\frac{1}{\sqrt{10}} & 1 & 0 & -\frac{1}{\sqrt{4}} & 0 \\ 0 & -\frac{1}{\sqrt{10}} & 0 & 1 & 0 & -\frac{1}{\sqrt{4}} \\ 0 & -\frac{1}{\sqrt{10}} & -\frac{1}{\sqrt{4}} & 0 & 1 & 0 \\ 0 & -\frac{1}{\sqrt{10}} & 0 & -\frac{1}{\sqrt{4}} & 0 & 1 \end{pmatrix} \end{array} \begin{array}{c} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{array}$$

(b) Matrice Laplacienne normalisée $\boldsymbol{\mathcal{L}}$

Figure 1.19 – Représentation d'un graphe grâce à sa matrice Laplacienne normalisée.

Cette approche « normalisée » permet également d'avoir des grandeurs extraites de cette matrice plus souvent comparables avec des invariants de graphes [113]. Il est possible de vérifier que cette matrice est semi-définie positive [8] : elle possède alors un spectre $(\chi_\ell)_{1\leq\ell\leq n}$ positif ou nul, appelé spectre Laplacien normalisé. En supposant que ce spectre soit rangé dans l'ordre croissant, les propriétés suivantes, issues de celles de la matrice Laplacienne $\mathbf{L}$, sont vérifiées : $\chi_1 = 0$ et la multiplicité de la valeur propre 0 est égale au nombre de composantes connectées du graphe étudié. Toutefois, un intérêt de la matrice Laplacienne normalisée est de voir son spectre borné par 2 avec $\chi_\ell = 2$ si et seulement si le

graphe G est biparti [8] :

$$0 = \chi_1 \leq \chi_2 \leq \cdots \leq \chi_{n-1} \leq \chi_n \leq 2. \tag{1.51}$$

Citons encore la propriété spectrale

$$\mathrm{Tr}(\boldsymbol{\mathcal{L}}) = \sum_{\ell=1}^{n} \chi_\ell \leq n \tag{1.52}$$

avec égalité si et seulement si le graphe ne possède pas de sommets isolés. Si tel est le cas, alors

$$\chi_n \geq \frac{n}{n-1}. \tag{1.53}$$

Par ailleurs, pour un graphe G avec $n \geq 2$ sommets, la deuxième plus petite valeur propre de $\boldsymbol{\mathcal{L}}$ vérifie

$$\chi_2 \leq \frac{n}{n-1} \tag{1.54}$$

avec égalité si et seulement G est le graphe complet $\mathcal{K}_n$ où, dans ce cas, le spectre Laplacien normalisée est constitué des valeurs propres 0 et $n/(n-1)$ de multiplicité respective 1 et $n-1$. Pour un graphe G qui n'est pas le graphe complet, alors $\chi_2 \leq 1$. Par ailleurs, la matrice Laplacienne normalisée d'un graphe k–régulier est égale à $\boldsymbol{\mathcal{L}} = \frac{1}{k}\mathbf{L} = \frac{1}{k}(\mathbf{D} - \mathbf{A}) = \mathbf{I}_n - \frac{1}{k}\mathbf{A}$ ce qui signifie que λ_ℓ est une valeur propre de $\mathbf{A}$ si et seulement si $1 - \frac{\lambda_\ell}{k}$ en est une de $\boldsymbol{\mathcal{L}}$.

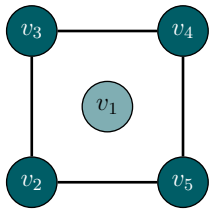
La matrice Laplacienne normalisée $\boldsymbol{\mathcal{L}}$ de spectre $(\chi_\ell)_{1\leq\ell\leq n}$ permet, elle aussi, de définir une énergie Laplacienne normalisée basée sur la définition générale de Bravo *et al.* [104, 130] :

$$E_{\boldsymbol{\mathcal{L}}}(G) = \sum_{\ell=1}^{n} \left|\chi_\ell - 1\right|. \tag{1.55}$$

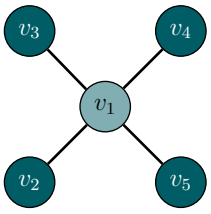
1.3.4 Étude du problème de cospectralité

Deux graphes différents peuvent être $\mathbf{M}$–équiénergétiques [131], c'est-à-dire posséder la même énergie ($E_{\mathbf{M}}(G_1) = E_{\mathbf{M}}(G_2)$ avec pourtant $G_1 \neq G_2$). C'est particulièrement le cas pour des graphes dits $\mathbf{M}$–cospectraux. Deux graphes distincts G_1 et G_2 sont dits $\mathbf{M}$–cospectraux si les matrices de représentation $\mathbf{M}(G_1)$ et $\mathbf{M}(G_2)$ des deux graphes possèdent les mêmes spectres. Lorsque la matrice $\mathbf{M}$ est la matrice d'adjacence $\mathbf{A}$, cela signifie que les deux graphes possèdent le même spectre d'adjacence. Les premiers travaux sur les graphes cospectraux datent des années 1970 [132, 133] et sont suivis par ceux de Godsil et McKay en 1982 qui se sont penchés sur une construction de graphes de ce type [134].

Puis sont arrivés les travaux plus récents de Haemers *et al.* qui forment un cadre théorique à l'étude de ces graphes [49, 135]. Des conjectures quant à ce type de graphes existent telles que celle de Schwenk statuant que presque tous les arbres sont cospectraux [136]. La figure 1.20 montre un exemple, introduit par Collatz et Sinogowitz en 1957 [88], de deux graphes non isomorphes $\mathbf{A}$–cospectraux : le graphe cycle auquel un sommet isolé est ajouté $\mathcal{C}_4 \cup \mathcal{K}_1$ et le graphe étoile $\mathcal{S}_4 = \mathcal{K}_{1,4}$. Ces deux graphes possèdent un spectre d'adjacence égal à $(-2, 0, 0, 0, 2)$. Cela illustre en outre une propriété sur le spectre d'adjacence des graphes bipartis : si G est un graphe biparti et que λ est une valeur propre de sa matrice d'adjacence $\mathbf{A}$, alors $-\lambda$ en est nécessairement une aussi, avec la même multiplicité que celle de λ [89, 91]. La figure 1.21 présente quant à elle un exemple de deux graphes $\mathbf{L}$–cospectraux.

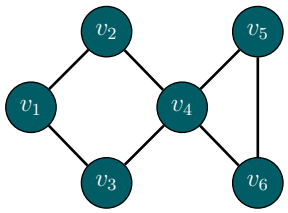


(a) Graphe $\mathcal{C}_4 \cup \mathcal{K}_1$ vérifiant
$\mathrm{spec}(\mathbf{A}) = (-2, 0, 0, 0, 2)$
$\mathrm{spec}(\mathbf{L}) = (0, 0, 2, 2, 4)$

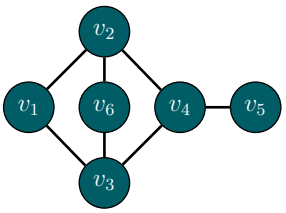


(b) Graphe $\mathcal{K}_{1,4}$ vérifiant
$\mathrm{spec}(\mathbf{A}) = (-2, 0, 0, 0, 2)$
$\mathrm{spec}(\mathbf{L}) = (0, 1, 1, 1, 5)$

Figure 1.20 – Deux graphes $\mathbf{A}$–cospectraux non $\mathbf{L}$–cospectraux. Les nuances de vert révèlent la bipartition.



(a) Graphe vérifiant
$\mathrm{spec}(\mathbf{A}) = (-2.1, -1, -0.5, 0, 1.2, 2.5)$
$\mathrm{spec}(\mathbf{L}) = (0, 0.7, 2, 3, 3, 5.2)$



(b) Graphe vérifiant
$\mathrm{spec}(\mathbf{A}) = (-2.5, -0.7, 0, 0, 0.7, 2.5)$
$\mathrm{spec}(\mathbf{L}) = (0, 0.7, 2, 3, 3, 5.2)$

Figure 1.21 – Deux graphes $\mathbf{L}$–cospectraux non $\mathbf{A}$–cospectraux.

Par ailleurs, un axe d'étude issu de celui de la recherche de graphes cospectraux est la recherche de graphes déterminés par leurs spectres, c'est-à-dire des graphes G tels que tout autre graphe admettant le même spectre que G lui est isomorphe. Néanmoins, c'est une tâche ardue à laquelle se prêtent bien des conjectures, notamment celle statuant que « presque tous » les graphes sont déterminés par leurs spectres [48, 137, 138], appuyée par des simulations numériques qui ont permis de calculer la proportion de graphes déterminés par leurs spectres pour un nombre de sommets inférieur ou égal à 12 [138]. Les résultats, issus de différents travaux [134, 135, 139] sont listés dans le tableau 1.1. À la vue de ces

résultats, la proportion du nombre de graphes déterminés par leurs spectres diminue jusqu'à $n = 11$ sommets puis augmente passé ce seuil, ce qui appuie la conjecture précédente.

Nombre de sommets n	Nombre de graphes N	Proportion
1	1	1
2	2	1
3	4	1
4	11	1
5	34	0.941
6	156	0.936
7	1 044	0.895
8	12 346	0.861
9	274 668	0.814
10	12 005 168	0.787
11	1 018 997 864	0.789
12	165 091 172 592	0.812

Tableau 1.1 – Proportion de graphes déterminés par leurs spectres (avec un ordre $n \in [\![1, 12]\!]$).

Dans ce chapitre, un nombre conséquent de matrices de représentation ont été rappelées. Une des raisons pour lesquelles certaines de ces matrices ont été introduites puis utilisée est due, entre autres, au fait qu'elles produisent justement moins de graphes cospectraux. C'est le cas, par exemple, de la matrice Laplacienne sans-signe $\mathbf{Q}$ [9]. Pour le constater, le tableau 1.2 s'appuie sur des travaux de la théorie spectrale de graphes pour lister le nombre de graphes $\mathbf{M}$–cospectraux ayant des ordres allant de 1 à 11 avec $\mathbf{M} \in \{\mathbf{A}, \mathbf{L}, \mathbf{Q}, \boldsymbol{\mathcal{L}}\}$ [135,140]. Il peut être constaté avec ce tableau, que c'est la matrice Laplacienne normalisée qui possède le moins de graphes cospectraux (pour des graphes possédant 9 sommets ou moins), suivie de la matrice Laplacienne sans-signe. Prenons par exemple l'intégralité des graphes ayant 9 sommets : il y en a 18.6% qui sont $\mathbf{A}$–cospectraux, 15.5% qui sont $\mathbf{L}$–cospectraux, 6.9% qui sont $\mathbf{Q}$–cospectraux et 0.4% qui sont $\boldsymbol{\mathcal{L}}$–cospectraux.

Le véritable problème de la cospectralité est que, dans le cas d'une classification de graphes basée sur le spectre de ces derniers (qualifiée alors de classification spectrale), il est impossible de construire une mesure de similarité basée uniquement sur les spectres d'une seule matrice de représentation puisque plusieurs graphes peuvent avoir le même. Il est alors nécessaire de construire une mesure qui combine plusieurs spectres, à l'image de la Similarité Spectrale Conjointe définie par Bay-Ahmed *et al.* [53]. Pour savoir quelle combinaison il serait judicieux de choisir, le tableau 1.3 donne le nombre de graphes cospectraux en testant différentes combinaisons de spectres [47]. Il apparait alors qu'une réduction considérable du nombre de graphes cospectraux est possible en considérant des paires de spectres plutôt qu'un seul spectre (tableau 1.2), en particulier pour des graphes ayant un nombre de sommets inférieur à 10. Par exemple, les deux matrices traditionnelles que sont les matrices d'adjacence $\mathbf{A}$ et Laplacienne $\mathbf{L}$ ne constituent pas nécessairement la meilleure paire à considérer lors d'une classification

		Matrices de représentation			
Nombre de sommets n	**Nombre de graphes N**	$\mathbf{A}$	$\mathbf{L}$	$\mathbf{Q}$	$\boldsymbol{\mathcal{L}}$
1	1	0 (0)	0 (0)	0 (0)	0 (0)
2	2	0 (0)	0 (0)	0 (0)	0 (0)
3	4	0 (0)	0 (0)	0 (0)	0 (0)
4	11	0 (0)	0 (0)	2 (0.182)	2 (0.182)
5	34	2 (0.059)	0 (0)	4 (0.118)	4 (0.118)
6	156	10 (0.064)	4 (0.026)	16 (0.103)	14 (0.090)
7	1 044	110 (0.105)	130 (0.125)	102 (0.098)	52 (0.050)
8	12 346	1 722 (0.139)	1 767 (0.143)	1 201 (0.097)	201 (0.016)
9	274 668	51 039 (0.186)	42 595 (0.155)	19 001 (0.069)	1 092 (0.004)
10	12 005 168	2 560 516 (0.213)	1 412 438 (0.118)	636 607 (0.053)	–
11	1 018 997 864	215 264 372 (0.211)	91 274 836 (0.090)	38 966 935 (0.038)	–

Tableau 1.2 – Nombre et proportion de graphes cospectraux selon différentes représentations (avec un ordre $n \in [\![1, 12]\!]$) [135, 140].

spectrale. Il faudra peut-être plus se reposer sur la paire $\mathbf{Q} \wedge \mathbf{A}$, $\boldsymbol{\mathcal{L}} \wedge \mathbf{A}$ ou encore $\boldsymbol{\mathcal{L}} \wedge \mathbf{L}$. La pire paire de matrices à considérer est $\mathbf{L} \wedge \mathbf{Q}$ pour lesquelles il peut être conjecturé qu'au vu de la similarité de leurs expressions, leurs spectres soient similaires dans bien des cas. Il est à noter également ce fait contre-intuitif : combiner toutes les matrices ne permet pas de réduire drastiquement le nombre de graphes cospectraux. En effet, il y en a presque autant qu'avec la paire $\boldsymbol{\mathcal{L}} \wedge \mathbf{L}$.

		Combinaisons de matrices de représentation					
Nombre de sommets n	**Nombre de graphes connectés N**	$\mathbf{A} \wedge \mathbf{L}$	$\mathbf{L} \wedge \mathbf{Q}$	$\mathbf{Q} \wedge \mathbf{A}$	$\boldsymbol{\mathcal{L}} \wedge \mathbf{A}$	$\boldsymbol{\mathcal{L}} \wedge \mathbf{L}$	$\mathbf{A} \wedge \mathbf{L} \wedge \mathbf{Q} \wedge \boldsymbol{\mathcal{L}}$
6	112	0	0	0	0	0	0
7	853	0	16	0	0	0	0
8	11 117	0	232	0	6	0	0
9	261 080	82	4139	8	14	6	2
10	11 716 571	13 864	107 835	10 716	10 281	10 256	10 124

Tableau 1.3 – Nombre de graphes ($n \in [\![6, 10]\!]$) cospectraux en combinant les spectres de différentes matrices [47].

La prise en compte de ce problème de cospectralité est très important lors de tâches telles que la classification spectrale de graphes. Or, dans ce travail de thèse, nous souhaitons également classer des séries temporelles, si possible vues comme des graphes. Ainsi, la section suivante rappelle les différentes méthodes existantes pour y parvenir.

1.4 Transformation d'un signal en un graphe de visibilité

Représenter dans un nouvel espace l'information contenue dans un objet permet de la mettre différemment en valeur car sa nouvelle représentation induit l'utilisation d'outils alternatifs. C'est le cas d'un réseau électrique vu comme un graphe, un graphe vu comme une matrice ou encore d'un signal représenté sous la forme d'un spectrogramme, d'un scalogramme ou d'un graphe. En effet, l'idée d'utiliser une transformation d'un signal en un autre objet permettant une extraction d'informations autre que celles traditionnellement analysées dans le domaine temporel n'est pas nouvelle. Ce dernier peut être vu comme un spectre en fréquences *via* une transformée de Fourier, il peut être représenté dans le domaine temps-fréquence grâce à une transformée de Fourier à court terme donnant naissance à un spectrogramme ou encore dans le domaine temps-échelle avec une transformée en ondelettes donnant lieu à un scalogramme. L'objectif de cette section est de présenter une méthode transformant une série temporelle en un réseau complexe : le graphe de visibilité. Parmi, toutes les méthodes qui existent et qui permettent une telle transformation, le graphe de visibilité est sûrement celle qui combine la simplicité mathématique et algorithmique tout en possédant des propriétés structurelles intéressantes ainsi qu'une interprétabilité géométrique accessible. Par ailleurs, ce n'est pas l'objet de cette section mais l'inverse est possible : voir un graphe comme un signal et l'étudier comme tel [141].

Les premiers à avoir transformé un signal en réseau complexe sont Zhang et Small en 2006 en utilisant ce qu'ils ont appelé les cycles [142]. Ils se sont, pour cela, intéressés à l'analyse de caractéristiques topologiques de signaux pseudo-périodiques. Leur méthode est divisée en plusieurs tâches : la série temporelle étudiée est segmentée en cycles et chaque cycle est considéré comme un sommet du réseau, puis les corrélations entre cycles sont calculées et une arête existe entre deux sommets si la corrélation entre les deux cycles correspondants dépasse un certain seuil. Une variante de cette méthode, appelée TSCN (pour *Time Series Complex Network*), calcule une distance euclidienne entre cycles qui doit avoir cette fois le même nombre d'échantillons et si cette distance est inférieure à un seuil, alors les sommets représentant les cycles sont reliés dans le graphe correspondant [143]. Une autre méthode, appelée méthode par récurrence est introduite par Marwan, Dönner *et al.* [144,145]. Dans le cadre de l'analyse de séries temporelles, la récurrence d'un état $\vec{x}_i$ au temps $t = i\Delta t$ (où $i \in \mathbb{N}$, Δt est le temps d'échantillonage et $\vec{x} \in \mathbb{R}^p$ un état dans l'espace des phases[22] de dimension p) réfère à n'importe quel état $\vec{x}_j$ du système à un temps $j\Delta t$ qui serait similaire[23] à l'état initial $\vec{x}_i$. La représentation de la récurrence (*Recurrence Plot*) stocke toutes les récurrences sous la forme d'une matrice $\mathbf{R}$ où $R_{ij} = 1$ si l'état $\vec{x}_j$ est similaire à l'état $\vec{x}_i$ dans l'espace des phases et $R_{ij} = 0$ sinon [144–146]. La matrice $\mathbf{R}$ peut alors être vue comme une matrice d'adjacence d'un graphe dit de récurrence de la série tempo-

22. Espace abstrait dont les coordonnées sont les variables dynamiques du système étudié.
23. Qui dit similarité, dit mesure de similarité. En général, on considère une simple distance euclidienne.

relle étudiée. Un algorithme permettant de transformer un signal en graphe qui a donné naissance à plusieurs travaux de recherche est le graphe de visibilité, initialement introduit par Lacasa *et al.* en 2008 [21]. Depuis, cet algorithme a connu une quantité importante de variantes, aussi bien dans sa construction que dans les éventuelles pondérations des arêtes. Le graphe de visibilité d'un signal a autant de sommets que ce dernier possède d'échantillons et les arêtes sont construites grâce à un critère géométrique de visibilité entre les échantillons. Le principal avantage que possède cette méthode réside dans le fait que le graphe obtenu hérite de nombreuses propriétés intéressantes du signal, et révèle des informations non négligeables sur la série temporelle elle-même. Le graphe de visibilité est devenu, par sa simplicité d'implémentation et toutes les informations qu'on peut en extraire, une technique très utilisée pour l'identification de propriétés dynamiques des séries temporelles, pour l'analyse de la dépendance à long terme ou encore de la fractalité. Parmi ces applications, il est normal de penser à la classification de signaux vus comme des graphes (et donc l'extraction d'attributs dans le domaine des graphes plutôt que dans les domaines classiques du temps-fréquence) ainsi qu'à la caractérisation de processus stochastiques (qui sera l'objet du chapitre suivant). Cet algorithme de visibilité possède plusieurs versions et variantes qui sont présentées dans les sous-sections suivantes.

1.4.1 Graphe de visibilité naturelle

Soit une série temporelle $\mathbf{x} = (x_i)_{1 \leq i \leq n}$ possédant n échantillons indexés sur une échelle de temps discrète $i \in [\![1, n]\!]$. Le graphe de visibilité naturelle associé à la série temporelle $\mathbf{x}$ est un graphe possédant n sommets où une arête existe entre les sommets i et j s'il y a visibilité entre les échantillons x_i et x_j, c'est-à-dire si et seulement si, pour tout $i \leq k \leq j$, l'échantillon x_k vérifie

$$x_k < x_j + (x_i - x_j)\frac{j-k}{j-i}. \tag{1.56}$$

Un exemple de construction d'un graphe de visibilité naturelle associé à une série temporelle est donné en figure 1.22. Sur cette figure, une série temporelle de 20 échantillons est considérée et la visibilité géométrique (équation (1.56)) entre ces échantillons est représentée par des lignes vertes. Enfin, le graphe de visibilité naturelle est construit en considérant un sommet par échantillon et les arêtes existent en fonction de cette visibilité. Il est aisé de vérifier que le graphe de visibilité associé à une série temporelle est toujours connecté (chaque sommet voit au moins un voisin, celui de gauche ou celui de droite, avec une particularité sur les sommets intermédiaire $2, 3, \ldots, n-1$ qui voient toujours leurs deux voisins), non-orienté (à moins de vouloir le spécifier, si l'échantillon i peut voir l'échantillon j, alors l'échantillon j peut également voir l'échantillon i) et invariant par transformation affine de la série temporelle (le critère de visibilité est invariant par changement d'échelle soit de l'axe horizontal, soit

de l'axe vertical) [21]. Dans le cas de séries temporelles particulières, comme les signaux périodiques, le graphe de visibilité associé est régulier. Comme détaillé dans le chapitre suivant traitant de la classification des signaux, une information importante à extraire d'un graphe de visibilité est sa distribution des degrés et, dans le cas d'une série aléatoire extraite d'une loi uniforme, cette distribution suit une loi exponentielle de paramètre $1/k_0$: $p_k \sim \mathrm{e}^{-k/k_0}$ [21].

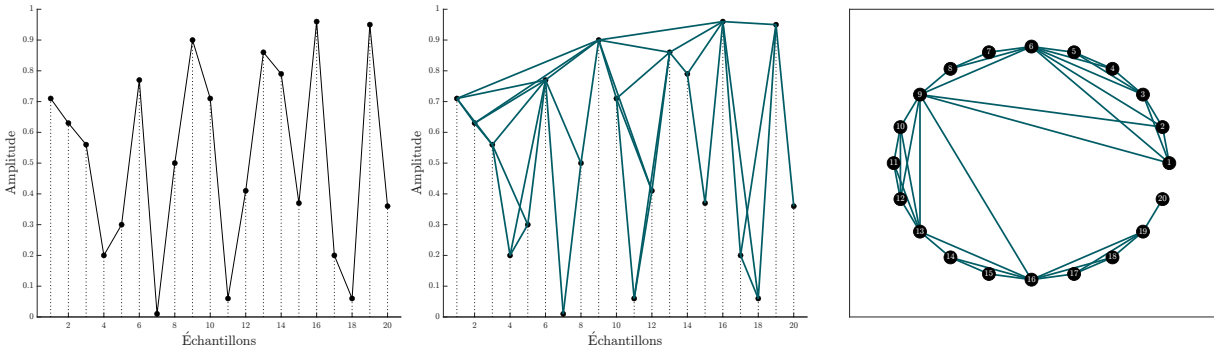


Figure 1.22 – De gauche à droite : Série temporelle ($n = 20$), algorithme de visibilité et graphe de visibilité naturelle associé ($m = 40$).

Il va de soi que, présentée comme telle, cette transformation n'est pas réversible : il est impossible de remonter au signal à partir de son graphe de visibilité. Une manière de permettre cette réversibilité est de pondérer les arêtes reliant les sommets i et j par des poids w_{ij}, ce qu'il est possible de définir assez facilement car la notion géométrique de visibilité permet d'introduire l'angle de visibilité, la portée de visibilité, etc. Ainsi, les différentes pondérations du tableau 1.4, toutes disponibles dans la librairie Python `ts2vg` de Carlos Bergillos [147], peuvent être définies. La raison pour laquelle il est question du package `ts2vg` plutôt qu'un autre est que ce dernier implémente une méthode de type *divide & conquer* pour construire le graphe de visibilité d'un signal, permettant alors de passer d'une complexité théorique en $O(n^2)$ à une complexité en $O(n \log n)$ [147,148].

1.4.2 Graphe de visibilité horizontale

En 2009, soit l'année suivant l'introduction du graphe de visibilité naturelle, une variante de cette méthode est introduite par Luque *et al.* [23] : le graphe de visibilité horizontale. Soit une série temporelle $\mathbf{x} = (x_i)_{1 \leq i \leq n}$ possédant n échantillons indexés sur une échelle de temps discrète $i \in [\![1, n]\!]$. Le graphe de visibilité horizontale (GVH) associé à la série temporelle $\mathbf{x}$ est un graphe possèdant n sommets où une arête existe entre les sommets i et j s'il y a visibilité horizontale entre les échantillons x_i et x_j, c'est-à-dire si et seulement si, pour tout $i \leq k \leq j$, l'échantillon x_k vérifie

$$x_k < \min(x_i, x_j) \tag{1.57}$$

Nom	Formule
Distance euclidienne	$w_{ij} = \sqrt{(j-i)^2 + (x_j - x_i)^2}$
Distance euclidienne au carré	$w_{ij} = (j-i)^2 + (x_j - x_i)^2$
Écart vertical	$w_{ij} = x_j - x_i$
Écart vertical absolu	$w_{ij} = \lvert x_j - x_i \rvert$
Écart horizontal	$w_{ij} = j - i$
Pente	$w_{ij} = \dfrac{x_j - x_i}{j - i}$
Pente absolue	$w_{ij} = \left\lvert \dfrac{x_j - x_i}{j - i} \right\rvert$
Angle de vue	$w_{ij} = \arctan\left(\dfrac{x_j - x_i}{j - i}\right)$
Angle de vue absolu	$w_{ij} = \arctan\left(\left\lvert \dfrac{x_j - x_i}{j - i} \right\rvert\right)$

Tableau 1.4 – Différentes pondérations possibles des arêtes d'un graphe de visibilité.

Un exemple de construction d'un graphe de visibilité horizontale associé à une série temporelle est donné en figure 1.23. Sur cette figure, la série temporelle de 20 échantillons de la figure 1.22 est considérée et la visibilité géométrique (équation (1.57)) entre ces échantillons est représentée par des lignes vertes. Enfin, le graphe de visibilité horizontale est construit en considérant un sommet par échantillon et les arêtes existent en fonction de cette visibilité. Bien entendu, par définition, le graphe de visibilité horizontale d'un signal est un sous-graphe de son graphe de visibilité naturelle. Par conséquent, toutes les propriétés structurelles du GVH sont héritées de celles du GVN, ainsi que la possibilité de pondération des arêtes du graphe de visibilité horizontale. Toutefois, son utilisation est bien souvent privilégiée vraisemblablement pour sa simplicité extrême qui, bien que contre-intuitif, soit une force dans sa capacité à classifier des graphes [149–151]. Par rapport au graphe de visibilité naturelle, d'autres propriétés ont été démontrées, notamment le fait que pour une série aléatoire extraite d'une loi uniforme, la distribution des degrés du graphe de visibilité horizontale associé suit une loi $p_k \sim \frac{1}{3}\left(\frac{2}{3}\right)^{k-2}$ et, par conséquent, son degré moyen est égale à $\bar{d} = \sum_{k=2}^{\infty} k p_k = \sum_{k=2}^{\infty} \frac{k}{3}\left(\frac{2}{3}\right)^{k-2} = 4$ [23]. Bien entendu, le degré minimal δ d'un graphe de visibilité horizontale est égal à 1 et le degré maximal Δ est inférieur ou égal à $n-1$. Si $\Delta = n-1$, alors il n'y a qu'un seul sommet qui possède ce degré [152].

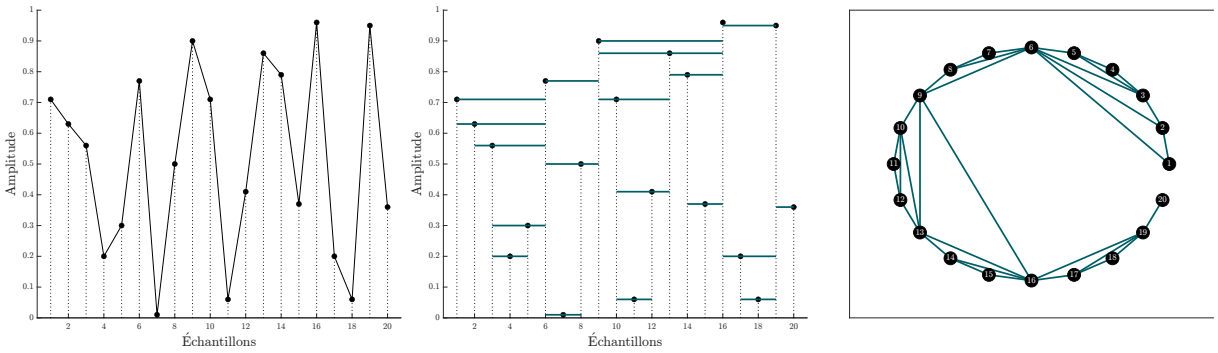


Figure 1.23 – De gauche à droite : Série temporelle ($n = 20$), algorithme de visibilité et graphe de visibilité horizontale associé ($m = 33$).

De plus, le nombre maximal d'arêtes est égal à $2n - 3$ [152]. Enfin, le degré moyen $\bar{d}$ d'un graphe de visibilité horizontale associée à une série infinie périodique de période T (sans valeur répétée dans une période) est égal à [153]

$$\bar{d} = 4 - \frac{2}{T}. \tag{1.58}$$

Une conséquence intéressante du résultat précédent est que pour toute série temporelle associée à son graphe de visibilité horizontale, alors ce dernier a un degré moyen $2 \leq \overline{k} \leq 4$ avec la borne inférieure atteinte pour une série constante et la borne supérieure atteinte pour une série apériodique (aléatoire, chaotique).

Le graphe de visibilité horizontale filtré (f–GVH) est quant à lui une méthode de transformation d'une série temporelle en un graphe où le critère géométrique de visibilité dépend à présent d'un seuil f. Soit une série temporelle $\mathbf{x} = (x_i)_{1 \leq i \leq n}$ possédant n échantillons indexés sur une échelle de temps discrète $i \in [\![1, n]\!]$. Le graphe de visibilité horizontale filtré associé à la série temporelle $\mathbf{x}$ est un graphe possèdant n sommets où une arête existe entre les sommets i et j si et seulement si, pour tout $i \leq k \leq j$, l'échantillon x_k vérifie

$$x_k + f < \min(x_i, x_j). \tag{1.59}$$

Cette variante a été particulièrement utilisée pour estimer la périodicité d'une série temporelle [153]. En effet, en extrayant le degré moyen $\bar{d}$ du graphe de visibilité horizontale filtré de paramètre f avec f augmentant peu à peu, un plateau d^* apparaît sur un intervalle $f \in [f_1, f_2]$ dans lequel le graphe de visibilité horizontale filtré du signal périodique bruité sera équivalent au graphe de visibilité horizontale du signal périodique non bruité (qui a un degré moyen connu grâce à l'équation (1.58)), et alors l'estimation de la période T de la série temporelle d'origine se fait en calculant

$$T = \frac{2}{4 - d^*}. \tag{1.60}$$

La figure 1.24, inspirée du travail de Nuñez *et al.* [153], permet de constater ce fait : à gauche un signal périodique (de période 2) bruité et à droite l'évolution du degré moyen $\bar{d}$ du graphe de visibilité horizontale filtré de paramètre f en fonction de ce paramètre. Le palier se trouve à $d^* = 3$ soit une période estimée, grâce à l'équation (1.60), à $T = 2$.

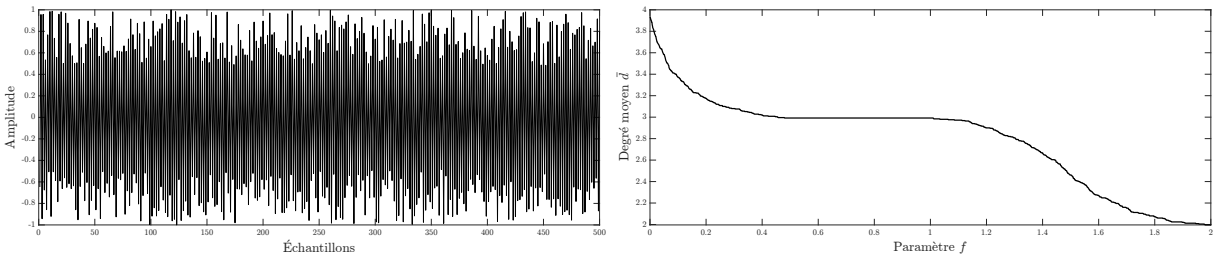


Figure 1.24 – À gauche : série temporelle périodique bruitée / À droite : évolution du degré moyen $\bar{d}$ du graphe de visibilité horizontale filtré en fonction du paramètre f.

D'autres variantes des graphes de visibilité existent : le graphe de visibilité différentielle (GVD) défini comme le graphe de visibilité naturelle auquel les arêtes du graphe de visibilité horizontale sont retirées [154], le graphe de visibilité signé défini comme l'union du graphe de visibilité associé à la série temporelle $\mathbf{x}$ et du graphe de visibilité associé à la série temporelle $-\mathbf{x}$ [155], ou encore le très récent graphe de visibilité image généralisant le graphe de visibilité naturelle aux images matricielles [156].

1.5 Conclusion

Dans ce chapitre, qui est une entame pour les suivants, nous avons établi le périmètre global de la thèse en rappelant les concepts fondamentaux de la théorie des graphes et des transformations possibles de séries temporelles en graphes. Ainsi, ces rappels s'articulent autour de trois axes. Dans un premier temps, des notions qui ont trait à la structure de graphes sont exposées : distribution des degrés, plus courte chaîne, force moyenne et divers attributs structurels. Ensuite, des généralités relatives aux matrices de représentation et à la théorie spectrale de graphes sont abordées. L'accent est mis en particulier sur les matrices d'adjacence, Laplacienne et Laplacienne sans-signe ainsi que sur leurs propriétés spectrales respectives. Ces dernières permettent également d'aborder le problème de cospectralité. Enfin, nous avons presenté les éléments nécessaires à la bonne compréhension de ce travail de thèse en introduisant l'algorithme de visibilité. Ce dernier permet de transformer des séries temporelles en réseaux complexes, ouvrant alors la voie à une analyse de signaux basée sur des outils de la théorie des graphes. Cette revue de la littérature et ce rappel du contexte constituent une base essentielle pour que le manuscrit soit aussi « auto-suffisant » que possible.

Chapitre 2

Classification et caractérisation de signaux par graphes de visibilité

« La Science remplace du visible compliqué par de l'invisible simple. »

Jean Perrin

2.1 Introduction

La théorie des graphes et le traitement du signal ont longtemps été des domaines scientifiques coexistant simultanément sans jamais interférer. À première vue, il n'y a pas de lien évident entre les deux. D'un côté la théorie des graphes, née au XVIII^{e} siècle, étudie, à l'aide d'outils mathématiques rappelés au chapitre précédent, des systèmes physiques modélisés sous la forme de graphes. De l'autre, le traitement de signal, pour lequel fixer le commencement serait sans doute omettre bon nombre de découvertes passées, a pour objectif notamment l'analyse et l'extraction d'informations de signaux. Ce n'est que récemment qu'un lien a été créé entre ces deux disciplines. Un axe de recherche appelé traitement du signal sur graphe a vu le jour ces dernières années avec la généralisation d'outils classiques du traitement de signal (translation, modulation, convolution, transformée de Fourier, etc.) adaptés aux signaux définis sur les sommets d'un graphe [10, 11]. Depuis le début des années 2000, il est possible de créer des graphes à partir de séries temporelles en appliquant des algorithmes présentés précédemment [142, 144, 145], notamment le graphe de visibilité naturelle et toutes ses variantes [21, 22] qui sont les outils essentiels au travail mené dans ce chapitre. À noter que l'inverse est également possible, à savoir construire une série temporelle à partir d'un graphe [141]. Mais quel est l'apport d'une

transformation comme le graphe de visibilité ? Est-ce bénéfique de considérer cette représentation sous forme de graphes plutôt que les signaux d'origine pour en effectuer une classification ? Est-ce que cette représentation permet de mettre en valeur l'information contenue dans un signal à l'image d'une représentation temps-fréquence (spectrogramme) ou temps-échelle (scalogramme) ? Nous répondons à ces questions tout au long de ce chapitre. Pour le processus de classification, nous utilisons les séparateurs à vaste marge (SVM), principalement pour leur simplicité d'implémentation, l'explicabilité mathématique de son problème d'optimisation sous-jacent, son astuce du noyau (*kernel trick* en anglais) [27] et ses performances défiant souvent les méthodes récentes d'apprentissage profond. Après avoir rappelé son fonctionnement, nous testons deux méthodes de classification de séries temporelles vues comme des graphes de visibilité : une première basée sur une extraction d'attributs issus de la théorie des graphes qui sont mis en entrée d'un SVM et une deuxième sur la construction d'un noyau de SVM à l'aide de distances statistiques [157] (distance en variation totale, distance de Hellinger, distance de Jensen-Shannon, etc.) permettant de comparer des distributions d'attributs des graphes de visibilité considérés, telles que celle des degrés. Deux cas d'usage sont alors étudiés : la détection d'épilepsie dans des signaux EEG et la détection d'anomalies magnétiques. Les résultats obtenus mettent en avant la performance de ces deux méthodes, comparativement à d'autres stratégies issues de la littérature qui abordent le problème de classification de séries temporelles vues comme des graphes de visibilité [37–39,158–160]. L'intérêt initial du graphe de visibilité mis en avant dans l'article original [21] est d'être capable, grâce à sa distribution de degrés, de caractériser un processus stochastique tel qu'un mouvement Brownien fractionnaire (ou de manière équivalente, un bruit Gaussien fractionnaire), souvent paramétré par un coefficient appelé exposant de Hurst H [22,161,162]. Nous montrons, dans ce chapitre, qu'une estimation de ce paramètre est possible grâce à la théorie des graphes. Pour cela, nous utilisons un outil récent appelé plan informationnel de Fisher-Shannon [42] qui, comme son nom l'indique, combine deux quantifieurs de la théorie de l'information, à savoir l'entropie de Shannon et l'information de Fisher. À notre connaissance, l'utilisation conjointe de la théorie des graphes et des outils de la théorie de l'information est rare, ce qui renforce l'originalité de ces travaux. Notre méthode d'estimation est appliquée à des signaux réels et les résultats obtenus sont cohérents avec des estimateurs plus traditionnels. De plus, une étude de l'erreur d'estimation sur des signaux synthétiques est proposée. Ainsi, ce chapitre se veut être une exploration des possibilités qu'offrent les graphes de visibilité pour représenter, classifier et caractériser des séries temporelles.

2.2 Rappel sur les séparateurs à vaste marge (SVM)

Aujourd'hui présent dans beaucoup de domaines comme la reconnaissance d'images, la traduction automatique, la recommandation de produits, le diagnostic médical et bien plus encore, l'apprentissage automatique (*machine learning* en anglais) vise à permettre aux ordinateurs d'apprendre à partir des données, c'est-à-dire d'utiliser des algorithmes pour analyser de grandes quantités de données, en extraire des règles permettant de faire des prédictions ou de prendre des décisions basées sur ces règles. Les données collectées peuvent être d'origines très variées : des images, du texte, des vidéos, des signaux ou encore des graphes. Supposons que ces données aient été préparées et nettoyées. L'objectif suivant est de choisir un modèle qui sera par la suite entraîné c'est-à-dire que ses paramètres seront ajustés pour réduire les erreurs entre ce qu'il prédit et la vérité terrain. Ce choix de modèle dépend nécessairement du type de données et de l'objectif de la tâche. Par exemple, il existe l'apprentissage non-supervisé pour lequel le modèle explore des données non étiquetées pour en extraire des groupes basés sur une mesure de similarité (par exemple, l'algorithme des K-moyennes). Il existe également l'apprentissage supervisé pour lequel le modèle apprend à partir de données étiquetées : chaque donnée est associée à une valeur scalaire (régression) ou à une catégorie connue (classification) et fait des prédictions basées sur ces étiquettes. Des exemples de modèles appartenant à cette catégorie : les réseaux de neurones, les arbres de décision ou encore les séparateurs à vaste marge (SVM) qui font l'objet de cette section.

2.2.1 Principe

Dans les années 1990, sont apparues les Machines à Vecteurs de Support[1] (SVM pour *Support Vector Machine*), algorithmes d'apprentissage supervisé destinés principalement à la classification binaire [163]. C'est un outil extrêmement robuste, facile à implémenter et dont la capacité de généralisation et les performances sont équivalentes, dans beaucoup de cas, à celles de réseaux de neurones simples comme les perceptrons multicouches qui ont pris une place importante dans le paysage de l'apprentissage automatique. Soit un jeu de données $\mathbf{X} \in \mathbb{R}^{N \times p}$ contenant N observations $(\mathbf{x}_i)_{1 \leq i \leq N}$, chacune de dimension p (c'est-à-dire p attributs), d'étiquettes (*labels* en anglais) $\mathbf{y} = (y_i)_{1 \leq i \leq N}$ avec $y_i \in \{-1, 1\}$ pour tout i : en effet, dans le cas d'une classification binaire, seules les classe -1 et 1 sont considérées. L'objectif principal du SVM est de déterminer un hyperplan séparateur optimal

$$\mathcal{H} : \langle \mathbf{w}, \mathbf{x} \rangle + b = 0 \tag{2.1}$$

1. En français, il est commun de les nommer Séparateurs à Vaste Marge pour garder le sens de l'acronyme

où $\langle \cdot, \cdot \rangle$ désigne le produit scalaire[2], $\mathbf{w} \in \mathbb{R}^p$ est le vecteur normal au plan $\mathcal{H}$ et $b \in \mathbb{R}$ est appelé biais du plan. Cet hyperplan séparateur optimal doit, comme son nom l'indique, séparer linéairement le jeu de données en deux catégories selon leurs étiquettes. Il doit être optimal car, dans le cas où les données sont linéairement séparables, il existe une infinité d'hyperplans séparateurs. Soient les deux hyperplans $\mathcal{H}_- : \langle \mathbf{w}, \mathbf{x} \rangle + b = -1$ et $\mathcal{H}_+ : \langle \mathbf{w}, \mathbf{x} \rangle + b = 1$ définis de part et d'autre de l'hyperplan $\mathcal{H}$ comme illustré par la figure 2.1.

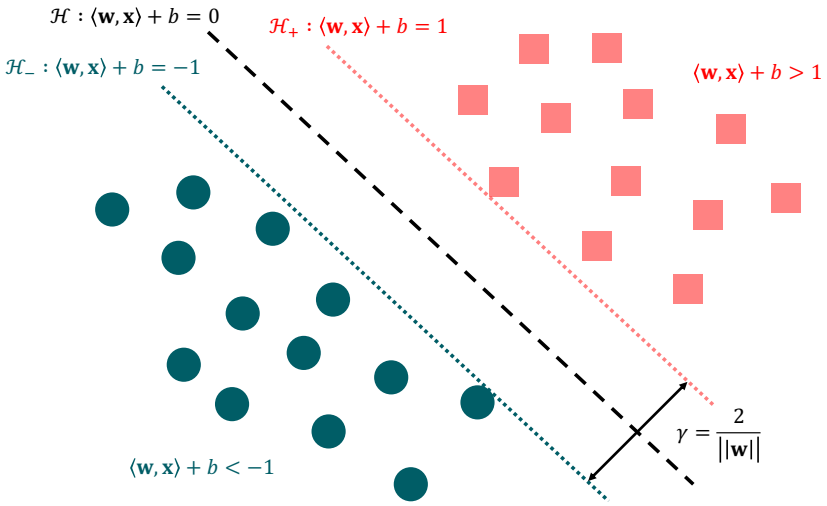


Figure 2.1 – Principe du SVM pour une classification binaire (« cercle vert » ou « carré rouge ») : trouver un hyperplan $\mathcal{H}$ séparateur optimal, c'est-à-dire qui maximise la marge γ.

Plaçons-nous toujours dans le cas où les données sont linéairement séparables : les points « positifs » doivent tous être à l'extérieur de $\mathcal{H}_+$, c'est-à-dire $\langle \mathbf{w}, \mathbf{x}_i \rangle + b \geq 1$ si $y_i = 1$ et les points « négatifs » doivent tous être à l'extérieur de $\mathcal{H}_-$, c'est-à-dire $\langle \mathbf{w}, \mathbf{x}_i \rangle + b \leq -1$ si $y_i = -1$. Il est alors possible de combiner ces deux contraintes en une seule : $y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b \right) \geq 1$. En ce qui concerne la grandeur à maximiser, il s'agit de la marge, c'est-à-dire de la distance entre les deux hyperplans $\mathcal{H}_-$ et $\mathcal{H}_+$ égale à $2/\|\mathbf{w}\|$. Ainsi, le problème d'optimisation sous contraintes résolu par un SVM est

$$\underset{\mathbf{w} \in \mathbb{R}^p, b \in \mathbf{R}}{\operatorname{argmax}} \frac{2}{\|\mathbf{w}\|}, \quad \text{avec } y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b \right) \geq 1, \forall i \in [\![1, N]\!]. \tag{2.2}$$

Un problème de minimisation, appelé problème primal du SVM, strictement équivalent au précédent

2. On rappelle que le produit scalaire entre deux vecteurs colonnes $\mathbf{a} = (a_i)_{1 \leq i \leq p}$ et $\mathbf{b} = (b_i)_{1 \leq i \leq p}$ est défini par

$$\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^\top \mathbf{b} = \sum_{i=1}^{p} a_i b_i.$$

est le suivant :

$$\underset{\mathbf{w}\in\mathbb{R}^p, b\in\mathbf{R}}{\operatorname{argmin}} \frac{\|\mathbf{w}\|^2}{2}, \quad \text{avec } y_i\left(\langle \mathbf{w}, \mathbf{x}_i\rangle + b\right) \geq 1, \forall i \in [\![1, N]\!]. \tag{2.3}$$

En calculant le Lagrangien de ce problème, le problème dual peut être reformulé en

$$\underset{\boldsymbol{\alpha}\in\mathbb{R}^N}{\operatorname{argmax}} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j\rangle, \quad \text{avec } \sum_{i=1}^{N} \alpha_i y_i = 0 \tag{2.4}$$

qui admet pour avantage principal de voir l'équation de l'hyperplan séparateur optimal s'écrire comme

$$\mathcal{H}^* : \sum_{i=1}^{N} \alpha_i^* y_i \langle \mathbf{x}_i, \mathbf{x}\rangle + b \tag{2.5}$$

avec $\boldsymbol{\alpha}^* = (\alpha_i^*)_{1\leq i\leq N}$ solution de ce problème. Ainsi, l'hyperplan peut être vu comme une fonction de décision permettant de classer une nouvelle observation $\mathbf{x}$, ce dernier s'écrivant comme produits scalaires entre $\mathbf{x}$ et les données initiales.

Bien souvent, les données ne sont pas linéairement séparables. Dans le cas où elles le sont « presque », c'est-à-dire dans le cas où il existe quelques points empêchant la séparabilité linéaire, il est possible de spécifier un hyperparamètre[3] de régularisation C qui permet d'autoriser des empiètements de marge tout en s'assurant que cette dernière soit la plus large possible. Dans le cas où les données ne sont pas du tout linéairement séparable, comme celles en dimension 1 à gauche de la figure 2.2, une solution peut être d'introduire des variables polynomiales (ici, $x_2 = x_1^2$) pour retrouver un cas linéairement séparable. Il est également possible d'introduire des variables de similarité. Pour ces dernières, une fonction de similarité qui calcule la ressemblance entre chaque observation et des points de repère est calculée (sur la figure 2.3, ce sont les points -2 et 1 qui servent de repère). La plus classique est une fonction de base radiale (RBF pour *Radial Basis Function*) : $\phi_\gamma(\mathbf{x}, \ell) = \exp\left(-\gamma\|\mathbf{x} - \ell\|^2\right)$. Les fonctions de base radiale valent 1 au point de repère et tendent vers 0 à mesure que les points s'en éloignent. Ainsi, sur la figure 2.3, deux variables de similarité sont créées (x_2 et x_3). Il devient alors possible de trouver un hyperplan permettant de séparer linéairement les données (courbe bleue en tirets sur la figure 2.3). Dans les faits, la question du choix des points de repère ne se pose pas car toutes les observations servent de points de repère.

3. Un hyperparamètre d'un modèle d'apprentissage automatique est un paramètre fixé par l'utilisateur à l'avance et qui n'est pas appris par le modèle lui-même. Contrairement aux paramètres du modèle qui sont optimisés automatiquement (comme $\mathbf{w}$ et b dans notre cas), les hyperparamètres contrôlent la manière dont le modèle est entraîné et influencent directement ses performances.

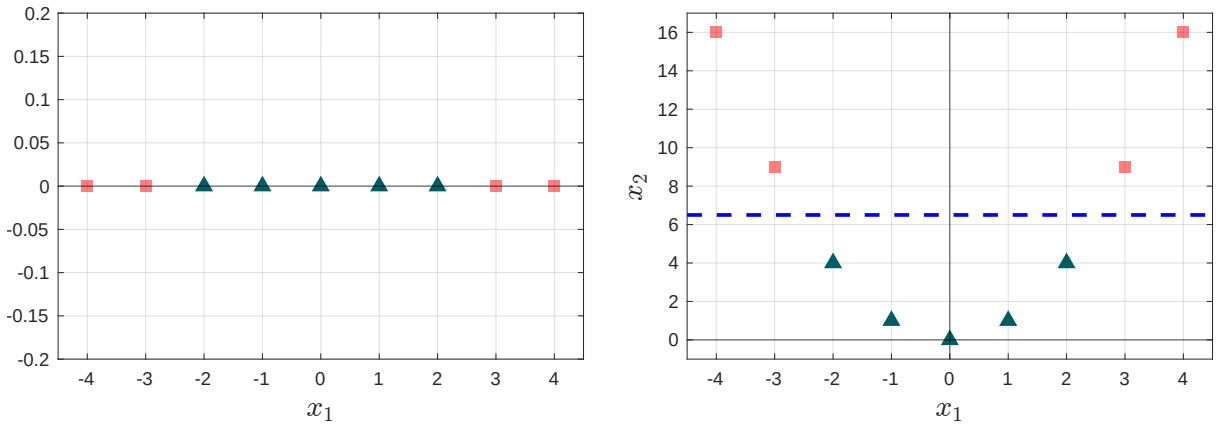


Figure 2.2 – Ajout d'une variable polynomiale $x_2 = x_1^2$ afin de rendre les données linéairement séparables, alors qu'elles ne l'étaient pas avec la variable x_1 seule [164].

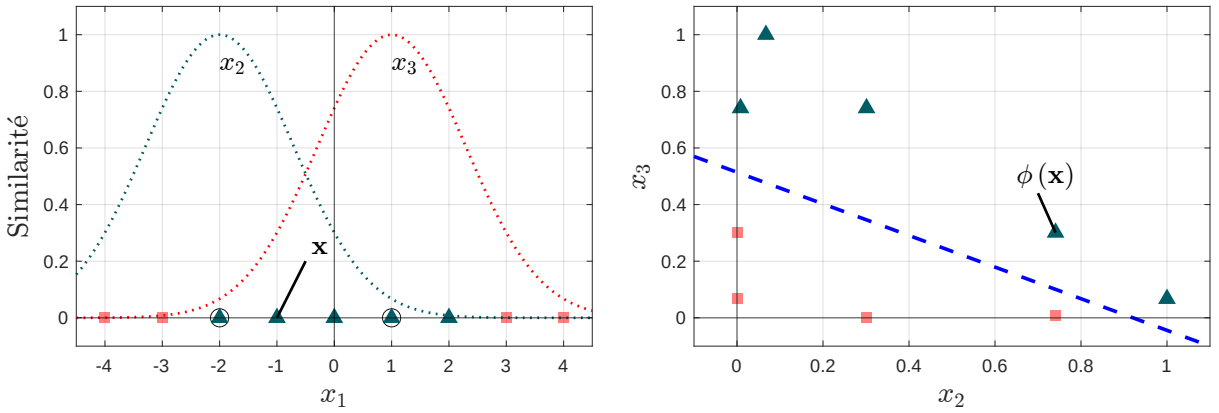


Figure 2.3 – Création de deux variables de similarité, ici $x_2 = \exp(-0.3\|x_1 + 2\|^2)$ et $x_3 = \exp(-0.3\|x_1 - 1\|^2)$, afin de rendre les données linéairement séparables [164]. Si ϕ désigne l'application permettant de passer de l'espace de gauche à l'espace de droite, un exemple de projection $\phi(\mathbf{x})$ d'un point $\mathbf{x}$ est proposé.

Pour des jeux de données conséquents, cette opération coûte en temps de calcul. L'astuce du noyau (*kernel trick* en anglais) est alors envisagée [27]. Cette astuce permet d'obtenir le même résultat qu'en considérant de nouvelles variables mais sans réellement les créer. L'objectif en adjoignant des variables est de passer d'un espace de dimension p à un espace $\mathcal{F}$ de dimension supérieure dans lequel l'hyperplan séparateur optimal sera recherché. Il est possible d'exprimer cette transformation à l'aide d'une fonction $\Phi : \mathbb{R}^p \longrightarrow \mathcal{F}$ qui, à $\mathbf{x}$, associe $\Phi(\mathbf{x})$. La solution du problème d'optimisation précédent est alors

$$\mathcal{H}^* : \sum_{i=1}^{N} \alpha_i^* y_i \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}) \rangle + b \tag{2.6}$$

qui ne dépend ainsi que des produits scalaires $\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}) \rangle$. Ce n'est donc pas le choix de la fonction Φ qui est primordial mais bien celui d'une fonction $K : \mathbb{R}^p \times \mathbb{R}^p \longrightarrow \mathbb{R}$, appelée fonction noyau,

qui, à un couple $(\mathbf{x}, \mathbf{y})$ associe

$$K(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle. \tag{2.7}$$

Il existe des fonctions noyaux traditionnelles, implémentées dans toutes les librairies d'apprentissage automatique :

- Noyau linéaire : $K(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$;
- Noyau polynomial : $K(\mathbf{x}, \mathbf{y}) = (\gamma \langle \mathbf{x}, \mathbf{y} + r \rangle)^{\alpha}$ où $\alpha, \gamma, r \in \mathbb{R}$;
- Noyau gaussien (ou RBF) : $K(\mathbf{x}, \mathbf{y}) = \exp\left(-\gamma \|\mathbf{x} - \mathbf{y}\|^2\right)$ avec le coefficient d'échelle $\gamma \in \mathbb{R}$ (constante inversement proportionnelle à l'écart-type de la gaussienne).

Ces fonctions noyau matérialisent également une notion de distance entre les points $\mathbf{x}$ et $\mathbf{y}$ c'est pourquoi nous allons, dans la suite de ce manuscrit, utiliser plusieurs distances, notamment en remplacement de la distance euclidienne dans le noyau RBF. En effet, nos points $\mathbf{x}$ et $\mathbf{y}$ sont, dans ce chapitre, des distributions de probabilité et il est admis que la distance euclidienne n'est pas le meilleur des outils pour pouvoir comparer ces dernières. Pour utiliser une fonction noyau K, celle-ci doit vérifier le théorème de Mercer [165] dont une condition à respecter est que la matrice de Gram $\mathbf{K} = [K(\mathbf{x}_i, \mathbf{x}_j)]_{i,j}$ associée à la fonction noyau K soit semi-définie positive. Bien entendu, cela est difficile à montrer de manière théorique sur des données expérimentales : une façon intuitive de le faire est de tester si cette condition est vérifiée sur des sous-ensembles raisonnablement grands de la base de données initiale. Si l'approche empirique révèle l'existence d'une seule valeur propre négative, alors la fonction noyau considérée n'est pas valide.

2.2.2 Métriques de performance

Mesurer la performance d'un algorithme dédié à une tâche de régression (c'est-à-dire que ce qui est prédit est scalaire et continu) est généralement simple. Imaginons l'entraînement d'un algorithme dédié à la prédiction de la température du lendemain à partir de différentes variables météorologiques journalières issues d'un jeu de données conséquent. La performance pourra être quantifiée par une simple erreur moyenne entre les températures prédites et celles mesurées. En revanche, pour un algorithme de classification, ce n'est pas si facile. En effet, cela dépend grandement du contexte et de la tâche à effectuer. En considérant un algorithme qui doit détecter une maladie à partir d'une échographie, un détecteur de courriers indésirables ou encore un détecteur d'anomalies magnétiques (MAD pour *Magnetic Anomaly Detector*), les objectifs de performances sont bien différents car les contextes le sont également. Prenons l'exemple du MAD qui a pour objectif d'effectuer une classification binaire : présence (classe 1) ou non (classe -1) d'objets ferromagnétiques dans une série temporelle, sous-entendu

un sous-marin ou une mine dans un cadre militaire. Dans le cas d'une classification binaire, une matrice de confusion, dont une représentation est donnée en tableau 2.1, est souvent utilisée. Cette matrice mesure la qualité d'un algorithme de classification et donne une image complète de la façon dont le modèle fonctionne. Elle contient le nombre de vrais positifs VP (nombre de fois où la classe prédite est 1 et que son étiquette actuelle est 1 aussi), le nombre de vrais négatifs VN (nombre de fois où la classe prédite est -1 et que son étiquette actuelle est -1 aussi), le nombre de faux positifs FP (nombre de fois où la classe prédite est 1 tandis que son étiquette actuelle est -1) et le nombre de faux négatifs FN (nombre de fois où la classe prédite est -1 tandis que son étiquette actuelle est 1). Dans le cas précis du MAD, il est clair que c'est le taux de faux négatifs qu'il faudra chercher à minimiser.

		Classe actuelle	
		Négatif : -1	Positif : 1
Classe prédite	Négatif : -1	**VN (Vrais négatifs)**	**FN (Faux négatifs)**
	Positif : 1	**FP (Faux positifs)**	**VP (Vrais positifs)**

Tableau 2.1 – Matrice de confusion construite pour évaluer un classifieur binaire (classe 1 ou -1).

En l'état, cette matrice ne sert qu'à l'affichage de ces grandeurs. Pour que les résultats soient plus interprétables, les métriques suivantes sont fréquemment calculées : l'exactitude (*accuracy*) égale à $(\mathrm{VP}+\mathrm{VN})/(\mathrm{VP}+\mathrm{VN}+\mathrm{FP}+\mathrm{FN})$, la sensibilité égale à $\mathrm{VP}/(\mathrm{VP}+\mathrm{FN})$ ainsi que la spécificité égale à $\mathrm{VN}/(\mathrm{VN}+\mathrm{FP})$. Il existe également la courbe nommée ROC (pour *Receiver Operating Characteristic*), tracé du taux de vrais positifs (qu'il faut maximiser) en fonction du taux de faux positifs (qu'il faut minimiser). Ainsi, une bonne métrique de performance est l'aire sous cette courbe (appelée AUC pour *Area Under the Curve*), qu'il convient de maximiser. Une illustration de cette métrique est proposée en figure 2.4 dans laquelle sont représentées les courbes ROC de deux classifieurs. Le classifieur 1 est meilleur que le classifieur 2 au sens de l'AUC car cette dernière a une valeur plus proche de 1.

2.2.3 Validation croisée

Dans le cadre d'une tâche d'apprentissage automatique supervisée, c'est-à-dire quand les données mises en entrée sont étiquetées, il est commun de séparer la base de données en deux sous-ensembles : un consacré à l'entraînement et un associé au test du modèle entraîné qui doit être complètement indépendant du premier tout en étant représentatif. Les données d'entraînement doivent être indépendantes mais représentatives des données de test. Reprenons l'exemple de la prédiction de température

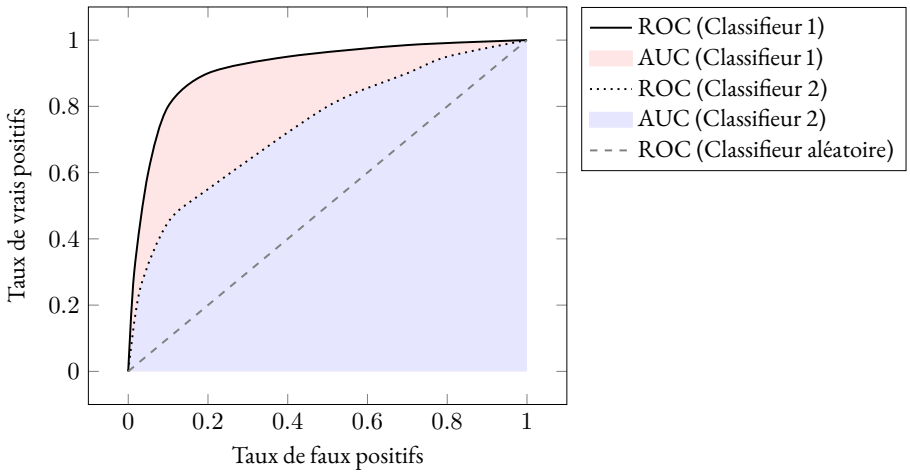


Figure 2.4 – Courbes ROC de deux classifieurs et leurs AUC respectives.

pour illustrer ce point : si, dans le jeu de données d'entraînement, il n'y a que des observations relevées au printemps, à l'été et à l'automne, il ne faut pas espérer que l'algorithme soit performant sur un jeu de test composé uniquement de données relevées en hiver. Dans ce cas, les données d'entraînement ne sont pas représentatives des données de test. À présent, pour entraîner un modèle, les hyperparamètres de l'algorithme doivent être optimisés de sorte à obtenir le meilleur résultat sur le jeu de données d'entraînement (c'est intrinsèque à son entraînement) mais aussi faire en sorte que l'algorithme entraîné ait une bonne capacité de généralisation, c'est-à-dire qu'il soit performant également sur le jeu de données de test. Le sur-apprentissage est associé à la situation où un modèle prédictif serait très performant sur les données d'entraînement mais mauvais sur les données de test, le sous-apprentissage correspondant alors à la situation où un modèle serait mauvais sur les deux jeux de données. Si une succession manuelle d'entraînement-test est effectuée pour avoir un modèle performant, il est aussi naturel de penser que les données de test ne sont plus si indépendantes des données d'entraînement car elles auront été prises en compte lors de l'optimisation de l'algorithme. Pour éviter ce phénomène, une validation croisée est souvent employée. Il s'agit alors de quantifier les performances d'un modèle en moyennant des métriques calculées sur des jeux de données variés. Pour la validation croisée à κ couches, $(\kappa-1)/\kappa$ de toutes les données servent à l'entraînement et le $1/\kappa$ restant au test. Cette manipulation est effectuée κ fois de telle sorte que toutes les observations opèrent $\kappa-1$ fois à l'entraînement et 1 fois au test. Pour avoir une idée générale de la qualité de l'algorithme sur le jeu de données complet, il faut alors moyenner les performances au cours des κ itérations, sous-entendu moyenner les exactitudes, les sensibilités, etc. La figure 2.5 illustre le principe d'une validation croisée à 6 couches. La base de données est séparée en $\kappa = 6$ sous-ensembles. Puis à chaque itération, 5 sous-ensembles servent à l'entraînement et 1 sous-ensemble sert au test.

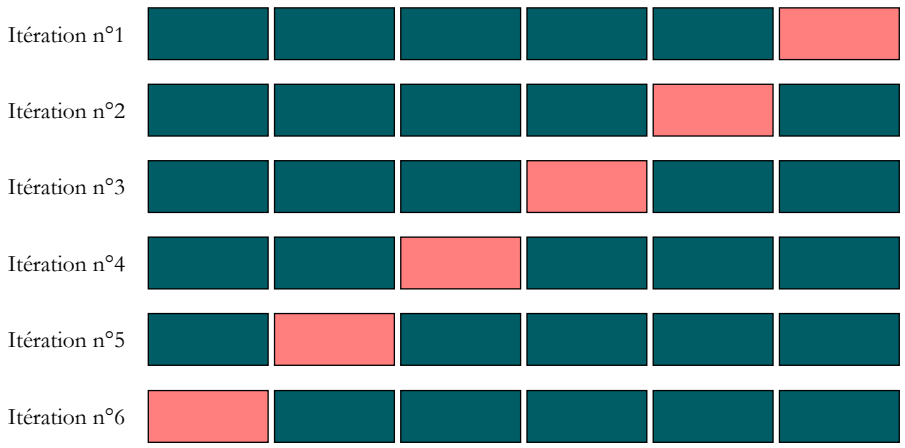


Figure 2.5 – Principe de fonctionnement de la validation croisée (ici à $\kappa = 6$ couches).

2.3 Classification de séries temporelles par graphe de visibilité

2.3.1 Motivations et méthodes existantes

L'idée d'analyser et de classer des séries temporelles est ancienne. En effet, des signaux astronomiques ont été utilisés dès l'époque des babyloniens pour prédire la position des astres, jusqu'à ce que les lois de Johannes Kepler aient été établies [166]. Les signaux astronomiques sont encore largement utilisés aujourd'hui, avec notamment l'observation récente des ondes gravitationnelles [167]. Bien entendu, l'analyse de séries temporelles est présente dans bien d'autres domaines comme l'économie, la météorologie, la biologie, la chimie, etc [168]. Les outils tels que la statistique et l'informatique ont permis le développement de modèles mathématiques permettant l'analyse de ces séries chronologiques, la particularité de ces données, par rapport à d'autres sources, étant que ce sont des données séquentielles ordonnées suivant le temps. Par ailleurs, des méthodes comme l'analyse de Fourier et les modèles statistiques (par exemple, ARIMA pour *AutoRegressive Integrated Moving Average*) ont été introduites pour décomposer et prédire le comportement futur de ces séries temporelles. Enfin, plus récemment, les méthodes issues de l'apprentissage automatique ont révolutionné le domaine, avec des techniques telles que les réseaux neuronaux récurrents (RNN) et les *transformers*, capables de capturer de manière automatique les dépendances complexes existantes entre les données séquentielles [168, 169]. Ces méthodes ont, en réalité, bénéficié de l'augmentation des moyens de calculs et de la disponibilité grandissante de quantité considérable de données dans tous les domaines scientifiques précédemment mentionnées.

Pour faire écho aux rappels de la section précédente, la classification de N séries temporelles $\mathbf{x}_i$ peut traditionnellement se faire de deux manières. Dans un premier temps, il est possible d'extraire

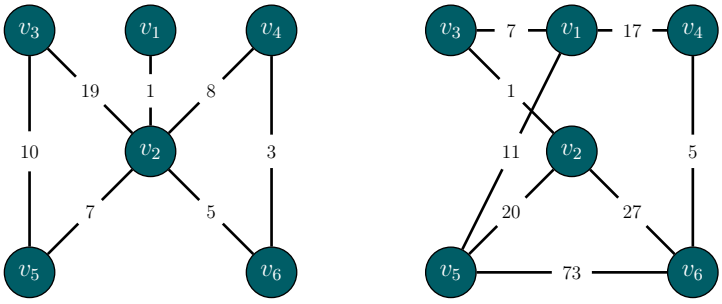
des attributs, c'est-à-dire que pour les N signaux, p attributs sont calculés (moyenne, variance, puissance moyenne, etc.) : nous avons listé bon nombre de ces attributs en annexe A afin de proposer un catalogue fourni (non exhaustif, bien entendu). Ce sont alors ces vecteurs d'attributs qui sont mis en entrée d'un algorithme comme le SVM présenté précédemment. Pour compléter, il est possible de faire appel à un algorithme de réduction de dimension comme l'ACP (Analyse en Composantes Principales) lorsque le nombre d'attributs est trop important et que certains d'entre eux sont véritablement corrélés. Cette technique a un aspect très intéressant : les signaux de la base de données n'ont pas l'obligation de disposer d'un même nombre d'échantillons puisqu'ils sont tous soumis à l'extraction des p attributs. Cependant, cela signifie que leur nouvelle représentation par des attributs se doit d'être extrêmement fidèle. Dans le cas où les signaux de la base de données possèdent tous le même nombre d'échantillons, il est possible de considérer chaque échantillon comme un attribut et ce sont alors directement les signaux qui sont mis en entrée de l'algorithme choisi. Il existe bien entendu des algorithmes permettant de considérer directement les séries temporelles bien qu'elles ne soient pas de même taille comme le DTW (pour *Dynamic Time Warping*), les LSTM (pour *Long-Short Term Memory*), etc. Toutefois, ces méthodes ne seront pas l'objet de cette section.

Dans le chapitre 1, des méthodes de transformation de signaux en graphes ont été présentés, notamment le graphe de visibilité. Il est alors légitime de se poser la question de la classification de séries temporelles vues comme des graphes. Or, la classification de graphes est également un sujet largement étudié depuis les années 1980. Au même titre que pour les séries temporelles, une manière de classifier des graphes est d'en extraire des attributs mis en entrée d'un algorithme de classification. Ces attributs peuvent être toutes les grandeurs rencontrées dans le chapitre 1 et bien d'autres encore. Ainsi, un exemple d'attributs calculés pour deux graphes G_1 ($n = 6$ sommets / $m = 7$ arêtes) et G_2 ($n = 6$ sommets / $m = 8$ arêtes), dont les structures et pondérations sont bien distinctes, est proposé en figure 2.6.

Par ailleurs, lorsqu'il est question de classification de graphes, il est commun de parler de noyaux sur graphes. Ces derniers sont en réalité des mesures de similarité entre graphes et peuvent être stockées sous la forme d'une matrice de Gram $\mathbf{K}$, définie par ses coefficients

$$[\mathbf{K}]_{ij} = K(G_i, G_j) \tag{2.8}$$

où G_i et G_j désignent deux graphes distincts de la base de données. L'intérêt est de pouvoir intégrer, dans un noyau SVM, des mesures de similarité adaptées aux graphes. Il existe un certain nombre de ces noyaux dans la littérature. Ainsi, le noyau *Random Walk* de Gärtner *et al.* [170] a pour objectif de



Attributs	Références	Valeurs pour G_1	Valeurs pour G_2
Nombre de sommets n	–	6	6
Nombre d'arêtes m	–	7	8
Degré minimal δ	page 28	1	2
Degré moyen $\bar{d}$	équation (1.2)	2.333	2.667
Degré maximal Δ	page 28	5	3
Force moyenne $\bar{s}$	équation (1.17)	17.667	53.667
Déviation des degrés $\mathrm{dev}(G)$	équation (1.4)	5.333	2.667
Variance des degrés $\mathrm{var}(G)$	équation (1.5)	1.556	0.22
Indice de Zagreb $Z_1(G)$	équation (1.8)	42	44
Indice de Zagreb $Z_2(G)$	équation (1.9)	53	60
Indice hyper-Zagreb $Z_h(G)$	équation (1.10)	264	244
Indice de Randić $R(G)$	équation (1.11)	2.712	2.966
Indice harmonique $H(G)$	équation (1.12)	1.238	1.467
Inverse Sum Indeg index $\mathrm{ISI}(G)$	équation (1.13)	8.548	10.8
Rayon spectral λ_n	page 39	2.709	2.741
Écart spectral $\lambda_n - \lambda_{n-1}$	page 41	1.709	2.03
Énergie $E_{\mathbf{A}}(G)$	équation (1.26)	7.806	8.139
Indice d'Estrada $\mathrm{EE}(G)$	équation (1.28)	19.836	20.494
Connectivité naturelle $\bar{\lambda}(G)$	équation (1.29)	1.196	1.228
Valeur de Fiedler μ_2	page 45	1	1.586
Rayon spectral Laplacien μ_n	page 44	6	5
Indice de Kirchhoff $K(G)$	équation (1.44)	17	11.343
Énergie $E_{\mathbf{L}}(G)$	équation (1.45)	10	8.828
Laplacian Energy Like $\mathrm{LEL}(G)$	équation (1.46)	7.914	8.743
Énergie $E_{\mathbf{Q}}(G)$	équation (1.48)	9.411	8.27
Énergie $E_{\mathcal{L}}(G)$	équation (1.55)	3.306	3.064

Figure 2.6 – Exemples d'attributs extraits de deux graphes simples et pondérés G_1 et G_2.

compter le nombre de marches communes[4] entre deux graphes. Le noyau *Shortest Path* [30], quant à lui, compare les plus courts chemins de deux graphes G_i et G_j. Le noyau *Graphlet Count* consiste à compter les sous-graphes connectés non isomorphes de taille limite k, appelés *graphlets*, de deux graphes G_i et G_j [31,171]. D'autres méthodes peuvent être utilisées pour comparer des graphes comme le test d'isomorphisme de Weisfeiler-Lehman ou encore des mesures de similarité comme la distance

4. Deux marches sont dites communes si elles ont la même longueur et que les propriétés de voisinages des sommets visités sont préservées.

d'édition [172, 173] (*graph edit distance* en anglais) qui dénombre le nombre minimal d'opérations nécessaires (ajout ou suppression d'arêtes, ajout ou suppression de sommets, etc.) pour passer d'un graphe G_1 à un graphe G_2.

Nous avons pour objectif, dans ce travail, d'étudier le potentiel des graphes de visibilité pour la classification de séries temporelles dans le cadre de deux tâches : la détection d'épilepsie dans des signaux EEG et la détection d'anomalies magnétiques. Pour cela, nous allons tester deux méthodes : la première étant l'extraction d'attributs de ces graphes de visibilité, appliqués en entrée d'un SVM, la deuxième étant la construction d'un noyau adapté de SVM où la distance euclidienne du noyau RBF est remplacée par une distance entre distributions de probabilités. Ces deux cas d'études ont été choisis mais il est tout à fait possible de transposer le travail effectué ici à d'autres problèmes de classification.

2.3.2 Détection d'épilepsie dans des signaux EEG

L'électroencéphalogramme (EEG) constitue l'enregistrement de l'activité électrique du cerveau établi à partir de mesures de courants délivrés par les connexions nerveuses. Il s'agit du moyen le plus courant pour détecter les crises d'épilepsie, ces dernières étant issues d'un dysfonctionnement passager d'un groupe de neurones [174]. Récemment, des études ont montré l'apport de l'apprentissage automatique pour la détection et la classification de signaux EEG chez les sujets épileptiques, soit en passant d'abord par l'extraction d'attributs pertinents du signal EEG, comme les caractéristiques temporelles et fréquentielles [175] ou les coefficients d'ondelettes [176], soit en manipulant directement ces signaux dans des réseaux de neurones dit récurrents [177]. Par ailleurs, nous nous sommes penchés sur la problématique de détection d'épilepsie dans des signaux EEG car il existe des travaux analysant ces signaux au travers de leurs représentations sous forme de graphes de visibilité [37–39,158–160]. Il est également possible de profiter de nombreuses bases de données de signaux EEG disponibles comme celle de l'université de Bonn [178]. Cette base de données est composée de 5 ensembles d'EEG monocapteur provenant de

— Personnes saines avec les yeux ouverts (ensemble A);
— Personnes saines avec les yeux fermés (ensemble B);
— Personnes malades sans crise hippocampique (ensemble C);
— Personnes malades sans crise d'épilepsie (ensemble D);
— Personnes malades durant une crise d'épilepsie (ensemble E).

Chaque ensemble est constitué de 100 signaux chacun contenant 4097 échantillons. Pour augmenter la taille de la base de données sans perte d'information, chaque signal EEG peut être segmenté en 4 tranches de 1024 échantillons [38]. La base de données est ainsi constituée de 5 ensembles de 400

signaux possédant 1024 échantillons chacun. De plus, étant donné que la classification qui nous intéresse est de déterminer si une personne est dans une crise d'épilepsie ou pas, seuls les tests A vs. E, B vs. E, C vs. E et D vs. E seront à l'étude dans ce chapitre.

Méthode n° 1 : Extraction d'attributs des graphes de visibilité

Dans les travaux sus-cités, les entrées des différents algorithmes de classification correspondent à des attributs extraits des graphes de visibilité comme le degré moyen combiné à des grandeurs comme le poids moyen ou la modularité [38] ou encore la complexité [158]. Nous pensons qu'en augmentant ce nombre d'attributs, les résultats de classification n'en seront que meilleurs, quitte à effectuer une réduction de dimension si nécessaire, c'est-à-dire si certains attributs sont corrélés (ce qui ne va pas être fait dans cette partie). La méthode développée consiste à extraire des attributs des graphes de visibilité horizontale pondérés selon l'angle de vue, c'est-à-dire $w_{ij} = \arctan \frac{x_j - x_i}{j-i}$. Les 26 attributs considérés sont ceux listés en figure 2.6. Ces vecteurs d'attributs sont mis en entrée d'un SVM avec noyau RBF. Le principe de cette méthode est résumé en figure 2.7. Ainsi, pour N séries temporelles de la base de données, après construction de leurs graphes de visibilité horizontale pondérés correspondants, 26 attributs en sont extraits, formant alors une matrice mise en entrée de la SVM de taille $N \times 26$.

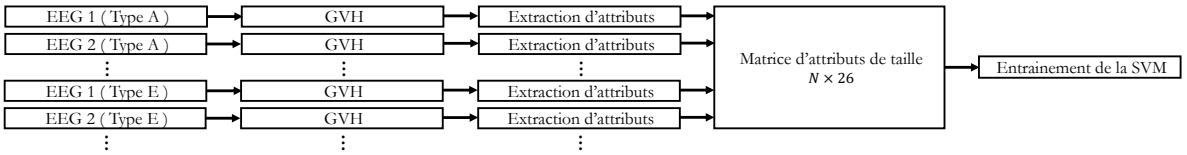


Figure 2.7 – Principe de la méthode n° 1 (cas A vs. E)

Méthode n° 2 : Construction d'un noyau de SVM adapté

Cette deuxième méthode repose sur une intuition quant aux degrés des graphes de visibilité horizontale. Prenons l'exemple de deux signaux EEG (l'un de type A et l'autre de type E) et leurs distributions des degrés associées comme présenté en figure 2.8. Une observation directe des séries temporelles montre une plus forte activité électrique liée à la crise épileptique pour le signal appartenant à l'ensemble E (selon les catégories comparées, les différences ne sont pas toujours aussi évidentes). Il serait pertinent de quantifier cette différence à partir de leurs distributions de degrés associés. Ces dernières, qui peuvent être vues comme des distributions de probabilités, sont comparées à partir de distances statistiques. Soient deux distributions discrètes $\mathbf{p} = (p_i)_{1 \leq i \leq n}$ et $\mathbf{q} = (q_i)_{1 \leq i \leq n}$. Parmi les distances statistiques, citons la distance en variation totale donnée par

$$d_{\mathrm{VT}}(\mathbf{p}, \mathbf{q}) = 2 \sup_{1 \leq i \leq n} |p_i - q_i|, \tag{2.9}$$

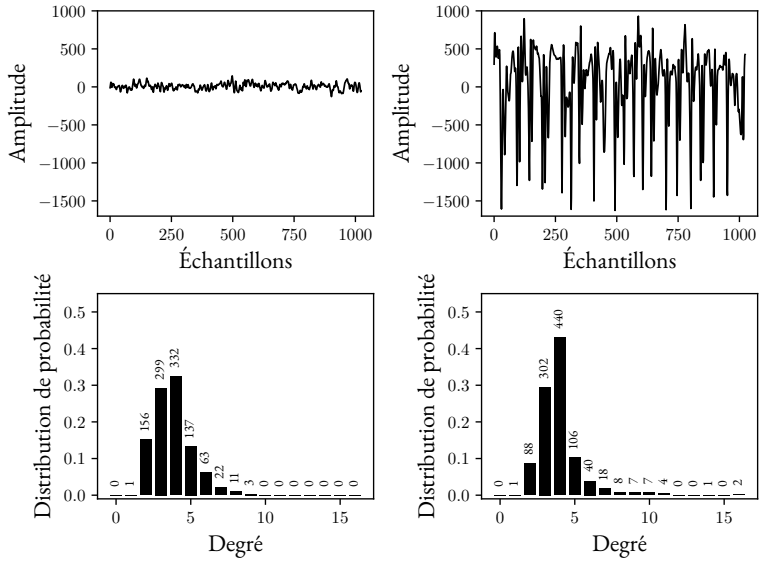


Figure 2.8 – Deux exemples de signaux EEG (1024 échantillons) et la distribution des degrés de leurs graphes de visibilité horizontale associés (1024 sommets) : Type A à gauche (1935 arêtes) / Type E à droite (2013 arêtes).

la distance[5] de Bhattacharyya [157]

$$d_{\mathrm{B}}(\mathbf{p},\mathbf{q}) = \sum_{i=1}^{n} \sqrt{p_i q_i}, \tag{2.10}$$

la distance de Hellinger [157]

$$d_{\mathrm{H}}(\mathbf{p},\mathbf{q}) = \sqrt{\frac{1}{2}\sum_{i=1}^{n} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2}, \tag{2.11}$$

ou encore la distance de Jensen-Shannon qui, basée sur une symétrisation de la divergence de Kullback-Leibler[6], peut être définie par [180]

$$d_{\mathrm{JS}}(\mathbf{p},\mathbf{q}) = S(\mathbf{m}) - \frac{S(\mathbf{p}) + S(\mathbf{q})}{2}, \qquad \mathbf{m} = \frac{\mathbf{p}+\mathbf{q}}{2} \tag{2.13}$$

5. C'est un abus de langage car la distance de Bhattacharyya ne vérifie pas l'inégalité triangulaire.
6. La divergence de Kullback-Leibler entre deux distributions $\mathbf{p} = (p_i)_{1\leq i\leq n}$ et $\mathbf{q} = (q_i)_{1\leq i\leq n}$ est donnée par [179]

$$\mathrm{KL}(\mathbf{p}\|\mathbf{q}) = \sum_{i=1}^{n} p_i \ln\left(\frac{p_i}{q_i}\right). \tag{2.12}$$

où $S(\mathbf{p})$ désigne l'entropie de Shannon de la distribution discrète $\mathbf{p}$ donnée par [181]

$$S(\mathbf{p}) = -\sum_{i=1}^{n} p_i \ln(p_i). \tag{2.14}$$

Bien que ce ne soit pas véritablement un outil adapté, il est également possible de comparer les distributions de probabilités discrètes en utilisant des distances traditionnelles comme la distance euclidienne

$$d(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_2 = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}. \tag{2.15}$$

Les distributions de probabilités sont, dans ce travail, les distributions des degrés des graphes de visibilité horizontale des signaux considérés. Ainsi, en notant $\mathbf{p}_i$ (resp. $\mathbf{p}_j$) la distribution des degrés du graphe de visibilité horizontale construit à partir du i^{e} signal (resp. j^{e}), il est alors possible de créer la matrice de Gram $\mathbf{K}$ comme suit :

$$[\mathbf{K}]_{ij} = \exp\left(-\gamma d(\mathbf{p}_i, \mathbf{p}_j)\right), \quad 1 \leq i, j \leq N. \tag{2.16}$$

où d peut être choisie parmi les distances précédemment exposées. Ce noyau est donc une simple variante du noyau RBF (*cf.* page 67) avec, en lieu et place de la distance euclidienne, des distances statistiques adaptées à la comparaison de distributions de probabilités. La deuxième méthode proposée dans ce travail est schématisée en figure 2.9. Après avoir transformé les signaux EEG en graphe de visi-

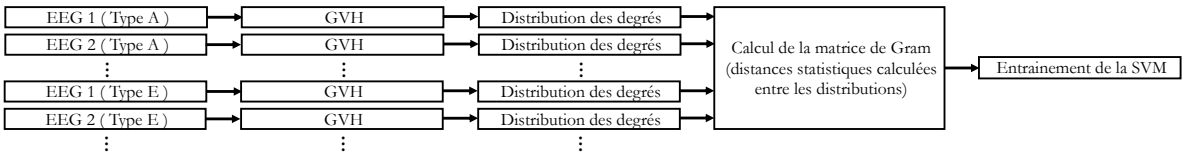


Figure 2.9 – Principe de la méthode n° 2 (cas A vs. E)

bilité horizontale (GVH), les distributions des degrés sont calculées et la matrice de Gram constituée des distances statistiques est construite. La SVM est alors entraînée grâce à cette matrice.

Résultats de classification

Pour plus d'homogénéité des résultats, nous avons ré-implémenté les méthodes proposées par les travaux antérieurs étudiant, eux aussi, cette problématique de détection d'épilepsie dans des signaux EEG (issus de la même base de données) à l'aide de graphes de visibilité [37–39]. Parmi ces derniers, citons

— le travail de Zhu *et al.* [37] : après avoir calculé les graphes de visibilité horizontale pondérés des signaux EEG (avec des poids d'arêtes égaux à $w_{ij} = |(x_i - x_j)(i - j)| + 1$), ils mettent en entrée de l'algorithme de classification le degré moyen et la force moyenne.

— le travail de Supriya *et al.* [38] : après avoir calculé les graphes de visibilité naturelle pondérés des signaux EEG (avec des poids d'arêtes égaux à $w_{ij} = \arctan \frac{x_j - x_i}{j - i}$), ils mettent en entrée de l'algorithme de classification la modularité[7] [182,183] et la force moyenne.

— le travail de Rajadurai *et al.* [39] : après avoir calculé les graphes de visibilité naturelle pondérés des signaux EEG (avec des poids d'arêtes égaux à $w_{ij} = \arctan \frac{x_j - x_i}{j - i}$), ils mettent en entrée de l'algorithme de classification le degré moyen, la force moyenne et l'entropie moyenne des nœuds définie comme une variante de l'entropie de Shannon (2.14) par

$$\bar{S} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} p_{ij} \log_2 p_{ij}, \qquad p_{ij} = \frac{w_{ij}}{\sum_{k=1}^{n} w_{ik}} \tag{2.17}$$

Pour toutes ces méthodes, une validation croisée à 10 couches est opérée pour entraîner la SVM. Par ailleurs, une recherche d'hyperparamètres optimaux est effectuée pour le paramètre de régularisation $C \in \{1, 10, 50, 100, 250, 500, 750, 1000\}$ et le coefficient d'échelle $\gamma \in \{0.1, 0.5, 1, 5, 10, 50\}$, que ce soit pour notre noyau adapté ou le noyau RBF utilisé pour les autres méthodes. Plusieurs métriques de performance sont considérées pour valider la pertinence des différentes stratégies : l'exactitude, la sensibilité, la spécificité et l'aire sous la courbe (AUC). Toutefois, avant de comparer les différentes méthodes entre elles, une question demeure : quelle est la meilleure distance statistique à prendre en compte pour cette méthode n° 2 dans le cas de la détection d'épilepsie ? Le tableau 2.2 liste les résultats de détection avec les différentes distances statistiques. À la lecture de ce tableau, il est clair que deux distances se placent en tête des performances en termes d'exactitude, de sensibilité et d'AUC : la distance de Jensen-Shannon et la distance d'Hellinger. Cette dernière semble bien adaptée à la détection d'épilepsie dans des signaux EEG avec, dans le cas du test D vs. E, une meilleure sensibilité que celle obtenue avec la distance de Jensen-Shannon. En revanche, si la métrique de performance à maximiser est la spécificité, alors c'est la distance de Bhattacharyya qu'il faudra considérer. En effet, c'est celle qui maximise cette métrique pour les quatre tests. Pour les futurs résultats exposés, seuls ceux de la meilleure distance sont retenus pour chaque test et métrique de performance.

Les résultats, recensés dans le tableau 2.3, montrent que les deux méthodes proposées dans ce travail sont compétitives comparativement aux autres stratégies issues de la littérature, et ceci, quelle que soit la métrique de performance considérée, surtout pour les classifications les plus complexes, à sa-

7. La modularité d'un réseau mesure la qualité d'un partitionnement de ce dernier en plusieurs communautés.

Tests	Distances statistiques	Exactitude	Sensibilité	Spécificité	AUC
A vs. E	Jensen-Shannon	**99.50 %**	**99.25 %**	99.75 %	**1.00**
	Variation totale	99.00 %	98.50 %	99.50 %	0.99
	Euclidienne	99.00 %	98.50 %	99.50 %	0.99
	Bhattacharyya	99.25 %	98.75 %	**100 %**	0.99
	Hellinger	**99.50 %**	**99.25 %**	99.75 %	**1.00**
B vs. E	Jensen-Shannon	**96.88 %**	**96.00 %**	**100 %**	**0.97**
	Variation totale	92.12 %	88.25 %	**100 %**	0.92
	Euclidienne	95.12 %	93.50 %	99.75 %	0.95
	Bhattacharyya	96.62 %	95.00 %	**100 %**	**0.97**
	Hellinger	**96.88 %**	**96.00 %**	99.75 %	**0.97**
C vs. E	Jensen-Shannon	**99.75 %**	99.50 %	**100 %**	**1.00**
	Variation totale	98.50 %	98.25 %	99.25 %	0.99
	Euclidienne	98.75 %	98.00 %	99.50 %	0.99
	Bhattacharyya	99.88 %	**99.75 %**	**100 %**	**1.00**
	Hellinger	**99.75 %**	**99.75 %**	**100 %**	**1.00**
D vs. E	Jensen-Shannon	96.62 %	97.50 %	97.00 %	**0.97**
	Variation totale	94.88 %	95.75 %	94.75 %	0.95
	Euclidienne	96.12 %	96.50 %	96.75 %	0.96
	Bhattacharyya	96.62 %	97.25 %	**97.25 %**	**0.97**
	Hellinger	**97.00 %**	**98.25 %**	97.00 %	**0.97**

Tableau 2.2 – Comparaisons des résultats de la méthode n° 2 avec différentes distances statistiques dans le noyau (2.16).

voir (C vs. E) et (D vs. E). Seule la méthode de Zhu *et al.* se montre efficace, surtout en sensibilité où cette méthode obtient toujours un score proche de 100 %. Les résultats répertoriés dans le tableau 2.3 permettent de tirer un certain nombre de conclusions. En effet, comme le montrent les résultats de la méthode n° 1, prendre en compte plus d'attributs des graphes de visibilité permet d'augmenter les performances de classification. Cette méthode pourrait vraisemblablement être améliorée en extrayant davantage d'attributs et en effectuant une ACP pour ôter les variables corrélées. Il existe, de surcroît, de nombreuses variantes de graphes de visibilité (naturelle, horizontale, différence, signée, etc.) et pour toutes ces variantes, il existe une quantité considérable de pondérations possibles (angle, angle absolu, écart horizontal, vertical, etc.). Peut-être existe-t-il une combinaison particulièrement adaptée à la détection d'épilepsie dans des signaux EEG ? Les résultats de la méthode n° 2 appuient, en un certain sens, notre conjecture quant au fait que les distributions des degrés seules des graphes de visibilité horizontale permettent, à condition d'être comparées à l'aide d'une distance adéquate, une bonne classification. Pour le cas de la détection d'épilepsie, la distance d'Hellinger semble être la plus adaptée pour comparer les distributions des degrés. Les deux méthodes présentées dans ce travail s'avèrent plus performantes que les autres stratégies issues de la littérature qui extraient un nombre contenu d'attributs de ces graphes de visibilité (modularité, degré moyen, etc.), preuve que cette manière de représenter les séries temporelles regorge d'informations utiles relatives à ces dernières.

Tests	Méthodes	Exactitude	Sensibilité	Spécificité	AUC
A vs. E	Zhu *et al.* [37]	99.00 %	**100 %**	**99.50 %**	**0.99**
	Supriya *et al.* [38]	84.88 %	86.25 %	90.75 %	0.85
	Rajadurai *et al.* [39]	99.12 %	**99.75 %**	98.50 %	**0.99**
	Méthode n° 1 (attributs)	**99.62 %**	**100 %**	**100 %**	**1.00**
	Méthode n° 2 (noyau adapté)	**99.50 %**	99.25 %	**100 %**	**1.00**
B vs. E	Zhu *et al.* [37]	93.25 %	**99.25 %**	95.00 %	0.93
	Supriya *et al.* [38]	81.75 %	81.75 %	93.50 %	0.82
	Rajadurai *et al.* [39]	93.25 %	91.50 %	95.25 %	0.93
	Méthode n° 1 (attributs)	**96.75 %**	**100 %**	**98.00 %**	**0.97**
	Méthode n° 2 (noyau adapté)	**96.88 %**	96.00 %	**100 %**	**0.97**
C vs. E	Zhu *et al.* [37]	98.50 %	**100 %**	98.50 %	**0.99**
	Supriya *et al.* [38]	75.12 %	74.75 %	90.75 %	0.75
	Rajadurai *et al.* [39]	92.00 %	92.75 %	91.75 %	0.92
	Méthode n° 1 (attributs)	**99.62 %**	**100 %**	**100 %**	**1.00**
	Méthode n° 2 (noyau adapté)	**99.88 %**	**99.75 %**	**100 %**	**1.00**
D vs. E	Zhu *et al.* [37]	96.12 %	**100 %**	95.25 %	0.96
	Supriya *et al.* [38]	68.62 %	73.50 %	75.25 %	0.69
	Rajadurai *et al.* [39]	82.62 %	86.75 %	80.00 %	0.83
	Méthode n° 1 (attributs)	**97.12 %**	**100 %**	**97.25 %**	**0.97**
	Méthode n° 2 (noyau adapté)	**97.00 %**	**98.25 %**	**97.25 %**	**0.97**

Tableau 2.3 – Comparaison des résultats avec des méthodes de la littérature à partir de graphes de visibilité.

2.3.3 Détection d'anomalies magnétiques

Les analyses, menées auparavant, sont peut-être spécifiques à la détection d'épilepsie, les signaux EEG étant une classe de signaux relativement particulière. Dans cette sous-section, nous nous intéressons à la détection d'anomalies magnétiques (MAD pour *Magnetic Anomaly Detector*). Le MAD est une technique utilisée pour identifier des variations ou des perturbations dans le champ magnétique terrestre. Ce champ étant généralement stable et uniforme dans une région donnée, une perturbation peut indiquer la présence d'objets contenant des matériaux ferromagnétiques comme un sous-marin, une épave ou des infrastructures métalliques enfouies. Le MAD peut avoir des applications dans le domaine de la géologie, l'archéologie sous-marine ou encore dans le domaine militaire, où le MAD est principalement utilisée pour détecter des sous-marins [184, 185]. Pour des raisons évidentes de confidentialité, il n'est pas permis d'analyser des signaux réels issus de patrouilleurs maritimes, munis de capteurs magnétiques. Toutefois, il est possible de travailler sur des anomalies magnétiques simulées de manière relativement fidèle. La base de données utilisée est composée de 10000 signaux synthétiques d'une durée de 1 minute et échantillonnés à 5 Hz, séparés en deux classes : bruit (classe -1) ou bruit avec signature (classe 1). Un exemple de ces signaux est présenté en figure 2.10.

Au sein de chacune de ces classes, les signaux sont répartis selon 5 profondeurs d'immersion de l'ob-

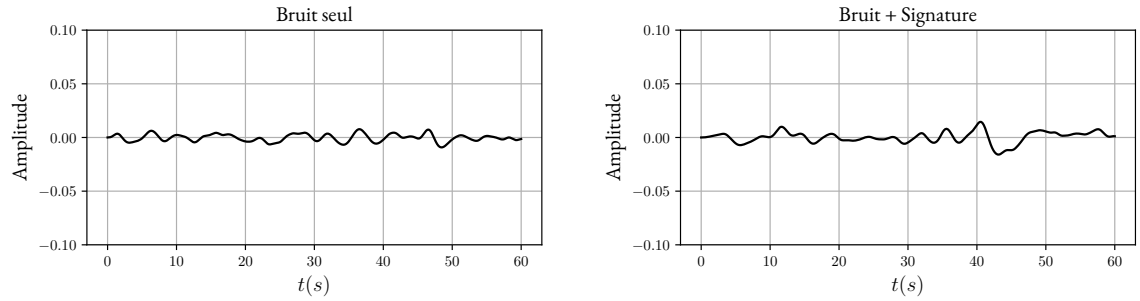


Figure 2.10 – Exemple de séries temporelles étudiées : bruit seul à gauche et bruit avec une anomalie magnétique à droite.

jet d'intérêt signant (pour feindre une profondeur d'immersion de l'objet d'intérêt de plus en plus grande, le rapport signal à bruit (SNR pour *Signal to Noise Ratio*) est diminuée). Ainsi, nous testons toutes les méthodes de classification précédemment présentées afin de comparer les résultats obtenus. Pour toutes ces méthodes, une validation croisée à 10 couches est opérée pour entraîner la SVM. Par ailleurs, une recherche d'hyperparamètres optimaux est effectuée pour le paramètre de régularisation $C \in \{1, 10, 50, 100, 250, 500, 750, 1000\}$ et le coefficient d'échelle $\gamma \in \{0.1, 0.5, 1, 5, 10, 50\}$, que ce soit pour notre noyau adapté ou le noyau RBF utilisé pour les autres méthodes. Comme précédemment, la question se pose concernant la distance statistique la plus appropriée pour comparer les distributions de degrés. Les résultats de classification de la méthode n° 2 avec différentes distances sont présentés en tableau 2.4. Les exactitudes sont toutes proches les unes des autres, tout comme les sensibilités. Toutefois, les meilleurs résultats en termes de spécificité reviennent à la distance de Bhattacharyya qui, affichant un résultat 8 % supérieur aux autres distances, semble alors plus adaptée dans le cas de la détection d'anomalies magnétiques.

Distances statistiques	Exactitude	Sensibilité	Spécificité	AUC
Jensen-Shannon	73.00 %	66.10 %	90.50 %	0.73
Variation totale	70.30 %	**66.70 %**	85.2 %	0.70
Euclidienne	72.95 %	65.8 %	87.30 %	0.73
Bhattacharyya	**73.85 %**	66.00 %	**98.90 %**	**0.74**
Hellinger	**73.10 %**	**67.40 %**	90.70 %	0.73

Tableau 2.4 – Comparaisons des résultats de la méthode n° 2 avec différentes distances statistiques dans le noyau (2.16).

Rappelons néanmoins que, pour des tâches comme la détection d'anomalies magnétiques, c'est surtout la sensibilité qu'il faut maximiser car c'est la mesure qui révèle la capacité de l'algorithme à détecter un maximum de cibles (c'est-à-dire à avoir le moins de faux négatifs possible). En la matière, c'est la méthode n° 1 basée sur une extraction d'attributs qui est la meilleure parmi toutes les méthodes

basées sur les graphes de visibilité. En effet, même si ces méthodes ont été développées pour la détection d'épilepsie et ne sont pas les plus appropriées pour la détection d'anomalies magnétiques, nous les avons testées pour cette tâche car ce sont des stratégies qui traitent de la classification de séries temporelles à l'aide de graphes de visibilité. Les différents résultats, recensés dans le tableau 2.5, montrent que la méthode n° 1 est meilleure pour l'exactitude d'au moins 3 %, pour la sensibilité d'au moins 32 % et pour l'AUC d'au moins 0.04. La méthode de Zhu *et al.* [37] se distingue encore une fois parmi celles de la littérature.

Méthodes	Exactitude	Sensibilité	Spécificité	AUC
Zhu *et al.* [37]	**74.05 %**	63.3 %	**89.30 %**	**0.74**
Supriya *et al.* [38]	58.00 %	58.90 %	61.10 %	0.58
Rajadurai *et al.* [39]	71.95 %	55.20 %	89.30 %	0.72
Méthode n° 1 (attributs)	**77.50 %**	**99.80 %**	87.90 %	**0.78**
Méthode n° 2 (noyau adapté)	73.85 %	67.40 %	**98.90 %**	0.74

Tableau 2.5 – Comparaison des résultats avec des méthodes de la littérature à partir de graphes de visibilité.

Des remarques identiques à celles formulées lors de la détection d'épilepsie concernant l'information que fournissent les graphes de visibilité peuvent être réitérées. Pour apporter du crédit à cette assertion, la section suivante s'intéresse à la caractérisation de processus stochastiques (mouvements Browniens fractionnaires et bruits Gaussiens fractionnaires) en se basant sur la localisation de leurs graphes de visibilité associés dans un plan informationnel appelé plan de Fisher-Shannon.

2.4 Caractérisation de processus aléatoires dans un plan informationnel

Depuis des dizaines d'années, il est admis par la communauté scientifique qu'un nombre important de phénomènes physiques sont modélisables par des processus stochastiques possédant des propriétés algébriques et spectrales singulières. Ces phénomènes proviennent de domaines extrêmement variés comme la physique, la biologie ou encore la finance. Parmi ces derniers, citons le bruit électronique émis par le courant électrique passant à travers des composants [186], l'évolution de marchés financiers [187], la distribution des nombres premiers [188] ou encore le niveau d'eau du Nil[8] [41]. Il est en effet possible de modéliser mathématiquement ces phénomènes physiques en utilisant deux types de processus aléatoires intimement liés : les mouvements Browniens fractionnaires (fBm pour

8. Cette dernière application a d'ailleurs été mise en avant par l'hydrologue britannique Harold Edwin Hurst, précurseur de l'étude de ce type de signaux, repris par la suite par de grands noms de l'analyse fractale de séries temporelles comme le mathématicien franco-américain Benoit Mandelbrot.

fractionnal Brownian motion en anglais), non-stationnaires par construction, ainsi que les bruits Gaussiens fractionnaires (fGn pour *fractional Gaussian noise* en anglais), stationnaires par construction n'étant ni plus ni moins que les incréments des précédents. Ces deux types de processus aléatoires peuvent, eux aussi, être approchés par des signaux appelés bruits colorés, caractérisés par une DSP (ou densité spectrale de puissance[9]) qui suit une loi de puissance $S(f) \sim f^{-\beta}$ où le paramètre β varie généralement entre -1 et 3. L'étude de tous ces signaux se fait traditionnellement avec des outils classiques du traitement de signal tels que la transformée de Fourier ou la transformée en ondelettes, mais le travail effectué dans le cadre de cette thèse se penche plutôt sur l'utilisation conjointe d'une représentation de ces signaux sous la forme de graphes, construits grâce à l'algorithme de visibilité, et de grandeurs de la théorie de l'information telles que l'information de Fisher, l'entropie de Shannon ou encore celle de Rényi. Il ne serait pas possible d'aller plus loin sans parler du travail de Bruna Amin Gonçalves qui, dans sa thèse [24], s'est intéressée à ces problématiques, c'est-à-dire l'analyse des processus stochastiques à l'aide de graphes de visibilité horizontale projetés dans un plan informationnel, à savoir le plan Fisher-Shannon. Cette section débute par la description du travail de Gonçalves [24,25] où il est démontré qu'une localisation de ces processus stochastiques selon leurs paramètres respectifs (coefficient de Hurst H pour les fBm et fGn et pente β pour les bruits colorés) est aussi possible avec des graphes de visibilité naturelle. Ceux-ci, par construction, contiennent plus d'informations quant à la structure du signal sous-jacent, et également possible dans un plan informationnel généralisant celui de Fisher-Shannon : celui de Fisher-Rényi. Dans cette section, une méthode d'estimation du coefficient de Hurst est par ailleurs présentée. De nombreuses méthodes, issues du traitement de signal, permettent d'estimer ce coefficient. Ce coefficient joue un rôle primordial dans l'étude de tels signaux tant il permet de mettre en lumière des propriétés de dépendance à court ou long-terme pour en étudier le comportement et la prédictibilité (très utile dans le cadre de cours financiers par exemple). C'est la raison pour laquelle la méthode proposée dans cette section est bien entendu comparée avec ces outils classiques pour en montrer la pertinence.

2.4.1 Bruits colorés, mouvements Browniens et bruits Gaussiens fractionnaires

« Signal généré par un processus aléatoire » constitue une définition parfois rencontrée pour désigner le bruit. Le bruit blanc est par exemple un processus aléatoire, stationnaire au second ordre, dont la densité spectrale de puissance est la même pour toutes les fréquences de la bande passante (la DSP d'un bruit blanc suit alors une loi en $1/f^{\beta}$ où $\beta = 0$). Lorsque des signaux aléatoires – des bruits –

9. La densité spectrale de puissance $S(f)$ d'un signal est le carré du module de sa transformée de Fourier.

sont étudiés, une information importante à extraire est la présence ou non d'une dépendance temporelle à court ou long-terme. En effet, dans le cas d'un cours financier, stochastique par essence, il peut être essentiel de savoir si des phénomènes tels que des *krachs* boursiers ou autre peuvent se reproduire dans un laps de temps donné. C'est pour caractériser les dépendances à long-terme dans des séries temporelles que Hurst a introduit un coefficient d'auto-similarité $H \in [0, 1]$ portant son nom, le coefficient de Hurst [41]. Son intérêt initial était de caractériser la dépendance à long-terme du niveau d'eau contenu dans des réservoirs situés le long du Nil, mais il a, par la suite, servi à développer de nouveaux types de processus stochastiques comme les mouvements Browniens fractionnaires [189]. Un mouvement Brownien fractionnaire de coefficient H est défini comme un processus Gaussien $B_H(t)$ auto-similaire, non stationnaire[10], centré pour tout $t \in [0, T]$ avec une covariance égale à

$$\mathbb{E}[B_H(t)B_H(s)] = \frac{1}{2}\left(|t|^{2H} + |s|^{2H} - |t-s|^{2H}\right) \tag{2.18}$$

Dans le cas où $H \in [0, 1/2[$, le processus présente une courte dépendance et ses incréments sont corrélés négativement, tandis que la dépendance est longue et ses incréments sont positivement corrélés dans le cas où $H \in]1/2, 1]$. Lorsque $H = 1/2$, le mouvement Brownien standard est obtenu [190]. Comme cela a été évoqué ci-dessus, il est possible d'approcher les mouvement Browniens fractionnaires par des bruits colorés, c'est-à-dire des séries temporelles dont la DSP suit une loi en $f^{-\beta}$. Dans le cas des mouvements Browniens fractionnaires, le lien entre le coefficient de Hurst H et la pente β de la DSP est donné par $\beta = 2H + 1$, ce qui implique que $\beta \in [1, 3]$ car $H \in [0, 1]$. Le problème avec les mouvements Browniens fractionnaires est qu'ils sont non stationnaires alors que bon nombre de phénomènes physiques le sont. Ainsi, pour pouvoir les modéliser, il est possible de se servir des incréments $G_H(t) = B_H(t) - B_H(t-1)$ d'un mouvement Brownien fractionnaire $B_H(t)$ [191]. Ce processus stochastique, appelé bruit Gaussien fractionnaire, est en effet stationnaire. Cette fois, le lien entre la pente β de la DSP et le coefficient de Hurst H est $\beta = 2H - 1$ de sorte que, lorsque H varie de 0 à 1, la pente varie de -1 à 1, permettant d'aller du bruit rose ($\beta = 1$) au bruit bleu ($\beta = -1$) en passant par le bruit blanc Gaussien ($\beta = 0$) [190]. La figure 2.11 apporte une illustration de tous ces processus stochastiques ainsi que du lien existant entre coefficient de Hurst et pente de DSP. Pour $\beta = -1$, le bruit obtenu est un bruit bleu, pour $\beta = 0$, le bruit obtenu est un bruit blanc Gaussien, pour $\beta = 1$, le bruit obtenu est un bruit rose, pour $\beta = 2$, le bruit obtenu est un bruit Brownien (parce que c'est aussi un mouvement Brownien standard) et pour $\beta = 3$, il est parfois évoqué la notion de bruit noir.

Que ce soit les bruits Gaussiens fractionnaires ou leurs intégrations, les mouvements Browniens fractionnaires, ils sont tous utilisés pour modéliser une large gamme de phénomènes physiques et la

10. La non-stationnarité est une non-propriété. Un processus est dit stationnaire lorsque ses propriétés statistiques sont indépendantes de l'origine des temps.

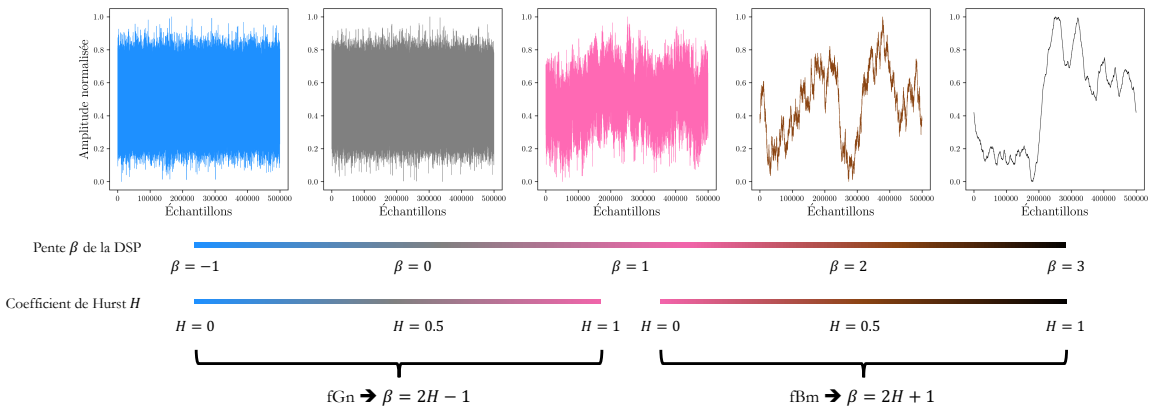


Figure 2.11 – Différents bruits colorés (*i.e.* des bruits ayant une DSP suivant une loi en $f^{-\beta}$) de 5×10^5 échantillons normalisés entre 0 et 1. Figure reproduite depuis celle de Marmelat *et al.* [192]

plupart de leurs propriétés statistiques sont caractérisées par le paramètre H. Pour construire un fGn de paramètre H, il est possible de construire un fBm de paramètre H et d'en calculer sa dérivée discrète d'ordre 1 [193]. De même, pour étudier un fGn de paramètre H, il peut être plus facile de l'intégrer (en calculer la somme cumulée) pour le voir comme un fBm de paramètre H : en effet, la plupart des outils classiques du traitement de signal qui étudie le coefficient de Hurst le font à partir de fBm uniquement. Il faut toutefois veiller à retrancher la moyenne du fGn avant de l'intégrer sinon le résultat ne sera pas de type fBm [193]. Dans la suite, nous allons avoir besoin de fBm (ou fGn) réels, c'est-à-dire issus de vrais phénomènes physiques, mais aussi de signaux simulés. Pour générer des mouvements Browniens fractionnaires, nous allons utiliser la librairie Python `fbm`[11] qui implémente la méthode de Davies et Harte [194], dont le principe est rappelé en annexe B. Cette dernière a l'avantage d'être une simulation exacte et très rapide de mouvements Browniens fractionnaires. Cette librairie est préférée à la fonction `wfbm` de Matlab implémentant la méthode introduite par Fabrice Sellan, Patrice Abry et Yves Meyer basée sur une transformée en ondelettes d'un bruit blanc Gaussien [195]. En effet, un débat est apparu sur le fait qu'elle n'approximait pas un mouvement Brownien fractionnaire, mais un autre processus Gaussien n'ayant pas d'incréments stationnaires [196]. Elle fut réaccréditée notamment par Pipiras qui a proposé certaines modifications dans l'algorithme de simulation [197]. Ne sachant si ces modifications ont été prises en compte dans la fonction Matlab, nous avons alors privilégié la librairie Python. Quant aux bruits colorés, une manière efficiente d'en générer avec la bonne pente de DSP peut être la méthode proposée par Timmer et Koenig [198], implémentée dans la librairie Python `colorednoise`[12]. Cette méthode utilise deux tirages aléatoires Gaussiens dans le domaine

11. `https://pypi.org/project/fbm/`
12. `https://pypi.org/project/colorednoise/`

fréquentiel qui, après normalisation dans le but d'obtenir la loi en puissance souhaitée, correspondent à la partie réelle et la partie imaginaire de la transformée de Fourier. Une transformée de Fourier inverse permet alors d'obtenir le bruit coloré dans le domaine temporel. Bien que moins intuitive qu'une simple manipulation spectrale d'un bruit blanc Gaussien comme peut le proposer Zhivomirov [199], cette méthode permet d'obtenir le résultat souhaité, c'est-à-dire un signal dont la DSP suit une loi en $f^{-\beta}$.

2.4.2 Distributions extraites des graphes de visibilité

L'idée initiale derrière la méthode de caractérisation de processus stochastiques développé ici est de comparer un signal réel par rapport à un nombre important de signaux générés pour en déduire soit une pente de DSP, soit un coefficient de Hurst. Mais que compare-t-on et où les compare-t-on ? Avant d'apporter une réponse à la deuxième question dans la sous-section suivante, penchons-nous sur la première. Nous allons comparer des informations extraites des graphes de visibilité, à savoir les distributions de degrés. En effet, comme cela a été évoqué en chapitre 1, les degrés et leurs distributions représentent un outil extrêmement rapide à calculer, robuste et dont l'information permet une caractérisation relativement efficace de graphes. De plus, le rôle que joue la distribution des degrés dans l'estimation du coefficient de Hurst est assez importante car, après l'introduction du graphe de visibilité naturelle par Lacasa *et al.* en 2008 [21], de nombreux articles se sont succédés présentant la relation affine entre la pente (estimée) γ de la distribution des degrés et le coefficient de Hurst H (ou encore la pente β de DSP) [22,161,162]. En effet, dans le cadre de l'étude de mouvements Browniens fractionnaires de paramètre H, la distribution des degrés de leurs graphes de visibilité naturelle associés suit une loi en $p_k \sim k^{-\gamma}$. En d'autres termes, le graphe de visibilité naturelle d'un mouvement Brownien fractionnaire est un graphe sans échelle [22]. Cette « pente » γ peut être estimée en utilisant l'estimateur du maximum de vraisemblance [200] :

$$\gamma = 1 + n \left(\sum_{i=1}^{n} \ln \frac{k_i}{k_{\min}} \right)^{-1} \tag{2.19}$$

où les $(k_i)_{1 \leq i \leq n}$ sont les degrés existants dans le graphe et $k_{\min}$ correspond au plus petit degré pour lequel la loi de puissance existe. Par suite, il est possible de mettre en évidence la relation $\gamma(H) = 3 - 2H$. Pour des bruits colorés de pente β, la relation devient $\gamma(\beta) = 4 - \beta$, ce qui est tout à fait normal puisque $\beta = 2H + 1$. Quant aux bruits Gaussiens fractionnaires de paramètre H, la relation est $\gamma(H) = 5 - 2H$. La figure 2.12 appuie les assertions précédentes dans le cas de mouvements Browniens fractionnaires admettant 0.2, 0.5 et 0.8 pour coefficients de Hurst et dont les tirages possèdent

10^6 échantillons.

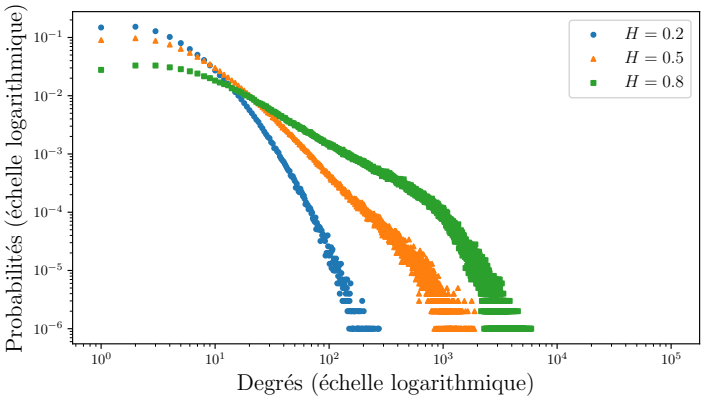


Figure 2.12 – Distribution des degrés des graphes de visibilité construits à partir de trois mouvements Browniens fractionnaires de coefficient de Hurst égal à 0.2 (points bleus), 0.5 (triangles oranges) et 0.8 (carrés jaunes).

Le travail de Lacasa *et al.* est le premier qui met en évidence un lien clair entre coefficient de Hurst et graphes de visibilité, ce qui solidifie grandement le pont créé entre la théorie des graphes et le traitement de signal. Cette méthode permet d'ores et déjà une estimation du coefficient de Hurst mais ne semble pas très robuste car il est nécessaire de disposer d'un autre estimateur pour le coefficient γ de la loi de puissance. D'où la volonté dans cette thèse de définir une autre stratégie avec les mêmes outils. La distribution des degrés contient bien des informations importantes sur le coefficient de Hurst des processus sous-jacents, comme illustré par la figure 2.13 où des tirages de fBm ayant 100000 échantillons pour différentes valeurs de H et la distribution des degrés de leurs graphes de visibilité naturelle associée sont représentés. Visuellement, ces distributions sont bien différentes les unes des autres. Pour des petites valeurs de H, c'est-à-dire lorsque le processus est plus rugueux (ou que les dépendances se présentent à court terme), la distribution des degrés est très étroite autour des degrés 2 et 3, ce qui s'explique assez simplement : de par la structure temporelle du processus, il y a peu d'échantillons qui « voient » un nombre important d'échantillons voisins. Pour des coefficients de Hurst élevé, autrement dit lorsque le processus est plus lisse, il est clair qu'un nombre plus important d'échantillons voit un grand nombre de leurs voisins, ce qui se traduit par une queue de distribution plus épaisse. Toutes ces observations peuvent faire penser que cette distribution des degrés est un outil adéquat pour discriminer deux processus stochastiques ayant des coefficients de Hurst différents.

Bien entendu, comme mentionné en section 1.4, les graphes de visibilité sont pondérables en ce sens où des poids peuvent être appliqués à ses arêtes (pente de visibilité, angle de visibilité, etc.) : nous pourrions alors utiliser la distribution des pondérations. C'est d'ailleurs ce que font Gonçalves *et al.*

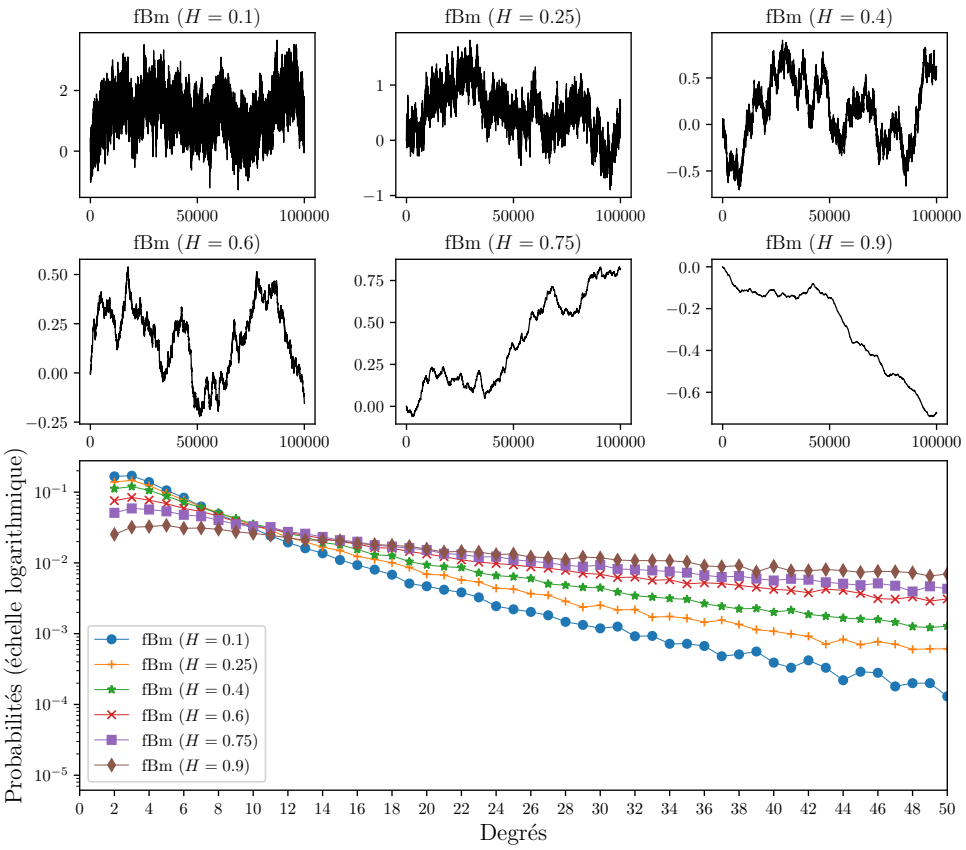


Figure 2.13 – Tirages contenant 100000 points de mouvements Browniens fractionnaires pour un coefficient de Hurst $H \in \{0.1, 0.25, 0.4, 0.6, 0.75, 0.9\}$ et les distributions des degrés de leurs graphes de visibilité naturelle associés.

dont ce travail s'inspire grandement [25]. L'inconvénient est que le calcul de la distribution des pondérations implique alors le choix du nombre de classes ainsi que de leurs bornes ce qui relève majoritairement de la subjectivité. L'avantage, pour n'en présenter qu'un, est qu'une fois ce choix fait, les graphes étudiés n'ont plus besoin d'avoir le même nombre de sommets pour comparer des distributions de même taille puisque ce ne sont plus les degrés qui sont comparés mais les poids des arêtes. Pour la distribution des degrés, le problème du choix subjectif ne se pose pas car elle est calculée pour toutes les valeurs de degré, de 1 au nombre de sommets. Mais en cas de comparaison des distributions de même taille, il faut que les graphes étudiés soient de même ordre, ce qui n'est pas toujours vrai. À partir de l'observation des distributions représentées en figure 2.13, il est raisonnable de se dire que l'information importante se situe dans les premiers degrés, et ce même pour des signaux ayant un nombre très important d'échantillons. Dans la suite, les distributions seront ainsi calculées que jusqu'à un degré fixé à $\varepsilon = 100$. Bien entendu, il est quand même nécessaire que les signaux soient de taille suffisante pour que la distribution des degrés soit ressemblante à son asymptote. Il est à noter la possibilité de

calculer d'autres distributions non conventionnelles comme la distribution des distances [201] ou la distribution des centralités [202]. Ces distributions sont toutefois plus complexes à calculer pour le travail mené ici, à savoir la caractérisation de processus stochastiques en vue d'une estimation de la pente β de DSP ou du coefficient de Hurst H.

2.4.3 Construction d'un squelette dans un plan informationnel

Plan informationnel de Fisher-Shannon et variantes

Maintenant que nous sommes en mesure de simuler des processus stochastiques tels que des mouvements Browniens fractionnaires, des bruits Gaussiens fractionnaires ou encore des bruits colorés, d'en calculer leurs graphes de visibilité associés et d'extraire la distribution des degrés de ces derniers, il est question dans cette sous-section d'introduire les plans dits informationnels, en d'autres termes définis par deux métriques informationnelles, dans lesquels nous allons représenter ces distributions. L'idée originelle revient à Vignat et Bercher qui, en 2003, définissent le plan informationnel de Fisher-Shannon [42]. Ce plan, comme son nom l'indique, combine deux grandeurs informationnelles : une globale (entropie de Shannon) et une locale (information de Fisher). Leur ambition était de pouvoir analyser des séries temporelles après en avoir extrait leurs distributions d'amplitudes [42]. L'introduction de ce plan fut motivée car l'information de Fisher rend possible la caractérisation du comportement non-stationnaire de signaux alors que l'entropie de Shannon peine à le faire. Ils ont étudié ce plan dans le cas où les variables aléatoires étaient continues et ils ont mis en avant l'existence d'une frontière dans ce plan, atteinte lorsque les variables aléatoires sont Gaussiennes [42]. Pour nous retrouver dans le cas continu, considérons une variable aléatoire X dont la densité de probabilité est $f_X(x)$. Bien entendu, dans le cas qui nous concerne pour ce travail, les distributions des degrés $\mathbf{p} = (p_i)_{1\leq i\leq n}$ sont discrètes. Les grandeurs informationnelles considérées par Vignat et Bercher sont l'entropie différentielle de Shannon [181]

$$H_X = -\int f_X(x) \ln\left(f_X(x)\right)\, \mathrm{d}x \tag{2.20}$$

dont la discrétisation a été rappelée plus tôt dans ce chapitre (équation (2.14)) et l'information de Fisher [203]

$$I_X = \int \left(\frac{\partial}{\partial x} f_X(x)\right)^2 \frac{\mathrm{d}x}{f_X(x)} \tag{2.21}$$

dont une discrétisation est donnée par [204]

$$F(\mathbf{p}) = \frac{1}{2}\sum_{i=1}^{n-1} \left(\sqrt{p_{i+1}} - \sqrt{p_i}\right)^2 . \tag{2.22}$$

Si ce plan est intéressant, une généralisation de celui-ci peut l'être également. Il est ainsi possible de proposer un plan informationnel de Fisher-Rényi où l'entropie de Shannon est remplacée par l'entropie de Rényi de paramètre $\alpha \in [0, \infty]$ définie par

$$R_\alpha(\mathbf{p}) = \frac{1}{1-\alpha} \ln \left(\sum_{i=1}^{n} p_i^\alpha \right) \tag{2.23}$$

En effet, lorsque α tend vers 1, l'entropie de Rényi tend vers l'entropie de Shannon. Pour $\alpha = 0$, l'entropie de Hartley $R_0(\mathbf{p}) = \ln n$ est obtenue ce qui revient à n'étudier que l'information de Fisher car les distributions considérées sont toutes de même taille. Quand $\alpha = 2$, l'équation (2.23) conduit à l'entropie de collision. Enfin, lorsque α tend vers l'infini, l'entropie min égale à $R_\infty(\mathbf{p}) = -\ln(\max_i p_i)$ est atteinte. La prochaine étape va consister à projeter des signaux réels dans le plan informationnel de Fisher-Rényi. Ces signaux n'auront pas nécessairement le même nombre de points que ceux générés pour créer les squelettes qui vont suivre. Ainsi, pour éviter que le nombre de points ne soit un problème, et ce même si nos distributions font la même taille car calculées jusqu'à un degré fixé à 100, seule l'entropie de Rényi normalisée définie par $\widehat{R}_\alpha(\mathbf{p}) = R_\alpha(\mathbf{p}) / \ln n$ est considérée. Si une généralisation de l'entropie de Shannon est recherchée, il peut être naturel de penser à l'entropie de Tsallis ou son équivalent en théorie de l'information, l'entropie de Havrda-Charvát.

Bruits colorés

Dans le but de représenter des bruits colorés simulés dans les plans informationnels susmentionnés, nous utilisons l'algorithme de Zhivomirov [199] avec une pente variant de -1 à 2.75 par pas de 0.25 où chaque signal possède 50000 échantillons. Pour ce faire, l'algorithme 1 suivant sera utilisé pour concevoir les squelettes de référence.

La figure 2.14 représente ainsi, pour quatre valeurs différentes du paramètre α (0, 1, 2 et ∞), les projections de ces bruits générés dans les plans de Fisher-Rényi. Les points affichés correspondent aux centroïdes de 10 tirages pour chaque pente. Ces figures montrent bien que la localisation précise de bruits colorés dans des plans de Fisher-Rényi est également possible, pour des paramètres α différents de 1, étendant de fait le travail de Gonçalves *et al.* [25]. Bien entendu, d'un point de vue tout à fait objectif, l'intérêt que présente l'introduction de l'entropie de Rényi est maigre, *a minima* pour la construction des squelettes. Il est tout de même à noter que c'est uniquement l'entropie de Rényi qui provoque la déformation du squelette selon les paramètres α, puisque l'information de Fisher ne dépend pas de ce dernier. Ceci est d'autant plus intéressant que l'entropie de Rényi a une propriété de décroissance monotone : $R_0(\mathbf{p}) \geq R_1(\mathbf{p}) \geq \ldots \geq R_\infty(\mathbf{p})$. Dans ces plans informationnels où

Algorithme 1 Construction d'un squelette $\mathbf{B}$ dans un plan Fisher-Rényi pour des bruits colorés

Entrée : Nombre d'échantillons n, nombre de tirages N, vecteur des L pentes de DSP $\boldsymbol{\beta} = (\beta_i)_{1\leq i\leq L}$, seuil de troncature ε et paramètre α de l'entropie de Rényi.
Sortie : Squelette $\mathbf{B}$.

for $i = 1 : L$ **do**
  $\mathbf{r} \leftarrow \mathbf{0}$
  $\mathbf{f} \leftarrow \mathbf{0}$
  for $t = 1 : N$ **do**
    $\mathbf{x} \leftarrow \text{bruit_colore}(n, \beta_i)$
    $G \leftarrow \text{NVG}(\mathbf{x})$ ▷ Équation (1.56)
    $\mathbf{p} \leftarrow \text{distribution_degres}(G, 1:1:\varepsilon)$ ▷ Équation (1.7)
    $\mathbf{r}_t \leftarrow \widehat{R}_\alpha(\mathbf{p})$ ▷ Équation (2.23)
    $\mathbf{f}_t \leftarrow F(\mathbf{p})$ ▷ Équation (2.22)
  $\mathbf{B}_{i,1} \leftarrow \text{mean}(\mathbf{r})$
  $\mathbf{B}_{i,2} \leftarrow \text{mean}(\mathbf{f})$

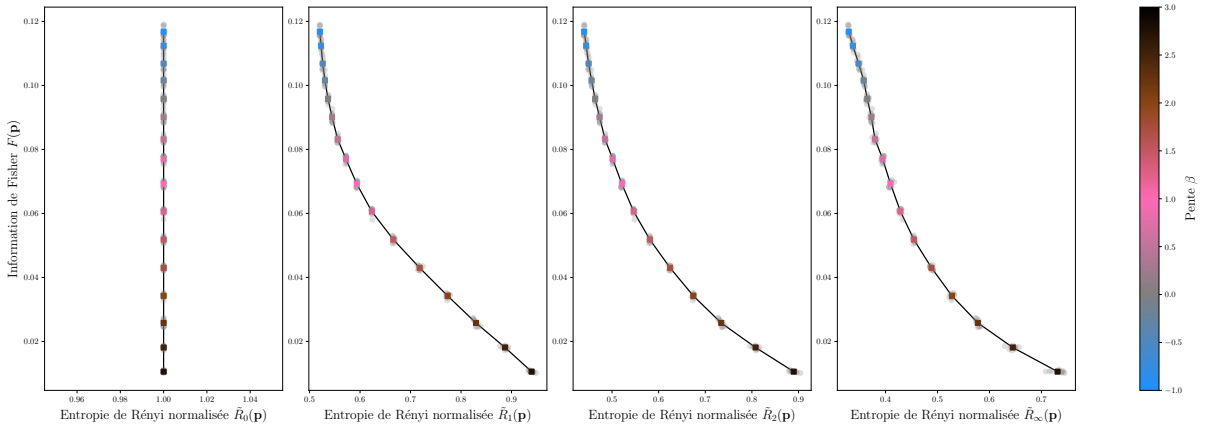


Figure 2.14 – Plans informationnels de Fisher-Rényi pour localiser des bruits colorés en utilisant la **distribution des degrés tronquée** de leurs **graphes de visibilité naturelle** associés. Les pentes β de DSP varient de -1 à 2.75 et les signaux sont constitués de 5×10^4 échantillons. La ligne continue reliant les carrés colorés (aux couleurs des bruits colorés), centroïdes de 10 tirages pour chaque pente (cercles grisés en transparence), forme un squelette de référence. Il y a quatre plans pour quatre valeurs du paramètre α de l'entropie de Rényi : 0, 1, 2 et ∞.

les squelettes sont localisés précisément, pouvoir représenter, à partir d'une même méthode de calcul, une série temporelle réelle (modélisable par un bruit coloré de pente β^*) par un point dans ces plans de Fisher-Rényi et, par une simple projection orthogonale sur le squelette de référence, en déduire une estimation de la pente β^* est donc envisageable. Des méthodes traditionnelles de traitement du signal existent bien sûr pour estimer la pente de DSP, mais il ne faut pas oublier que la DSP elle-même est issue d'estimateurs. En s'affranchissant de cette estimation, il semble alors possible d'obtenir une méthode plus robuste.

Intérêt de la troncature de la distribution des degrés du graphe de visibilité naturelle

Beaucoup de travaux récents, notamment celui de Gonçalves *et al.* [24,25], optent pour le graphe de visibilité horizontale. Outre l'argument de simplicité de construction mathématique et informatique, la raison est généralement liée au fait que cette variante met mieux en avant les informations locales du signal, alors que le graphe de visibilité naturelle est plus adapté pour des signaux dont l'étude des fluctuations doit être menée sur un temps d'observation plus long. Il parait donc intuitif que ce soit ce dernier qui soit utilisé pour l'analyse de processus stochastiques comme les mouvements Browniens ou les bruits fractionnaires. Cela peut se traduire également par le fait qu'un nombre d'échantillons très important soit requis pour faire converger la distribution des degrés vers sa forme « théorique ». C'est le cas dans les travaux de Gonçalves dans lesquels il y a nécessité d'avoir des signaux générés à plus de 10^6 échantillons pour avoir des distributions des degrés asymptotiques et donc des projections dans le plan de Fisher-Shannon interprétables. Si seuls 5×10^4 échantillons de processus stochastiques sont tirés, que les distributions des degrés tronquées des graphes de visibilité horizontale sont extraites avant d'être projetées dans les plans informationnels, les squelettes correspondants apparaissent en figure 2.15. Il est clair, grâce aux tirages des processus stochastiques représentés par les cercles grisés, que lorsque la pente devient grande, au moins supérieure à 2, alors le nombre de points des signaux générés est insuffisant pour obtenir des localisations précises.

Quant à la troncature des distributions des degrés, au-delà du fait que ce soit une intuition par rapport à la figure 2.13, elle permet réellement d'avoir des squelettes plus précis dans les plans informationnels. La figure 2.16 montre que, lorsque sont considérées les distributions des degrés entières (*i.e.* calculées pour des degrés allant de 1 à 5×10^4) des graphes de visibilité naturelle, alors les squelettes sont moins bien définis que sur la figure 2.14 pour laquelle les distributions étaient tronquées.

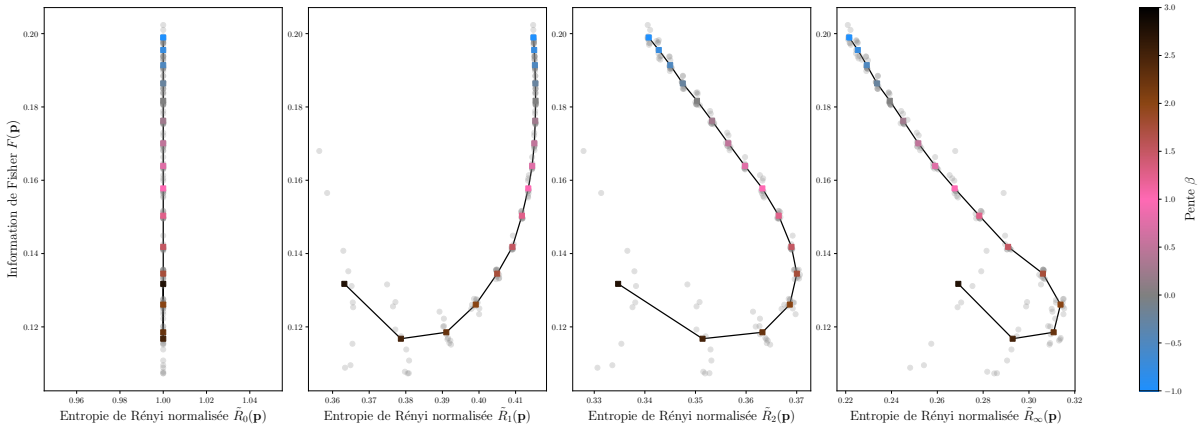


Figure 2.15 – Même dispositif expérimental que pour la figure 2.14 où ce sont les **distributions des degrés tronquées** des **graphes de visibilité horizontale** qui sont cette fois extraites.

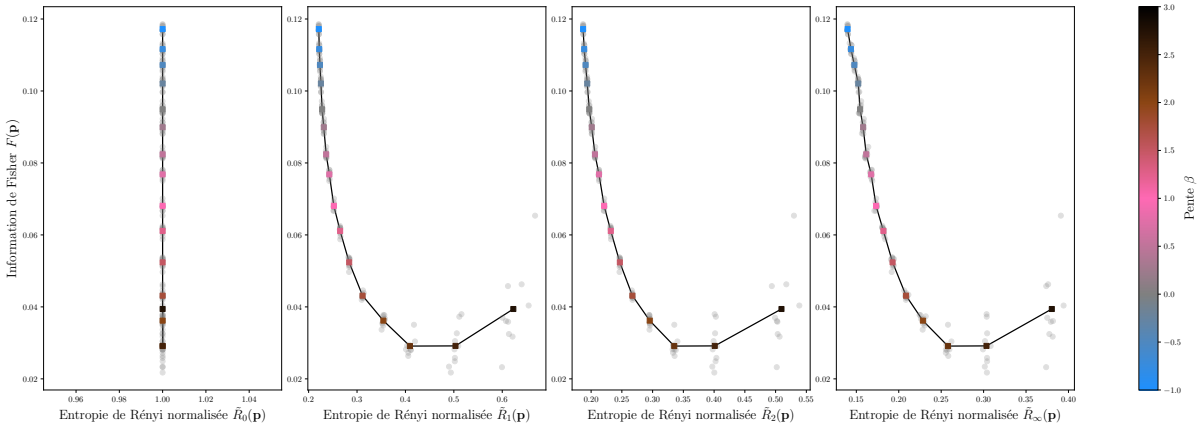


Figure 2.16 – Même dispositif expérimental que pour la figure 2.14 où ce sont les **distributions des degrés entières** des **graphes de visibilité naturelle** qui sont cette fois extraites.

Mouvements Browniens et bruits Gaussiens fractionnaires

Dans le but de représenter des mouvements Browniens fractionnaires dans les plans informationnels de Fisher-Rényi, nous utilisons la librairie `FBM` de Python, implémentant la méthode de Davies et Harte précédemment mentionnée, avec un coefficient de Hurst variant de 0.125 à 0.875 par pas de 0.125 où chaque série temporelle possède 50000 échantillons. Cet intervalle de coefficient de Hurst, non conventionnel, sert à se comparer aux squelettes de la sous-section précédente, car il correspond en réalité à $H = (\beta - 1)/2$ pour un β variant de 1.25 à 2.75 par pas de 0.25. Pour concevoir les squelettes de référence, l'algorithme 2 sera utilisé.

La figure 2.17 représente ainsi, pour quatre valeurs différentes du paramètre α (0, 1, 2 et ∞), les

Algorithme 2 Construction d'un squelette $\mathbf{B}$ dans un plan Fisher-Rényi pour des fBm

Entrée : Nombre d'échantillons n, nombre de tirages N, vecteur des L coefficients de Hurst $\mathbf{h} = (h_i)_{1 \leq i \leq L}$, seuil de troncature ε et paramètre α de l'entropie de Rényi.
Sortie : Squelette $\mathbf{B}$.

for $i = 1 : L$ **do**
  $\mathbf{r} \leftarrow \mathbf{0}$
  $\mathbf{f} \leftarrow \mathbf{0}$
  for $t = 1 : N$ **do**
    $\mathbf{x} \leftarrow \mathrm{FBM}(n, h_i)$
    $G \leftarrow \mathrm{NVG}(\mathbf{x})$ ▷ Équation (1.56)
    $\mathbf{p} \leftarrow \mathrm{distribution_degres}(G, 1:1:\varepsilon)$ ▷ Équation (1.7)
    $\mathbf{r}_t \leftarrow \widehat{R}_\alpha(\mathbf{p})$ ▷ Équation (2.23)
    $\mathbf{f}_t \leftarrow F(\mathbf{p})$ ▷ Équation (2.22)
  $\mathbf{B}_{i,1} \leftarrow \mathrm{mean}(\mathbf{r})$
  $\mathbf{B}_{i,2} \leftarrow \mathrm{mean}(\mathbf{f})$

projections de 10 tirages de ces processus stochastiques dans les plans de Fisher-Rényi (représentés par des points grisés sur les figures). Les carrés affichés correspondent aux centroïdes de ces 10 tirages pour chaque valeur de H. Reliés, ils forment le squelette de référence pour les mouvements Browniens fractionnaires. Les squelettes obtenus en figure 2.17 sont évidemment très similaires à ceux de la figure 2.14 puisqu'il a été rappelé dans la sous-section 2.4.1 qu'un mouvement Brownien fractionnaire de coefficient de Hurst H pouvait être approché par un bruit coloré dont la pente de DSP vaut

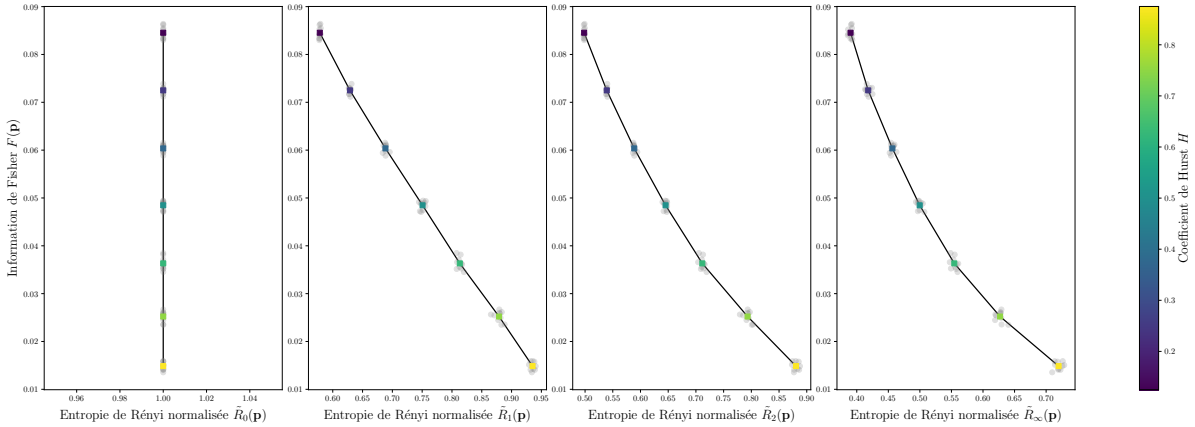


Figure 2.17 – Plans informationnels de Fisher-Rényi pour localiser des mouvements Browniens fractionnaires en utilisant la **distribution des degrés tronquée** de leurs **graphes de visibilité naturelle** associés. Les coefficients de Hurst H varient de 0.1 à 0.9. et les signaux sont constitués de 5×10^4 échantillons. La ligne continue reliant les carrés noirs, centroïdes de 10 tirages pour chaque valeur de H, forme un squelette de référence. Il y a quatre plans pour quatre valeurs du paramètre α de l'entropie de Rényi : 0, 1, 2 et ∞.

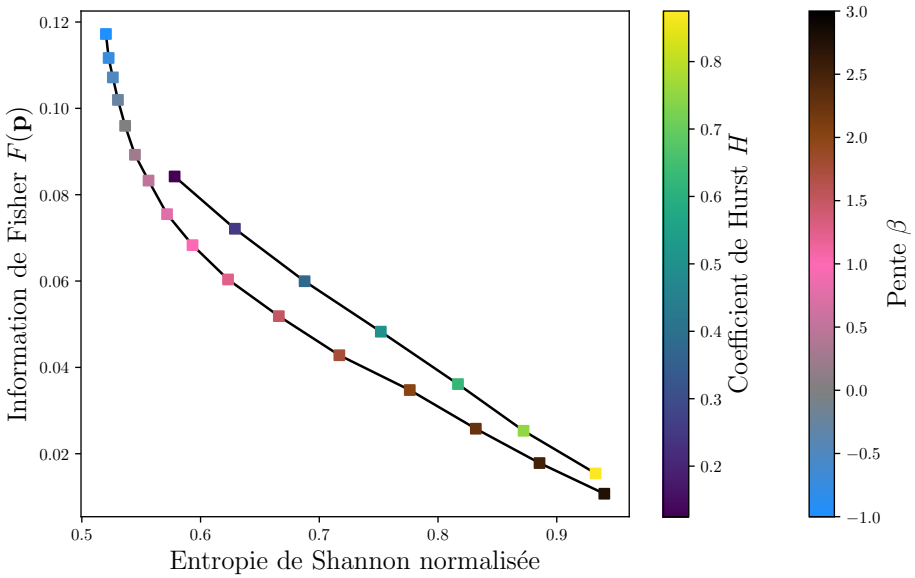


Figure 2.18 – Superposition des squelettes de la figure 2.14 et de la figure 2.17 pour $\alpha = 1$, *i.e.* pour l'entropie de Shannon.

$\beta = 2H + 1$. Aussi, en superposant en figure 2.18 les squelettes calculés pour l'entropie de Shannon (*i.e.* un paramètre α égal à 1), il est visible que, numériquement, les bruits colorés ne sont pas strictement équivalents à des mouvements Browniens fractionnaires. Une fois de plus, l'intérêt du graphe de visibilité naturelle plutôt que sa variante horizontale ainsi que le fait de tronquer la distribution des degrés est renforcé par l'observation de la figure 2.19 et de la figure 2.20 qui présentent des squelettes bien moins précis dans les plans Fisher-Rényi. Les figures correspondantes aux squelettes construits avec les distributions entières des degrés des graphes de visibilité horizontale sont volontairement omises ici car elles font l'objet d'une étude approfondie dans le travail de Gonçalves [25]. Dans ce cas, pour avoir un squelette précis, il est nécessaire d'avoir des séries temporelles d'au moins 10^6 échantillons.

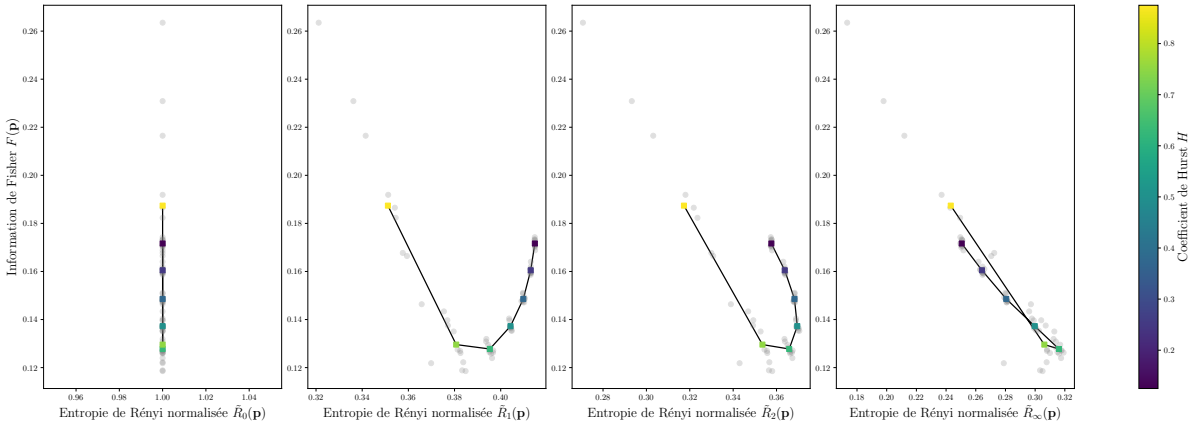


Figure 2.19 – Même dispositif expérimental que pour la figure 2.17 où ce sont les **distributions des degrés tronquées** des **graphes de visibilité horizontale** qui sont cette fois extraites.

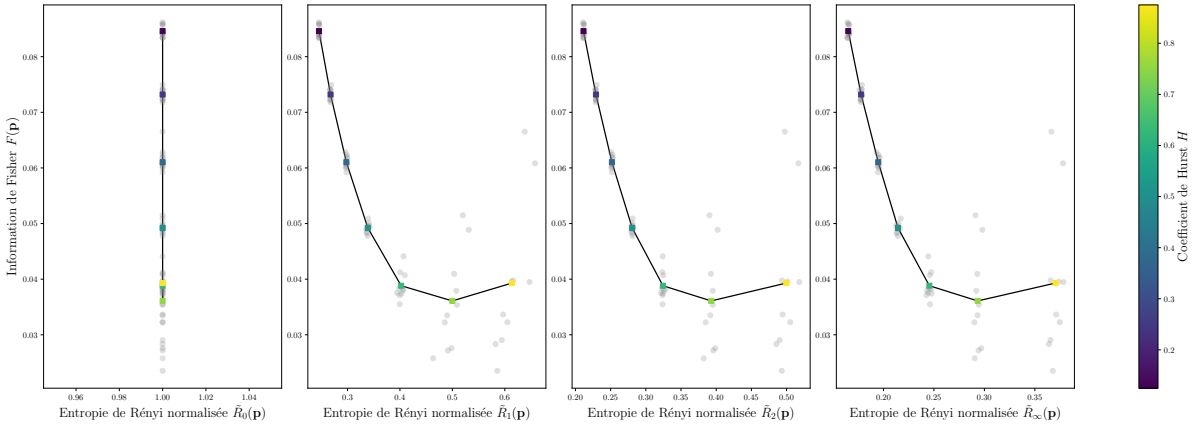


Figure 2.20 – Même dispositif expérimental que pour la figure 2.17 où ce sont les **distributions des degrés entières** des **graphes de visibilité naturelle** qui sont cette fois extraites.

2.4.4 Estimation du coefficient de Hurst

La sous-section précédente a montré qu'une localisation précise, aussi bien de bruits colorés que de mouvements Browniens fractionnaires, était possible dans les plans informationnels de Fisher-Rényi. Pour que les figures et les analyses de cette sous-section soient bien perçues, nous allons nous concentrer sur la projection des mouvements Browniens fractionnaires, dont la simulation est exacte grâce à la méthode de Davies et Harte, dans le plan Fisher-Shannon.

Dans le but de caractériser l'évolution du comportement fractal dans des séries temporelles, Gonçalves *et al.* ont délaissé la distribution des degrés, à juste titre dans le cadre de leur travail, au profit de la distribution des écarts verticaux (*cf.* tableau 1.4), dont le calcul relève toutefois d'un choix important de l'utilisateur. La localisation dans le plan Fisher-Shannon de cette distribution leur a permis de détecter d'un point de vue qualitatif des changements dans le phénomène climatique « El Niño ». Notre travail se veut être une analyse quantitative : nous allons utiliser la localisation de signaux réels, modélisables par des processus stochastiques type fBm ou fGn, de concert avec le squelette dûment construit dans le plan Fisher-Shannon, pour extraire une estimation du coefficient de Hurst. Pour juger de la pertinence de notre estimateur, nous allons le comparer à d'autres, issus du traitement de signal, et dont un rappel succinct est effectué ci-après.

Estimateurs existants du coefficient de Hurst

Outre la méthode présentée précédemment basée sur l'estimation du paramètre de la loi de puissance que suit la distribution des degrés d'un graphe de visibilité naturelle (illustrée par la figure 2.12), il existe bien d'autres méthodes permettant d'estimer le coefficient de Hurst d'un mouvement Brownien fractionnaire. Pour en rappeler les rudiments, nous allons nous baser sur les notations de l'article de Serinaldi [193]. La méthode **PSD** désigne l'utilisation de la densité spectrale de puissance de la série temporelle $\mathbf{x}$ [192,193]. Comme évoqué au début de cette section, les fBm et fGn peuvent être approchés par des bruits colorés, c'est-à-dire des signaux dont la densité spectrale de puissance suit une loi de puissance de paramètre β. Ainsi, un fGn d'exposant de Hurst H peut être modélisé par un bruit coloré avec $\beta \in [-1, 1]$ (la relation entre β et H étant $\beta = 2H - 1$) alors qu'un fBm peut être modélisé par un bruit coloré avec $\beta \in [1, 3]$ (la relation entre β et H étant $\beta = 2H + 1$). La méthode **WAV** [205] désigne la méthode basée sur la transformée en ondelettes. En effet, Simonsen *et al.* ont montré que la variance de l'ondelette $W_{a,b}$ suit une loi de puissance a^β (les mêmes relations susmentionnées entre H et β peuvent alors être utilisées). La méthode **R/S** [206] fait référence à la méthode « *rescaled-range analysis* ». La série temporelle $\mathbf{x} = (x_i)_{1 \leq i \leq N}$ de N échantillons et de moyenne $\bar{x}$ est centrée : $\mathbf{y} = \mathbf{x} - \bar{x}$. La série $\mathbf{z}$ est alors construite en effectuant la somme cumulée des éléments de $\mathbf{y}$, c'est-à-dire que pour tout $n = 1, 2, \ldots, N$, on a $z_n = \sum_{i=1}^{n} y_i$. Soient les séries temporelles

suivantes : une série temporelle de variances

$$s_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - m_n)^2}, \qquad m_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad n = 1, 2, \ldots, N$$

et une série d'étendues

$$r_n = \max_{1\leq i\leq n} z_i - \min_{1\leq i\leq n} z_i, \qquad n = 1, 2, \ldots, N.$$

Alors, il peut être prouvé que

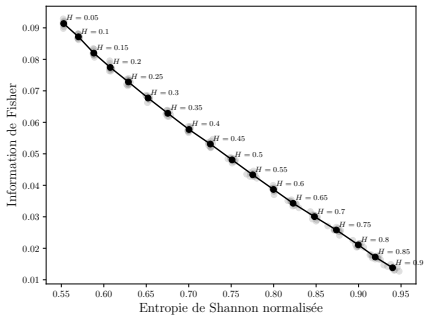
$$\mathbb{E}\left[\frac{r_n}{s_n}\right] \propto n^H$$

lorsque n tend vers l'infini. Enfin, l'analyse des fluctuations redressées (**DFA** pour *Detrentred Fluctuation Analysis* en anglais) [207] divise la somme cumulée de la série temporelle centrée (la même série z_n que pour la méthode R/S) en segments consécutifs de longueur n sur lesquels des régressions linéaires sont effectuées. La moyenne des erreurs quadratiques moyennes des régressions, c'est-à-dire des différences entre les échantillons du segment et leurs approximations linéaires associées, notée $\sigma(n)$, suit une loi de puissance n^α où α est un équivalent de l'exposant de Hurst. Avant d'appliquer ces méthodes à des signaux réels, le paragraphe suivant est dédié à la présentation de notre stratégie d'estimation du coefficient de Hurst.

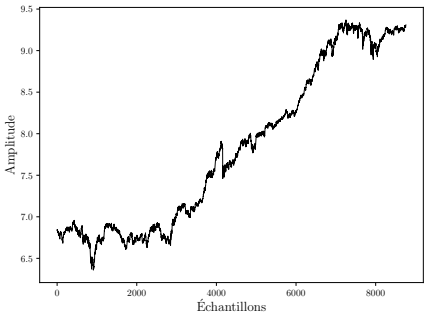
Application sur des signaux réels

Nous proposons, dans cette partie, une méthode d'estimation du coefficient de Hurst basée sur le travail qui a été mené jusqu'alors, à savoir l'utilisation de squelettes dans un plan informationnel (ici, celui de Fisher-Shannon) construits à partir des distributions des degrés tronquées des graphes de visibilité naturelle de signaux générés. Le principe est simple : nous appliquons le même traitement à une série temporelle réelle qui est ainsi représentée par un point (S, F) dans le plan informationnel considéré. Ce point est projeté orthogonalement sur le squelette de référence pour donner un point $(S_{\text{proj}}, F_{\text{proj}})$ et, par interpolation, il est possible d'obtenir H^*, estimation de l'exposant de Hurst. Cette méthode présente l'avantage d'être très efficiente. En effet, si le squelette de référence est déjà construit (avec autant de points que souhaité pour obtenir une précision satisfaisante), il ne reste qu'à l'importer. De plus, la mémoire que requiert le stockage du squelette est insignifiante puisqu'il est possible de ne stocker que les centroïdes. La méthodologie est illustrée par la figure 2.21. La partie projection et estimation par interpolation fait l'objet de l'algorithme 3.

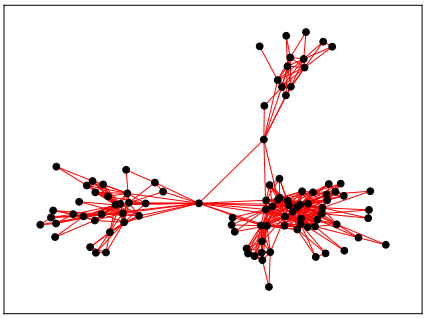
① Importer (ou construire) le squelette de référence dans le plan informationnel



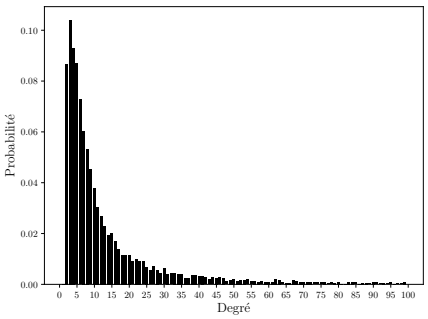
② Lire la série temporelle $\mathbf{x}$ à étudier



③ Construire le graphe de visibilité naturelle G associé à $\mathbf{x}$

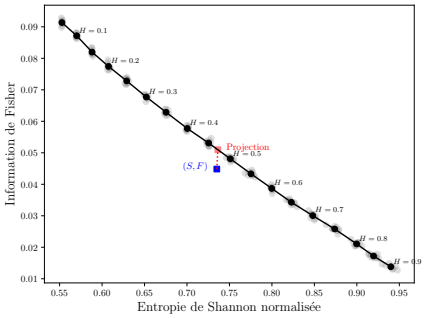


④ Extraire la distribution des degrés tronquée $\mathbf{p}$ du graphe G



⑤ Calculer le point (S, F) où F est l'information de Fisher et S est l'entropie de Shannon normalisée de la distribution de degrés tronquée $\mathbf{p}$

⑥ Obtenir la projection orthogonale[a] $(S_{\text{proj}}, F_{\text{proj}})$ de (S, F) sur le squelette



a. La projection ne semble pas orthogonale car les axes ne sont pas à la même échelle

Figure 2.21 – Diagramme illustrant notre stratégie d'estimation du coefficient de Hurst.

Algorithme 3 Estimation du coefficient de Hurst du processus modélisant une série temporelle $\mathbf{x}$

Entrée : Squelette $\mathbf{B} = \{(S_i^c, F_i^c)\}_{1 \leq i \leq L}$ contenant les L centroïdes dans le plan informationnel de Fisher-Shannon correspondant aux coefficients de Hurst $(h_i)_{1 \leq i \leq L}$ et seuil de troncature ε
Sortie : Coefficient de Hurst H estimé.

1: **if** $\mathbf{x}$ est modélisable par un fGn **then**
2: $\quad \mathbf{x} \leftarrow \mathrm{cumsum}(\mathbf{x})$
3: $G \leftarrow \mathrm{NVG}(\mathbf{x})$ ▷ Équation (1.56)
4: $\mathbf{p} \leftarrow \mathrm{distribution_degres}(G, 1:1:\varepsilon)$ ▷ Équation (1.7)
5: $S \leftarrow S(\mathbf{p})$ ▷ Équation (2.14)
6: $F \leftarrow F(\mathbf{p})$ ▷ Équation (2.22)
7: Trouver les deux centroïdes (S_j^c, F_j^c) et (S_k^c, F_k^c) (dont les coefficients de Hurst respectifs sont h_j et h_k) les plus proches du point (S, F).
8: $X \leftarrow \dfrac{(S_k^c - S_j^c)(S - S_j^c) + (F_k^c - F_j^c)(F - F_j^c)}{(S_k^c - S_j^c)^2 + (F_k^c - F_j^c)^2}$
9: $S_{\mathrm{proj}} \leftarrow S_j^c + X(S_k^c - S_j^c)$
10: $F_{\mathrm{proj}} \leftarrow F_j^c + X(F_k^c - F_j^c)$
11: $d \leftarrow \sqrt{(S_j^c - S_{\mathrm{proj}})^2 + (F_j^c - F_{\mathrm{proj}})^2}$
12: $d_{\mathrm{tot}} \leftarrow \sqrt{(S_j^c - S_k^c)^2 + (F_j^c - F_k^c)^2}$
13: $H \leftarrow h_j + (h_k - h_j)\dfrac{d}{d_{\mathrm{tot}}}$

Dans le but de tester notre méthode sur des séries temporelles réelles, nous l'appliquons sur celles décrites, listées et étudiées par Serinaldi qui, dans son article précédemment mentionné, présente les résultats des estimateurs classiques du coefficient de Hurst calculées sur ces dernières. Parmi elles, les observations quotidiennes des prix à terme provenant du NYMEX pour trois produits énergétiques : le pétrole brut, le fioul de chauffage (tous deux observés du 2 juillet 1990 au 1er novembre 2006) et le gaz naturel (du 3 janvier 1994 au 31 août 2009). Des observations quotidiennes de deux indices boursiers américains seront également utilisées : le Dow Jones Industrial Average (DJI) et le New-York Stock Exchange (NYSE), tous deux établis du 5 février 1971 au 1er décembre 2006. Ces séries temporelles issues de marchés financiers sont de type fBm et leurs exposants de Hurst se situent aux alentours de $H \approx 0.5$ [193, 208]. Pour montrer que notre stratégie fonctionne également sur des signaux pouvant être modélisés par un processus fGn, les prix horaires du marché de l'électricité de l'Alberta (couvrant la période du 1er janvier 2000 au 30 juin 2008) sont pris en compte [209]. Cette série temporelle est en effet régie par un processus fGn avec un exposant de Hurst se situant autour de $H \approx 0.9$. Serinaldi l'explique en écrivant que « puisque les prix horaires au comptant sont caractérisés par de fortes périodicités quotidiennes, hebdomadaires et mensuelles, l'analyse des données horaires permet

de mettre en évidence l'impact de la dynamique cyclique sur les estimations de H » et apporte une analyse sur l'anti-persistance des marchés de l'électricité [193]. Pour estimer l'exposant de Hurst de cette série temporelle en utilisant notre stratégie, il est nécessaire de la centrer et de l'intégrer (en calculer la somme cumulée) afin d'obtenir un signal de type fBm. Grâce à l'approche présentée dans cette partie, toutes ces séries temporelles sont plongées dans le plan informationnel de Fisher-Shannon dans lequel a été construit un squelette plus précis à l'aide de tirages de mouvements Browniens fractionnaires de 100 000 points pour des coefficients de Hurst variant de 0.05 à 0.9 par pas de 0.05. Une visualisation est proposée en figure 2.22. À partir de ces projections, il est alors possible d'estimer les coefficients de Hurst des processus sous-jacents.

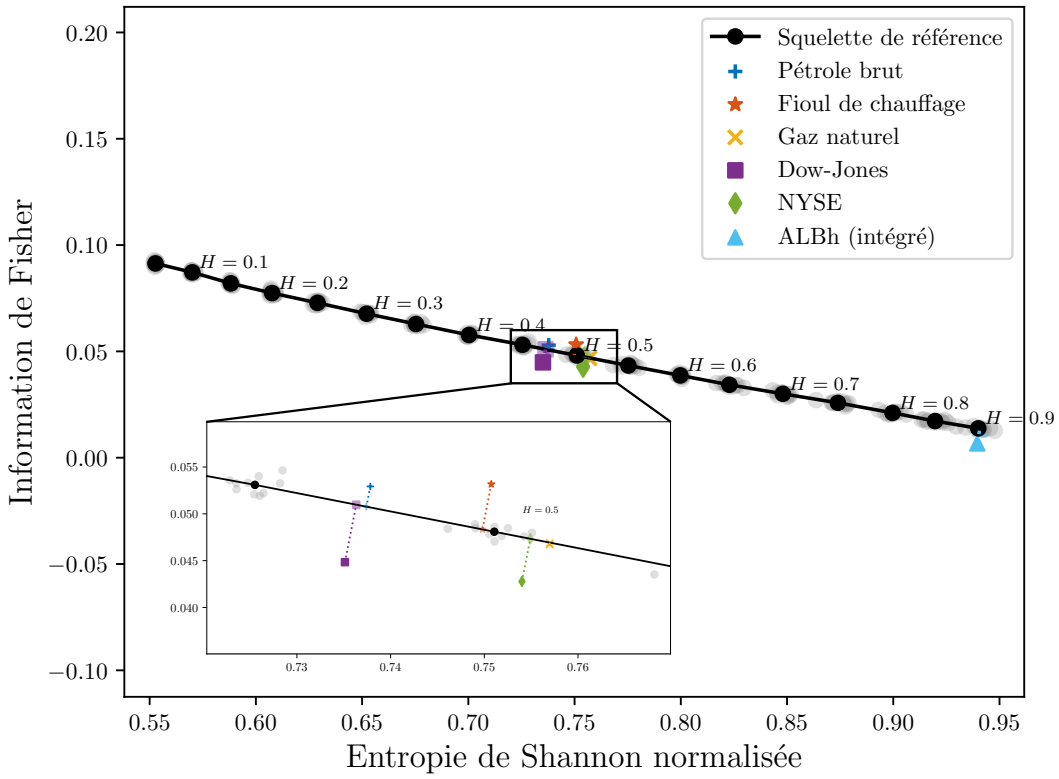


Figure 2.22 – Dans le plan informationnel de Fisher-Shannon, squelette de référence constitué des centroïdes (points noirs) de 10 tirages de mouvements Browniens fractionnaires de 100 000 points (points gris) avec un coefficient de Hurst H variant de 0.05 à 0.9. Les plongements des séries temporelles réelles sont représentés par des marqueurs colorés pour lesquels les projections orthogonales sur le squelette sont affichées en légère transparence.

Les résultats de notre estimateur, **VG/FS**, ainsi que les quatre méthodes classiques de la littérature sont listés dans le tableau 2.6. Bien que la vérité terrain n'existe pas dans ce cas d'étude, une comparaison avec les autres estimateurs est toujours possible. Ce faisant, il est clair que les valeurs estimées de H par la méthode **VG/FS** introduites dans ce travail sont cohérentes avec celles issues des méthodes **PSD**,

WAV, **R/S** et **DFA**. Comme cela a déjà été mentionné, il n'existe pas de méthode « étalon » à laquelle comparer nos estimations. Ainsi, pour fournir une analyse qualitative des différentes estimations, le graphique en figure 2.23 est proposé. Ce dernier montre, parmi tous les estimateurs traditionnels, l'estimation minimale, maximale et moyenne du coefficient de Hurst H. Les estimations issues de notre méthode **VG/FS** sont représentées avec les mêmes marqueurs que ceux de la figure 2.22. Pour la série temporelle représentant le cours du pétrole brut, notre méthode surestime le paramètre H par rapport aux méthodes classiques, tandis que pour la série temporelle du Dow Jones, elle le sous-estime. Pour tous les autres signaux, nos estimations sont proches de la moyenne, étant presque identiques dans le cas des séries temporelles du NYSE et du gaz naturel. Ces résultats permettent d'apporter un élément supplémentaire en faveur de l'utilisation conjointe d'un plan informationnel et du graphe de visibilité pour caractériser la persistance d'une série temporelle, fonction de son exposant de Hurst.

Base de données / Méthodes	**VG/FS**	**PSD**	**WAV**	**R/S**	**DFA**
Pétrole brut (fBm)	0.472	0.423	0.452	0.498	0.466
Fioul de chauffage (fBm)	0.498	0.441	0.443	0.423	0.458
Gaz naturel (fBm)	0.512	0.433	0.465	0.654	0.481
Dow-Jones (fBm)	0.470	0.491	0.480	0.554	0.481
NYSE (fBm)	0.508	0.491	0.487	0.545	0.505
ALBh (fGn)	0.899	0.923	0.829	0.838	0.855

Tableau 2.6 – Estimations du coefficient de Hurst avec notre méthode **VG/FS** ainsi qu'avec d'autres estimateurs classiques du traitement de signal. Les résultats de ces estimateurs classiques sont tirés de l'article de Serinaldi [193].

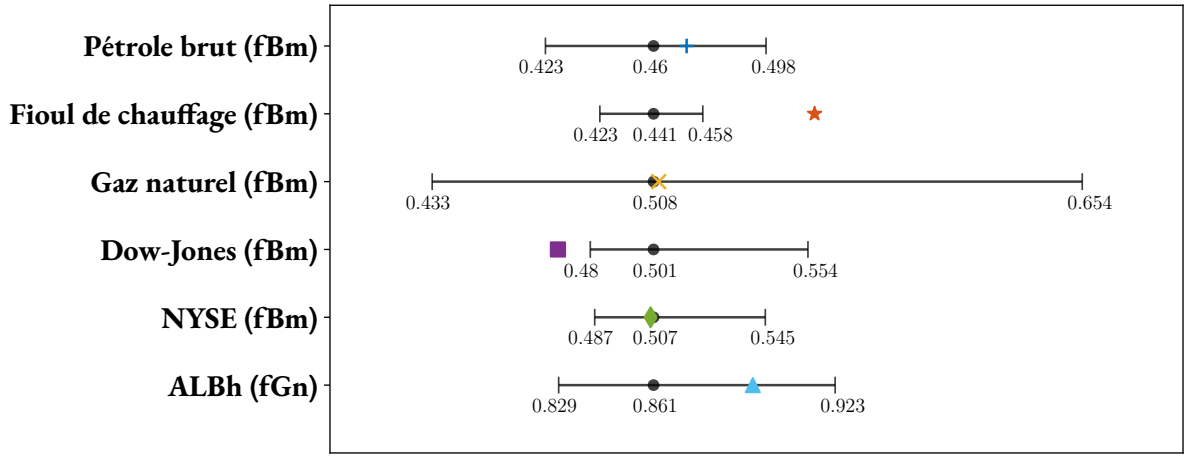


Figure 2.23 – Minimums, maximums et moyennes des estimations du coefficient de Hurst (stockées dans le tableau 2.6). Nos estimations (par la méthode **VG/FS**) sont représentées par des marqueurs colorés (ceux de la figure 2.22).

Erreurs d'estimation sur des données synthétiques

Bien que la méthode **VG/FS** proposée dans ce manuscrit soit basée sur l'utilisation de mouvements Browniens fractionnaires synthétiques représentés dans le plan informationnel de Fisher-Shannon, formant alors un squelette de référence, il est important de quantifier *a posteriori*, *i.e.* après la construction de ce squelette, l'erreur que notre estimateur commet sur d'autres signaux synthétiques. Pour en effectuer une analyse quantitative, considérons 20 tirages de fBm, ayant un nombre croissant d'échantillons et étant régis par des coefficients de Hurst H variant entre 0.1 à 0.9. Chacune de ces séries temporelles synthétiques est introduite dans notre estimateur **VG/FS** et, pour chaque valeur H cible, la moyenne et l'écart-type des 20 estimations sont calculés. Nous procédons ainsi pour des fBm composés de 1 000, 2 500, 5 000, 10 000, 25 000 et 50 000 échantillons. Les résultats sont présentés en tableau 2.7 (pour plus de lisibilité, les valeurs de H affichées le sont par pas de 0.1). Ainsi, le nombre

		Nombre d'échantillons dans les séries temporelles synthétiques					
		1 000	**2 500**	**5 000**	**10 000**	**25 000**	**50 000**
Coefficient de Hurst	$H = 0.1$	0.13 ± 0.02	0.12 ± 0.01	0.11 ± 0.01	0.10 ± 0.00	0.10 ± 0.00	0.10 ± 0.00
	$H = 0.2$	0.16 ± 0.02	0.17 ± 0.01	0.19 ± 0.01	0.19 ± 0.01	0.20 ± 0.00	0.20 ± 0.00
	$H = 0.3$	0.25 ± 0.02	0.28 ± 0.02	0.29 ± 0.01	0.29 ± 0.01	0.30 ± 0.01	0.30 ± 0.00
	$H = 0.4$	0.35 ± 0.03	0.39 ± 0.02	0.39 ± 0.02	0.40 ± 0.01	0.40 ± 0.01	0.40 ± 0.01
	$H = 0.5$	0.45 ± 0.04	0.49 ± 0.03	0.50 ± 0.01	0.51 ± 0.01	0.50 ± 0.01	0.50 ± 0.01
	$H = 0.6$	0.56 ± 0.04	0.60 ± 0.03	0.60 ± 0.02	0.60 ± 0.01	0.60 ± 0.01	0.60 ± 0.01
	$H = 0.7$	0.67 ± 0.05	0.71 ± 0.03	0.72 ± 0.02	0.71 ± 0.02	0.71 ± 0.01	0.70 ± 0.01
	$H = 0.8$	0.77 ± 0.04	0.81 ± 0.03	0.82 ± 0.02	0.81 ± 0.02	0.81 ± 0.02	0.79 ± 0.01
	$H = 0.9$	0.85 ± 0.06	0.87 ± 0.02	0.87 ± 0.02	0.89 ± 0.01	0.89 ± 0.01	0.89 ± 0.01

Tableau 2.7 – Moyennes et écarts types des estimations données par la méthode **VG/FS** selon différentes valeurs cibles de l'exposant de Hurst H utilisées pour générer des fBm synthétiques. Le nombre d'échantillons dans ces séries temporelles synthétiques varie de 1 000 à 50 000.

d'échantillons a une influence évidente sur la qualité de l'estimation. Pour un nombre de points égal à 50 000, l'estimation moyenne est presque toujours égale à la valeur cible H avec un écart-type très faible. Mais la méthode donne également de bons résultats avec des séries temporelles plus courtes. Si nous considérons les fBm tirés avec 2 500 échantillons, les moyennes ne s'éloignent pas trop des valeurs cibles. Une analyse globale permettant de constater l'effet du nombre d'échantillons dans les séries temporelles sur l'erreur absolue moyenne (MAE) de l'estimation de H est proposée en figure 2.24. Ainsi, la MAE décroît drastiquement à mesure que les séries temporelles considérées s'allongent. En fait, elle tombe en dessous de 0.01 lorsque les signaux dépassent 20 000 échantillons. De plus, la MAE de 0.025 pour des signaux de 2 500 échantillons est déjà relativement correcte, comparée aux différences qui peuvent exister entre l'estimateur minimum et maximum parmi les estimateurs traditionnels proposés dans tableau 2.6 et illustrés dans figure 2.23, qui sont dans tous les cas supérieures à 0.03.

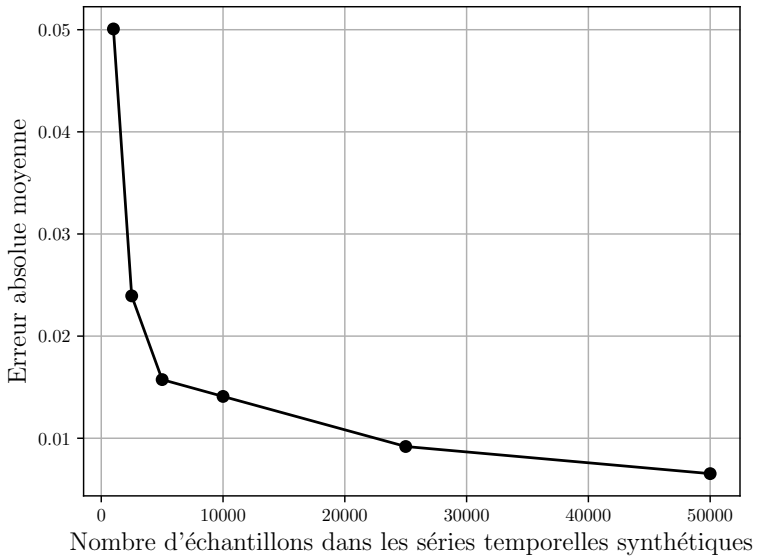


Figure 2.24 – Erreur absolue moyenne de l'estimateur **VG/FS** sur des fBm synthétiques (20 tirages pour des coefficients de Hurst variant de 0.1 à 0.9, *i.e.* 340 séries temporelles) en fonction du nombre d'échantillons dans les tirages.

2.5 Conclusion

Dans ce chapitre, nous avons montré les possibilités apportées par les graphes de visibilité (qu'elle soit naturelle ou horizontale) dans leur capacité à classifier des séries temporelles et même à les caractériser. Cet algorithme de visibilité, développé par Lacasa *et al.* [21, 23], permet de construire un graphe à partir d'un signal. L'avantage que présente cette méthode est sa simple interprétation géométrique et sa complexité algorithmique contenue en $O(n \log n)$, où n désigne le nombre d'échantillons du signal, c'est-à-dire autant qu'une transformée de Fourier rapide (FFT pour *Fast Fourier Transform*). Un intérêt de représenter des séries temporelles sous la forme de graphes est de pouvoir disposer d'une quantité considérable d'outils de la théorie des graphes comme des attributs structurels, spectraux, ou encore d'éléments provenant des différentes matrices de représentation.

Initialement, ce travail repose sur des tâches étudiant la détection d'épilepsie dans des signaux EEG en considérant ces derniers à travers leurs graphes de visibilité [37–39, 158–160]. Ces travaux étant exploratoires et basés sur un nombre réduit d'attributs extraits des graphes, nous pensions qu'extraire plus d'attributs ou encore comparer, à l'aide de distances statistiques, les distributions de degrés des graphes, permettrait d'obtenir de meilleurs résultats de classification. Ainsi, nous avons développé deux méthodes. La première consiste à appliquer 26 attributs extraits de chaque graphe de visibilité horizontale (et donc de chaque signal) en entrée d'un SVM à l'image d'une classification classique de

séries temporelles. La deuxième repose sur la construction d'un noyau de SVM adapté dans la mesure où la distance euclidienne du traditionnel noyau RBF est remplacée par les distances statistiques précédemment évoquées. Nos deux stratégies présentent des résultats performants comparativement aux autres méthodes de la littérature, la distance statistique d'Hellinger se distinguant plus particulièrement. Attention toutefois à ne pas en faire une généralisation. En effet, cette méthode est basée sur la comparaison des distributions de degrés des graphes de visibilité horizontale : il est tout à fait possible qu'avec un graphe de visibilité naturelle, une autre distance se distingue. Les stratégies évoquées ont, dans un premier temps, été appliquées pour la détection d'épilepsie. Nous avons cherché un autre cas applicatif pour les tester. Ainsi, dans le cadre d'une détection d'anomalies magnétiques, la méthode basée sur l'extraction de 26 attributs s'est révélée être la plus performante de toutes, surtout en termes de sensibilité surpassant de près de 33 % les autres stratégies. Le noyau adapté permet d'obtenir d'excellents résultats en termes de spécificité avec, de surcroît, une distance statistique optimale qui diffère de la détection d'épilepsie. En effet, dans ce cas, c'est la distance de Bhattacharyya qui est la meilleure en termes d'exactitude et d'AUC. Une perspective pourrait être de comparer les performances à d'autres méthodes de classification telles que l'utilisation de réseaux de neurones récurrents ou encore de SNN (pour *Spiking Neural Network*). De plus, une grande variété de pondérations et de variantes aux graphes de visibilité sont possibles, d'où l'existence probable d'une configuration adaptée à chaque problème.

Ces graphes de visibilité renferment un nombre important d'informations relatives aux séries temporelles initiales. Nous nous sommes alors attachés à caractériser des processus stochastiques à partir de ce type de graphe. En effet, les distributions de degrés $\mathbf{p} = (p_k)_{1 \leq k \leq n}$ des graphes de visibilité, naturelle cette fois-ci, construits à partir de mouvements Browniens fractionnaires (fBm), caractérisés par un exposant de Hurst H compris entre 0 et 1, suivent une loi en puissance $p_k \sim k^{-\gamma}$ où $\gamma(H) = 3 - 2H$. Des propriétés similaires existent pour les bruits Gaussiens fractionnaires (fGn), incréments des fBm, ou encore pour les bruits colorés, séries temporelles dont la densité spectrale de puissance suit une loi en puissance dont la pente est égale à un coefficient β (une relation existant avec l'exposant de Hurst [192]). Ce travail est basé sur ces propriétés et sur les résultats de Gonçalves *et al.* [24, 25] qui localisent les distributions de degrés des graphes de visibilité horizontale dans le plan informationnel Fisher-Shannon, introduit récemment par Vignat et Bercher [42]. En effet, les distributions de degrés permettent de caractériser les processus stochastiques initiaux mais il est plus aisé de manipuler ces dernières à travers deux quantificateurs de la théorie de l'information : l'entropie de Shannon qui mesure une information globale et l'information de Fisher qui mesure une information locale. Nous avons commencé par montrer que ces localisations étaient également possibles avec un quantificateur généralisant l'entropie de Shannon, à savoir l'entropie de Rényi. Néanmoins, pour

y aboutir, les processus stochastiques nécessitent beaucoup d'échantillons afin que les localisations soient précises. Nous avons montré qu'il est possible d'arriver au même résultat avec le graphe de visibilité naturelle et non plus horizontale et en tronquant les distributions des degrés. Tout ce travail donne lieu à la création, à l'aide de signaux synthétiques, d'un squelette de référence dans le plan informationnel, selon différentes valeurs, soit de la pente β des bruits colorés, soit des coefficients de Hurst H. Ainsi, nous avons développé une méthode d'estimation du coefficient H. Les résultats obtenus sur quelques signaux réels, principalement issus du monde financier dont nous savons leurs modélisations possibles par des fBm ou des fGn, sont cohérents avec les estimateurs plus classiques du traitement de signal, ce qui appuie la pertinence de cette méthode.

Publications scientifiques relatives à ce chapitre

- ✍ **T. Averty**, D. Daré-Emzivat et A.-O. Boudraa. Détection d'épilepsie dans les signaux EEG par graphe de visibilité et un noyau de SVM adapté. *GRETSI*, pages 1–4, 2022
- ✍ **T. Averty**, A.-O. Boudraa et D. Daré-Emzivat. Hurst exponent estimation using natural visibility graph embedding in Fisher–Shannon plane. *Signal Processing*, 230 :109884, 2025

Chapitre 3

Vulnérabilité informationnelle d'un graphe

« Quiconque est un jour la cible d'une rumeur devient ensuite vulnérable à toutes les autres. »

Jacques Attali

3.1 Introduction

3.1.1 Mise en contexte

Aujourd'hui, nos sociétés dépendent de plus en plus d'infrastructures telles que les réseaux d'électricité, d'eau ou encore de gaz, de systèmes de transport, et bien d'autres. De tels systèmes font fonctionner divers agents (physiques ou numériques) en collaboration pour faciliter les interactions sociales ou répondre à des besoins essentiels. Par exemple, un approvisionnement stable en électricité est nécessaire pour bon nombre d'infrastructures afin qu'elles puissent opérer normalement [5–7]. Si l'apport en électricité venait à s'arrêter brutalement, il est possible que les systèmes cessent de fonctionner ou, *a minima*, qu'ils fonctionnent en mode dégradé. Les réseaux électriques sont des exemples de réseaux physique type qui sont soumis à bien des risques : perturbations naturelles, défauts physiques, sabotages, etc. Il est alors normal de parler de « vulnérabilités » [210, 211]. Mais comment évaluer ces vulnérabilités ? Il n'existe pas, à notre connaissance, de définition arrêtée de la vulnérabilité. Freitas *et al.* ont écrit [62] : « La vulnérabilité d'un réseau admet pour notion contraire la robustesse, définie comme la capacité d'un réseau à continuer à fonctionner lorsqu'une partie du réseau est naturelle-

ment endommagée ou ciblée par une attaque ». La vulnérabilité peut alors être considérée comme une mesure de sensibilité suite à des incidents pouvant entraîner des réductions importantes de fonctionnement du réseau. Un réseau (ou une infrastructure) est un système qui peut être représenté par un graphe dont les sommets sont les principaux composants du réseau et les arêtes constituent les connexions physiques entre eux. Par exemple, le réseau électrique nord-américain peut être vu comme un graphe dans lequel les 14 099 sommets correspondent aux centrales électriques ou aux sous-stations de transmission, tandis que les 19 657 arêtes représentent les lignes électriques [3, 4]. Il est clair qu'un tel réseau joue un rôle crucial dans le quotidien de millions d'américains. Que se passerait-il si les lignes électriques étaient sabotées ? Certaines connexions sont-elles plus vulnérables que d'autres et comment quantifier ces vulnérabilités ? C'est précisément la question à l'origine de ce chapitre.

3.1.2 Travaux sur la vulnérabilité de graphes

La topologie d'un réseau met en avant des structures précises entre les agents le composant *via* des connexions entre eux [212]. Les outils mathématiques disponibles en théorie des graphes permettent à la fois la représentation et l'analyse de ces réseaux, notamment les vulnérabilités des agents ou des connexions, qu'elles soient globales ou individuelles. Différentes mesures pour les quantifier ont été développées ces dernières années. Ces mesures peuvent être classées en trois catégories[1] : utilisation de quantifieurs classiques de la théorie des graphes, du spectre d'adjacence ou du spectre Laplacien [62]. En outre, dans chaque catégorie, les quantités extraites peuvent fournir des informations globales sur l'ensemble du réseau ou des informations locales basées sur les sommets et/ou les arêtes.

Un exemple de mesure globale de la vulnérabilité basée sur l'utilisation directe de la structure du graphe est la connectivité, en d'autres termes le nombre minimal d'arêtes à supprimer pour déconnecter le graphe, une valeur élevée étant naturellement synonyme d'une faible vulnérabilité globale. Il y a également les différentes définitions de centralités qui peuvent jouer ce rôle de mesures de vulnérabilité. En effet, la centralité étant une mesure permettant de quantifier l'importance relative des arêtes (ou des sommets) dans la structure globale du graphe, elle peut être capable d'identifier les éléments clés dans ce dernier, c'est-à-dire ceux qui jouent un rôle central dans la connectivité, la diffusion d'information ou les interactions au sein du réseau. Parmi ces notions, introduites en grande partie pour caractériser l'importance des sommets, certaines ont une définition adaptée pour les arêtes, notamment la centralité d'intermédiarité [213, 214] (*betweenness centrality*), cette dernière comptant le nombre de fois que l'arête étudiée se trouve sur un plus court chemin reliant deux sommets. D'autres stratégies, basées quant à elles sur le spectre d'adjacence, ont été développées dans le but de mesurer la vulnérabilité du graphe et de déterminer les sommets à immuniser en priorité (ou, de manière équiva-

1. Cette sous-section est, en réalité, une synthèse de l'étude exhaustive menée par Freitas *et al.* [62].

lente, à supprimer) afin que les sommets restants soient les plus résistants aux attaques de virus [58]. Dans la littérature, ces attaques sont souvent simulées comme la suppression d'un certain pourcentage de sommets [3, 57, 215–217] ou d'arêtes [57, 218] du graphe. L'immunisation des sommets est essentielle pour protéger les réseaux d'une propagation de virus, par exemple. Pour ce faire, Chen *et al.* proposent une mesure simple pour juger de la vulnérabilité globale du graphe, appelée *Shield Value* [58], qui n'est autre que la plus grande valeur propre λ_n de la matrice d'adjacence $\mathbf{A}(G)$ du graphe G étudié [58, 219]. Plus cette valeur est élevée, plus le graphe est vulnérable [58]. Mais si cette mesure peut être employée pour analyser efficacement la propagation des virus, elle peut être interprétée différemment dans la théorie spectrale des graphes. En effet, comme discutée au chapitre 1, une valeur élevée de λ_n signifie que le graphe est mieux connecté structurellement, et donc moins vulnérable aux éventuelles pannes puisqu'il existe toujours des chemins secondaires. Les mêmes outils peuvent alors conduire à des conclusions différentes, d'où la subjectivité évidente liée à la notion de vulnérabilité évoquée précédemment. Enfin, l'indice de Kirchhoff $K(G)$ (équation (1.44)), basé sur la résistance effective du graphe vu comme un réseau électrique et pouvant être exprimé à partir du spectre Laplacien [120,122], peut également être interprété comme une mesure de vulnérabilité : plus cet indice est faible, moins le réseau est considéré comme vulnérable.

Certaines mesures de robustesse peuvent être mieux adaptées que d'autres selon les types de graphes étudiés [62]. De plus, les notions de connectivité ne peuvent pas permettre, à elles seules, une mesure de la vulnérabilité du réseau car elles n'identifient pas les arêtes vulnérables et, d'autre part, certaines mesures ne peuvent être utilisées que pour comparer la vulnérabilité de graphes du même ordre ou du même type. L'objectif de ce travail est alors de définir une mesure fiable de la vulnérabilité d'un graphe prenant en compte sa structure topologique.

3.1.3 Motivations

Nous faisons appel, dans ce chapitre, à des notions de théorie de l'information afin d'explorer le concept de vulnérabilité. En effet, à notre connaissance, il existe peu de travaux qui ont eu recours aux outils de la théorie de l'information pour décrire la vulnérabilité d'un réseau. Nous faisons le choix ici de considérer l'entropie de von Neumann d'un graphe [45] comme élément principal sur lequel se base ce travail. En effet, c'est une mesure relativement récente [45, 220] qui permet de quantifier la complexité[2] d'un graphe en évaluant précisément son contenu informationnel en fonction de sa structure. Cet outil a été introduit en considérant le graphe comme un système physique interprété dans le monde quantique et potentiellement caractérisé par sa matrice de densité. Il est ainsi possible

2. Tout comme la vulnérabilité ou la robustesse, la complexité (ici de la structure) d'un graphe est une notion très subjective. Des travaux ont tout de même fait état des différentes méthodes permettant de l'évaluer objectivement [221].

de définir la matrice de densité d'un graphe à partir de sa matrice Laplacienne, en tirant profit de cette analogie [46]. Dans ce contexte, nous introduisons la « vulnérabilité informationnelle » d'un graphe. Cette notion est construite à partir de l'évolution du contenu informationnel lorsque le réseau initial est attaqué, une attaque pouvant être, par exemple, la suppression d'une arête du graphe modélisant le réseau [61, 222]. Le contenu informationnel du graphe d'origine, dont nous supposons le fonctionnement normal, constitue une référence qui doit être préservée et dont tout écart peut être critique pour la performance du réseau. Une perturbation locale, que ce soit la suppression ou la modification du poids d'une arête, induit inévitablement des variations dans le contenu informationnel du graphe. Ces variations dépendent immanquablement de la position de l'arête perturbée dans la structure du graphe (centrale, périphérique, etc.) ainsi que de sa pondération. Au même titre qu'une notion de centralité, l'ampleur de la variation dans le contenu informationnel peut alors révéler le rôle de l'arête dans la stabilité et le fonctionnement du réseau représenté par le graphe. Nous montrons dans ce chapitre que les arêtes dont les perturbations font le plus décroître l'entropie de von Neumann, et donc le contenu informationnel, sont responsables d'une dégradation de la stabilité du réseau. Ces arêtes sont alors celles à protéger en priorité car ce sont elles qui augmentent le risque d'incidents en cas d'attaque et rendent ainsi le réseau plus vulnérable. Nous constatons également que cette vulnérabilité informationnelle des arêtes a un lien fort avec leurs connectivités. En effet, une corrélation sera ultérieurement analysée entre la mesure de vulnérabilité d'une arête et les degrés des sommets liés par cette dernière. Après avoir rappelé la définition de l'entropie de von Neumann d'un graphe, vu comme un système, nous introduisons la vulnérabilité informationnelle d'une arête, que nous nommons par ailleurs « saillance ». Nous développons et utilisons alors un algorithme permettant de construire une carte de vulnérabilité du réseau étudié, à l'image de celui proposé par Bay-Ahmed *et al.* [82, 223]. Cet algorithme étant relativement gourmand dans le cas de grands graphes, nous en proposons des versions optimisées, principalement permises grâce à des approximations de l'entropie de von Neumann. Parmi celles-ci, nous citons le travail de Han *et al.* [224] avec leur approximation par « entropie quadratique », de Chen *et al.* [225] avec leur approximation FINGER (pour *Fast Incremental von Neumann EntRopy*) ou encore de Choi *et al.* [226] avec leur travail complet recensant notamment des méthodes basées sur des développements limités. Nous introduisons également notre propre approximation basée sur la théorie de perturbation matricielle, où comment les valeurs propres de la matrice Laplacienne évoluent suite à une perturbation du graphe (et donc de sa matrice d'adjacence). Après avoir testé l'algorithme initial ainsi que ses versions optimisées sur des graphes synthétiques et réels, nous confirmons la pertinence de cette mesure permettant de caractériser la vulnérabilité des liens dans un réseau.

3.2 Graphe en tant que système : matrice de densité et entropie

3.2.1 Représentation d'un graphe dans le domaine quantique

Les notions d'ordre, de taille, de degrés d'un graphe révèlent certes, des propriétés relatives au graphe mais s'avèrent souvent limitées pour en établir la complexité [227]. Une idée est de considérer le graphe comme un système physique qui, mis en correspondance avec des états quantiques, permet d'avoir recours aux outils développés en mécanique quantique afin de fournir des mesures significatives de la complexité [76]. Rappelons qu'en mécanique quantique, un système peut se trouver dans un état dit mixte, c'est-à-dire un mélange statistique d'états purs. Ainsi, une correspondance entre graphe et système physique peut être de voir les sommets du graphe comme ces états quantiques purs, les interactions entre eux pouvant être représentées par les arêtes [228]. Une autre manière d'étudier le lien graphe / système physique est de définir une matrice de densité du graphe à partir de sa Laplacienne [46], permettant alors une interprétation fidèle en mécanique quantique. En effet, la matrice de densité d'un système est la pierre angulaire nécessaire pour son étude. L'introduction de cette matrice de densité permet de définir l'entropie de von Neumann du graphe, notion essentielle pour mesurer l'information quantique d'un système [229, 230] et donc le contenu informationnel du graphe permettant de le modéliser. L'entropie de von Neumann d'un graphe, parfois simplement appelée entropie d'un graphe, a été largement utilisée pour caractériser les structures saillantes dans des réseaux issus de divers domaines tels que la biologie, la physique ou encore les sciences sociales. Cette entropie, rappelée dans une prochaine sous-section, est un exemple de mesure permettant de quantifier l'information quantique et de décrire l'incertitude d'un état quantique [231]. Appliquée aux graphes, cette mesure peut être utilisée pour distinguer différentes structures de graphes. Effectivement, elle est maximale pour les graphes complets, minimale pour les graphes vides, et prend des valeurs intermédiaires pour les graphes en étoile [45, 46, 227].

3.2.2 Matrice de densité

L'état d'un système physique est représenté par une matrice semi-définie positive de trace unitaire appelée matrice de densité (ou opérateur de densité). Cette matrice, notée $\boldsymbol{\rho}$ décrit un système physique dont l'état est une combinaison d'états quantiques purs $|\Psi_i\rangle$, formant une base orthonormée de l'espace d'état (espace de Hilbert complexe de dimension n), chacun avec une probabilité p_i [232]. L'opérateur de densité $\boldsymbol{\rho}$ est alors défini par

$$\boldsymbol{\rho} \coloneqq \sum_{i=1}^{n} p_i \, |\Psi_i\rangle\langle\Psi_i| \, . \tag{3.1}$$

Les coefficients sur la diagonale de la matrice $\boldsymbol{\rho}$ sont alors les probabilités p_i de chaque état pur $|\Psi_i\rangle$, et les coefficients en dehors de la diagonale représentent la cohérence quantique n'admettant pas d'interprétation classique [233]. Les valeurs propres ν_i de cette matrice $\boldsymbol{\rho}$ étant équivalentes aux probabilités p_i, l'expression (3.1) est en réalité la décomposition spectrale classique de $\boldsymbol{\rho}$. Les conditions nécessaires et suffisantes pour que $\boldsymbol{\rho}$ soit une matrice de densité sont les suivantes : la trace de $\boldsymbol{\rho}$ doit être égale à 1, $\boldsymbol{\rho}$ doit être semi-définie positive et $\boldsymbol{\rho}$ doit être hermitienne, c'est-à-dire que $\boldsymbol{\rho} = \boldsymbol{\rho}^{\mathsf{H}}$.

Avec tous ces éléments, une correspondance fidèle entre système quantique et graphe peut alors être établie [46, 228]. En effet, Braunstein *et al.* [46] ont introduit une matrice de densité $\boldsymbol{\rho}$ d'un graphe G en normalisant sa matrice laplacienne $\mathbf{L}$ par sa trace :

$$\boldsymbol{\rho} \coloneqq \frac{\mathbf{L}}{\operatorname{Tr}(\mathbf{L})}. \tag{3.2}$$

3.2.3 Entropie de von Neumann d'un graphe

L'entropie a été introduite initialement pour mesurer l'incertitude associée aux probabilités p_i attribuées aux états purs $|\Psi_i\rangle$ présentées précédemment, mais elle permet également depuis quelques années de caractériser la structure des graphes ou des réseaux complexes grâce au fait que le graphe soit interprété comme un système physique. C'est pourquoi de nombreuses définitions de l'entropie d'un graphe ont été introduites [220, 234, 235]. Dans le monde quantique, l'entropie classique est celle de von Neumann, introduite par le mathématicien américano-hongrois John von Neumann dans les années 30 pour prouver l'irréversibilité de la mesure quantique [231]. Cette entropie donne un renseignement précieux quant à la quantité d'information qui peut être extraite d'une observation sur un état mixte, c'est-à-dire un état combinant des états purs. En d'autres termes, il s'agit d'une mesure quantitative du mélange de la matrice de densité $\boldsymbol{\rho}$ du système définie à une constante κ près (ici la constante de Boltzmann) par

$$S(\boldsymbol{\rho}) \coloneqq -\kappa \operatorname{Tr}(\boldsymbol{\rho} \ln \boldsymbol{\rho}) = -\kappa \sum_{i=1}^{n} p_i \ln p_i. \tag{3.3}$$

Par conséquent, l'interprétation de la matrice Laplacienne (normalisée par sa trace) d'un graphe comme un opérateur de densité ouvre la possibilité de caractériser ce dernier en utilisant l'entropie de von Neumann [236]. C'est la raison pour laquelle la définition de l'entropie de von Neumann du graphe G que nous allons retenir est celle de Passerini et Severini [45] définie en utilisant les valeurs

propres[3] $0 = \nu_1 \leq \nu_2 \leq \cdots \leq \nu_n$ de la matrice de densité $\boldsymbol{\rho}$ du graphe G :

$$S(G) \coloneqq -\sum_{\ell=1}^{n} \nu_\ell \ln \nu_\ell. \tag{3.4}$$

Comment interpréter cette entropie ? De manière naturelle tout d'abord : elle peut être vue comme une entropie calculée à partir de la distribution de probabilités constituée des valeurs propres de la matrice de densité du graphe. Ainsi, par analogie avec l'entropie de Shannon d'un ensemble statistique en théorie de l'information, elle fournit un moyen de caractériser le contenu en information du graphe étudié [46, 237]. Une interprétation plus structurelle (bien que sa définition soit uniquement basée sur un aspect spectral), illustrée par le tableau 3.1, est que l'entropie $S(G)$ d'un graphe est une mesure de sa régularité ou encore de sa complexité [45, 224, 238]. En effet, l'entropie $S(G)$ décroît à mesure que le graphe G devient de moins en moins régulier (*cf.* page 29) : un graphe G d'ordre n atteint une entropie maximale égale à $\ln(n-1)$ si et seulement s'il est égal au graphe complet $\mathcal{K}_n$ et atteint une entropie minimale[4] égale à 0 si et seulement s'il est vide (c'est-à-dire si $\mathcal{E} = \emptyset$) [46].

G	Complet $\mathcal{K}_5$ 	Roue $\mathcal{W}_5$ 	Cycle $\mathcal{C}_5$ 
$S(G)$	$\ln(4) \approx 1.39$	1.35	1.28
G	Chemin $\mathcal{P}_5$ 	Étoile $\mathcal{S}_5$ 	Vide 
$S(G)$	1.17	1.07	0

Tableau 3.1 – Entropie de von Neumann de graphes d'ordre 5.

3.3 Vulnérabilité informationnelle d'une arête

Soit $G = (\mathcal{V}, \mathcal{E})$ un graphe éventuellement pondéré, non-orienté, possédant n sommets et m arêtes, $\mathbf{A} = [a_{ij}]_{1 \leq i,j \leq n}$ désignant sa matrice d'adjacence, $\mathbf{W} = [w_{ij}]_{1 \leq i,j \leq n}$ sa matrice de poids, $\mathbf{L} = [l_{ij}]_{1 \leq i,j \leq n}$ sa matrice Laplacienne et $\boldsymbol{\rho} = \mathbf{L}/\operatorname{Tr}(\mathbf{L})$ sa matrice de densité. L'ajout ultérieur d'un symbole $\widetilde{\cdot}$ fait référence au fait que ces notations concernent un graphe perturbé.

3. Avec les notions de théorie spectrale rappelées au chapitre 1, la matrice de densité $\boldsymbol{\rho}$ d'un graphe G admet 0 pour valeur propre minimale de multiplicité le nombre de composantes connectées du graphe.

4. L'égalité $0 \ln 0 = 0$ pouvant être prouvée avec un argument de continuité.

Évaluer la vulnérabilité d'une arête est crucial lorsque sont étudiés des réseaux réels comme des réseaux électriques, informatiques ou sociaux. Toutefois, il existe une multitude de méthodes afin d'y parvenir [3, 57, 215–218]. L'idée initiale que nous proposons dans ce travail est de perturber l'arête étudiée et quantifier l'évolution de l'entropie de von Neumann du graphe suite à cette perturbation. Cette évolution sera inévitablement différente car les modifications apportées à structure du graphe induisent des changements dans le spectre Laplacien. L'impact diffère vraisemblablement d'une arête à l'autre car la sensibilité de l'entropie de von Neumann à la perturbation varie en fonction du voisinage et des propriétés locales de la zone dans laquelle se trouve l'arête perturbée [61]. Ce travail constitue en réalité une extension à celui proposé par Bay-Ahmed *et al.* [82, 223].

3.3.1 Saillance d'une arête

Comme rappelé dans le paragraphe précédent, nous avons décidé de mesurer la saillance d'une arête dans un graphe G grâce à l'impact de sa perturbation sur l'entropie de von Neumann du graphe. Soit $\widetilde{G}_{ij,\xi}$ le graphe G après qu'une arête $\{i, j\}$ ait été perturbée par un coefficient $\xi \in [0, 1]$. D'un point de vue matriciel, cela signifie que l'entrée w_{ij} (resp. w_{ji}) de la matrice de poids $\mathbf{W}$ du graphe G devient $\widetilde{w}_{ij} = (1 - \xi)w_{ij}$ (resp. $\widetilde{w}_{ji} = (1 - \xi)w_{ji}$) dans la matrice de poids $\widetilde{\mathbf{W}}$ du graphe perturbé. Pour définir la saillance d'une arête, nous allons dans un premier temps calculer l'évolution de l'entropie avant et après perturbation comme suit :

$$\Delta_{ij,\xi} = S(G) - S(\widetilde{G}_{ij,\xi}), \tag{3.5}$$

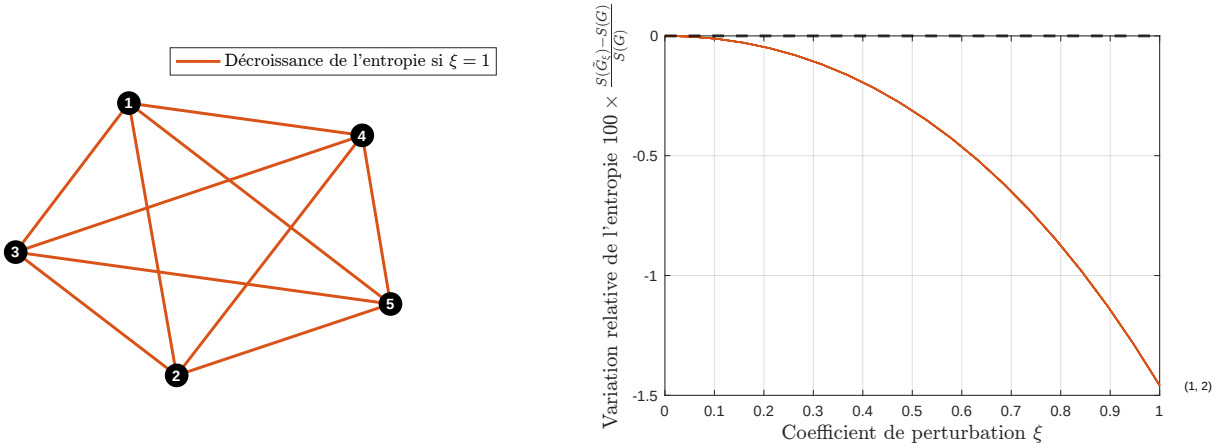
où

$$S(\widetilde{G}_{ij,\xi}) \coloneqq -\sum_{\ell=1}^{n} \widetilde{\nu}_{\ell}^{(ij,\xi)} \ln \widetilde{\nu}_{\ell}^{(ij,\xi)} \tag{3.6}$$

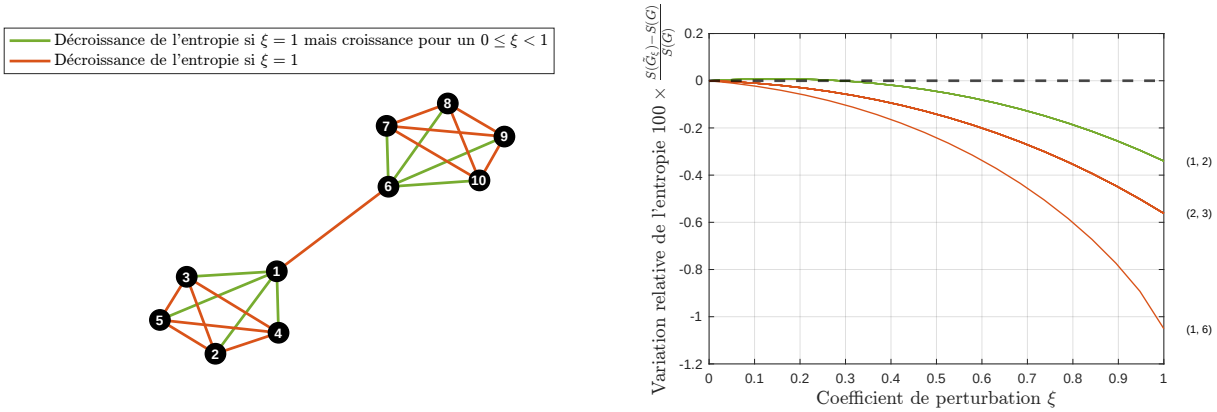
désigne l'entropie de von Neumann du graphe perturbé $\widetilde{G}_{ij,\xi}$ avec $(\widetilde{\nu}_{\ell}^{(ij,\xi)})_{1\leq\ell\leq n}$ les valeurs propres de la matrice de densité de $\widetilde{G}_{ij,\xi}$. Puis, pour quantifier les effets de la perturbation, nous allons utiliser la variation relative de l'entropie définie par :

$$\eta_{ij,\xi} \coloneqq 100 \times \frac{\Delta_{ij,\xi}}{S(G)} = 100 \times \frac{S(\widetilde{G}_{ij,\xi}) - S(G)}{S(G)}. \tag{3.7}$$

La valeur $\eta_{ij,\xi}$ peut être interprétée comme une mesure de vulnérabilité de l'arête $\{i, j\}$ et donne une information sur l'influence de cette arête dans le fonctionnement du système, représenté par le graphe G. Par conséquent, l'arête ayant la plus petite saillance $\eta_{ij,\xi}$, pour un paramètre de perturbation ξ fixé, causera le moins d'effet sur l'entropie du graphe et donc sur son contenu informationnel.



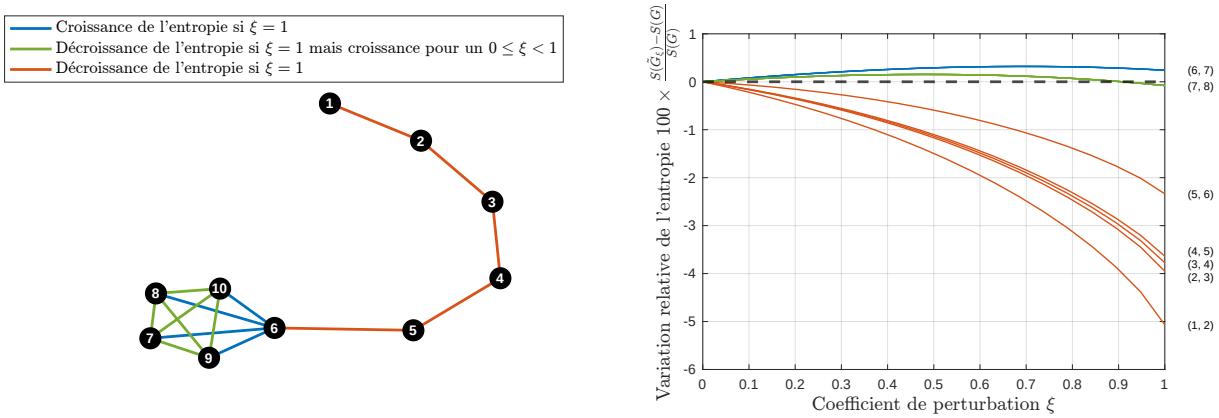
(a) Graphe complet (5 sommets / 10 arêtes)



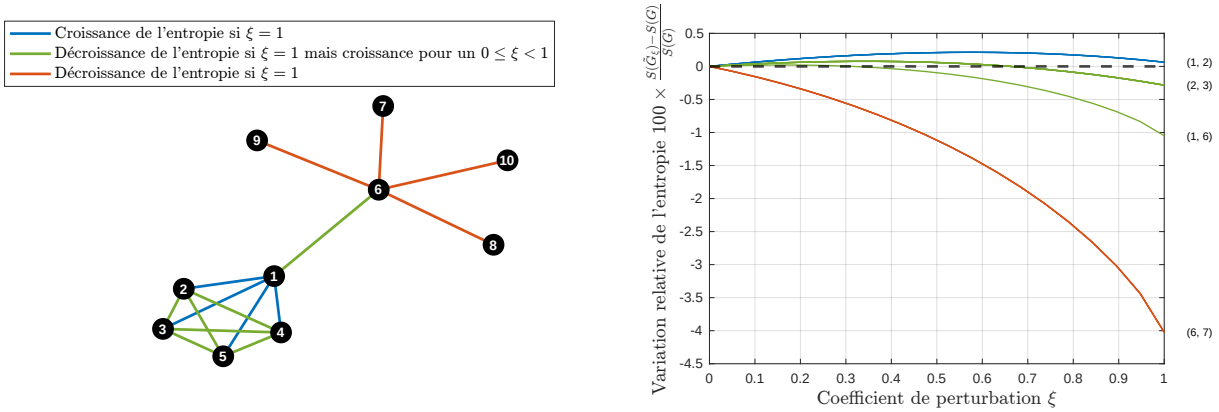
(b) Graphe *Barbell* (10 sommets / 21 arêtes)

Figure 3.1 – Colonne de gauche : graphes dont la couleur des arêtes correspond à l'impact sur l'entropie de von Neumann si ces dernières sont perturbées : le **bleu** signifie que l'entropie augmente pour toute valeur du coefficient de perturbation ξ entre 0 et 1, le **rouge** signifie que l'entropie diminue pour toute valeur du coefficient de perturbation ξ entre 0 et 1 et le **vert** signifie que l'entropie diminue si l'arête est supprimée ($\xi = 1$) mais qu'il existe une perturbation $0 \leq \xi < 1$ telle que l'entropie augmente / **Colonne de droite** : courbes représentant la variation relative de l'entropie en fonction du paramètre de perturbation ξ variant de 0 à 1. Chaque courbe correspond à une arête du graphe et pour certaines arêtes particulières, leurs libellés sont affichés sur la droite des courbes. La ligne noire en pointillés correspond à une variation relative de l'entropie égale à 0. *(La figure se poursuit en page suivante.)*

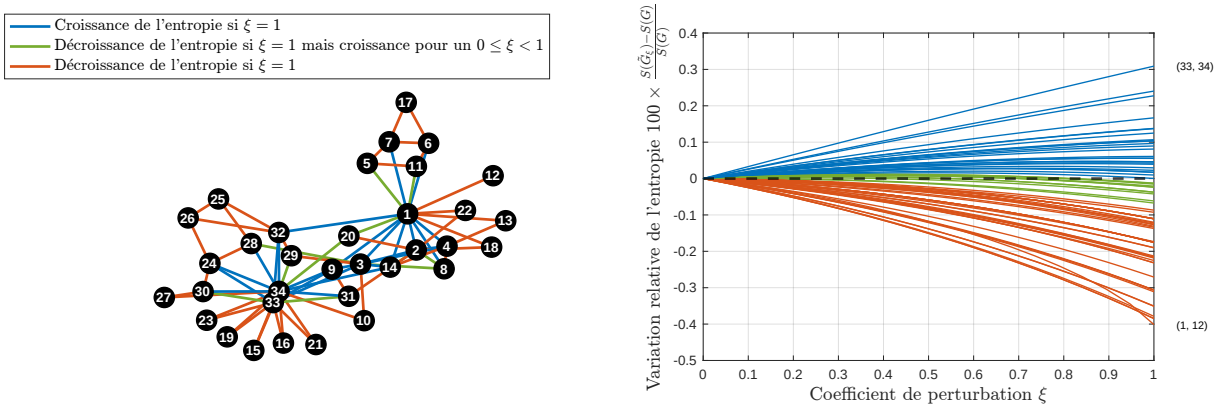
Pour prendre la mesure de ce que représente la saillance d'une arête définie par l'équation (3.7), nous avons considéré 5 graphes (le graphe complet $\mathcal{K}_5$, le graphe *Barbell* $\mathcal{B}_5$, le graphe comète $\mathcal{K}_5 \cup \mathcal{P}_5$, le graphe Complet+Étoile $\mathcal{K}_5 \cup \mathcal{S}_5$ et le graphe KarateClub [239]) et avons calculé, pour toutes les arêtes $\{i, j\}$ de ces graphes, la valeur $\eta_{ij,\xi}$ pour un paramètre de perturbation ξ variant de 0 (pas de perturbation de l'arête) à 1 (suppression de l'arête) par pas de 0.1. Ces résultats font l'objet de la figure 3.1. De manière plus détaillée, la figure 3.1a présente les résultats relatifs au graphe complet $\mathcal{K}_5$ (5 sommets / 10 arêtes). Il est clair que toutes les arêtes de ce graphe jouent le même rôle. Ainsi, il n'y



(c) Graphe comète (10 sommets / 15 arêtes)



(d) Graphe Complet+Étoile connecté (10 sommets / 15 arêtes)



(e) Graphe KarateClub (34 sommets / 78 arêtes)

Figure 3.1 – *(Suite.)*

a qu'une seule courbe de variation relative de l'entropie et, pour toute valeur du coefficient de perturbation ξ, l'entropie de von Neumann décroît suite à la perturbation de n'importe quelle arête. La

figure 3.1b permet d'étudier le graphe *Barbell* $\mathcal{B}_5$ (10 sommets / 21 arêtes) et les variations relatives de l'entropie de ses arêtes. La particularité de ce graphe est qu'il est composé de deux régions fortement connectées (sous-graphes complets de taille 5) reliées entre elles par l'arête $\{1, 6\}$, garantissant ainsi l'accès à tous les sommets du graphe. Intuitivement, cette arête doit se révéler la plus sensible aux perturbations et la plus vulnérable par rapport aux autres arêtes. Les courbes de droite sur la figure 3.1b étayent ce propos car celle qui correspond à l'arête $\{1, 6\}$ montre une décroissance plus importante de l'entropie. Comme il n'existe que trois types d'arêtes dans ce graphe, il n'y a que trois courbes de variation relative de l'entropie. La courbe correspondant à l'arête reliant le sommet 1 aux autres (comme celle reliant le sommet 6 aux autres) est représentée en vert, ce qui dénote un comportement différent : l'entropie diminue si cette arête est supprimée, mais il existe une perturbation ξ comprise entre 0 et 1 telle que l'entropie augmente. À notre connaissance, aucune explication évidente n'a pu être donnée à ce comportement. Le graphe comète $\mathcal{K}_5 \cup \mathbf{P}_5$ (10 sommets / 15 arêtes) fait, quant à lui, l'objet de la figure 3.1c. Un premier constat est que plus les arêtes sont situées à l'extrémité de la queue de la comète, plus elles sont responsables d'une diminution de l'entropie si elles sont perturbées. Les arêtes formant cette queue ont toutes un rôle différent : si l'arête $\{1, 2\}$ est supprimée, alors un sommet se retrouve isolé tandis que si c'est l'arête $\{5, 6\}$ qui est supprimée, alors le graphe se retrouve séparé en deux sous-graphes plus équilibrés. Une première force de cette variation relative de l'entropie comme mesure de vulnérabilité est d'attribuer une plus grande responsabilité aux arêtes qui isolent des sommets si elles sont supprimées. Toutefois, sur la partie complète de la comète, il est à noter de nouveau que certaines arêtes provoquent une augmentation de l'entropie de von Neumann quel que soit le paramètre ξ de perturbation. Au même titre que pour l'analyse précédente concernant le graphe *Barbell* $\mathcal{B}_5$, ce phénomène n'admet pas d'explication évidente. Nous retenons que la variation relative de l'entropie, bien qu'elle soit définie de manière purement spectrale, prend véritablement en compte la structure du graphe étudié. En effet, comme nous le montrons dans la suite de ce chapitre, il y a une relation entre la variation relative de l'entropie d'une arête suite à sa perturbation et le degré moyen des deux sommets connectés par cette dernière. Nous considérons ensuite, sur la figure 3.1d, le graphe connecté Complet+Étoile $\mathcal{K}_5 \cup \mathcal{S}_5$ (10 sommets / 15 arêtes). Il n'y a que quatre courbes différentes à droite de la figure ce qui signifie que seuls quatre rôles sont joués par les arêtes. En particulier, les arêtes formant le graphe étoile provoquent une diminution de l'entropie pour tout coefficient de perturbation ξ, ce qui est logique compte tenu de l'interprétation précédente car leurs suppressions conduisent à l'isolement de sommets. Pour la partie « graphe complet », les mêmes rôles que pour le graphe précédent sont retrouvés. Cependant, l'arête $\{1, 6\}$ a un comportement singulier dans ce graphe. Celle-ci devrait adopter le même comportement que l'arête $\{1, 6\}$ du graphe *Barbell* $\mathcal{B}_5$, puisque ces arêtes relient des sommets de même degré. Toutefois, dans le cas du graphe Complet+Étoile, il existe un

paramètre ξ qui provoque une augmentation de l'entropie. Une conjecture pourrait être établie en étudiant le 2-voisinage (les voisins des voisins), c'est-à-dire les degrés des sommets reliés aux sommets connectés par l'arête étudiée. Enfin, le dernier graphe étudié est le graphe **KarateClub** (34 sommets / 78 arêtes) [239]. Il s'agit d'un graphe non-pondéré décrivant les relations sociales entre les 34 membres d'un club de karaté dans une université américaine. Les nœuds représentent les individus et les liens les interactions entre eux. Sur la figure 3.1e, il est facile de voir que toute arête joue un unique rôle, ce qui semble logique car nous sommes face à un graphe réel, de sorte qu'il y a autant de courbes à droite de la figure que d'arêtes. Il peut être constaté que plusieurs arêtes provoquent une augmentation de l'entropie si elles sont supprimées. Ces arêtes sont souvent situées dans les zones assez denses du graphe. Leurs suppressions ne paraissent donc pas poser de problème du point de vue de l'entropie et, par conséquent, du point de vue du contenu informationnel du système. Une interprétation naïve est ainsi de se dire que ces arêtes sont probablement des informations redondantes dans le sens où il peut y avoir des chemins alternatifs pour relier les sommets. Deux arêtes méritent une attention particulière : l'arête $\{33, 34\}$ provoquant la plus forte augmentation de l'entropie est celle qui relie les deux sommets les plus connectés du graphe, et l'arête $\{1, 12\}$ qui provoque la plus forte décroissance de l'entropie est la seule qui, si elle est supprimée, isole un sommet (le sommet 12).

Trois observations peuvent être tirées de ces résultats concernant la saillance $\eta_{ij,\xi}$ d'une arête $\{i, j\}$ suite à une perturbation de paramètre ξ telle que définie en équation (3.7). Cette saillance, prenant la forme d'une variation relative de l'entropie, présente un comportement cohérent en ce sens que l'entropie varie de manière monotone en fonction du paramètre ξ. Dans un second temps, une même perturbation n'affecte pas de la même manière l'entropie (ou le contenu informationnel du graphe) selon les arêtes perturbées. En effet, certaines arêtes présentent une saillance beaucoup plus élevée que d'autres : il peut être conjecturé que c'est en raison de leur emplacement et de la topologie du graphe. Il est toutefois possible de rencontrer des graphes pour lesquels il existe des arêtes jouant le même rôle et ayant alors la même saillance. C'est le cas notamment pour le graphe complet où toutes les arêtes sont similaires. La troisième et dernière information déduite est que la perturbation ou la suppression d'une arête ne diminue pas toujours l'entropie, dans certains cas elle l'augmente, donnant des valeurs de saillance $\eta_{ij,\xi}$ supérieures à 0. Ce comportement peut sembler contre-intuitif mais il est possible de supposer qu'il est dû à la présence d'informations redondantes au sein du graphe. Pour prolonger, il est souvent considéré que l'augmentation de l'entropie dénote une création d'information. Or, dans le cas de graphes, ce n'est pas toujours le cas : la suppression d'une arête peut entraîner une augmentation de l'entropie (voir figure 3.1). Toutes ces observations confirment la capacité de l'entropie de von Neumann à être un outil approprié pour mesurer l'impact graduel de changements locaux se produisant dans la topologie du graphe.

3.3.2 Algorithme EIVP (*Edge Informational Vulnerability to Perturbation*)

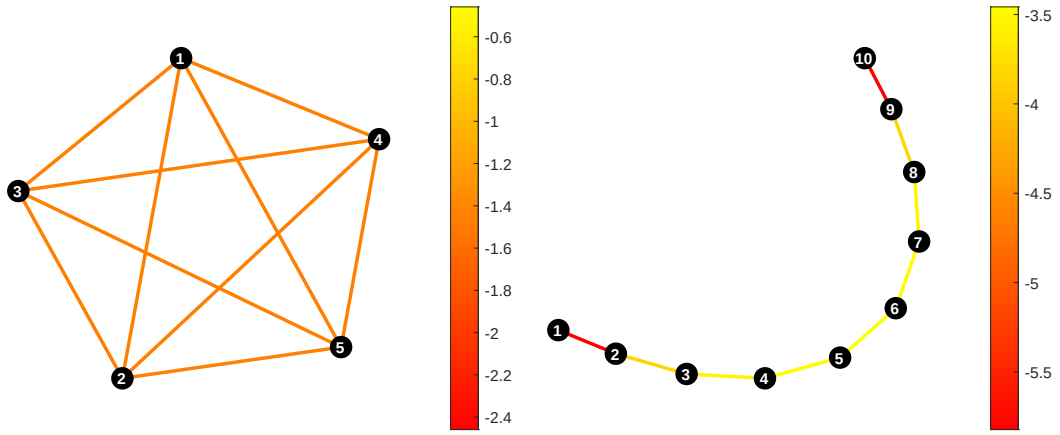
Ainsi, la vulnérabilité d'une arête peut être mesurée à partir de l'évolution de l'entropie de von Neumann du réseau si cette arête venait à être perturbée. Les résultats obtenus montrent la pertinence de cette méthode de mesure de la vulnérabilité des liens dans le réseau. Pour proposer une cartographie des vulnérabilités de ces arêtes, une nouvelle version pondérée du graphe est calculée grâce à un algorithme qui sera appelé **EIVP** (en anglais ***E**dge **I**nformational **V**ulnerability to **P**erturbation*), variante de l'algorithme **VPV** proposé par Bay-Ahmed *et al.* [82, 223]. L'idée est d'attribuer la variation relative de l'entropie $\eta_{ij,\xi}$ comme poids à l'arête $\{i, j\}$, tout en prenant le poids initial w_{ij} éventuel. Cet algorithme de repondération, dont un pseudo-code est donné en algorithme 4, permet d'afficher aisément les structures vulnérables, éventuellement dissimulées, du réseau.

Algorithme 4 Algorithme **EIVP** (***E**dge **I**nformational **V**ulnerability to **P**erturbation*)

Entrée : Graphe $G = (\mathcal{V}, \mathcal{E}, \mathbf{W})$ et coefficient de perturbation $\xi \in [0, 1]$
Sortie : Graphe $G_S^{\text{EIVP}} = (\mathcal{V}, \mathcal{E}, \mathbf{W}_S^{\text{EIVP}})$

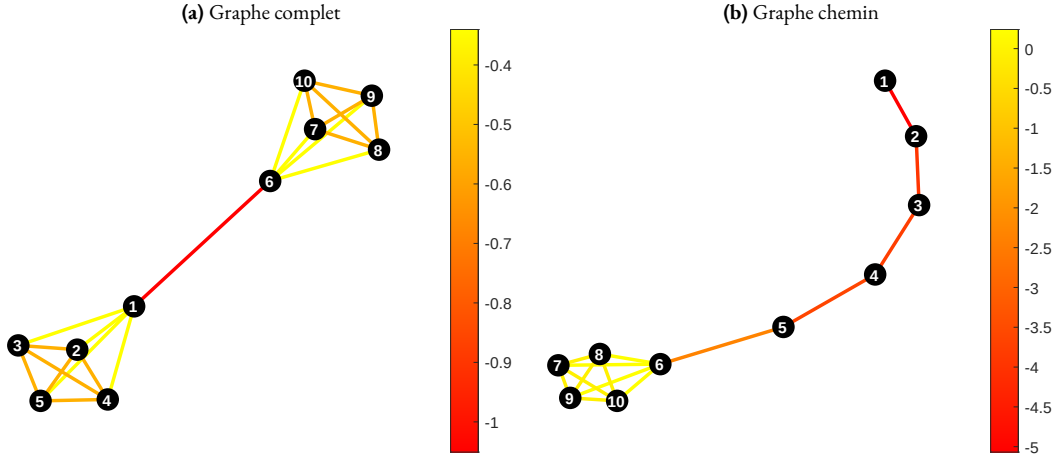
1: Calcul de l'entropie $S(G)$ du graphe initial ▷ Équation (3.4)
2: **for** $\{i, j\} \in \mathcal{E}$ **do**
3: $\quad [\widetilde{\mathbf{W}}]_{kl} = \begin{cases} w_{kl}(1-\xi), & \text{si } \{k,l\} = \{i,j\} \\ w_{kl}, & \text{sinon} \end{cases}$
4: $\quad \widetilde{G}_{ij,\xi} \leftarrow \left(\mathcal{V}, \mathcal{E}, \widetilde{\mathbf{W}}\right)$
5: $\quad$ Calcul de l'entropie $S(\widetilde{G}_{ij,\xi})$ du graphe perturbé ▷ Équation (3.4)
6: $\quad [\mathbf{W}_S^{\text{EIVP}}]_{ij} \leftarrow 100 \times \frac{S(\widetilde{G}_{ij,\xi}) - S(G)}{S(G)}$ ▷ Équation (3.7)

La figure 3.2 illustre l'application de l'algorithme **EIVP** à quelques graphes avec un coefficient de perturbation égal à 1 pour chaque arête perturbée (signifiant leurs suppressions). L'exemple le plus simple est le graphe complet $\mathcal{K}_5$ (5 sommets / 10 arêtes) qui est représenté dans sa version pondérée (sous-entendu par l'algorithme **EIVP**) dans la figure 3.2a. Toutes les arêtes jouant le même rôle, elles sont toutes aussi vulnérables les unes que les autres, *a minima* du point de vue du contenu informationnel du graphe. La figure 3.2b présente un graphe chemin $\mathcal{P}_{10}$ (10 sommets / 9 arêtes). L'algorithme **EIVP**, calculé sur ce graphe, apporte un élément interprétable relatif à l'information contenue dans son spectre Laplacien. En effet, les arêtes les plus vulnérables sont celles dont la suppression isolerait des sommets (ici, les sommets 1 et 10) ou, dans une moindre mesure, diviserait le graphe en deux sous-graphes inégalement répartis. C'est pourquoi les arêtes sont de moins en moins vulnérables à mesure qu'elles sont situées au milieu du chemin. La version pondérée du graphe *Barbell* $\mathcal{B}_5$ (10 sommets / 21 arêtes) est quant à elle proposée en figure 3.2c. Comme attendu, l'arête $\{1, 6\}$ est la plus vulnérable de la structure. En effet, cette dernière relie deux communautés importantes du graphe et sa suppression



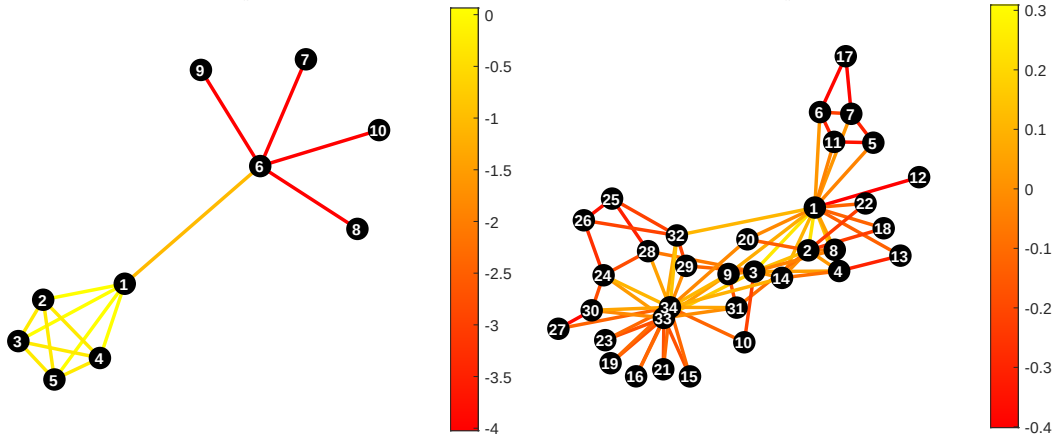
(a) Graphe complet

(b) Graphe chemin



(c) Graphe *Barbell*

(d) Graphe comète



(e) Graphe Complet+Étoile connecté

(f) Graphe KarateClub

Figure 3.2 – Après être passés par l'algorithme **EIVP**, graphes dont les couleurs des arêtes représentent leurs saillances lorsqu'elles sont supprimées ($\xi = 1$ pour toutes les arêtes $\{i, j\}$). Plus les arêtes sont rouges, plus elles sont responsables d'une diminution de l'entropie et, par conséquent, considérées comme vulnérables d'un point de vue informationnel.

entraînerait une déconnexion du graphe. Il est également possible de voir les trois niveaux de vulnérabilité correspondant aux trois rôles distincts joués par les arêtes. La figure 3.2d représente le graphe comète (10 sommets / 15 arêtes) mis en entrée de l'algorithme **EIVP**. Ce graphe renforce les remarques faites pour les graphes chemin et *Barbell*. En effet, plus les arêtes sont situées à l'extrémité de la comète, plus elles sont vulnérables car ce sont celles dont la suppression conduirait à un isolement de sommets. L'arête $\{5, 6\}$, reliant le chemin à la partie complète, n'est pas la plus vulnérable car sa suppression divise simplement le graphe en deux sous-graphes relativement équilibrés. Ce point d'attention montre, et ce sera le sujet d'une discussion future, que notre mesure de vulnérabilité basée sur une variation relative de l'entropie n'est pas corrélée avec la centralité d'intermédiarité des arêtes qui prend en compte le nombre de plus courts chemins passant par les arêtes en question. Effectivement, pour ce graphe, cette mesure de centralité aurait désigné l'arête $\{5, 6\}$ comme la plus vulnérable de toutes ce qui n'est pas le cas ici. La figure 3.2e permet de montrer, grâce au graphe Complet+Étoile (10 sommets / 15 arêtes), la cohérence des résultats que peut fournir l'algorithme **EIVP**. Ce dernier étant très sensible aux discontinuités potentielles de la structure, les arêtes de la partie Étoile se trouvent toutes de couleur rouge car leurs suppressions entraîneraient une déconnexion du graphe et un isolement de certains sommets (ici le 7, 8, 9 et 10). Une fois de plus, il peut être noté que l'algorithme met bien en avant les différents niveaux de vulnérabilité correspondant aux différents rôles joués par les arêtes dans le graphe (visibles également sur la figure 3.1d). L'arête $\{1, 6\}$ de ce graphe Complet+Étoile mérite une attention particulière car sa saillance permet de la classer comme relativement vulnérable. C'était un événement attendu car cette arête garantit la communication entre tous les sommets du graphe. Néanmoins, les arêtes de la partie Étoile restent plus vulnérables que cette arête, leurs suppressions provoquant une rupture brutale de la structure en isolant des sommets. La suppression de l'arête $\{1, 6\}$ est synonyme de déconnexion du graphe, mais en deux sous-graphes de tailles comparables, l'algorithme **EIVP** arrive à bien différencier ces deux niveaux de risques. Enfin, un graphe plus complexe, à savoir le graphe KarateClub [239], a été étudié et sa version pondérée par l'algorithme **EIVP** est représentée en figure 3.2f. Il semble plus complexe d'analyser arête par arête les saillances. Toutefois, les poids attribués paraissent cohérents avec les propriétés de la structure. En effet, les arêtes les plus vulnérables (c'est-à-dire les arêtes les plus rouges) sont celles dont la suppression risque de déconnecter le graphe et d'isoler certains sommets (par exemple les arêtes $\{1, 12\}$, $\{6, 17\}$, $\{7, 17\}$). Il est également possible d'observer que les arêtes les moins vulnérables (ainsi que celles responsables d'une augmentation de l'entropie) sont situées dans la zone la plus connectée de la structure, là où l'accès aux sommets peut se faire par de nombreux chemins secondaires. Par ailleurs, l'algorithme se comporte d'une manière particulière avec les sommets de degré 2. En effet, ces derniers peuvent être classés en deux groupes : $13, 27$ et $15, 16, 19, 21, 23$. Dans le premier cas, les deux arêtes permettant d'accéder aux sommets

13 et 27 n'ont pas la même saillance, en fonction de l'importance des sommets à l'autre extrémité de l'arête. L'arête $\{27, 30\}$ est alors plus vulnérable que l'arête $\{27, 34\}$, car le sommet 34 est plus central dans le réseau (son degré est plus élevé) relativement au sommet 30. Ainsi, l'algorithme **EIVP** met en évidence la vulnérabilité des arêtes permettant l'accès à des sommets de degré inférieur afin d'éviter le risque d'isolement de ces derniers. De même, en ce qui concerne le sommet 13, le lien $\{4, 13\}$ semble plus vulnérable que $\{1, 13\}$, car le sommet 4 est moins important dans le réseau que le sommet 1. Pour les sommets du second groupe, ils sont connectés au réseau grâce à des arêtes qui partagent un niveau de vulnérabilité similaire, parce que les sommets aux extrémités de ces arêtes ont des centralités et des degrés comparables. À titre d'exemple, le sommet 21 est connecté aux sommets 33 et 34, ayant les mêmes voisinages et des degrés tous deux importants, *via* les arêtes $\{21, 33\}$ et $\{21, 34\}$, qui présentent alors le même niveau de vulnérabilité informationnelle. Il est possible d'observer la même chose pour les sommets $15, 16, 19$ et 23.

3.3.3 Corrélation avec d'autres attributs

En résumé des analyses précédentes, la mesure de la vulnérabilité d'une arête définie par sa mesure de saillance (3.7) peut être vue comme un compromis degré-centralité. La figure 3.3 illustre, pour le graphe KarateClub, la relation étroite entre la saillance $\eta_{ij,1}$ d'une arête $\{i, j\}$ définie par (3.7) et la moyenne géométrique $\sqrt{\deg(i)\deg(j)}$ des degrés des sommets connectés par cette arête. Cette relation entre le degré d'un sommet et la valeur propre correspondante n'est pas nouvelle [227] mais, grâce à ce travail, nous confirmons, d'une manière différente, l'existence d'une relation entre degré et évolution de l'entropie, mise en avant dans [61]. En rappelant que le coefficient de corrélation entre deux vecteurs $\mathbf{x} = (x_i)_{1\leq i\leq n}$ et $\mathbf{y} = (y_i)_{1\leq i\leq n}$ est défini par

$$r_{\mathbf{x},\mathbf{y}} \coloneqq \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \qquad \text{avec } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ et } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{3.8}$$

cette valeur étant toujours comprise entre 1 et -1 avec égalité si les vecteurs sont parfaitement (inversement) correlés, le coefficient de corrélation entre le vecteur contenant les moyennes géométriques des degrés et celui contenant les saillances $\eta_{ij,1}$ est égal à 0.97.

Les résultats que met en avant l'algorithme **EIVP** sont les suivants : les arêtes les plus vulnérables sont celles qui isolent des sommets en cas de suppression, celles qui sont classées comme un peu moins vulnérables sont celles qui provoquent la division du graphe en sous-graphes déséquilibrés et les arêtes classées comme non vulnérables sont celles qui se trouvent dans des zones très denses ce qui conduit à la conjecture qu'elles constituent une redondance d'information. Ce sont des constats très intéressants

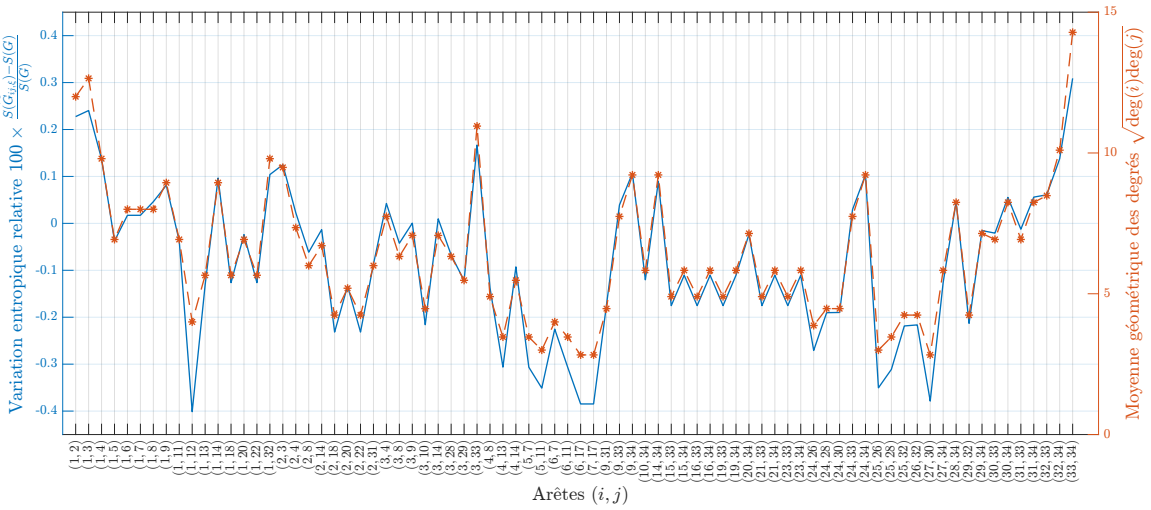


Figure 3.3 – Étude pour le graphe KarateClub : relation entre la variation relative de l'entropie d'une arête si elle est retirée du graphe (courbe **bleue**) et la moyenne géométrique des degrés des sommets connectés par cette arête (courbe **orange**).

car cet algorithme, basé essentiellement sur des attributs spectraux (valeurs propres de la matrice de densité), caractérise structurellement et topologiquement le graphe étudié. Un premier exemple est la relation existante entre la variation relative de l'entropie due à la suppression d'une arête et la moyenne géométrique des degrés des sommets connectés par cette arête, précédemment illustrée et expliquée. Afin d'approfondir cette analyse, l'objectif de cette section est d'étudier d'autres attributs structurels ou spectraux avec lesquels notre mesure de vulnérabilité pourrait être corrélée. Pour ce faire, nous allons calculer les corrélations entre le vecteur $\boldsymbol{\eta} = (\eta_{ij,1})_{\{i,j\}\in\mathcal{E}}$ contenant les variations relatives de l'entropie et les cinq vecteurs suivants contenant des mesures pour chaque arête :

— Vecteur $\mathbf{fv}$ contenant l'évolution de la valeur de Fiedler [8] (*i.e.* la deuxième plus petite valeur propre μ_2 de la matrice Laplacienne) lorsque les arêtes sont supprimées :

$$[\mathbf{fv}]_{ij} = 100 \times \frac{\widetilde{\mu}_2^{(ij,1)} - \mu_2}{\mu_2}, \tag{3.9}$$

où $\widetilde{\mu}_2^{(ij,1)}$ désigne la valeur de Fiedler du graphe dont l'arête $\{i, j\}$ a été supprimée.

— Vecteur $\mathbf{sv}$ contenant l'évolution de la *Shield Value* [58] (*i.e.* la plus grande valeur propre λ_n de la matrice d'adjacence) lorsque les arêtes sont supprimées :

$$[\mathbf{sv}]_{ij} = 100 \times \frac{\widetilde{\lambda}_n^{(ij,1)} - \lambda_n}{\lambda_n}, \tag{3.10}$$

où $\widetilde{\lambda}_n^{(ij,1)}$ désigne la *Shield Value* du graphe dont l'arête $\{i, j\}$ a été supprimée.

— Vecteur $\mathbf{ki}$ contenant l'évolution de l'indice de Kirchhoff (équation (1.44)) [120, 122] (*i.e.* la somme des inverses des valeurs propres de la matrice Laplacienne) lorsque les arêtes sont supprimées :

$$[\mathbf{ki}]_{ij} = 100 \times \frac{K(\widetilde{G}_{ij,1}) - K(G)}{K(G)}. \tag{3.11}$$

— Vecteur $\mathbf{bc}$ contenant les centralités d'intermédiarité des arêtes [213,214] (*i.e.* le nombre de plus courts chemins passant par l'arête étudiée) :

$$[\mathbf{bc}]_{ij} = \sum_{1 \leq k,l \leq n} \frac{\sigma(k,l|\{i,j\})}{\sigma(k,l)}, \tag{3.12}$$

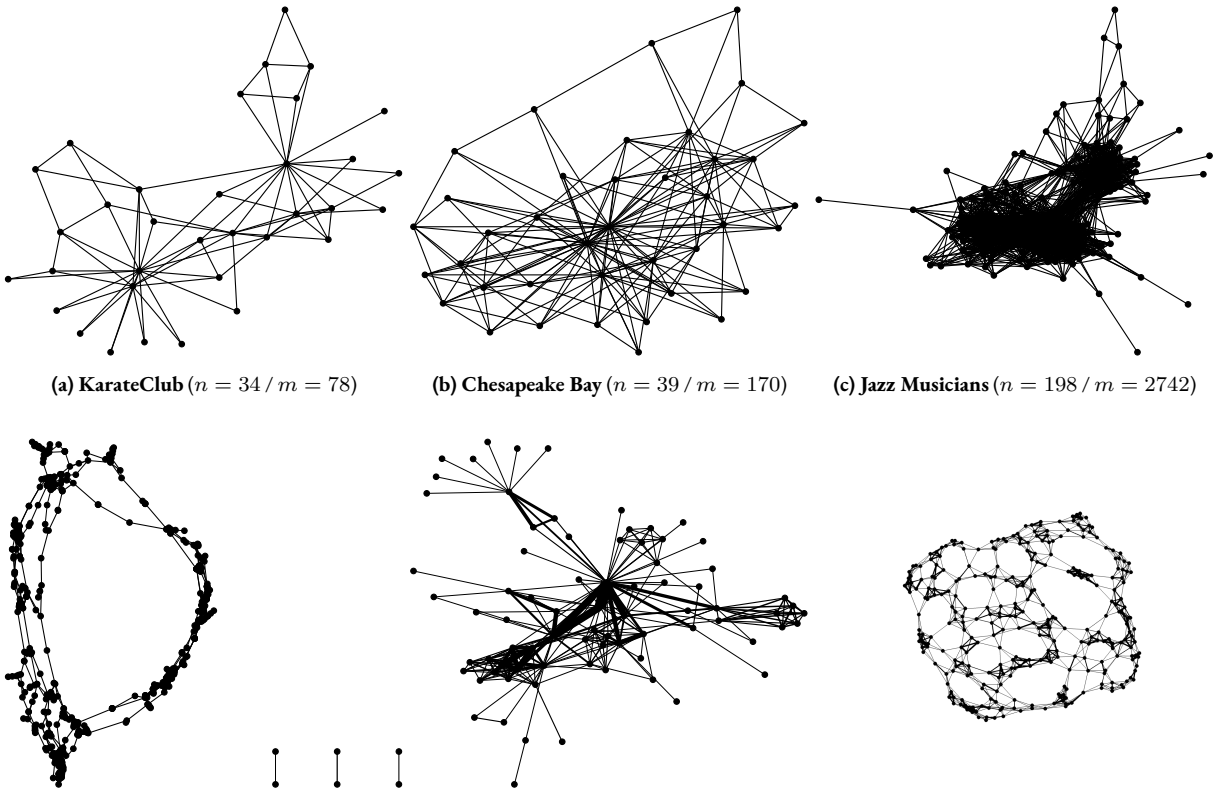
où $\sigma(k,l)$ désigne le nombre de plus courts chemins entre les sommets k et l et $\sigma(k,l|\{i,j\})$ désigne le nombre de plus courts chemins entre les sommets k et l passant par l'arête $\{i,j\}$.

— Vecteur $\mathbf{gm}$ contenant les moyennes géométriques des degrés des sommets reliés par les arêtes :

$$[\mathbf{gm}]_{ij} = \sqrt{\deg(i)\deg(j)}. \tag{3.13}$$

Cette étude s'est portée sur 6 graphes de différents types connus de la littérature et pouvant être aisément trouvés en ligne. Parmi ces derniers, représentés en figure 3.4, le graphe **KarateClub** [239] rencontré plus tôt dans ce chapitre, le graphe et non pondéré **Chesapeake Bay** [240] (figure 3.4b) illustrant le réseau trophique de la baie de Chesapeake où les sommets sont des groupes d'organismes et les arêtes des échanges de carbone, le graphe non pondéré **Jazz Musicians** [241] (figure 3.4c) modélisant un réseau de collaboration dans le monde du jazz où chaque sommet est un musicien et une arête traduit le fait que deux musiciens aient déja joué ensemble, le graphe pondéré **SciGrid** [242] (figure 3.4d) modélisant le réseau électrique allemand où chaque sommet est une station et les arêtes sont les lignes électriques entre ces stations, le graphe pondéré **Les Misérables** [243] (figure 3.4e) contenant les co-occurrences intra-chapitres (arêtes) des personnages (sommets) du roman « Les Misérables » de Victor Hugo et enfin un graphe pondéré **Random Sensor** [244] qui, comme son nom l'indique, est un réseau aléatoire de capteurs.

Toutes les corrélations sont listées dans le tableau 3.2. Avant toute chose, il convient de noter que certains des graphes étudiés sont pondérés mais, bien que les mesures précédemment mentionnées (en particulier celle introduite dans ce chapitre) puissent toutes prendre en compte d'éventuelles pondérations, nous avons choisi de ne pas les considérer (en réalité, nous avons considéré les poids comme tous égaux à 1 s'ils étaient non nuls dans le graphe d'origine) afin d'obtenir des résultats homogènes entre les différents graphes, qu'ils soient pondérés ou non.



(a) KarateClub ($n = 34$ / $m = 78$) **(b) Chesapeake Bay** ($n = 39$ / $m = 170$) **(c) Jazz Musicians** ($n = 198$ / $m = 2742$)

(d) SciGrid ($n = 511$ / $m = 679$) **(e) Les Misérables** ($n = 77$ / $m = 254$) **(f) Random Sensor** ($n = 300$ / $m = 1070$)

Figure 3.4 – Graphes issus de la littérature et utilisés dans cette étude de corrélation.

	$r_{\boldsymbol{\eta},\mathbf{fv}}$	$r_{\boldsymbol{\eta},\mathbf{sv}}$	$r_{\boldsymbol{\eta},\mathbf{ki}}$	$r_{\boldsymbol{\eta},\mathbf{bc}}$	$r_{\boldsymbol{\eta},\mathbf{gm}}$
KarateClub [239]	0.17	−0.90	−0.28	0.38	0.97
Chesapeake Bay [240]	0.41	−0.93	−0.67	0.44	0.97
Jazz Musicians [241]	0.31	−0.76	−0.10	−0.16	0.93
SciGrid [242]	−0.02	−0.27	0.27	0.34	0.93
Les Misérables [243]	0.64	−0.70	0.33	0.01	0.91
Random Sensor [244]	0.25	−0.28	−0.42	−0.19	0.98

Tableau 3.2 – Coefficients de corrélation entre le vecteur $\boldsymbol{\eta}$ contenant les variations relatives de l'entropie et les vecteurs $\mathbf{fv}$, $\mathbf{sv}$, $\mathbf{ki}$, $\mathbf{bc}$ et $\mathbf{gm}$. Ceci pour les 6 graphes de la littérature présentés ci-dessus.

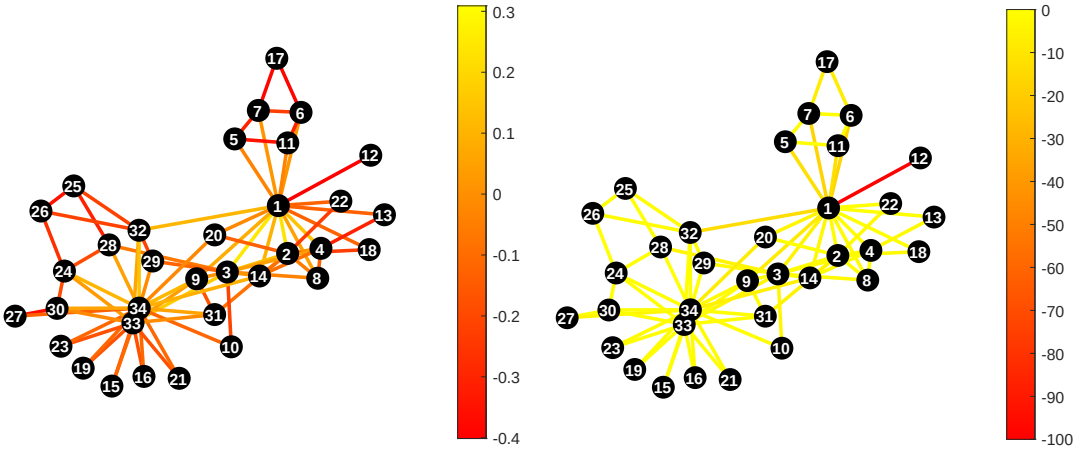
À la lecture du tableau 3.2, il est clair que la valeur de Fiedler, l'indice de Kirchhoff ainsi que la centralité d'intermédiarité des arêtes ne sont pas très liés linéairement avec notre mesure basée sur la variation relative de l'entropie. Il peut sembler étrange que la valeur de Fiedler ne le soit pas car ce sont deux mesures basées sur le spectre Laplacien. Néanmoins, cela peut s'expliquer par le fait que la suppression d'une arête d'un graphe fait évoluer toutes les valeurs propres de sa matrice Laplacienne, et pas seulement la deuxième plus petite. Nous reviendrons sur ce point dans la section suivante en

faisant appel à la théorie de perturbation matricielle. D'autre part, et nous l'avions déjà remarqué à la figure 3.3, la variation relative de l'entropie d'une arête ôtée du graphe et la moyenne géométrique des degrés des sommets connectés par cette arête sont corrélées. En effet, les valeurs de corrélation correspondantes sont proches de 1. Enfin, un autre point qui mérite d'être noté est la corrélation qui semble exister entre le vecteur $\boldsymbol{\eta}$ et le vecteur $\mathbf{sv}$. C'est un résultat intéressant car la *Shield Value*, bien que de nature spectrale, est basée sur la matrice d'adjacence. Ainsi, nous avons une mesure issue de la matrice Laplacienne qui est corrélée avec une autre reposant sur le spectre d'adjacence.

Pour mieux visualiser ces résultats et ces liens linéaires, le graphe KarateClub, pondéré avec les différents vecteurs $\boldsymbol{\eta}$, $\mathbf{fv}$, $\mathbf{sv}$, $\mathbf{ki}$, $\mathbf{bc}$ et $\mathbf{gm}$, fait l'objet de la figure 3.5. La figure 3.5a représente le graphe **KarateClub** pondéré grâce à l'algorithme **EIVP** et jouant le rôle de référence. Il est aisé de constater que l'évolution de la *Shield Value* (figure 3.5c) et la moyenne géométrique des degrés (figure 3.5f) sont bien corrélées avec la référence. Toutefois, l'évolution de la *Shield Value* est inversement corrélée à la variation relative de l'entropie, résultat pouvant également être lu dans le tableau 3.2. En effet, plus la *Shield Value* est importante, plus le graphe considéré est « vulnérable » [58, 62]. Compte-tenu de cette interprétation, la diminution de la *Shield Value* suite à la perturbation d'une arête révèle alors une meilleure robustesse du graphe. Ces deux mesures montrent par ailleurs que les arêtes en périphérie sont très vulnérables et que les arêtes les plus résistantes se trouvent au centre du graphe, dans les zones où la connectivité est élevée. Quant à la valeur de Fiedler (figure 3.5b) et l'indice de Kirchhoff (figure 3.5d), les évolutions de ces attributs spectraux montrent l'arête $\{1, 12\}$ comme la plus vulnérable, loin devant les autres arêtes. Ce constat est tout à fait normal car ces deux mesures ont pour intérêt principal de caractériser la connectivité du graphe. Or, la suppression de cette arête provoquerait une déconnexion du graphe. Cela montre les limites de ces mesures dans le cadre de ce travail car, dans notre vision de la vulnérabilité informationnelle, la prise en compte de la seule connectivité n'est pas suffisante.

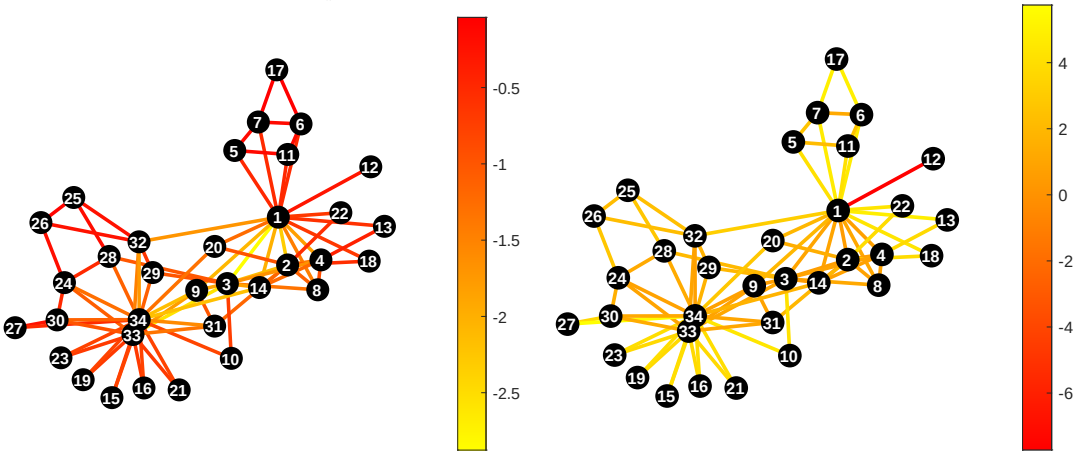
3.4 Approximations de l'entropie

Dans ce chapitre, les graphes traités avec l'algorithme **EIVP** sont, jusqu'à présent, de taille relativement restreinte, ce dernier relevant d'une complexité en $O(mn^3)$. En effet, il nécessite le recalcul du spectre Laplacien dans son intégralité (complexité approximativement cubique en le nombre de sommets n) pour les m arêtes perturbées voire supprimées. Cette complexité est rédhibitoire pour de nombreux graphes réels qui sont bien souvent de taille beaucoup plus importante. Pour pallier cette contrainte, il est possible d'envisager une approximation de l'entropie de von Neumann de graphes. À chaque étape de l'algorithme **EIVP**, il est question de calculer l'entropie du graphe perturbé donnée par



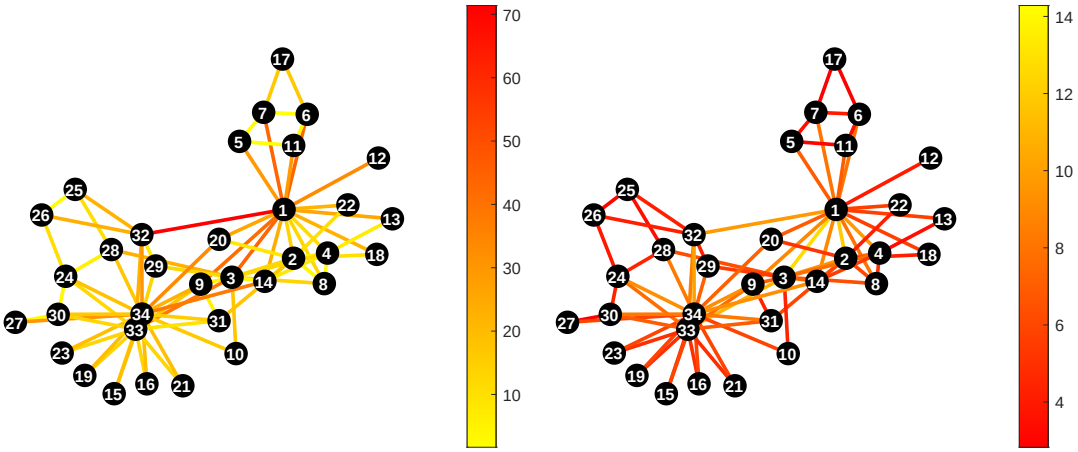
(a) Variation relative de l'entropie (vecteur η)

(b) Évolution de la valeur de Fiedler (vecteur $\mathbf{fv}$)



(c) Évolution de la *Shield Value* (vecteur $\mathbf{sv}$)

(d) Évolution de l'indice de Kirchhoff (vecteur $\mathbf{ki}$)



(e) Centralité d'intermédiarité des arêtes (vecteur $\mathbf{bc}$)

(f) Moyenne géométrique des degrés (vecteur $\mathbf{gm}$)

Figure 3.5 – Pondérations du graphe **KarateClub** avec les différents vecteurs η, $\mathbf{fv}$, $\mathbf{sv}$, $\mathbf{ki}$, $\mathbf{bc}$ et $\mathbf{gm}$.

l'équation (3.6). Deux stratégies vont alors être explorées dans cette section visant à approximer cette entropie. Dans un premier temps, il est possible d'approcher la fonction $f(\widetilde{\nu}_\ell^{(ij,\xi)}) \coloneqq \widetilde{\nu}_\ell^{(ij,\xi)} \ln \widetilde{\nu}_\ell^{(ij,\xi)}$ avec un développement limité d'ordre 1 ou 2. Les travaux relatifs à cette réflexion et présentés dans la section suivante sont ceux de Han *et al.* [224], de Chen *et al.* [225] et de Choi *et al.* [226]. Dans un second temps, en remarquant que $\widetilde{\nu}_\ell^{(ij,\xi)}$ est une variable singulière puisque c'est une valeur propre, la théorie de perturbation matricielle est utilisée pour écrire une expression analytique de $\widetilde{\nu}_\ell^{(ij,\xi)}$ en fonction d'attributs du graphe initial G, accélérant grandement l'algorithme **EIVP** car le calcul du spectre ne sera plus requis à chaque étape.

3.4.1 Approche « approximations quadratiques »

Comme introduit précédemment, une première manière d'approcher l'entropie de von Neumann est d'approximer directement la fonction $f(\nu_\ell) \coloneqq \nu_\ell \ln \nu_\ell$. Ce point a fait l'objet de nombreux travaux [224–226]. Parmi les objectifs de ces travaux, le but principal est de retrouver $\mathrm{Tr}(\boldsymbol{\rho})$, égal à 1, ou la quantité $\mathrm{Tr}(\boldsymbol{\rho}^2)$, appelée pureté de l'état quantique $\boldsymbol{\rho}$ [245]. Cette dernière quantité peut être facilement calculée grâce aux attributs structurels du graphe G avec l'égalité du lemme suivant.

Lemme 3.1 (Lemme 1, [226])**.** *Soit un graphe $G = (\mathcal{V}, \mathcal{E})$ de matrice de densité $\boldsymbol{\rho}$ et dont les arêtes $\{i,j\} \in \mathcal{E}$ sont potentiellement pondérées par des poids w_{ij}. Alors*

$$\mathrm{Tr}(\boldsymbol{\rho}^2) = \frac{1}{\mathrm{Tr}(\mathbf{L})^2} \left(Z_1(G) + 2 \sum_{\{i,j\}\in\mathcal{E}} w_{ij}^2 \right) \tag{3.14}$$

où $Z_1(G)$ est l'indice de Zagreb du graphe G (équation (1.8)*).*

Grâce à ce résultat, il est clair que la complexité du calcul de $\mathrm{Tr}(\boldsymbol{\rho}^2)$ est alors en $O(n+m)$ qui est toujours inférieure à la complexité initiale en $O(n^2)$. En effet, pour tout graphe, l'inégalité $m < n(n-1)/2 < n^2$ est vérifiée. De plus, dans le cas où le graphe est non pondéré (c'est-à-dire $w_{ij} = 1$ pour toute arête $\{i,j\} \in \mathcal{E}$), la complexité chute encore, atteignant $O(n)$, car un corollaire du lemme précédent nous donne

$$\mathrm{Tr}(\boldsymbol{\rho}^2) = \frac{1}{(2m)^2} \left(Z_1(G) + 2m \right). \tag{3.15}$$

Nous allons également avoir besoin de la quantité $\mathrm{Tr}(\widetilde{\boldsymbol{\rho}}^2)$ dans les futurs calculs d'approximations, où $\widetilde{\boldsymbol{\rho}}$ est la matrice de densité du graphe perturbé. Une expression permettant de la calculer rapidement est obtenue par la proposition suivante.

Proposition 3.2. *Soit un graphe* $G = (\mathcal{V}, \mathcal{E}, \mathbf{W})$ *perturbé par une matrice* $\mathbf{E} = [e_{ij}]_{1 \leq i,j \leq n}$. *Un graphe perturbé* $\widetilde{G}$ *admettant* $\widetilde{\mathbf{W}} \coloneqq \mathbf{W} + \mathbf{E}$ *comme matrice de poids et* $\widetilde{\boldsymbol{\rho}}$ *comme matrice de densité est alors obtenu. La quantité* $\mathrm{Tr}(\widetilde{\boldsymbol{\rho}}^2)$ *du graphe perturbé* $\widetilde{G}$ *peut être exprimée par*

$$\mathrm{Tr}(\widetilde{\boldsymbol{\rho}}^2) = \frac{Z_1(G) + \sum_{i=1}^{n} \left[\left(\sum_{j=1}^{n} e_{ij} \right)^2 + \sum_{j=1}^{n} 2s(i)e_{ij} + (w_{ij} + e_{ij})^2 \right]}{\left(\mathrm{Tr}(\mathbf{L}) + \sum_{1 \leq i,j \leq n} e_{ij} \right)^2}. \tag{3.16}$$

Preuve. Après perturbation, le i^{e} sommet admet une force (équation (1.16)) égale à $\sum_{j=1}^{n} w_{ij} + e_{ij}$. Ainsi, la matrice Laplacienne $\widetilde{\mathbf{L}}$ du graphe perturbé s'écrit, grâce à ses coefficients, comme suit :

$$\widetilde{\mathbf{L}} = (\widetilde{l}_{ij})_{1 \leq i,j \leq n} = \begin{cases} l_{ii} + \sum_{j=1}^{n} e_{ij}, & \text{si } i = j \\ l_{ij} - e_{ij}, & \text{si } i \neq j. \end{cases} \tag{3.17}$$

Soit $\widetilde{\boldsymbol{\rho}} = \widetilde{\mathbf{L}} / \mathrm{Tr}(\widetilde{\mathbf{L}})$ la matrice de densité associée au graphe perturbé. Ainsi, il vient

$$\mathrm{Tr}(\widetilde{\boldsymbol{\rho}}^2) = \frac{\mathrm{Tr}(\widetilde{\mathbf{L}}^2)}{\mathrm{Tr}(\widetilde{\mathbf{L}})^2}. \tag{3.18}$$

Or,

$$\mathrm{Tr}(\widetilde{\mathbf{L}}) = \mathrm{Tr}(\mathbf{L}) + \sum_{1 \leq i,j \leq n} e_{ij} \tag{3.19}$$

et

$$\begin{aligned}
\mathrm{Tr}(\widetilde{\mathbf{L}}^2) &= \sum_{1 \leq i,j \leq n} \widetilde{l}_{ij}^2 \\
&= \sum_{\substack{1 \leq i,j \leq n \\ i=j}} \widetilde{l}_{ij}^2 + \sum_{\substack{1 \leq i,j \leq n \\ i \neq j}} \widetilde{l}_{ij}^2 \\
&= \sum_{i=1}^{n} \left(l_{ii} + \sum_{j=1}^{n} e_{ij} \right)^2 + \sum_{\substack{1 \leq i,j \leq n \\ i \neq j}} (l_{ij} - e_{ij})^2 \\
&= \sum_{i=1}^{n} l_{ii}^2 + 2 \sum_{1 \leq i,j \leq n} l_{ii} e_{ij} + \sum_{i=1}^{n} \left(\sum_{j=1}^{n} e_{ij} \right)^2 + \sum_{\substack{1 \leq i,j \leq n \\ i \neq j}} l_{ij}^2 - 2 l_{ij} e_{ij} + e_{ij}^2 \\
&= Z_1(G) + \sum_{1 \leq i,j \leq n} 2s(i)e_{ij} + w_{ij}^2 + 2w_{ij}e_{ij} + e_{ij}^2 + \sum_{i=1}^{n} \left(\sum_{j=1}^{n} e_{ij} \right)^2
\end{aligned}$$

$$= Z_1(G) + \sum_{i=1}^{n} \left[\left(\sum_{j=1}^{n} e_{ij} \right)^2 + \sum_{j=1}^{n} 2s(i)e_{ij} + (w_{ij} + e_{ij})^2 \right]. \tag{3.20}$$

Le résultat (3.16) souhaité est obtenu en insérant (3.19) et (3.20) dans (3.18). ■

Ainsi, $\mathrm{Tr}(\widetilde{\boldsymbol{\rho}}^2)$ ne dépend que des éléments de la matrice de poids du graphe initial ainsi que des valeurs de perturbations.

Revenons aux approximations de l'entropie de von Neumann $S(G)$ d'un graphe G (équation (3.4)). Une première idée pour l'approcher autour d'un point $0 < \nu_* \leq 1$ serait de considérer le développement limité à l'ordre 1 de $\ln \nu_\ell$ au voisinage de ν_* :

$$\ln \nu_\ell \underset{\nu_*}{\sim} \ln \nu_* + \frac{1}{\nu_*}(\nu_\ell - \nu_*). \tag{3.21}$$

En multipliant ce développement limité par ν_ℓ, on obtiendrait

$$\nu_\ell \ln \nu_\ell \underset{\nu_*}{\sim} \nu_\ell(\ln \nu_* - 1) + \frac{1}{\nu_*}\nu_\ell^2. \tag{3.22}$$

Pour obtenir l'entropie approchée, il suffirait alors de calculer

$$\begin{aligned} -\sum_{\ell=1}^{n} \nu_\ell \left(\ln \nu_* - 1\right) + \frac{1}{\nu_*}\nu_\ell^2 &= (1 - \ln \nu_*) \sum_{\ell=1}^{n} \nu_\ell - \frac{1}{\nu_*} \sum_{\ell=1}^{n} \nu_\ell^2 \\ &= (1 - \ln \nu_*)\, \mathrm{Tr}(\boldsymbol{\rho}) - \frac{1}{\nu_*}\, \mathrm{Tr}(\boldsymbol{\rho}^2) \\ &= 1 - \ln \nu_* - \frac{1}{\nu_*}\, \mathrm{Tr}(\boldsymbol{\rho}^2). \end{aligned} \tag{3.23}$$

C'est, en substance, ce qu'ont proposé Han *et al.* [224] dans leur article bien qu'ils aient considéré la matrice Laplacienne normalisée. Dans [224], la valeur ν_* qu'ils choisissent est égale à 1, donnant lieu à ce qu'ils appellent l'entropie quadratique, approchant l'entropie de von Neumann $S(G)$:

$$S_{\mathrm{Han}}(G) = 1 - \mathrm{Tr}(\boldsymbol{\rho}^2). \tag{3.24}$$

Toutefois, approximer $\nu_\ell \ln \nu_\ell$ n'est pas équivalent à multiplier par ν_ℓ le développement limité de $\ln \nu_\ell$, qui plus est autour d'un point $\nu_* = 1$. En effet, le voisinage du point d'approximation dans lequel le développement limité est calculé est l'endroit où l'erreur entre la fonction et son développement limité est minimale. Si ce point est égal à 1, cela revient à penser que les valeurs propres (variables ν_ℓ) sont concentrées autour de 1, ce qui ne peut pas être le cas car la somme de ces dernières vaut 1.

Une variante de cette approximation, appelée FINGER (pour ***F**ast **IN**cremental von Neumann **G**raph **Ent**__R__opy*, a été proposée par Chen *et al.* [225]. Dans leur article, à la place de l'approximation (3.22) de la fonction $\nu_\ell \ln \nu_\ell$, les auteurs en définissent une autre à partir de la formule suivante

$$\nu_\ell \ln \nu_\ell \approx \ln(\nu_{\max})\nu_\ell(1-\nu_\ell). \tag{3.25}$$

Cette expression donne ainsi lieu à l'approximation de l'entropie définie comme suit :

$$S_{\text{Chen}}(G) = -\ln(\nu_{\max})\left[1 - \operatorname{Tr}(\boldsymbol{\rho}^2)\right]. \tag{3.26}$$

Cette variante nécessite le calcul de la valeur propre maximale qui, bien que moins coûteuse à calculer que l'entièreté du spectre, en l'occurrence $O(m+n)$ opérations grâce à une méthode de puissance itérée [246, 247], n'est pas nécessairement requise pour avoir une bonne approximation.

Toutes ces approximations ont fait l'objet de l'article complet et détaillé de Choi *et al.* [226]. Parmi elles, la solution permettant de pallier les difficultés mentionnées que nous allons considérer dans ce travail est basée sur un développement limité à l'ordre 2 de la fonction $\nu_\ell \ln \nu_\ell$ autour d'un point $0 < \nu_* \leq 1$ et fait l'objet de la proposition suivante [248].

Proposition 3.3. *Soit un graphe $G = (\mathcal{V}, \mathcal{E})$ admettant n sommets et m arêtes. Une approximation de l'entropie de von Neumann de G est donnée par*

$$S_{\text{DL},\nu_*}(G) = \frac{n\nu_*}{2} - \ln \nu_* - \frac{1}{2\nu_*}\operatorname{Tr}(\boldsymbol{\rho}^2). \tag{3.27}$$

Preuve. Soit le développement limité à l'ordre 2 de la fonction $\nu_\ell \ln \nu_\ell$ autour d'un point $0 < \nu_* \leq 1$:

$$\nu_\ell \ln \nu_\ell \underset{\nu_*}{\sim} \nu_* \ln \nu_* + (\nu_\ell - \nu_*)(\ln \nu_* + 1) + \frac{1}{2\nu_*}(\nu_\ell - \nu_*)^2 = -\frac{\nu_*}{2} + \nu_\ell \ln \nu_* + \frac{1}{2\nu_*}\nu_\ell^2. \tag{3.28}$$

Ainsi, ce développement limité permet d'obtenir l'approximation de l'entropie suivante :

$$\begin{aligned} S_{\text{DL},\nu_*}(G) &= -\sum_{\ell=1}^{n}\left(-\frac{\nu_*}{2} + \nu_\ell \ln \nu_* + \frac{1}{2\nu_*}\nu_\ell^2\right) \\ &= \frac{n\nu_*}{2} - \ln \nu_* \sum_{\ell=1}^{n}\nu_\ell - \frac{1}{2\nu_*}\sum_{\ell=1}^{n}\nu_\ell^2 \\ &= \frac{n\nu_*}{2} - \ln \nu_* - \frac{1}{2\nu_*}\operatorname{Tr}(\boldsymbol{\rho}^2). \end{aligned}$$

■

Rappelons que la somme des valeurs propres vaut 1 et que, par conséquent, leur moyenne vaut $1/n$. Pour de grands graphes, la moyenne des valeurs propres va donc tendre vers 0 et la plupart des valeurs propres sera entre 0 et $1/n$. Il semble donc logique de calculer le développement limité (3.28) autour d'un point $\nu_* = 1/n$. De plus, les valeurs propres de graphes possédant la propriété de petit-monde (*cf.* page 34) sont souvent concentrées autour de leur moyenne $1/n$, ce qui apporte une raison de plus de considérer cette valeur particulière. Dans ce cas, l'approximation de l'entropie est alors

$$S_{\mathrm{DL},1/n}(G) = \frac{1}{2} + \ln n - \frac{n}{2}\operatorname{Tr}(\boldsymbol{\rho}^2). \tag{3.29}$$

Les trois approximations de la fonction $f(\nu_\ell) = \nu_\ell \ln \nu_\ell$, à savoir les équations (3.22), (3.25) et (3.28), sont représentées en figure 3.6 en supposant que le graphe considéré a 10 sommets et une valeur propre maximale $\nu_{\max} = 0.2$. Notant que l'approximation de Han *et al.* (équation (3.22)) étant indépendante d'attributs structurels et spectraux tels que le nombre de sommets ou la valeur propre maximale, elle sera vraisemblablement moins pertinente. L'approximation de Chen *et al.* (équation (3.25)) prend en compte la valeur propre la plus grande mais si le spectre est bien distribué autour de $1/n$, cette approximation commet une erreur significative. Il peut être constaté grâce aux courbes de la figure 3.6 que la meilleure approximation présentée dans ce travail est celle basée sur un développement limité (équation (3.28)) car, par construction, elle est tangente à la fonction $f(\nu_\ell)$ autour de la valeur propre moyenne $1/n$ et donc plus précise si les valeurs propres sont autour de cette valeur moyenne. Or, dans le cas où des réseaux réels sont étudiés, cette assertion n'est pas déraisonnable [226].

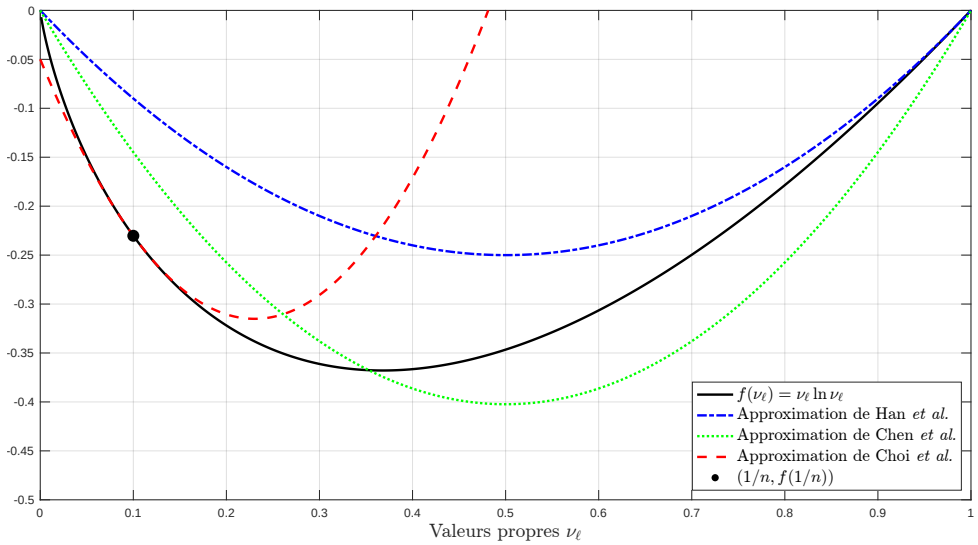


Figure 3.6 – Différentes approximations de la fonction $f(\nu_\ell) = \nu_\ell \ln \nu_\ell$ avec G possédant $n = 10$ sommets et une valeur propre maximale $\nu_{\max} = 0.2$.

Ces approximations provenant de ces différents travaux peuvent alors être considérées en lieu et place de l'entropie de von Neumann dans l'algorithme **EIVP** (algorithme 4) pour accélérer le calcul de la carte de vulnérabilité du graphe étudié. Dans le cas de l'approximation de Han *et al.*, la matrice de poids en sortie de cet algorithme serait donc notée $\mathbf{W}^{\text{EIVP}}_{S_{\text{Han}}}$.

Cette sous-section s'achève enfin en apportant un premier élément de réponse à la question suivante : « À quelle(s) condition(s) l'entropie décroît-elle suite à la suppression d'une arête ? » Dans leurs travaux, Passerini et Severini ont étudié numériquement l'évolution de l'entropie $S(G)$ lorsque des arêtes étaient ajoutées au graphe initial [45]. Toutefois, il ne peut rien advenir de l'expression

$$S(G) - S(\widetilde{G}) = \sum_{\ell=1}^{n} \widetilde{\nu}_\ell \ln(\widetilde{\nu}_\ell) - \nu_\ell \ln(\nu_\ell). \tag{3.30}$$

Pour effectuer une analyse plus détaillée, il semble alors judicieux de faire appel aux approximations précédentes. C'est ce qu'ont proposé Minello *et al.* pour construire de manière itérative des graphes d'entropie maximale [61], en utilisant l'approximation de Han *et al.* [224]. Or, dans ce chapitre, nous nous intéressons non pas à l'ajout, mais bien à la perturbation d'une arête allant jusqu'à sa suppression. Alors, à la manière de Minello *et al.* dans [61], étudions l'évolution de l'entropie lorsqu'une arête $\{i, j\}$ est retirée à l'aide de l'approximation (3.29). Les coefficients de la matrice de perturbation $\mathbf{E}$ sont donc tous égaux à 0 sauf $e_{ij} = e_{ji} = -1$. Sachant cela, en posant $s_{ij} \coloneqq \deg(i) + \deg(j)$, calculons

$$\begin{aligned}
S_{\text{DL},1/n}(G) - S_{\text{DL},1/n}(\widetilde{G}) &= \ln\left(\frac{\operatorname{Tr}(\widetilde{\boldsymbol{\rho}}^2)}{\operatorname{Tr}(\boldsymbol{\rho}^2)}\right) \\
&= \ln\left(\frac{\dfrac{1}{2(m-1)} + \dfrac{1}{4(m-1)^2}\left(Z_1(G) - 2s_{ij} + 2\right)}{\dfrac{1}{2m} + \dfrac{1}{4m^2} Z_1(G)}\right) \\
&= \ln\left(\frac{m^2\left(2(m-1) + Z_1(G) - 2s_{ij} + 2\right)}{(m-1)^2\left(2m + Z_1(G)\right)}\right) \\
&= \ln\left(\frac{m^2}{(m-1)^2}\left(1 - \frac{2s_{ij}}{2m + Z_1(G)}\right)\right) \\
&= \ln\left(\frac{m^2}{(m-1)^2}\left(1 - \frac{s_{ij}}{2m^2 \operatorname{Tr}(\boldsymbol{\rho}^2)}\right)\right).
\end{aligned}$$

Par conséquent, en considérant l'approximation de l'entropie basée sur un développement limité, une condition pour observer une diminution de l'entropie lorsque l'arête $\{i, j\}$ est retirée est :

$$S_{\mathrm{DL},1/n}(G) - S_{\mathrm{DL},1/n}(\widetilde{G}) \geq 0 \Longrightarrow \frac{m^2}{(m-1)^2}\left(1 - \frac{s_{ij}}{2m^2\,\mathrm{Tr}(\boldsymbol{\rho}^2)}\right) > 1$$
$$\Longrightarrow \frac{s_{ij}}{2} \leq (2m-1)\,\mathrm{Tr}(\boldsymbol{\rho}^2). \tag{3.31}$$

La relation entre l'évolution de l'entropie approchée et la moyenne des degrés des sommets reliés par les arêtes testées, qui a été relevée et analysée dans l'étude de corrélation précédente, est alors de nouveau retrouvée.

3.4.2 Approche « théorie de perturbation matricielle »

Rappels sur la théorie de perturbation matricielle

Étudier l'évolution des valeurs et vecteurs propres d'une matrice $\mathbf{M}$ suite à des perturbations de ses coefficients constitue un domaine très actif depuis de nombreuses années. Ce problème, appelé analyse de sensibilité, a beaucoup d'intérêt en physique et en ingénierie [249]. En effet, les valeurs et vecteurs propres sont des quantités fondamentales pour étudier le comportement d'un système. Or, le système étant le graphe lui-même dans ce présent chapitre, les valeurs et vecteurs propres sont ceux de la matrice de densité du graphe. Soit une matrice $\mathbf{M} \in \mathcal{M}_n$ une matrice carrée de taille n et une fonction Φ qui agit sur $\mathbf{M}$. Cette section concerne l'étude du comportement de la fonction Φ suite à une faible perturbation $\mathbf{E}$ (où $\mathbf{E}$ est en réalité une matrice de perturbations) de la matrice $\mathbf{M}$, en d'autres termes comparer $\Phi(\mathbf{M})$ et $\Phi(\mathbf{M}+\mathbf{E})$ [250]. La perturbation visée étant celle d'une arête, pouvant aller jusqu'à sa suppression, le système est qualifié de robuste si la perturbation ne nuit pas au fonctionnement de ce dernier. Or, les notions de robustesse et de vulnérabilité, bien que subjectives, sont belles et bien antagonistes l'une de l'autre [62]. Soient $(\lambda_\ell)_{1\leq\ell\leq n}$ les n valeurs propres distinctes de $\mathbf{M}$. Nous évaluons la modification apportée à la valeur propre λ_ℓ suite à une perturbation de ses coefficients $(m_{kl})_{1\leq k,l\leq n}$. Soit $(\lambda_i, \mathbf{u}_i)$ (resp. $(\lambda'_j, \mathbf{v}_j)$) l'*eigenpair* de droite (resp. de gauche) de la matrice $\mathbf{M}$. Les vecteurs $\mathbf{u}_i$ et $\mathbf{v}_j$ sont bi-orthogonaux, c'est-à-dire que $\langle \mathbf{u}_i, \mathbf{v}_j \rangle = \delta_{ij}$ où δ_{ij} est le symbole de Kronecker (égal à 1 si $i = j$ et 0 sinon) et que $\mathbf{u}_i = \mathbf{v}_i$ si $\mathbf{M}$ est symétrique. Grâce au théorème de Hellmann-Feynman[5] [251] appliqué à un paramètre β, il suit

$$\frac{\partial \lambda_\ell}{\partial \beta} = \mathbf{v}_\ell^\top \frac{\partial \mathbf{M}}{\partial \beta}\mathbf{u}_\ell. \tag{3.33}$$

5. Soit H_λ l'hamiltonien d'un système quantique dépendant d'un paramètre λ. La fonction d'onde normalisée ψ_λ et l'énergie (valeur propre) E_λ de cette fonction dépendent également de ce paramètre. En mécanique quantique, le théorème de Hellmann-Feynman [251] stipule que

$$\frac{\partial E_\lambda}{\partial \lambda} = \langle \psi_\lambda | \frac{\partial H_\lambda}{\partial \lambda} | \psi_\lambda \rangle \tag{3.32}$$

Cette équation montre que la sensibilité d'une valeur propre de la matrice $\mathbf{M}$ par rapport à un paramètre β dépend bien de changements effectués sur la matrice $\mathbf{M}$ relativement à ce paramètre. Ainsi, et c'est ce qui doit être étudié ici, le paramètre β n'est autre qu'une entrée m_{kl} particulière de la matrice $\mathbf{M}$. En sachant que la dérivée de $\mathbf{M}$ par rapport à une entrée m_{kl} est égale à 1 au coefficient correspondant et à 0 sinon, soit

$$\frac{\partial \mathbf{M}}{\partial m_{kl}} = [\delta_{ik}\delta_{jl}], \tag{3.34}$$

alors la sensibilité de la valeur propre λ_ℓ est donnée par

$$\frac{\partial \lambda_\ell}{\partial m_{kl}} = \mathbf{v}_\ell^\top \frac{\partial \mathbf{M}}{\partial m_{kl}} \mathbf{u}_\ell = [\mathbf{v}_\ell]_k [\mathbf{u}_\ell]_l, \quad k, l \in \{1, \dots, n\} \tag{3.35}$$

où $[\mathbf{v}_\ell]_k$ désigne la k^{e} composante du vecteur propre de gauche $\mathbf{v}_\ell$ et où $[\mathbf{u}_\ell]_l$ désigne la l^{e} composante du vecteur propre de droite $\mathbf{u}_\ell$. La sensibilité de la valeur propre λ_ℓ par rapport à une entrée m_{kl} est dépendante d'un produit de composantes du vecteur propre de gauche et de droite correspondant. Il est alors possible de construire la matrice de sensibilité $\mathbf{\Gamma}_\ell$ dont les coefficients, définis par

$$[\mathbf{\Gamma}_\ell]_{kl} = \frac{\partial \lambda_\ell}{\partial m_{kl}} = [\mathbf{v}_\ell]_k [\mathbf{u}_\ell]_l, \qquad k, l \in \{1, \dots, n\}, \tag{3.36}$$

caractérisent la sensibilité par rapport auxdits coefficients. Si les entrées m_{kl} sont perturbées ($\widetilde{m}_{kl} = m_{kl} + \Delta m_{kl}$), la valeur propre λ_ℓ correspondante est perturbée comme suit :

$$\widetilde{\lambda}_\ell = \lambda_\ell + \Delta\lambda_\ell = \lambda_\ell + \sum_{k,l=1}^{n} \frac{\partial \lambda_\ell}{\partial m_{kl}} \Delta m_{kl} = \lambda_\ell + \sum_{k,l=1}^{n} [\mathbf{v}_\ell]_k [\mathbf{u}_\ell]_l \Delta m_{kl}. \tag{3.37}$$

Sensibilité des valeurs propres suite à la perturbation d'une arête

Supposons que la matrice $\mathbf{M}$ n'est autre que la matrice Laplacienne $\mathbf{L}$ du graphe G étudié. Ce graphe pouvant éventuellement être pondéré, partons de sa matrice de poids $\mathbf{W}$. La perturbation d'une arête $\{k, l\}$ induit la modification de deux coefficients de la matrice $\mathbf{W}$ (w_{kl} et w_{lk}). De manière à quantifier les effets d'une telle perturbation sur le spectre Laplacien, introduisons un paramètre $\xi \in [0, 1]$ tel que la matrice de poids du graphe perturbé soit définie à l'aide des coefficients suivants :

$$[\widetilde{\mathbf{W}}]_{ij} = \widetilde{w}_{ij} = \begin{cases} w_{ij}(1-\xi), & \text{si } \{i, j\} = \{l, k\} \\ w_{ij}, & \text{sinon.} \end{cases} \tag{3.38}$$

Ainsi, deux changements sont également introduits dans la matrice des degrés qui a maintenant des coefficients égaux à

$$[\widetilde{\mathbf{D}}]_{ii} = \begin{cases} \deg(i) - \xi w_{kl}, & \text{si } \{i,j\} = \{l,k\} \\ \deg(i), & \text{sinon.} \end{cases} \tag{3.39}$$

Naturellement, la matrice Laplacienne voit alors quatre de ses coefficients modifiés étant donné que la matrice Laplacienne du graphe perturbé est donnée par

$$\widetilde{\mathbf{L}} = \widetilde{\mathbf{D}} - \widetilde{\mathbf{W}}. \tag{3.40}$$

Un exemple de perturbation de la matrice Laplacienne d'un graphe peut être trouvé dans la Référence [252] où une évaluation des performances d'un réseau de distribution d'électricité est effectuée après perturbation des lignes électriques. La proposition suivante a pour sujet les changements dans le spectre Laplacien induits par la perturbation d'une seule arête $\{i,j\}$.

Proposition 3.4. *Supposons que le graphe G, de matrice Laplacienne $\mathbf{L}$ admettant $(\mu_\ell, \mathbf{u}_\ell)_{1\leq\ell\leq n}$ pour eigenpair, voit une de ses arêtes $\{i,j\}$ perturbée par un coefficient ξ, donnant lieu à un graphe perturbé $\widetilde{G}_{ij,\xi}$ de matrice Laplacienne $\widetilde{\mathbf{L}}$ et de matrice de densité $\widetilde{\boldsymbol{\rho}}$. Les valeurs propres $\widetilde{\nu}_\ell$ de $\widetilde{\boldsymbol{\rho}}$ sont alors données par*

$$\widetilde{\nu}_\ell = \frac{1}{\mathrm{Tr}(\mathbf{L}) - 2\xi w_{ij}} \left(\mu_\ell - \xi w_{ij} \left([\mathbf{u}_\ell]_i - [\mathbf{u}_\ell]_j\right)^2\right). \tag{3.41}$$

Preuve. Posons $\boldsymbol{\Delta} \coloneqq \widetilde{\mathbf{L}} - \mathbf{L}$ la différence entre la matrice Laplacienne du graphe initial et celle du graphe perturbé. L'arête $\{i,j\}$ étant perturbée, quatre coefficients de $\boldsymbol{\Delta}$ sont non-nuls :

$$[\boldsymbol{\Delta}]_{ij} = [\boldsymbol{\Delta}]_{ji} = \xi w_{ij}, \qquad [\boldsymbol{\Delta}]_{ii} = [\boldsymbol{\Delta}]_{jj} = -\xi w_{ij}.$$

Or, la matrice Laplacienne $\mathbf{L}$ est symétrique donc les vecteurs propres de gauche et de droite sont les mêmes. L'équation (3.37) nous donne ainsi

$$\begin{aligned} \widetilde{\mu}_\ell &= \mu_\ell + \sum_{k,l=1}^{n} [\mathbf{u}_\ell]_k [\mathbf{u}_\ell]_l [\boldsymbol{\Delta}]_{kl} \\ &= \mu_\ell - \xi w_{ij} \left([\mathbf{u}_\ell]_i^2 - 2[\mathbf{u}_\ell]_i [\mathbf{u}_\ell]_j + [\mathbf{u}_\ell]_j^2\right) \\ &= \mu_\ell - \xi w_{ij} \left([\mathbf{u}_\ell]_i - [\mathbf{u}_\ell]_j\right)^2 \end{aligned} \tag{3.42}$$

où les $(\widetilde{\mu}_\ell)_{1\leq\ell\leq n}$ désignent les valeurs propres de la matrice Laplacienne $\widetilde{\mathbf{L}}$.
La matrice de densité $\widetilde{\boldsymbol{\rho}}$ pouvant s'écrire $\widetilde{\mathbf{L}}/\,\mathrm{Tr}(\widetilde{\mathbf{L}})$ où $\mathrm{Tr}(\widetilde{\mathbf{L}}) = \mathrm{Tr}(\mathbf{L}) - 2\xi w_{ij}$, le résultat attendu est obtenu.

■

Comme la pondération w_{ij} est positive, il est clair que les valeurs propres sont toutes décroissantes suite à la perturbation d'une arête. Il faudra ainsi rester attentif pour que les valeurs propres $\widetilde{\nu}_\ell$ obtenues par (3.41) restent toutes positives. Toutefois, la question de la première valeur propre $\widetilde{\nu}_1$ ne se pose pas car la valeur propre ν_1 est nulle et correspond à un vecteur propre dont les coefficients sont constants. En effet, elle découle immédiatement de la première valeur propre μ_1 de la matrice Laplacienne dont une propriété a été rappelée en chapitre 1 (*cf* page 45).

Corollaire 3.5. *Dans le cas où le graphe est non pondéré, les vecteurs propres de sa matrice Laplacienne* $\mathbf{L}$ *vérifient, pour toute arête* $\{i,j\} \in \mathcal{E}$*,*

$$\sum_{\ell=1}^{n} \left([\mathbf{u}_\ell]_i - [\mathbf{u}_\ell]_j\right)^2 = 2. \tag{3.43}$$

Preuve. La proposition précédente et l'équation (3.42) nous donne

$$\widetilde{\mu}_\ell = \mu_\ell - \xi w_{ij} \left([\mathbf{u}_\ell]_i - [\mathbf{u}_\ell]_j\right)^2. \tag{3.44}$$

Supposons que l'arête $\{i,j\}$, non pondérée, soit supprimée (c'est-à-dire $\xi = 1$). On a alors

$$\widetilde{\mu}_\ell = \mu_\ell - \left([\mathbf{u}_\ell]_i - [\mathbf{u}_\ell]_j\right)^2. \tag{3.45}$$

Par conséquent, il peut être déduit grâce à l'équation (1.37) que

$$\begin{aligned}\sum_{\ell=1}^{n} \widetilde{\mu}_\ell = \sum_{\ell=1}^{n} \mu_\ell - \sum_{\ell=1}^{n} \left([\mathbf{u}_\ell]_i - [\mathbf{u}_\ell]_j\right)^2 &\Longrightarrow 2m - 2 = 2m - \sum_{\ell=1}^{n} \left([\mathbf{u}_\ell]_i - [\mathbf{u}_\ell]_j\right)^2 \\ &\Longrightarrow \sum_{\ell=1}^{n} \left([\mathbf{u}_\ell]_i - [\mathbf{u}_\ell]_j\right)^2 = 2.\end{aligned}$$

■

Algorithme fast–EIVP

Grâce au travail précédent, nous disposons d'une expression explicite permettant de calculer les valeurs propres du graphe perturbée à partir de celles du graphe initial, de sorte que le calcul du spectre Laplacien n'est pas requis à chaque étape de l'algorithme. Ainsi, l'algorithme **fast–EIVP** (algorithme 5) suivant peut être utilisé. Cet algorithme a une complexité de calcul en $O(n^3 + mn)$, réduisant significativement la complexité en $O(mn^3)$ de l'algorithme **EIVP** initial.

Algorithme 5 Algorithme **fast–EIVP**

Entrée : Graphe $G = (\mathcal{V}, \mathcal{E}, \mathbf{W})$ et coefficient de perturbation $\xi \in [0, 1]$
Sortie : Graphe $G_S^{\text{f-EIVP}} = \left(\mathcal{V}, \mathcal{E}, \mathbf{W}_S^{\text{f-EIVP}}\right)$

1: Calcul des *eigenpairs* de la matrice Laplacienne $(\mu_\ell, \mathbf{u}_\ell)$
2: Calcul de l'entropie $S(G)$ du graphe initial ▷ Équation (3.4)
3: **for** $\{i, j\} \in \mathcal{E}$ **do**
4: $\quad S(\widetilde{G}_{ij,\xi}) \leftarrow 0$
5: $\quad$ **for** $1 \leq \ell \leq n$ **do**
6: $\quad\quad \widetilde{\nu}_\ell = \left(\mu_\ell - \xi w_{ij} \left([\mathbf{u}_\ell]_i - [\mathbf{u}_\ell]_j\right)^2\right) / (\mathrm{Tr}(\mathbf{L}) - 2\xi w_{ij})$ ▷ Équation (3.41)
7: $\quad\quad S(\widetilde{G}_{ij,\xi}) \leftarrow S(\widetilde{G}_{ij,\xi}) - \widetilde{\nu}_\ell \ln \widetilde{\nu}_\ell$
8: $\quad \left[\mathbf{W}_S^{\text{f-EIVP}}\right]_{ij} \leftarrow 100 \times \frac{S(\widetilde{G}_{ij,\xi}) - S(G)}{S(G)}$ ▷ Équation (3.7)

3.4.3 Erreurs et temps de calcul des différentes approximations

L'objectif de cette section est de montrer que les approximations (notamment celle donnant lieu à l'algorithme **fast-EIVP**) peuvent être utilisées de manière efficace pour tous les graphes, quelles que soient leur forme et leur taille. Pour cela, les algorithmes suivants ont été mis à l'épreuve sur des graphes aléatoires d'Erdös-Renyi $\mathcal{G}_{n,m}$ [83] :

— Algorithme **EIVP** avec l'entropie de von Neumann exacte (équation (3.4)) noté simplement « EIVP ». La complexité théorique de cet algorithme est en $O(mn^3)$;

— Algorithme **EIVP** avec l'approximation par entropie quadratique de Han *et al.* (équation (3.24)) noté « Han-EIVP ». La complexité théorique de cet algorithme est en $O(mn + m^2)$;

— Algorithme **EIVP** avec l'approximation par développement limité de Choi *et al.* repris dans ce manuscrit (équation (3.29)) noté « DL-EIVP ». La complexité théorique de cet algorithme est en $O(mn + m^2)$;

— Algorithme **fast-EIVP** basé sur l'approximation des valeurs propres (équation (3.41)) noté « fast-EIVP ». La complexité théorique de cet algorithme est en $O(mn + n^3)$.

Afin de garantir des figures claires et lisibles, il a été décidé de ne pas tester l'algorithme **EIVP** avec l'approximation de Chen appelée FINGER (équation (3.26)) principalement car elle serait, de manière évidente, moins efficiente que les autres approximations dû au calcul requis de la valeur propre maximale. Il est néanmoins important de noter que des simulations ont été effectuées et que la fonction Matlab `eigs` permet d'obtenir la plus grande valeur propre plus rapidement que le spectre entier si n est suffisamment grand (approximativement $n > 200$). De plus, les complexités théoriques ont été rappelées mais, en pratique, le calcul des valeurs propres a une complexité inférieure à celle en $O(n^3)$ car les matrices manipulées sont creuses et symétriques.

Évaluons les résultats à l'aide des critères de performance classiques : temps de calcul[6] et erreur commise, ici par rapport à la sortie de l'algorithme **EIVP** avec l'entropie de von Neumann exacte. Les temps de calcul affichés ci-après en figure 3.7 et figure 3.8 sont le résultat d'une moyenne sur 10 essais et les erreurs sont les erreurs absolues moyennes calculées entre les matrices de poids que renvoient les algorithmes. Par exemple, l'« Erreur Han-EIVP vs EIVP » est calculée à partir de

$$\frac{1}{n^2} \sum_{1 \leq i,j \leq n} \left| \left[\mathbf{W}_S^{\text{EIVP}} \right]_{ij} - \left[\mathbf{W}_{S_{\text{Han}}}^{\text{EIVP}} \right]_{ij} \right|, \tag{3.46}$$

et l'« Erreur fast-EIVP vs EIVP » à l'aide de

$$\frac{1}{n^2} \sum_{1 \leq i,j \leq n} \left| \left[\mathbf{W}_S^{\text{EIVP}} \right]_{ij} - \left[\mathbf{W}_S^{\text{f-EIVP}} \right]_{ij} \right|. \tag{3.47}$$

Fixons dans un premier temps le nombre de sommets à $n = 100$ et faisons varier le nombre d'arêtes m de 100 à 4950 afin de passer d'un graphe particulièrement creux au graphe complet $\mathcal{K}_{100}$. Les courbes de temps de calcul et d'erreurs pour ce scénario sont tracées en figure 3.7. Considérons d'abord les temps de calcul. Dans ce cas d'étude, notons que le nombre d'arêtes admet $n(n-1)/2$ comme borne supérieure, atteinte pour un graphe complet. Si la borne inférieure est fixée à n afin d'éviter le cas où le graphe est trop creux, il est possible d'établir l'inégalité $mn + n^3 < mn + m^2 < mn^3$ relative aux complexités des différents algorithmes, qui peuvent être observées grâce aux différentes courbes orange sur la figure 3.7. L'algorithme **EIVP** avec l'entropie exacte a la charge de calcul la plus importante, contrairement à l'algorithme **fast–EIVP** qui est le plus performant en termes de temps de calcul. Quant à l'algorithme **EIVP** avec l'approximation de Han et l'algorithme **EIVP** avec l'approximation par développement limité, les temps de calcul sont similaires car les complexités le sont également. Il est d'ores et déjà clair que l'utilisation d'approximations de l'entropie de von Neumann est nécessaire pour atteindre des temps de calcul acceptables. En guise d'illustration, l'analyse de vulnérabilité d'un graphe possédant 100 sommets et 4950 arêtes prend, avec l'entropie exacte, plus d'une seconde, ce qui est contrariant pour des graphes plus grands pouvant être rencontrés dans le monde des réseaux réels.

Quant aux erreurs, elles diminuent toutes à mesure que le graphe devient dense, avec des valeurs d'erreur absolue moyenne autour de 10^{-3} à partir d'un petit nombre d'arêtes, ce qui est tout à fait satisfaisant. Une particularité à noter est l'erreur nulle qui est obtenue lorsque le graphe est complet avec l'algorithme **EIVP** basé sur un développement limité. En effet, cette approximation est construite de manière à être tangente à la fonction $f(\nu_\ell) = \nu_\ell \ln \nu_\ell$ au point $1/n$. Or, dans un graphe complet,

6. Code Matlab exécuté sur un PC doté d'un processeur Intel Core i9-10900F @ 2.80 GHz (10 cœurs, 20 threads).

toutes les valeurs propres sont égales à $1/n$, induisant alors une erreur nulle. Pour résumer, l'algorithme **fast–EIVP** est le plus efficace en termes d'erreurs commises. Il faut cependant garder à l'esprit que l'analyse de l'erreur proposée ici est globale, car c'est une erreur moyenne. Il faudrait donc veiller à ce que les saillances des arêtes proposées par les algorithmes basés sur des approximations gardent la même tendance que celui basé sur l'entropie exacte. Pour cela, rien de mieux qu'une étude exhaustive arête par arête.

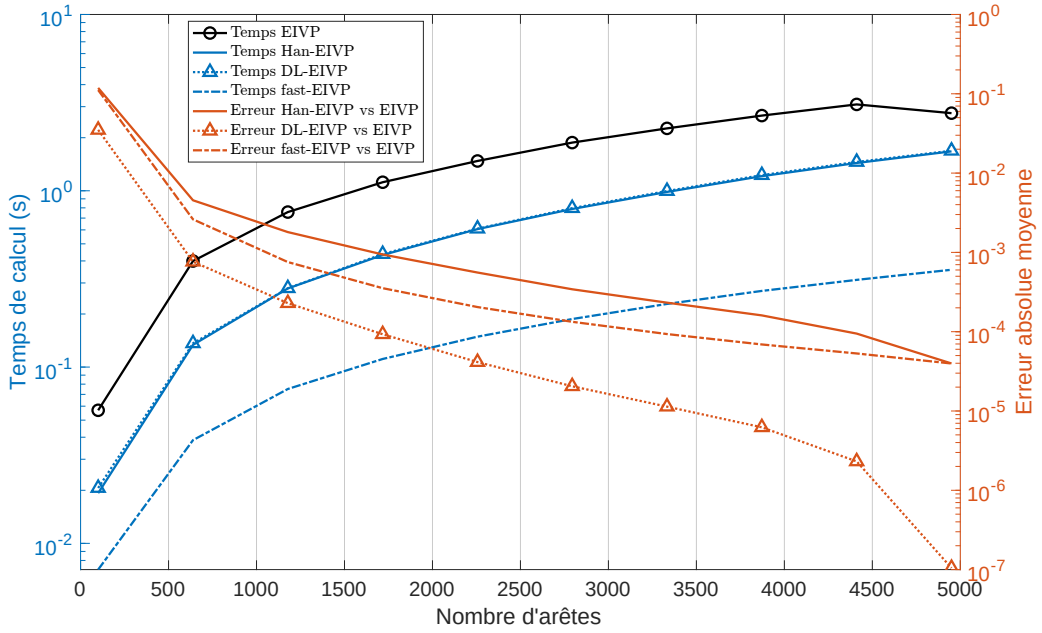


Figure 3.7 – Temps de calcul (courbes **bleues**) et erreurs (courbes **orange**) des algorithmes **EIVP** sur des graphes aléatoires d'Erdös-Renyi admettant un nombre de sommets égal à 100 et un nombre d'arêtes variant de 100 à 4950 (qui n'est autre que le graphe complet $\mathcal{K}_{100}$).

Bien que moins pertinentes d'un point de vue physique que les résultats précédents, des simulations ont également été réalisées en fixant le nombre d'arêtes (ici, à 500) et en faisant varier le nombre de sommets. Les courbes des temps de calcul et des erreurs sont représentées en figure 3.8. La dépendance « cubique » relative au calcul des valeurs propres peut être constatée pour l'algorithme **EIVP** avec l'entropie de von Neumann exacte. Il en est de même pour la dépendance linéaire des autres algorithmes **EIVP** mis en compétition. La complexité algorithmique en $O(n^3)$ requise au début de l'algorithme **fast–EIVP** ralentit le calcul à mesure que le nombre de sommets augmente. Toutefois, il ne semble pas pertinent de considérer des graphes avec un rapport n/m aussi grand. Une fois de plus, avec ce scénario, cet algorithme apparaît donc comme le plus performant en termes de temps de calcul. La remarque précédente sur les deux courbes superposées des approximations de Han *et al.* [224] et de Choi *et al.* [226] est la même dans ce cas. Quant aux erreurs, elles augmentent toutes comme prévu. En

effet, plus le graphe devient creux, plus le nombre d'arêtes jouant le même rôle augmente. En d'autres termes, le nombre d'arêtes qui seront responsables d'une isolation des sommets augmente nécessairement. L'erreur commise par les approximations sur ce type d'arête sera alors multipliée d'autant de fois qu'il y a d'arêtes de ce type.

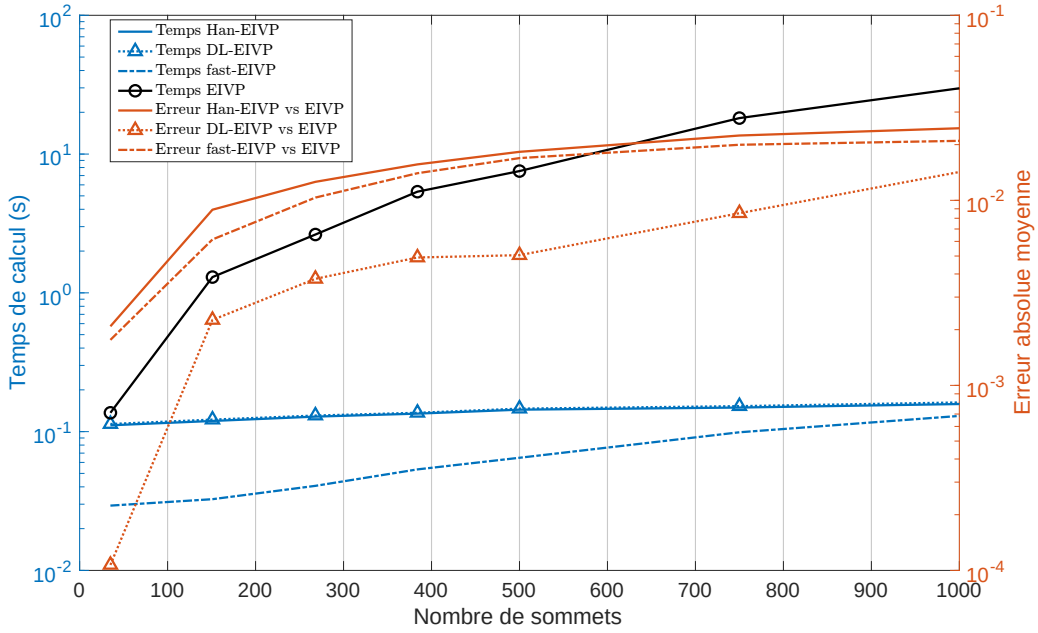


Figure 3.8 – Temps de calcul (courbes **bleues**) et erreurs (courbes **orange**) des algorithmes **EIVP** sur des graphes aléatoires d'Erdös-Renyi admettant un nombre d'arêtes égal à 500 et un nombre de sommets variant de 35 à 1000.

Bien entendu, ces mêmes analyses peuvent être effectuées avec les graphes réels présentés précédemment (*cf.* page 122). Les temps de calcul sont reportés dans le tableau 3.3. Il peut être constaté que, pour tous les graphes, l'algorithme **fast-EIVP** est le plus rapide d'un facteur au moins égal à 3, ce qui est logique étant donné sa complexité plus faible que les autres. De plus, **Han-EIVP** et **DL-EIVP** ont des temps de calcul similaires, observation qui avait déjà été faite lors de l'étude des graphes synthétiques précédents. Enfin, pour les graphes **Jazz Musicians** et **SciGrid**, ainsi que pour **Random Sensor**, la dépendance cubique en le nombre de sommets se révèle quantitativement être une contrainte. Examinons à présent les erreurs commises par ces différentes approximations et répertoriées dans le tableau 3.4. L'algorithme **DL-EIVP** n'est pas toujours le plus performant, bien que les résultats sur les graphes synthétiques aient pu le suggérer. En effet, l'algorithme **fast-EIVP** affiche de meilleurs résultats avec des erreurs au moins réduites de moitié, probablement parce que les graphes étudiés ne possèdent pas tous la propriété de petit monde, les valeurs propres n'étant pas toutes situées à proximité immédiate de la moyenne $1/n$. En conclusion, tous ces résultats montrent que les approximations introduites, en particulier celle de **fast-EIVP**, permettent d'obtenir des temps de cal-

cul véritablement réduits, avec quelques dizaines de millisecondes nécessaires au calcul de ce dernier, tout en commettant des erreurs moyennes convenables aux alentours de 10^{-2}.

	EIVP	Han-EIVP	DL-EIVP	fast-EIVP
KarateClub	26	20	21	**7**
Chesapeake Bay	55	42	41	**11**
Jazz musicians	10121	1054	1062	**199**
SciGrid	11599	232	235	**85**
Les Misérables	108	65	65	**15**
Random sensor	6832	364	370	**88**

Tableau 3.3 – Temps de calcul (en millisecondes) des différents algorithmes **EIVP** (avec ou sans approximations).

EIVP vs	Han-EIVP	DL-EIVP	fast-EIVP
KarateClub	0.1096	0.1288	**0.0632**
Chesapeake Bay	0.0367	0.0511	**0.0156**
Jazz musicians	0.0021	0.0011	**0.0002**
SciGrid	0.0132	**0.0091**	0.0093
Les Misérables	0.0538	0.1063	**0.0112**
Random sensor	0.0058	**0.0020**	0.0026

Tableau 3.4 – Erreur moyenne entre les saillances calculées grâce à l'algorithme **EIVP** et celles issues d'approximations

3.5 Conclusion

Dans ce chapitre, une notion subjective, n'admettant pas, à notre connaissance, de définition établie dans la littérature, a été abordée : la vulnérabilité des arêtes dans un graphe. Pour apporter un élément de réponse, l'entropie de von Neumann de graphes a été considérée pour quantifier la « vulnérabilité informationnelle » des liens dans un réseau. L'idée est simple : les liens sont perturbés individuellement (allant jusqu'à leurs suppressions) et l'impact sur l'entropie de von Neumann globale du graphe suite à cette perturbation est mesuré. Pour ce faire, nous avons introduit la notion de saillance d'une arête calculée comme la variation relative de l'entropie issue de cette perturbation. Ensuite, nous avons développé un algorithme, appelé **EIVP** (pour ***E**dge **I**nformational **V**ulnerability to **P**erturbation*), permettant de créer une carte de vulnérabilité du graphe étudié. Cette carte facilite l'analyse du réseau et permet de mettre en évidence visuellement la saillance des arêtes. Ainsi, cet algorithme de repondération fournit des informations sur les structures vulnérables cachées du réseau, au sens du contenu

informationnel. En effet, l'algorithme **EIVP** est très sensible aux risques de déconnexion qui pourraient survenir dans la structure, en classant comme très vulnérables les arêtes reliant des zones fragiles et faciles à isoler. Au contraire, il tend à caractériser des arêtes structurellement redondantes d'un point de vue informationnel. Il va sans dire que cet algorithme pourrait se révéler très utile pour les applications liées à la vulnérabilité des infrastructures et des réseaux logistiques en termes d'accessibilité et de résilience face aux attaques et aux défaillances. Toutefois, une limitation de cet algorithme est que le calcul de l'entropie de von Neumann sous-jacent relève d'une complexité en $O(n^3)$, ce qui peut être problématique dans le cas de grands graphes. C'est la raison pour laquelle nous avons utilisé des approximations classiques de l'entropie et avons introduit une approximation basée sur la théorie de perturbations matricielles. Cette dernière permet d'obtenir un algorithme appelé **fast-EIVP** ayant une complexité en $O(mn + n^3)$, permettant de fait d'accélérer sensiblement le calcul des cartes de vulnérabilité, le tout en commettant une erreur contenue par rapport à l'algorithme **EIVP** initial, de l'ordre de 10^{-2}. En effet, les résultats obtenus sur de multiples graphes, synthétiques et réels, ont montré l'efficacité de la stratégie proposée. En résumé, l'algorithme **EIVP** attribue les poids (saillances) de manière cohérente et adaptée aux propriétés locales de la structure du graphe analysé. Des tests supplémentaires sur un plus grand nombre de graphes réels permettraient de découvrir d'autres propriétés de cette mesure de vulnérabilité basée sur le contenu informationnel du graphe vu comme un système.

Publications scientifiques relatives à ce chapitre

- ✍ **T. Averty**, D. Daré-Emzivat, A.-O. Boudraa et Y. Préaux. Approximation de l'entropie de von Neumann de graphes pour une analyse de vulnérabilité. *GRETSI*, pages 1–4, 2022
- ✍ **T. Averty**, Hadj-Ahmed Bay-Ahmed, D. Daré-Emzivat, A.-O. Boudraa et C. Richard. Identifying vulnerable links in large networks using von Neumann graph entropy. *IEEE Transactions on Signal and Information Processing over Networks*, 2025 (rédigé)

Chapitre 4

Généralisation des représentations conventionnelles de graphes

« Penser c'est oublier des différences, c'est généraliser, c'est abstraire. »

Jorge Luis Borges

4.1 Introduction

Les mesures de similarité jouent un rôle fondamental dans de nombreuses applications comme la fouille de données [253], la recherche d'image par le contenu [254, 255], l'apprentissage automatique [256–258], le partitionnement [259, 260], la classification [261, 262], la recherche d'informations [263], la fouille de texte [264], la comparaison des graphes [265, 266], la reconnaissance de formes [172], [267], [264], le traitement d'images [268, 269] ou encore le traitement de langage naturel [270]. Pour chacune de ces applications, il faut une mesure de similarité bien adaptée. Le choix de cette mesure dépend des caractéristiques des données et du contexte de l'étude [253, 255]. Pour certaines applications comme le partitionnement, la distance euclidienne est utilisée comme mesure de similarité, car c'est un moyen simple et intuitif pour quantifier la similarité entre les observations associées aux objets. Par contre, cette mesure est moins performante dans le cas de la recherche d'image par le contenu [271]. En apprentissage automatique, les fonctions à noyau telles que le noyau polynomial ou le noyau RBF sont considérées comme des fonctions de similarité les plus adaptées [272, 273]. Ainsi, il n'existe pas de mesure de similarité universelle.

Le concept de mesure de similarité est un concept général qui s'applique à une paire d'objets qui peuvent être des points, des chaînes de caractères, des densités probabilités, des graphes, des images ou des signaux. Il n'y a pas de définition unique d'une mesure de similarité, mais cette quantité peut être vue comme l'inverse d'une mesure de distance entre deux objets. Analyser la similarité entre deux objets se fait toujours relativement à un ensemble d'attributs qui caractérisent ces objets. Lorsque les objets à comparer sont représentés par des graphes, ce problème se ramène à mesurer la similarité de graphes, et se pose ainsi le problème de leur représentation.

Établir une mesure de similarité entre graphes est une tâche essentielle pour les comparer ou les classifier. Pour cette étape de classification, il existe, dans la littérature du domaine, un grand nombre de noyaux comparant les structures des différents graphes considérés dont les premiers ont été présentés par Gärtner *et al.* [170] et n'ont cessé d'être enrichis. Ces noyaux se basent sur l'analyse d'invariants propres aux graphes comme le nombre de sous-graphes type (*graphlet*) contenus dans la structure des graphes [31, 171], le nombre de marches aléatoires[1] apparaissant au sein de ces graphes [170] ou encore la recherche des plus courts chemins [30]. La classification qui va nous intéresser plus particulièrement dans ce chapitre est la classification spectrale de graphes à partir d'une mesure de similarité entre les spectres de leurs matrices de représentation respectives. Cette mesure est souvent intégrée dans un noyau RBF de SVM en lieu et place de la distance euclidienne usuelle. Comme cela a été rappelé dès le premier chapitre, deux écoles de pensée trouvent leurs places aujourd'hui dans l'étude de graphes *via* leur représentation matricielle : celle qui prône la matrice d'adjacence $\mathbf{A}$ pour des raisons de simplicité, tant elle traduit de manière immédiate un graphe en matrice [11], et celle qui étudie la matrice Laplacienne $\mathbf{L}$ pour des considérations physico-mathématiques, notamment grâce à la forme quadratique (1.35) qui en découle [10]. Le problème majeur avec ces matrices de représentation, dont une présentation détaillée est proposée au chapitre 1, réside dans l'existence de graphes cospectraux, c'est-à-dire des graphes possédant le même spectre au regard d'une matrice particulières. Il est ainsi compréhensible qu'il ne soit pas immédiat d'introduire une mesure de similarité pour comparer des graphes *via* leurs spectres. Quelques travaux du domaine ont tenté d'apporter des solutions à ce problème [32, 47, 51, 52]. Parmi ces derniers, Bay-Ahmed *et al.* ont développé deux mesures de similarité spectrale entre graphes, présentées dans la suite de ce chapitre, qu'ils ont inséré dans un noyau RBF de SVM en remplacement de la distance euclidienne : la Covariance Spectrale (CS) [32] et la Similarité Spectrale Conjointe (SSC) [53]. Mais pourquoi vouloir étudier les spectres alors que ce sont les graphes, et donc leurs structures, qui sont à comparer ? Le chapitre 1 s'appuie sur le fait que l'utilisation

1. Une marche aléatoire de longueur T sur un graphe G avec un point de départ au sommet $v^{(1)}$ est un processus stochastique avec des variables aléatoires $X^{(1)}, X^{(2)}, \ldots, X^{(T)}$ telles que $X^{(1)} = v^{(1)}$ et $X^{(t+1)}$ est un sommet choisi uniformément au hasard parmi les voisins de $X^{(t)}$. La notion d'uniformément au hasard signifie que, pour atteindre un sommet j depuis le sommet i, la probabilité est égale à $a_{ij} / \deg(i)$ où les a_{ij} sont les coefficients de la matrice d'adjacence.

des spectres des matrices de représentation permet une mise en avant efficace de certaines propriétés énergétiques et structurelles des graphes étudiés. Les contraintes que rencontrent ces deux mesures sont les suivantes : la CS ne compare qu'un seul spectre (d'adjacence en l'occurrence) ce qui semble insuffisant dans bien des cas et la SSC requiert quant à elle le calcul de deux spectres (celui d'adjacence et celui Laplacien) induisant un temps de traitement plus long. S'ensuit alors une question générale : est-ce que les matrices de représentation usuelles sont les meilleures pour mettre en avant l'information spectrale de graphes, permettant ainsi de les comparer ? Nous apportons des éléments de réponse à cette question en introduisant deux nouvelles matrices d'adjacence généralisées, à savoir la matrice $\mathbf{T}_\alpha$ et le plan $\mathbf{P}_{\alpha,k}$ qui, de concert avec une nouvelle mesure de similarité par corrélation spectrale appelée $\mathrm{CorS}_{\mathbf{R}}$, opère une classification spectrale efficiente.

4.2 Matrices d'adjacence généralisées

4.2.1 État de l'art

Toutes les matrices de représentation introduites au chapitre 1 présentent des propriétés bien spécifiques et sont utiles à bien des égards. Le problème de l'existence de graphes cospectraux a conduit Van Dam *et al.* à combiner certaines matrices de représentation pour aboutir à une matrice de représentation généralisée donnant lieu à moins de graphes cospectraux avec des propriétés spectrales intéressantes quant au graphe sous-jacent. Ils ont, pour cela, proposé une définition de la matrice d'adjacence généralisée d'un graphe $G = (\mathcal{V}, \mathcal{E})$ d'ordre n et possédant m arêtes[2] [49] :

$$\mathbf{G} \coloneqq a\mathbf{A} + b\mathbf{I} + c\mathbf{J}, \qquad a, b, c \in \mathbb{R}, a \neq 0 \tag{4.1}$$

où $\mathbf{I}$ désigne la matrice identité et $\mathbf{J}$ la matrice formée que de un. Quelques années plus tard, Haemers et Omidi ont créé une matrice d'adjacence dite universelle, plus générale encore, définie cette fois comme [50] :

$$\mathbf{U} \coloneqq a\mathbf{A} + b\mathbf{I} + c\mathbf{J} + d\mathbf{D}, \qquad a, b, c, d \in \mathbb{R}, a \neq 0. \tag{4.2}$$

Ces matrices sont intéressantes car, à l'aide de paramètres, elles permettent d'évaluer graduellement l'importance des matrices de représentation classiques, notamment à travers leurs contenus spectraux. L'attrait pour ce type de matrices se révèle lorsque, quelques années plus tard, sont apparues des ma-

2. Dans la suite de ce chapitre, sauf cas particuliers, nous considérerons un tel graphe.

trices de représentation telles que la matrice Laplacienne déformée $\mathbf{\Delta}_\alpha$ définie par [54] :

$$\mathbf{\Delta}_\alpha \coloneqq \alpha^2(\mathbf{D}-\mathbf{I}) - \alpha\mathbf{A} + \mathbf{I}, \qquad \alpha \in \mathbb{R} \tag{4.3}$$

ou encore la matrice d'α–adjacence de Nikiforov, définie comme une combinaison convexe linéaire de $\mathbf{D}$ et $\mathbf{A}$ [55] :

$$\mathbf{A}_\alpha \coloneqq \alpha\mathbf{D} + (1-\alpha)\mathbf{A}, \qquad 0 \leq \alpha \leq 1. \tag{4.4}$$

Rappelons qu'un des objectifs de ce travail est de développer une méthode de classification spectrale de graphes : la question de la polarité des valeurs propres considérées est alors primordiale, et donc la notion de semi-définie positivité des différentes matrices étudiées. Dans ce même article [55], Nikiforov discute ainsi du caractère semi-défini positif de sa matrice $\mathbf{A}_\alpha$. Un des points saillants de la réflexion est de trouver le plus petit α noté α_0 pour lequel $\mathbf{A}_\alpha$ est semi-définie positive pour tout $\alpha \geq \alpha_0$. Il est facile de démontrer que $\alpha_0 \leq 1/2$ mais il n'existe, à ce jour, des valeurs de α_0 que pour des cas particuliers de graphes [274]. Par exemple, pour le graphe complet $\mathcal{K}_n$, la matrice $\mathbf{A}_\alpha$ est semi-définie positive si et seulement si $\alpha \geq 1/n$ [274]. Pour prouver ces différentes assertions quant à la semi-définie positivité, Nikiforov [55] a utilisé les écritures équivalentes suivantes de la forme bilinéaire symétrique associée à la matrice d'α-adjacence $\mathbf{A}_\alpha$ d'un graphe $G = (\mathcal{V}, \mathcal{E})$ d'ordre n avec $\mathbf{x} = (x_i)_{1\leq i\leq n} \in \mathbb{R}^n$:

$$\begin{aligned}
\langle \mathbf{A}_\alpha\mathbf{x}, \mathbf{x}\rangle &= \sum_{(i,j)\in\mathcal{E}} (\alpha x_i^2 + 2(1-\alpha)x_ix_j + \alpha x_j^2) && (4.5)\\
&= (2\alpha-1)\sum_{i\in\mathcal{V}} x_i^2 \deg(i) + (1-\alpha)\sum_{(i,j)\in\mathcal{E}} (x_i+x_j)^2 && (4.6)\\
&= \alpha\sum_{i\in\mathcal{V}} x_i^2\deg(i) + 2(1-\alpha)\sum_{(i,j)\in\mathcal{E}} x_ix_j. && (4.7)
\end{aligned}$$

La matrice d'α-adjacence possède également des propriétés spectrales pertinentes notamment le fait que la ℓ^{e} valeur propre $\lambda_\ell(\mathbf{A}_\alpha)$ puisse s'écrire

$$\begin{aligned}
\lambda_\ell(\mathbf{A}_\alpha) &= \alpha\lambda_\ell(\mathbf{D}) + (1-\alpha)\lambda_\ell(\mathbf{A}) && (4.8)\\
&= \alpha\deg(\ell) + (1-\alpha)\lambda_\ell(\mathbf{A}), \qquad \forall \ell \in [\![1,n]\!] && (4.9)
\end{aligned}$$

ou encore le fait que les spectres se superposent, c'est-à-dire si $0 \leq \beta < \alpha \leq 1$, alors

$$\lambda_\ell(\mathbf{A}_\beta) \leq \lambda_\ell(\mathbf{A}_\alpha), \quad \forall \ell \in [\![1,n]\!]. \tag{4.10}$$

Il est également possible de retrouver les matrices $\mathbf{A}$, $\mathbf{D}$ et $\mathbf{Q}$ de manière assez directe à partir de la matrice d'α-adjacence $\mathbf{A}_\alpha$. En effet, on a $\mathbf{A}_0 = \mathbf{A}$, $\mathbf{A}_1 = \mathbf{D}$, $2\mathbf{A}_{1/2} = \mathbf{Q}$. Néanmoins, bien que l'égalité $\mathbf{A}_\alpha - \mathbf{A}_\beta = (\alpha - \beta)\mathbf{L}$ soit vérifiée, la matrice d'α-adjacence ne fait pas apparaître la matrice Laplacienne $\mathbf{L}$. L'idéal serait d'avoir une matrice de représentation généralisée qui fait apparaître $\mathbf{A}$, $\mathbf{D}$ et $\mathbf{L}$, en particulier pour étudier leurs spectres de manière unifiée. La matrice « α–Laplacienne » $\mathbf{L}_\alpha$ a alors été introduite par Wang *et al.* [56] :

$$\mathbf{L}_\alpha \coloneqq \alpha\mathbf{D} + (\alpha - 1)\mathbf{A}, \qquad 0 \leq \alpha \leq 1. \tag{4.11}$$

La matrice $\mathbf{L}$ apparait bien lorsque $\alpha = 1/2$ et $\mathbf{L}_1 = \mathbf{D}$ mais c'est $-\mathbf{A}$ qui peut être retrouvée lorsque $\alpha = 0$, ce qui peut être problématique lorsque les N premières valeurs propres du spectre (exploitées ultérieurement) sont considérées. En effet, il n'y a pas symétrie du spectre d'adjacence comme cela a été évoqué dans le chapitre 1 : les N premières valeurs propres de $-\mathbf{A}$ ne sont pas les N premières de $\mathbf{A}$ en inversion de polarité. Toutefois, un résultat très intéressant sur la matrice $\mathbf{L}_\alpha$ concerne son domaine de semi-définie positivité et est exprimé par le théorème suivant.

Théorème 4.1 (Théorème 3.5, [56])**.** *Soit G un graphe connecté d'ordre n et soit $\alpha_0(G)$ le plus petit α tel que la matrice α-Laplacienne $\mathbf{L}_\alpha(G)$ soit semi-définie positive. Alors $\alpha_0 = 1/2$.*

Ces matrices ont suscité beaucoup d'intérêt, notamment à travers l'introduction d'une α-énergie $E_{\mathbf{A}_\alpha}$ d'un graphe G possédant n sommets et m arêtes définie par [275, 276]

$$E_{\mathbf{A}_\alpha} = \sum_{\ell=1}^{n} \left| \lambda_\ell(\mathbf{A}_\alpha) - \frac{2\alpha m}{n} \right| = \sum_{\ell=1}^{n} \left| \lambda_\ell(\mathbf{A}_\alpha) - \alpha\bar{d}(G) \right|. \tag{4.12}$$

Cette expression permet de décrire les énergies classiques : l'énergie d'un graphe (1.26) est retrouvée dans le cas où $\alpha = 0$, la moitié de l'énergie Laplacienne sans-signe (1.48) est identifiée dans le cas où $\alpha = 1/2$ et c'est l'écart de degrés (1.4) qui est observé dans le cas où $\alpha = 1$. Il demeure que ces matrices, bien qu'elles aient des propriétés intéressantes, ne respectent pas toutes les conditions souhaitées dans le cadre de ce travail telles que le fait de décrire les matrices usuelles $\mathbf{A}$, $\mathbf{D}$ et $\mathbf{L}$ et n'ont, de surcroît, pas trouvé d'applications concrètes. D'où l'intérêt de la matrice définie dans la section suivante.

4.2.2 Matrice $\mathbf{T}_\alpha$

Pour respecter le critère relatif au fait de vouloir introduire une matrice paramétrique permettant de retrouver les matrices de représentation $\mathbf{A}$, $\mathbf{D}$ et $\mathbf{L}$ et possédant des propriétés algébriques attendues, nous nous sommes inspirés des matrices précédentes pour définir une nouvelle matrice de

représentation d'un graphe, notée $\mathbf{T}_\alpha$, dont l'expression est donnée par

$$\mathbf{T}_\alpha := \alpha\mathbf{D} + (1-2\alpha)\mathbf{A}, \qquad 0 \leq \alpha \leq 1. \tag{4.13}$$

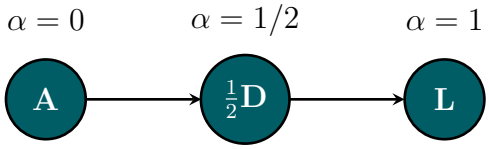


Figure 4.1 – Matrice $\mathbf{T}_\alpha$ décrivant les matrices classiques $\mathbf{A}$, $\mathbf{D}$ et $\mathbf{L}$.

Comme le montre le schéma de la figure 4.1, cette matrice vérifie en effet $\mathbf{T}_0 = \mathbf{A}$, $\mathbf{T}_1 = \mathbf{L}$ et $2\mathbf{T}_{1/2} = \mathbf{D}$. La positivité de cette matrice fait l'objet de la proposition suivante.

Proposition 4.2. *Si* $1/2 \leq \alpha \leq 1$*, alors la matrice* $\mathbf{T}_\alpha$ *est semi-définie positive.*

Preuve. La matrice $\mathbf{T}_\alpha$ peut être réécrite comme suit :

$$\mathbf{T}_\alpha = \alpha\mathbf{D} + (1-2\alpha)\mathbf{A} = \alpha\mathbf{D} + (1-2\alpha)(\mathbf{D}-\mathbf{L}) = (1+\alpha)\mathbf{D} + (2\alpha-1)\mathbf{L}.$$

En utilisant le fait que $\mathbf{D}$ et $\mathbf{L}$ soient semi-définies positives et qu'une somme de matrices semi-définies positives le soit également, il est nécessaire que $1+\alpha$ soit positif, ce qui est toujours le cas car $0 \leq \alpha \leq 1$, et que $2\alpha - 1$ soit positif ce qui implique que α soit supérieur à $1/2$ pour que $\mathbf{T}_\alpha$ soit semi-définie positive. ■

Cela étant, la réciproque n'est pas vraie : selon les graphes, il peut exister un $\alpha_0 < 1/2$ tel que $\mathbf{T}_\alpha$ est semi-définie positive pour $\alpha \geq \alpha_0$. En effet, comme le montre la figure 4.2, les trois graphes représentés (graphe chemin, graphe aléatoire d'Erdös-Rényi $\mathcal{G}_{50,0.1}$ et graphe comète) ont tous une matrice $\mathbf{T}_\alpha$ semi-définie positive pour $\alpha \geq 0.4$. Pour le graphe aléatoire, il est même possible de voir que $\mathbf{T}_\alpha$ est semi-définie positive pour $\alpha \geq 0.3$. Pour s'en convaincre, il est nécessaire de rappeler que la propriété de semi-définie positivité d'une matrice $\mathbf{M}$ est équivalente au fait que toutes les valeurs propres de $\mathbf{M}$ soient positives ou nulles. Cette figure permet également de constater que la matrice $\mathbf{T}_\alpha$ n'offre pas la propriété de superposition des spectres (4.10) comme pouvait en disposer la matrice d'α–adjacence de Nikiforov. En effet, considérons par exemple la 5$^{\text{e}}$ valeur propre $\lambda_5(\mathbf{T}_\alpha)$ des matrices $\mathbf{T}_\alpha$ du graphe chemin (à gauche), alors il peut être constaté que la propriété $\lambda_5(\mathbf{T}_\alpha) \leq \lambda_5(\mathbf{T}_\beta)$ pour $0 \leq \alpha \leq \beta \leq 1$ n'est pas vérifiée.
Parmi les propriétés algébriques que possèdent la matrice $\mathbf{T}_\alpha$, les égalités suivantes sont vérifiées :

$$\mathbf{T}_{\alpha+\beta} = \mathbf{T}_\alpha + \mathbf{T}_\beta - \mathbf{A}, \tag{4.14}$$

$$\mathbf{T}_{\alpha-\beta} = \mathbf{T}_\alpha - \mathbf{T}_\beta + \mathbf{A}. \tag{4.15}$$

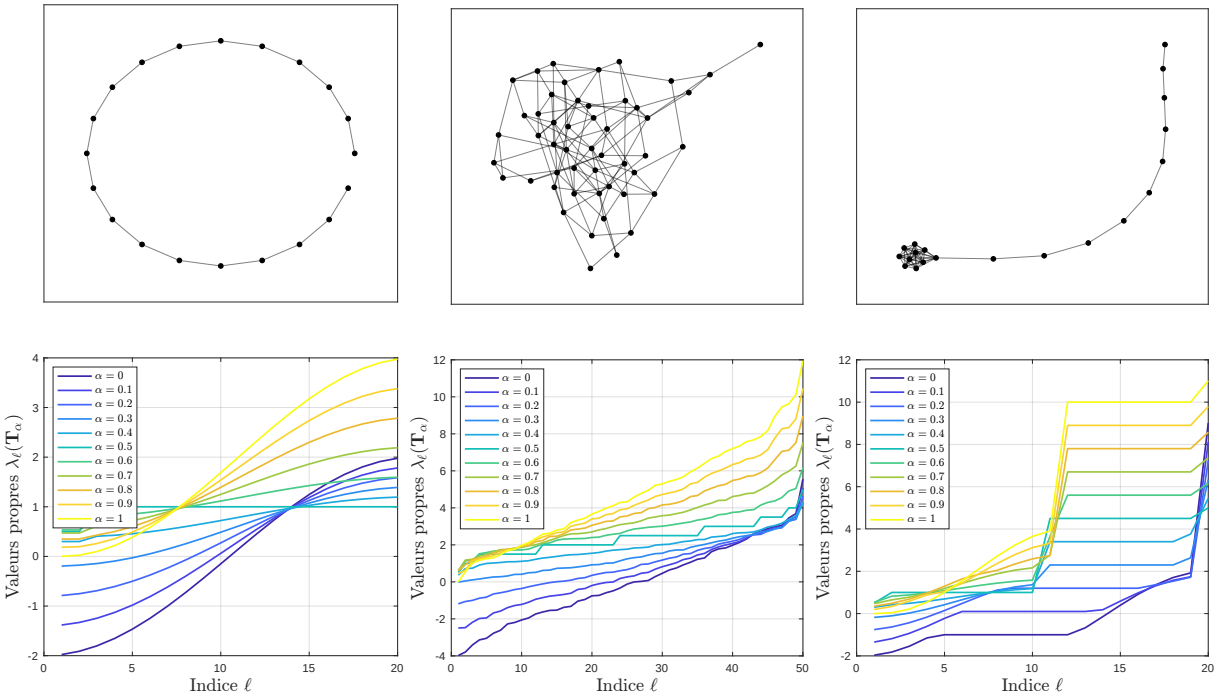


Figure 4.2 – Pour trois graphes différents (de gauche à droite : graphe chemin, graphe aléatoire d'Erdös-Rényi $\mathcal{G}_{50,0.1}$ et graphe comète), valeurs propres $\lambda_\ell(\mathbf{T}_\alpha)$ de leurs matrices $\mathbf{T}_\alpha$ respectives pour un paramètre α allant de 0 à 1 par pas de 0.1.

De plus, nous pouvons formuler les traces de cette matrice assez facilement comme suit :

$$\begin{aligned}
\operatorname{Tr}(\mathbf{T}_\alpha) &= \operatorname{Tr}(\alpha\mathbf{D} + (1-2\alpha)\mathbf{A}) \\
&= \alpha \operatorname{Tr}(\mathbf{D}) + (1-2\alpha)\operatorname{Tr}(\mathbf{A}) \\
&= \alpha \operatorname{Tr}(\mathbf{D}) \\
&= 2\alpha m \qquad (4.16) \\
\operatorname{Tr}\big(\mathbf{T}_\alpha^2\big) &= \operatorname{Tr}\big(\alpha^2\mathbf{D}^2 + 2\alpha(1-2\alpha)\mathbf{DA} + (1-2\alpha)^2\mathbf{A}^2\big) \\
&= \alpha^2 \operatorname{Tr}\big(\mathbf{D}^2\big) + 2\alpha(1-2\alpha)\operatorname{Tr}(\mathbf{DA})(1-2\alpha)^2 \operatorname{Tr}\big(\mathbf{A}^2\big) \\
&= \alpha^2 \operatorname{Tr}\big(\mathbf{D}^2\big) + (1-2\alpha)^2 \operatorname{Tr}\big(\mathbf{A}^2\big) \\
&= \alpha^2 \sum_{i\in\mathcal{V}} \deg(i)^2 + 2m(1-2\alpha)^2. \qquad (4.17)
\end{aligned}$$

Ces traces dépendent donc directement d'attributs structurels du graphe : le nombre de sommets, le nombre d'arêtes ou encore les degrés des sommets. D'après l'étude relative à la matrice de densité $\boldsymbol{\rho}$ du graphe G proposée dans le chapitre précédent, ces expressions s'avèrent également bien utiles dans l'utilisation d'un développement limité d'ordre 1 ou 2.

Une nouvelle α-énergie du graphe G basée sur les valeurs propres $(\lambda_\ell)_{1\leq\ell\leq n}$ de $\mathbf{T}_\alpha$ peut également être établie (à l'image de celle définie par l'équation (4.12)) :

$$E_{\mathbf{T}_\alpha} = \sum_{\ell=1}^{n} \left| \lambda_\ell(\mathbf{T}_\alpha) - \frac{2\alpha m}{n} \right| = \sum_{\ell=1}^{n} \left| \lambda_\ell(\mathbf{T}_\alpha) - \alpha \bar{d}(G) \right|, \tag{4.18}$$

où $\bar{d}(G) = 2m/n$ désigne le degré moyen du graphe G. Cette fois, pour $\alpha = 0$, nous retrouvons l'énergie (1.26) d'un graphe et pour $\alpha = 1$, c'est l'énergie Laplacienne (1.45). Pour $\alpha = 1/2$, nous obtenons la moitié de l'écart de degrés (1.4). En effet, comme $2\mathbf{T}_{1/2} = \mathbf{D}$, il vient

$$E_{\mathbf{T}_{1/2}} = \sum_{\ell=1}^{n} \left| \lambda_\ell \left(\frac{1}{2}\mathbf{D} \right) - \frac{m}{n} \right| = \frac{1}{2} \sum_{\ell=1}^{n} \left| \deg(\ell) - \frac{2m}{n} \right| = \frac{1}{2} \operatorname{dev}(G)$$

Une conjecture concernant l'énergie $E_{\mathbf{T}_\alpha}$ de cette matrice peut être établie.

Conjecture 4.3. *Soit G un graphe d'ordre n et $E_{\mathbf{T}_\alpha}$ l'énergie définie par* (4.18) *pour $\alpha \in [0, 1]$. Alors $E_{\mathbf{T}_\alpha}$ est décroissante pour $0 \leq \alpha \leq \alpha_{\min} \leq 1/2$ puis croissante pour $\alpha_{\min} \leq \alpha \leq 1$. Elle atteint son maximum en $\alpha = 1$.*

Pour quatre graphes (graphe comète $\mathcal{C}_{5,5}$, graphe cycle $\mathcal{C}_{10}$, graphe aléatoire d'Erdös-Rényi $\mathcal{G}_{10,0.1}$ et graphe complet $\mathcal{K}_{10}$) de même ordre pour que la comparaison soit interprétable, la figure 4.3 montre que l'énergie $E_{\mathbf{T}_\alpha}$ vérifie bien le comportement décrit par la conjecture 4.3. Pour le graphe complet ainsi que le graphe cycle, la valeur de α pour laquelle l'énergie est minimale est égale à 0.5. Cette valeur vaut 0.35 dans le cas du graphe comète, et 0.45 dans le cas du graphe aléatoire. Nous n'avons pas encore trouvé de contre-exemples à cette conjecture ce qui justifie alors sa présence dans cette section.

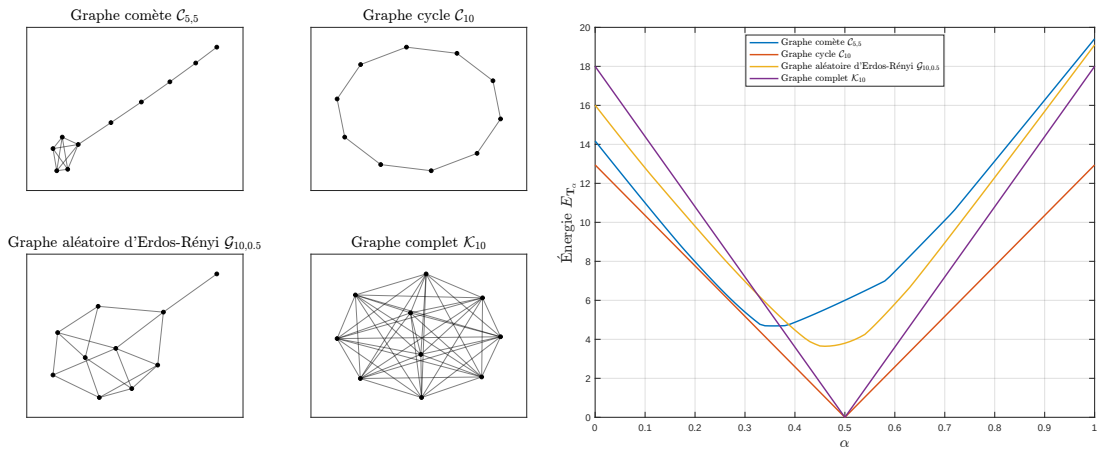


Figure 4.3 – Évolution de l'énergie $E_{\mathbf{T}_\alpha}$ en fonction de α pour quatre graphes d'ordre 10.

4.2.3 Plan de représentation $\mathbf{P}_{\alpha,k}$

Même si la matrice $\mathbf{T}_\alpha$ permet de décrire les matrices $\mathbf{A}$, $\mathbf{D}$ et $\mathbf{L}$, il serait intéressant d'ajouter la matrice Laplacienne sans-signe $\mathbf{Q}$ à cette liste car, comme évoqué dans le chapitre 1, la matrice $\mathbf{Q}$ a l'avantage de présenter moins de graphes cospectraux. Une première solution est de concaténer les matrices d'α-adjacence et α-Laplacienne en définissant une matrice $\mathbf{S}_\gamma \coloneqq (1-|\gamma|)\mathbf{D} + \gamma\mathbf{A}$ pour un paramètre γ variant de -1 à 1. Cette matrice, comme le montre la figure 4.4, décrit bien toutes les matrices classiques. Mais cette manière assez intuitive, n'étant définie qu'avec un seul paramètre, n'offre

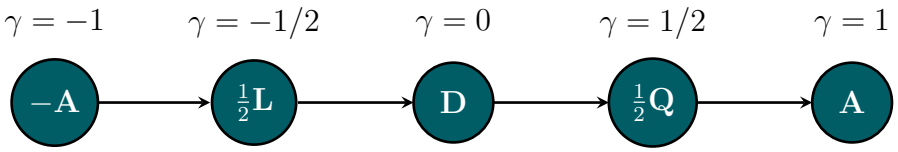
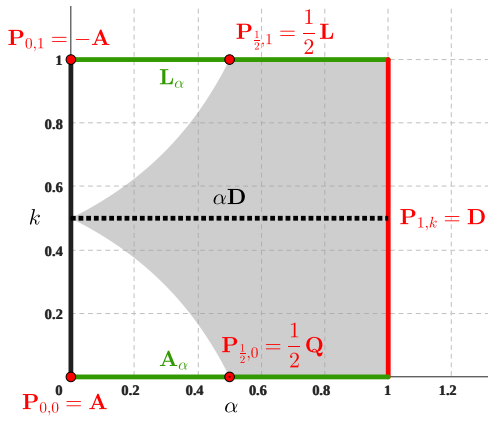


Figure 4.4 – Matrice $\mathbf{S}_\gamma$ qui décrit les matrices classiques $\mathbf{A}$, $\mathbf{D}$, $\mathbf{L}$ et $\mathbf{Q}$.

pas suffisamment de degré de liberté, car rappelons que l'objectif que nous nous sommes fixés est de savoir s'il existe une « meilleure » représentation en combinant toutes les matrices traditionnelles. Pour ce faire, il faut donc réfléchir à la construction d'une telle matrice avec plus de degrés de liberté, autrement dit de paramètres. La solution des matrices d'adjacence généralisée et universelle de Van Dam *et al.* présentant, elle, trop de paramètres (3 dans le cas de la matrice d'adjacence généralisée et 4 pour la matrice d'adjacence universelle), nous avons introduit en conséquence un plan $\mathbf{P}_{\alpha,k}$ de matrice de représentation, en d'autres termes une matrice basée sur deux paramètres α et k variant de 0 à 1 :

$$\mathbf{P}_{\alpha,k} \coloneqq \alpha\mathbf{D} + (2k-1)(\alpha-1)\mathbf{A}, \qquad \alpha, k \in [0,1]. \tag{4.19}$$



Matrices $\mathbf{P}_{\alpha,k}$	Matrices de représentation classiques	
$\mathbf{P}_{0,0}$	$\mathbf{A}$	Matrice d'adjacence
$2\mathbf{P}_{\frac{1}{2},1}$	$\mathbf{L}$	Matrice Laplacienne
$\mathbf{P}_{0,1}$	$-\mathbf{A}$	–
$2\mathbf{P}_{\frac{1}{2},0}$	$\mathbf{Q}$	Matrice Laplacienne sans-signe
$\mathbf{P}_{\alpha,\frac{1}{2}}$	$\alpha\mathbf{D}$	–
$\mathbf{P}_{1,k}$	$\mathbf{D}$	Matrice des degrés
$\mathbf{P}_{\alpha,0}$	$\mathbf{A}_\alpha$	Matrice d'α-adjacence
$\mathbf{P}_{\alpha,1}$	$\mathbf{L}_\alpha$	Matrice α-Laplacienne

Figure 4.5 – Plan de représentation $\mathbf{P}_{\alpha,k}$ pour $\alpha, k \in [0,1]$.

La majorité des matrices de représentation classiques peuvent être retrouvées grâce à cette expression, comme le montre la figure 4.5. Les matrices de représentation traditionnelles sont en **rouge** sur la figure tandis que les matrices d'α-adjacence [55] et α-Laplacienne [56] sont en **vert**. Ces égalités de matrices sont listées dans le tableau à droite. Les matrices définies dans la partie grisée sont nécessairement semi-définies positives comme le démontre la proposition 4.4. Il est alors envisageable d'étudier les changements graduels entre la matrice d'adjacence $\mathbf{A}$ et la matrice des degrés $\mathbf{D}$ en passant par la matrice Laplacienne $\mathbf{L}$ ou encore la matrice Laplacienne sans-signe $\mathbf{Q}$. La proposition suivante établit un domaine de semi-définie positivité de la matrice $\mathbf{P}_{\alpha,k}$ selon les paramètres α et k, propriété primordiale pour toute étude matricielle.

Proposition 4.4. *Soit G un graphe et $\mathbf{P}_{\alpha,k}$ sa matrice de représentation. Alors $\mathbf{P}_{\alpha,k}$ est semi-définie positive si $\frac{2\alpha-1}{2(\alpha-1)} \leq k \leq \frac{1}{2(1-\alpha)}$.*

Preuve. À l'image de ce qui est fait pour la proposition 4.2, il suffit de réécrire l'expression de $\mathbf{P}_{\alpha,k}$ comme suit :

$$\mathbf{P}_{\alpha,k} = \alpha\mathbf{D} + (2k-1)(\alpha-1)\mathbf{A} = [\alpha + (2k-1)(\alpha-1)]\,\mathbf{D} + (2k-1)(1-\alpha)\mathbf{L}.$$

Ainsi et dans un premier temps, comme la somme de deux matrices semi-définies positives l'est également, il est nécessaire d'avoir les conditions suivantes pour que $\mathbf{P}_{\alpha,k}$ soit semi-définie positive :

$$\begin{cases} 0 \leq \alpha + (2k-1)(\alpha-1) \\ 0 \leq (2k-1)(1-\alpha) \end{cases} \quad \Longrightarrow \quad \frac{1}{2} \leq k \leq \frac{1}{2(1-\alpha)} \tag{4.20}$$

Dans un second temps, montrer la semi-définie positivité directement à partir de sa définition, en d'autres termes montrer que $\langle \mathbf{P}_{\alpha,k}\mathbf{x}, \mathbf{x}\rangle \geq 0$ pour tout $\mathbf{x} = (x_i)_{1\leq i\leq n} \in \mathbb{R}^n$ est possible en écrivant

$$\langle \mathbf{P}_{\alpha,k}\mathbf{x}, \mathbf{x}\rangle = [\alpha - (2k-1)(\alpha-1)] \sum_{i\in\mathcal{V}} x_i^2 \deg(i) + (2k-1)(\alpha-1) \sum_{\{i,j\}\in\mathcal{E}} (x_i + x_j)^2,$$

c'est-à-dire, pour toute arête $\{i,j\} \in \mathcal{E}$,

$$\langle \mathbf{P}_{\alpha,k}\mathbf{x}, \mathbf{x}\rangle \geq [\alpha - (2k-1)(\alpha-1)]\,(x_i^2 + x_j^2) + (2k-1)(\alpha-1)(x_i + x_j)^2.$$

Cela implique que pour avoir la propriété de semi-définie positivité de la matrice $\mathbf{P}_{\alpha,k}$, il faut également avoir

$$\begin{cases} 0 \leq \alpha - (2k-1)(\alpha-1) \\ 0 \leq (2k-1)(\alpha-1) \end{cases} \quad \Longrightarrow \quad \frac{2\alpha-1}{2(\alpha-1)} \leq k \leq \frac{1}{2} \tag{4.21}$$

En groupant les conditions (4.20) et (4.21), alors suit la condition nécessaire de la proposition 4.4. ■

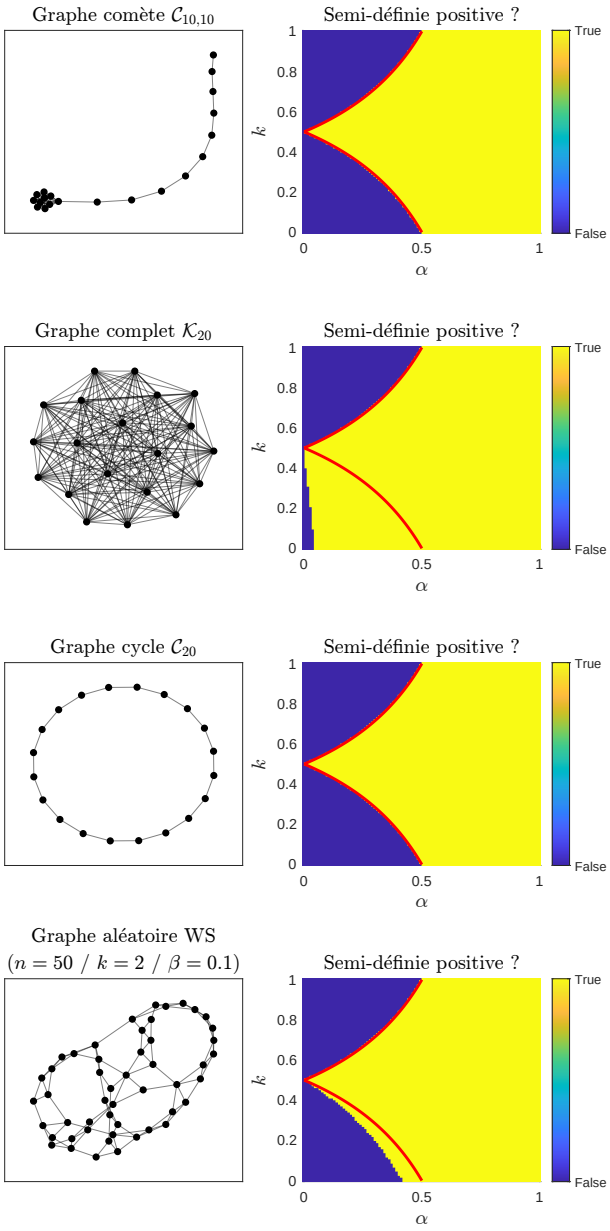


Figure 4.6 – Différents graphes et la semi-définie positivité de leurs matrices $\mathbf{P}_{\alpha,k}$ (jaune pour Vrai, bleu pour Faux). Les lignes rouges correspondent aux conditions de la proposition 4.4.

Une fois de plus, cette proposition met en évidence une condition nécessaire mais non suffisante. En effet, comme le montre la figure 4.6, il existe de nombreux graphes pour lesquels il existe un couple (α^*, k^*) en dehors des domaines suscités qui permet d'avoir une matrice $\mathbf{P}_{\alpha^*,k^*}$ semi-définie positive, notamment le graphe complet $\mathcal{K}_{20}$. Cependant, en rappelant le théorème 3.5 de Wang *et al.* [56] énoncé précédemment (Théorème 4.1), il peut être conjecturé que la condition ①, c'est-à-dire $\frac{1}{2} \leq k \leq \frac{1}{2(1-\alpha)}$, est bien nécessaire et suffisante. En d'autres termes, il est possible de conjecturer

qu'il n'existe pas de couple (α^*, k^*) dans le domaine $\left[0, \frac{1}{2}\right] \times \left[\frac{1}{2(1-\alpha)}, 1\right]$ tel que la matrice $\mathbf{P}_{\alpha^*,k^*}$ soit semi-définie positive : cela est bien illustré par la figure 4.6. Partout ailleurs, la condition n'est que nécessaire et la question naturelle suivante se pose : pour quel(s) graphe(s) les conditions posées dans la proposition 4.4 sont-elles nécessaires et suffisantes ? Cette question ouverte est également posée par Nikiforov pour sa matrice d'α-adjacence [55]. En l'occurrence, pour nos exemples, le graphe comète $\mathcal{C}_{10,10}$ ainsi que le graphe cycle $\mathcal{C}_{20}$ semblent avoir une matrice de représentation $\mathbf{P}_{\alpha,k}$ dont le domaine de semi-définie positivé est exactement celui de la proposition 4.4.

Par ailleurs, la propriété de superposition des spectres vérifiée par la matrice d'α-adjacence de Nikiforov [55] est également vérifiée par notre matrice $\mathbf{P}_{\alpha,k}$, selon le paramètre α. Cette assertion est formalisée par la proposition suivante.

Proposition 4.5. *Soit G un graphe. Si $\nu_\ell^{(\alpha,k)}$ (resp. $\nu_\ell^{(\alpha',k)}$) désigne la ℓ^e valeur propre de sa matrice $\mathbf{P}_{\alpha,k}$ (resp. $\mathbf{P}_{\alpha',k}$), alors, pour $\alpha \in [0,1]$ et $\alpha' \in [\alpha, 1]$,*

$$\nu_\ell^{(\alpha,k)} \leq \nu_\ell^{(\alpha',k)}, \quad \forall k \in [0,1]. \tag{4.22}$$

Preuve. Soit $\alpha \in [0,1]$ et $\alpha' \in [\alpha, 1]$. Pour utiliser le même schéma de preuve que dans [55], il nous faut écrire

$$\mathbf{P}_{\alpha',k} - \mathbf{P}_{\alpha,k} = (\alpha' - \alpha)\left[\mathbf{D} - (2k-1)\mathbf{A}\right].$$

Grâce à une version simplifiée du théorème de Weyl [277], il suit

$$\nu_\ell(\mathbf{P}_{\alpha',k}) - \nu_\ell(\mathbf{P}_{\alpha,k}) \geq (\alpha' - \alpha)\nu_1(\mathbf{D} - (2k-1)\mathbf{A}) \geq 0$$

où $\nu_1(\mathbf{D} - (2k-1)\mathbf{A})$, qui désigne la plus petite valeur propre de la matrice $\mathbf{M} := \mathbf{D} - (2k-1)\mathbf{A}$, est positive. Pour le prouver, montrons que $\mathbf{M}$ est semi-définie positive. Soit $\mathbf{x} \in \mathbb{R}^n$:

$$\begin{aligned}
\langle \mathbf{M}\mathbf{x}, \mathbf{x} \rangle &= \sum_{i \in \mathcal{V}} x_i^2 d_i - 2(2k-1) \sum_{\{i,j\} \in \mathcal{E}} x_i x_j \\
&= 2(1-k) \sum_{i \in \mathcal{V}} x_i^2 d_i + (2k-1) \sum_{\{i,j\} \in \mathcal{E}} (x_i - x_j)^2 \qquad (4.23) \\
&= 2k \sum_{i \in \mathcal{V}} x_i^2 d_i + (1-2k) \sum_{\{i,j\} \in \mathcal{E}} (x_i + x_j)^2. \qquad (4.24)
\end{aligned}$$

En utilisant l'expression (4.24) pour $0 \leq k \leq 1/2$ et l'expression (4.23) lorsque $1/2 \leq k \leq 1$, il suit que $\langle \mathbf{M}\mathbf{x}, \mathbf{x} \rangle \geq 0$. Par conséquent, toutes les valeurs propres de $\mathbf{M}$, en particulier ν_1, sont positives. ■

Ce résultat appuie l'introduction du plan $\mathbf{P}_{\alpha,k}$ d'un point de vue théorique et permet de faire des conjectures, puisqu'il est sûr que les spectres se superposent selon α pour tout k choisi entre 0 et 1.

4.3 Mesure de similarité par corrélation spectrale

L'objectif de cette section est de présenter une nouvelle méthode de classification de graphes à l'aide d'une mesure de similarité associée à un algorithme de classification comme le SVM. À cet effet, et comme cela a été évoqué en introduction de ce chapitre, Bay-Ahmed *et al.* [32] ont émis l'idée que le spectre d'adjacence $(\lambda_{1\ell})_{1\leq\ell\leq n_1}$ (resp. $(\lambda_{2\ell})_{1\leq\ell\leq n_2}$) d'un graphe G_1 (resp. G_2) d'ordre n_1 (resp. n_2) est une variable aléatoire générée par une loi de probabilité ayant une variance finie σ_1^2 (resp. σ_2^2). Une manière de caractériser la similarité entre les graphes G_1 et G_2 est alors de calculer la covariance spectrale $\mathrm{CS}(G_1, G_2)$ entre les deux spectres d'adjacence définie par

$$\mathrm{CS}(G_1, G_2) = \frac{1}{n-1} \sum_{\ell=1}^{n} \left(\lambda_{1\ell} - \bar{\lambda}_1\right) \left(\lambda_{2\ell} - \bar{\lambda}_2\right), \tag{4.25}$$

avec $n = \min(n_1, n_2)$ et $\bar{\lambda}_1$ (resp. $\bar{\lambda}_2$) désignant la moyenne statistique du spectre d'adjacence de G_1 (resp. G_2). Cette mesure a alors remplacé la distance euclidienne du traditionnel noyau Gaussien d'un SVM. Ils ont également introduit la Similarité Spectrale Conjointe $\mathrm{SSC}_\beta(G_1, G_2)$ entre G_1 et G_2, définie comme une combinaison convexe linéaire de paramètre $0 \leq \beta \leq 1$ d'une « distance euclidienne » calculée entre les spectres d'adjacence et Laplacien [53] :

$$\mathrm{SSC}_\beta\left(G_1, G_2\right) = \beta\, \mathrm{SS}_{\mathbf{L}}\left(G_1, G_2\right) + (1-\beta)\, \mathrm{SS}_{\mathbf{A}}\left(G_1, G_2\right), \tag{4.26}$$

avec

$$\mathrm{SS}_{\mathbf{A}}\left(G_1, G_2\right) = \sum_{\ell=1}^{n} \left(\lambda_{1\ell} - \lambda_{2\ell}\right)^2, \qquad \mathrm{SS}_{\mathbf{L}}\left(G_1, G_2\right) = \sum_{\ell=1}^{n} \left(\mu_{1\ell} - \mu_{2\ell}\right)^2, \tag{4.27}$$

où $(\lambda_{1\ell})_{1\leq\ell\leq n_1}$ (resp. $(\lambda_{2\ell})_{1\leq\ell\leq n_2}$) représente le spectre d'adjacence de G_1 (resp. G_2) et $(\mu_{1\ell})_{1\leq\ell\leq n_1}$ (resp. $(\mu_{2\ell})_{1\leq\ell\leq n_2}$) représente le spectre Laplacien de G_1 (resp. G_2). Bien entendu, à cause de l'existence du problème de cospectralité, ces deux mesures ne sont pas des distances : avoir $\mathrm{CS}(G_1, G_2)$ ou $\mathrm{SSC}_\beta\left(G_1, G_2\right)$ égal à 0 n'implique pas que G_1 soit égale à G_2.

Le problème avec ces mesures est qu'elles ne permettent pas réellement de comprendre comment évolue le spectre entre les matrices de représentation classiques ni de trouver une représentation permettant une meilleure discrimination spectrale entre graphes. En effet, la covariance spectrale ne compare que les spectres d'adjacence et la similarité spectrale conjointe est calculée entre les deux spectres (d'adjacence et Laplacien) et non pas entre des spectres issus de matrices variant entre $\mathbf{A}$ et $\mathbf{L}$. Cette dernière mesure nécessite, de surcroît, le calcul de deux spectres qui, pour de grands graphes, peut se

révéler coûteux en temps de calcul. Un deuxième point à relever concerne le fait que ces mesures utilisées pour comparer les spectres calculent des différences locales. Ces dernières, par définition, peuvent omettre beaucoup de valeurs propres en fonction du nombre de sommets des graphes comparés. Aussi, dans cette section, est introduite une mesure de similarité qui ne requiert qu'un spectre (sous-entendu une seule matrice de représentation), qui puisse comparer les spectres de manière globale par corrélation et qui permette de mettre en évidence que ce ne sont pas les matrices de représentation traditionnelles qui possèdent les informations spectrales les plus discriminantes. Les matrices $\mathbf{T}_\alpha$ ou encore $\mathbf{P}_{\alpha,k}$ sont alors des outils adaptés pour mener à bien ce travail.

Pour pallier toutes les difficultés susmentionnées, nous proposons de traiter le problème de classification de graphes en introduisant une mesure de similarité spectrale entre deux graphes G_1 et G_2 d'ordre n basée sur la corrélation entre les spectres de leurs matrices de représentation que ce soient celles qualifiées de conventionnelles ($\mathbf{A}, \mathbf{L}, \mathbf{D}$ ou $\mathbf{Q}$) ou bien celles présentées dans la section précédente ($\mathbf{T}_\alpha$ ou $\mathbf{P}_{\alpha,k}$). Si les graphes étudiés sont d'ordres différents, nous opérons une simple complétion par des nœuds (*node-padding*) car une corrélation ne peut être calculée qu'entre des vecteurs de même longueur [278]. Si $\widehat{\boldsymbol{\nu}}_1^{(\alpha)}$ (resp. $\widehat{\boldsymbol{\nu}}_2^{(\alpha)}$) représente le spectre standardisé[3] de la matrice de représentation $\mathbf{R}_1$ (resp. $\mathbf{R}_2$) du graphe G_1 (resp. G_2), alors la mesure de similarité entre les graphes G_1 et G_2 est définie grâce à une distance construite à partir de la corrélation de Pearson comme suit [279] :

$$\mathrm{CorS}_{\mathbf{R}}(G_1, G_2) \coloneqq \sqrt{1 - \left(\frac{1}{n}\left\langle \widehat{\boldsymbol{\nu}}_1^{(\alpha)}, \widehat{\boldsymbol{\nu}}_2^{(\alpha)} \right\rangle\right)^2}. \tag{4.28}$$

Plus les spectres standardisés sont corrélés, plus $\mathrm{CorS}_{\mathbf{R}}$ est minimisée, justifiant ainsi son interprétation de mesure de similarité. À l'image du travail précédent [32], cette mesure est intégrée dans un SVM en remplacement de la distance euclidienne dans le traditionnel noyau Gaussien. Ainsi, pour une base de données contenant N graphes $(G_i)_{1\leq i\leq N}$, la matrice de Gram $\mathbf{K}$ est définie par ses coefficients

$$[\mathbf{K}]_{ij} = \exp\left(-\gamma\, \mathrm{CorS}_{\mathbf{R}}(G_i, G_j)\right), \quad i, j \in [\![1, N]\!]. \tag{4.29}$$

4.4 Classification spectrale de graphes et de signaux

4.4.1 Bases de données

Dans le but de montrer l'intérêt de cette nouvelle famille $\mathbf{P}_{\alpha,k}$ de matrices de représentation (qui, en réalité, est une généralisation des matrices $\mathbf{A}_\alpha$, $\mathbf{L}_\alpha$ ou encore $\mathbf{T}_\alpha$) couplée avec la nouvelle mesure

3. On qualifie $\widehat{\mathbf{x}}$ de vecteur standardisé si on retranche au vecteur $\mathbf{x}$ sa moyenne et qu'on le divise par son écart-type.

de similarité par corrélation spectrale CorS présentée dans la sous-section précédente, nous avons sélectionné des bases de données de graphes et de signaux bien connues pour en effectuer une classification. Les bases de données **MUTAG** [280] et **PTC_MR** [281] contiennent des composés chimiques étiquetés respectivement en fonction de leur effet mutagène sur une bactérie et de leur cancérogénicité sur des rats mâles. La base de données **PROTEINS** [282] est constituée de protéines (les sommets étant les acides aminés reliés entre eux par des arêtes s'ils se situent à moins de 6 Å) classifiées selon qu'elles soient des enzymes ou non. Les bases de données de collaboration cinématographique **IMDB-BINARY** et **IMDB-MULTI** [283] sont quant à elles composées de réseaux *ego* de respectivement 1000 et 1500 acteurs/actrices qui ont joué des rôles dans des films listés dans la base de données IMDb, divisés en respectivement deux (Action / Romance) et trois genres (Comédie / Romance / SF). L'ensemble de données **ItalyPowerDemand** [284] contient des morceaux d'une série temporelle constituée de la demande d'électricité en Italie sur une période de douze mois. La tâche de classification consiste cette fois à distinguer les jours d'octobre à mars de ceux d'avril à septembre. Ces signaux sont convertis en graphes grâce à l'algorithme de visibilité horizontale présenté dans le chapitre 1. Les statistiques descriptives de ces bases de données sont rappelées dans le tableau 4.1.

	Nombre de graphes N	$\overline{n}$	$\overline{m}$
MUTAG	188	17.93	19.79
PTC_MR	344	14.29	14.69
PROTEINS	1113	39.06	72.82
IMDB-BINARY	1000	19.77	96.53
IMDB-MULTI	1500	13.00	65.94
ItalyPowerDemand	1096	24.00	35.51

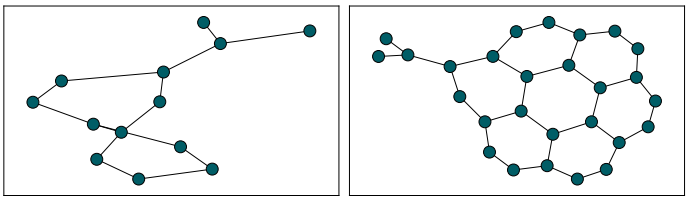
Tableau 4.1 – Statistiques des bases de données où $\overline{n}$ est le nombre moyen de sommets et $\overline{m}$ le nombre moyen d'arêtes.

La figure 4.7 présente quant à elle, pour chaque base de données, deux graphes extraits aléatoirement. La différence est sensible entre les types de graphes : les réseaux sociaux (IMDB-BINARY et IMDB-MULTI) présentent, par exemple, plus de communautés que les autres. Les graphes de visibilité contenus dans la base de données ItalyPowerDemand sont, eux aussi, singuliers et reconnaissables.

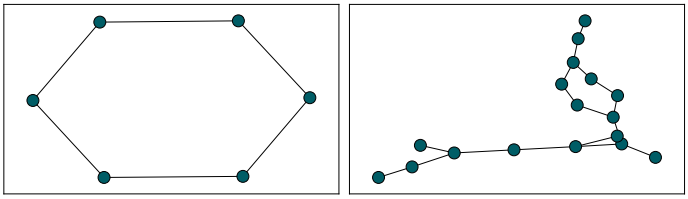
4.4.2 Méthodes & résultats

Les hyperparamètres de la SVM dont le noyau est donné par l'équation (4.29), c'est-à-dire le paramètre de normalisation γ comme le paramètre de régularisation C sont fixés à 1, et une validation croisée stratifiée à 10 couches est effectuée[4] dans le but de pouvoir comparer nos résultats avec ceux de la littérature, le cas échéant.

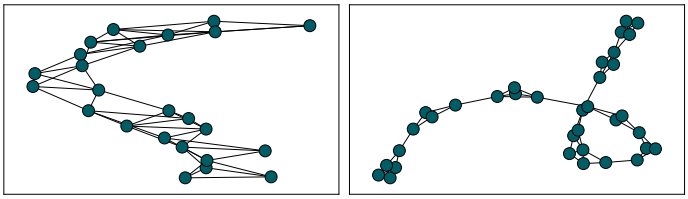
4. Le lecteur est invité à consulter la section 2.2 pour les rappels quant au SVM et la méthode de validation croisée.



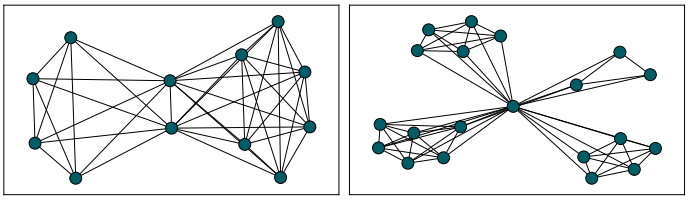
(a) Base de données **MUTAG**



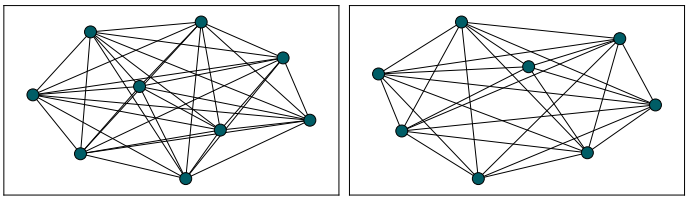
(b) Base de données **PTC_MR**



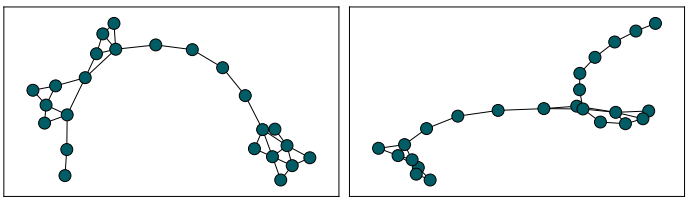
(c) Base de données **PROTEINS**



(d) Base de données **IMDB-BINARY**



(e) Base de données **IMDB-MULTI**



(f) Base de données **ItalyPowerDemand**

Figure 4.7 – Pour chaque base de données, exemples de deux graphes tirés aléatoirement.

Le premier intérêt de ce travail est de montrer que ce n'est pas le contenu spectral des matrices de représentation usuelles qui permette d'effectuer la meilleure classification de graphes avec le même noyau de corrélation spectrale CorS. Pour cela, nous avons considéré des paramètres α et k tirés par pas de 0.05 entre 0 et 1. Pour chaque combinaison, 10 itérations de la méthode établie ci-dessus, à savoir la validation croisée, sont effectuées. Ce procédé permet d'obtenir les « cartes » d'exactitudes présentées en figure 4.8. Ces cartes sont différentes d'une base de données à une autre, ce qui était attendu mais confirmé à la vue de cette figure. Une première affirmation qui peut être déduite : introduire une représentation unique ne convient pas pour toutes les bases de données, *a minima* pour des tâches de classification. En d'autres termes, ce n'est pas le contenu spectral d'une seule matrice qui permet la meilleure classification pour toutes les bases de données. De plus, les matrices optimales en termes d'exactitudes de classification diffèrent fortement d'une base de données à l'autre comme observé dans le tableau 4.2. En effet, c'est la matrice $\mathbf{P}_{1,k} = \mathbf{D}$ pour **MUTAG** avec une précision de 88.2% dépassant de près de 3% les autres combinaisons. Pour **PTC_MR**, c'est la matrice $\mathbf{P}_{0.1,0.75}$, avec une précision de 59.6% qui l'emporte légèrement ($+0.5\%$) sur les matrices classiques comme $\mathbf{L}$ ou $\mathbf{Q}$. Pour **PROTEINS**, c'est également une matrice particulière, $\mathbf{P}_{0.05,1}$, qui permet d'obtenir les meilleures exactitudes. C'est à nouveau des matrices non usuelles, $\mathbf{P}_{0.3,0.05}$ et $\mathbf{P}_{0.35,0.1}$ pour **IMDB-BINARY** et **IMDB-MULTI** respectivement qui atteignent des exactitudes de classification de 72.55% et 49.95% respectivement. Enfin, pour la base de données **ItalyPowerDemand** de signaux vus comme des graphes, c'est la matrice $\mathbf{P}_{0.35,0.25}$ qui donne les meilleurs résultats, légèrement supérieurs à ceux obtenus avec la matrice $\mathbf{Q}$.

	Meilleures matrices	**Matrices classiques**			
		$\mathbf{A} = \mathbf{P}_{0,0}$	$\mathbf{D} = \mathbf{P}_{1,k}$	$\mathbf{L} = 2\mathbf{P}_{0.5,1}$	$\mathbf{Q} = 2\mathbf{P}_{0.5,0}$
MUTAG	$88.20 \pm 0.37\%$ ($\mathbf{P}_{1,k}$)	$85.10 \pm 0.04\%$	$88.20 \pm 0.37\%$	$85.22 \pm 0.42\%$	$85.11 \pm 0.60\%$
PTC_MR	$59.60 \pm 0.66\%$ ($\mathbf{P}_{0.1,0.75}$)	$58.97 \pm 1.17\%$	$56.61 \pm 0.71\%$	$59.13 \pm 0.76\%$	$59.19 \pm 0.53\%$
PROTEINS	$74.54 \pm 0.12\%$ ($\mathbf{P}_{0.05,1}$)	$74.48 \pm 0.17\%$	$71.98 \pm 0.25\%$	$73.89 \pm 0.19\%$	$72.62 \pm 0.19\%$
IMDB-BINARY	$72.55 \pm 0.47\%$ ($\mathbf{P}_{0.3,0.05}$)	$70.01 \pm 0.75\%$	$71.20 \pm 0.54\%$	$71.14 \pm 0.56\%$	$70.46 \pm 0.55\%$
IMDB-MULTI	$49.95 \pm 0.51\%$ ($\mathbf{P}_{0.35,0.1}$)	$47.99 \pm 0.43\%$	$49.39 \pm 0.26\%$	$48.86 \pm 0.23\%$	$49.50 \pm 0.36\%$
ItalyPowerDemand	$86.65 \pm 0.22\%$ ($\mathbf{P}_{0.35,0.25}$)	$79.09 \pm 0.24\%$	$62.24 \pm 0.42\%$	$73.60 \pm 0.40\%$	$86.02 \pm 0.21\%$

Tableau 4.2 – Exactitudes (moyennés sur 10 itérations $\pm$ écart-type) pour la meilleure matrice $\mathbf{P}_{\alpha,k}$ et pour les matrices classiques ($\mathbf{A}$, $\mathbf{D}$, $\mathbf{L}$ et $\mathbf{Q}$).

Les raisons pour lesquelles ces matrices particulières permettent d'obtenir les meilleures exactitudes de classification méritent certainement d'être explorées. En analysant qualitativement la figure 4.8, il est par exemple possible de remarquer des similitudes entre les cartes des bases de données de même type. Pour les réseaux sociaux (IMDB-BINARY et IMDB-MULTI), l'information spectrale la plus discrimante semble se situer dans la partie inférieure du plan (*i.e.* pour $k < 0.5$), tandis que pour les graphes issus de la chimie moléculaire (MUTAG, PTC-MR et PROTEINS), elle semble se situer dans

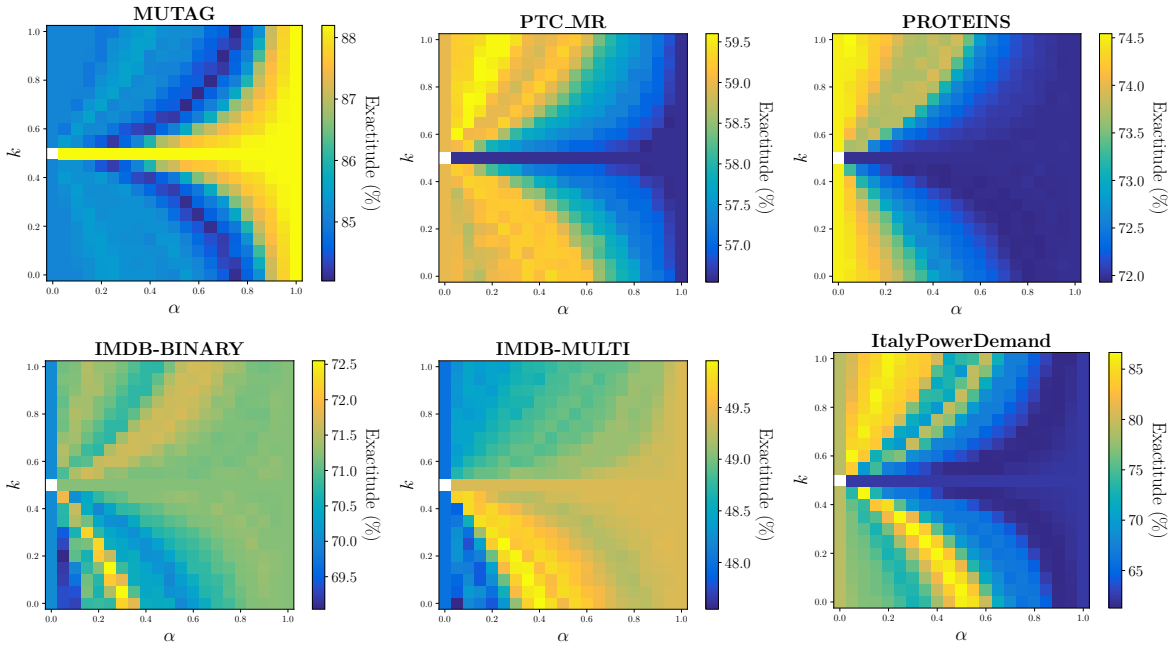


Figure 4.8 – Plan $\mathbf{P}_{\alpha,k}$, pour six bases de données de graphes, dont la couleur représente l'exactitude moyenne de classification pour des couples de paramètres (α, k).

la partie supérieure (*i.e.* pour $k > 0.5$). En tout état de cause et ce qu'il faut retenir à ce stade, c'est que les résulats exposés ici montrent que dans 5 cas sur 6, ce n'est pas une matrice de représentation standard ($\mathbf{A}$, $\mathbf{L}$, $\mathbf{D}$ ou $\mathbf{Q}$) qui permet d'effectuer la meilleure classification spectrale mais bien une matrice intermédiaire.

L'objectif de ce chapitre n'est pas d'introduire un noyau permettant de se placer parmi les meilleurs en termes de résultats de classification, notamment car les nouveaux algorithmes basés sur les réseaux de neurones sur graphes s'imposent [278], ou encore en termes de temps de calcul. Toutefois, il est important de pouvoir se comparer à d'autres travaux. Les autres noyaux avec lesquels nous allons comparer nos résultats sont les suivants :

— Le noyau **CS** défini à l'aide de la corrélation spectrale introduite par l'équation (4.25);
— Le noyau **SSC** défini à l'aide de la similarité spectrale conjointe introduite par l'équation (4.26);
— Le noyau **SP** comparant les plus courts chemins dans les graphes étudiés introduit dans [30];
— Le noyau **RW** comparant les marches aléatoires dans les graphes étudiés introduit dans [170];
— Le noyau **GK** dénombrant les sous-graphes types dans les graphes étudiés introduit dans [171];
— Le noyau **WL** basé sur le test d'isomorphisme de Weisfeiler-Lehman [40].

Il est à noter que c'est la librairie Python nommée GraKel [285] qui a été retenue pour effectuer ce travail car cette dernière permet l'utilisation très efficiente des noyaux structurels traditionnels et pré-

sente l'avantage d'un interfaçage aisé avec des noyaux personnalisés. Bien entendu, les noyaux structurels classiques ont vu leurs codes optimisés dans cette librairie ce qui impacte inévitablement les temps de calcul référencés dans le tableau 4.3. Pour chaque base de données, les temps de calcul affichés sont des moyennes sur 10 itérations des temps nécessaires aux calculs des matrices de Gram des différents noyaux sur l'intégralité de la base de données. Toutes ces actions ont été effectuées sur un Intel®Xeon Platinum 8180 @ 2.50 GHz avec 128 Go de RAM. Avant toute analyse, rappelons que les codes des noyaux personnalisés CorS, CS et SSC n'ont pas été optimisés. Au regard du tableau 4.3, le noyau le plus rapide dans toutes les situations est le noyau WL. En second, le noyau SP. Puis se classent les noyaux spectraux CS, SSC et CorS, ce dernier étant le plus rapide des trois dans 5 sur 6 bases de données. Ce résultat était attendu car le noyau CorS nécessite le calcul d'un unique spectre par rapport au noyau SSC qui en nécessite deux. La valeur relativement élevée dans le cas de la base de données PROTEINS est due au fait que tous les graphes subissent un *node-padding* de sorte à avoir le même nombre de sommets. En l'occurrence, dans cette base de données, tous les graphes sont ramenés à 2098 sommets, qui est une taille importante. Les noyaux CS ou SSC n'ayant pas besoin de complétion par des noeuds, ceci explique la différence de temps de calcul. Cela montre, une fois de plus, que les codes ont cruellement besoin d'une optimisation, ce qui sera fait prochainement. Les noyaux RW et GK sont connus pour être extrêmement chronophages ce qui les placent en dernière position avec même des problèmes de convergence pour quelques bases de données (PROTEINS et ItalyPowerDemand pour RW ainsi que IMDB-BINARY et IMDB-MULTI pour GK). Les noyaux SP et WL sont les plus performants car ils sont dépendants des sous-structures dans les graphes étudiés et sont donc extrêmement efficients dans le cas de graphes creux. Notre méthode, quant à elle, dépend uniquement du nombre de sommets, le calcul du spectre étant cubique en cette grandeur. Ainsi, l'utilisation du noyau CorS est plus adaptée à des graphes denses et d'un ordre restreint.

	CorS	**CS**	**SSC**	**SP**	**RW**	**GK**	**WL**
MUTAG	0.25	0.34	0.31	0.30	6.83	1.52	0.06
PTC_MR	0.68	0.97	0.68	0.45	19.52	1.76	0.09
PROTEINS	62.71	12.81	13.69	19.25	–	169	1.09
IMDB-BINARY	5.69	8.70	7.41	3.73	215	–	0.72
IMDB-MULTI	10.01	17.01	10.41	3.02	235	–	0.80
ItalyPowerDemand	4.93	12.35	13.04	2.95	–	24.79	0.62

Tableau 4.3 – Temps de calcul (en seconde) des différents noyaux considérés (moyennés sur 10 itérations).

La comparaison des exactitudes moyennes sur les 10 couches de validation croisée, elles-mêmes moyennées sur 10 itérations, obtenus avec ces différents noyaux, sont présentés en tableau 4.4. Dans ce tableau, sont affichés en gras les résultats des deux meilleurs noyaux. Avec cette information, il est

clair que le noyau WL, en plus d'être particulièrement efficient (au moins avec l'implémentation de la librairie GraKel), fait partie des deux meilleurs noyaux pour toutes les bases de données, accompagné du noyau CorS pour cinq bases de données sur six. Le noyau CS est celui avec les résultats les moins performants, ce qui était attendu compte tenu du postulat de ce chapitre : ce noyau ne comparant que les spectres d'adjacence des graphes étudiés, ce n'est vraisemblablement pas suffisant pour effectuer une classification performante. Le noyau SSC introduit dans un travail précédent permet d'obtenir de meilleurs résultats que le noyau CorS pour la base de données contenant des graphes de visibilité, qui sont des graphes tout à fait particuliers. Il semble alors intéressant de voir si cette observation se généralise à d'autres bases de données du même type.

Bien entendu, l'utilisation du noyau CorS n'est pas exclusif : nous avons montré, grâce à ce dernier, que le contenu spectral discriminant ne se situait pas dans les matrices de représentation classiques mais il est possible, sans aucun doute, de trouver un autre noyau permettant d'obtenir d'encore meilleures exactitudes de classification tout en mettant à profit ce résultat.

	CorS	**CS**	**SSC**	**SP**	**RW**	**GK**	**WL**
MUTAG	$\mathbf{88.2 \pm 0.4}\%$	$39.9 \pm 0.1\%$	$84.3 \pm 1.1\%$	$83.6 \pm 1.2\%$	$88.3 \pm 0.5\%$	$87.0 \pm 0.6\%$	$\mathbf{85.6 \pm 0.9}\%$
PTC_MR	$\mathbf{59.6 \pm 0.7}\%$	$58.1 \pm 0.1\%$	$57.8 \pm 2.1\%$	$58.2 \pm 1.9\%$	$54.2 \pm 1.8\%$	$55.6 \pm 0.3\%$	$\mathbf{62.3 \pm 1.0}\%$
PROTEINS	$\mathbf{74.5 \pm 0.1}\%$	$38.5 \pm 1.8\%$	$70.0 \pm 0.4\%$	$71.4 \pm 0.4\%$	–	$71.2 \pm 0.4\%$	$\mathbf{74.4 \pm 0.5}\%$
IMDB-BINARY	$\mathbf{72.6 \pm 0.5}\%$	$47.2 \pm 1.0\%$	$67.5 \pm 0.9\%$	$69.8 \pm 1.3\%$	$47.9 \pm 1.4\%$	–	$\mathbf{72.7 \pm 1.2}\%$
IMDB-MULTI	$\mathbf{50.0 \pm 0.5}\%$	$33.4 \pm 0.8\%$	$48.3 \pm 0.4\%$	$49.3 \pm 0.6\%$	$29.3 \pm 0.7\%$	–	$\mathbf{50.8 \pm 0.5}\%$
ItalyPowerDemand	$86.7 \pm 0.2\%$	$50.1 \pm 0.0\%$	$\mathbf{92.3 \pm 0.2}\%$	$91.6 \pm 0.5\%$	–	$67.9 \pm 0.4\%$	$\mathbf{96.1 \pm 0.1}\%$

Tableau 4.4 – Résultats de classification (exactitudes en % moyennées sur 10 itérations $\pm$ écarts-types) de différents noyaux sur des bases de données de la littérature. Les résultats des deux meilleurs noyaux pour chaque base de données sont affichés en gras.

4.5 Conclusion

Dans ce chapitre, nous avons introduit une nouvelle famille de représentation matricielle de graphes, généralisant les représentations classiques de la littérature du domaine. Motivés par le problème de cospectralité, Van Dam, Nikiforov ou encore Wang ont respectivement construit la matrice d'adjacence dite universelle, l'α–adjacence ou l'α-Laplacienne. Seulement, bien que les propriétés de ces matrices soient intéressantes, les applications pour les valoriser manquaient. C'est pourquoi nous avons introduit une première matrice de généralisation, appelée $\mathbf{T}_\alpha$, à laquelle sont associées certaines propositions démontrées. Nous avons également défini le plan de représentation $\mathbf{P}_{\alpha,k}$ qui, avec un certain nombre de couples de paramètres (α, k), permet de retrouver les matrices de représentation usuelles mais également d'étudier des matrices intermédiaires. Pour ce faire, nous avons utilisé les spectres de ces matrices constituant le plan $\mathbf{P}_{\alpha,k}$. En effet, la classification spectrale de graphes, que ce soit *via* l'ex-

traction d'attributs des spectres (rayon spectral, *spectral gap*, …) ou *via* l'utilisation des spectres dans leurs entièretés, est un domaine particulièrement riche. Nous avons construit un noyau de corrélation spectrale, appelé CorS, et l'avons illustré par la classification de différentes bases de données de graphes connues de la littérature, allant des structures moléculaires à des réseaux sociaux en passant par des signaux vus comme des graphes grâce à l'algorithme de visibilité. Les résultats en termes d'exactitudes sont comparables voire meilleurs à ceux pouvant être atteints avec des noyaux structurels classiques. Mais l'intérêt principal a été de montrer que les résultats de classification optimaux n'ont pas été obtenus grâce aux spectres des matrices de représentation conventionnelles mais bien grâce à des matrices qui sont des combinaisons de ces dernières, ces matrices optimales étant différentes pour chaque base de données. Cet élément, majeur en classification spectrale de graphes, permet de mettre en évidence que le contenu spectral optimal diffère d'une base de données à l'autre et ne réside pas dans les matrices classiques mais bien dans une combinaison de celles-ci. Il ressort de ces résultats que chaque base de données étudiée aura alors sa propre matrice de représentation optimale. Ainsi, une perspective est de trouver un critère permettant de pressentir quelle matrice $\mathbf{P}_{\alpha^*,k^*}$ sera la matrice de représentation optimale pour une base de données spécifiques.

Publications scientifiques relatives à ce chapitre

- ✍ **T. Averty**, D. Daré-Emzivat et A.-O. Boudraa. Sur la similarité spectrale des graphes par mesure de corrélation. *GRETSI*, pages 1–4, 2023
- ✍ **T. Averty**, A.-O. Boudraa et D. Daré-Emzivat. A New Family of Graph Representation Matrices : Application to Graph and Signal Classification. *IEEE Signal Processing Letters*, 31 :2935–2939, 2024

Conclusions & perspectives

« Il n'est pas de problème dont une absence de solution ne finisse par venir à bout. »

Henri Queuille

Rappel du contexte et des verrous scientifiques

Ce travail de recherche a mis en évidence la pertinence des graphes comme représentation de la donnée qui, associés à des outils mathématiques tels que l'algèbre linéaire, permettent d'analyser, de caractériser voire de classer les objets sous-jacents, que ce soient des réseaux ou des séries temporelles. En effet, certains réseaux disposent d'une structure topologique induisant une représentation naturelle sous forme de graphes tandis que des séries temporelles peuvent se présenter comme des graphes après application d'un algorithme tel que celui du graphe de visibilité. Il existe une quantité importante d'éléments extractibles de ces graphes permettant d'atteindre les objectifs cités, qu'ils proviennent de la multitude de matrices de représentation (matrice d'adjacence $\mathbf{A}$, matrice Laplacienne $\mathbf{L}$, matrice Laplacienne sans-signe $\mathbf{Q}$, ...), notamment *via* leurs valeurs et vecteurs propres, ou bien d'attributs structurels tels que la distribution des degrés. Dans ce cadre, cette thèse devait apporter des éléments de réponse aux verrous scientifiques reformulés comme suit :

— Dans le cadre de la classification de séries temporelles vues comme des graphes, il est possible de comparer ces derniers à l'aide de noyaux sur graphes. Néanmoins, ces noyaux basés sur des attributs structurels sont bien souvent complexes à calculer. Mais alors, comment comparer les graphes de visibilité d'une manière plus efficiente ?

— Lorsque les séries temporelles sont des processus stochastiques, type fBm ou fGn, les graphes de visibilité sont des outils particulièrement adaptés pour les caractériser, notamment à travers leur

coefficient de Hurst H, d'où la question suivante : est-il possible d'estimer, à l'aide de graphes de visibilité et de manière robuste, le coefficient de Hurst d'un processus stochastique ?

— Les spectres des différentes matrices de représentation révèlent des informations importantes (et éventuellement complémentaires) quant aux graphes qu'elles traduisent. Ainsi, identifier la matrice qui représente le mieux un graphe, particulièrement d'un point de vue spectral, est une question toujours ouverte et est sujet à de nombreux débats.

— Pour effectuer une classification spectrale de graphes, c'est-à-dire la classification de graphes selon le spectre de leurs matrices de représentation, il est nécessaire de tenir compte du problème de cospectralité (*i.e.* l'existence de graphes possédant le même spectre relativement à une matrice considérée). Comparer les spectres de plusieurs matrices de représentation semble alors être une démarche intéressante. Cependant, cette solution requiert un temps de calcul plus important. Existe-t-il alors une mesure de similarité spectrale, ne requérant qu'un seul spectre et permettant une bonne classification de graphes ?

— Enfin, un intérêt majeur de voir un réseau comme un graphe est de pouvoir développer des stratégies visant à étudier la vulnérabilité de ses composants. Une question que nous nous sommes posés est dans ce cas : comment quantifier efficacement cette vulnérabilité à l'aide de grandeurs issues de la théorie de l'information ?

Contributions principales

Les contributions principales de ce travail de thèse, reportées aux chapitres 2, 3 et 4, sont décrites comme suit :

✍ Dans la première partie du chapitre 2, nous étudions l'intérêt d'utiliser les graphes de visibilité dans le but de classer des séries temporelles. Pour cela, nous élaborons deux stratégies. La première est basée sur l'extraction d'attributs structurels et spectraux des graphes de visibilité mis en entrée d'un SVM doté d'un noyau RBF à l'image d'une classification conventionnelle. La deuxième est basée sur la définition d'un nouveau noyau de SVM qui, à l'aide de distances statistiques telles que la distance de Jensen-Shannon ou encore d'Hellinger, est capable de comparer de manière globale les distributions de degrés des graphes de visibilité. C'est, à notre connaissance, la première fois qu'un tel noyau est construit. Les résultats obtenus pour deux tâches, à savoir la détection d'épilepsie dans des signaux EEG et la détection d'anomalies magnétiques, appuient l'intérêt des graphes de visibilité dans leur capacité à véhiculer différemment l'information contenue dans les signaux initiaux. Dans la seconde partie du chapitre 2, nous développons

une méthode d'estimation du coefficient de Hurst H de fBm (et de fGn à une intégration près) basée sur la projection des distributions de degrés des graphes de visibilité qui leur sont associés dans un plan informationnel de Fisher-Shannon. En effet, à partir de processus stochastiques synthétiques formant dans ce plan un squelette de référence, l'estimation de H est possible grâce à une projection orthogonale sur ce dernier. Pour autant que nous sachions, c'est la première méthode d'estimation du coefficient de Hurst qui utilise la théorie des graphes. De plus, une étude menée dans ce chapitre montre que cette stratégie est rendue plus efficiente en utilisant la variante naturelle de l'algorithme de visibilité et en tronquant les distributions de degrés. En outre, une analyse des erreurs d'estimations sur des signaux synthétiques est proposée. Enfin, les estimations obtenues par cette méthode sur des séries temporelles réelles issues du monde de la finance sont cohérentes avec les estimateurs standards provenant du traitement de signal, appuyant de fait l'intérêt des graphes de visibilité combinés à une telle méthodologie.

✍ Dans le chapitre 3, nous approfondissons l'étude de l'algorithme **EIVP** proposé par Bay-Ahmed *et al.* [223], dont le point clé est de concevoir l'entropie de von Neumann comme une mesure du contenu informationnel du réseau, pour mettre en avant les vulnérabilités éventuelles des arêtes d'un graphe. En effet, l'entropie de von Neumann d'un graphe est calculée à l'aide des valeurs propres de sa matrice de densité $\boldsymbol{\rho}$. La perturbation d'une arête du graphe (et donc un lien du réseau sous-jacent) induit une modification du spectre et donc une évolution de l'entropie. Ainsi, la vulnérabilité d'une arête est définie comme la variation relative de l'entropie de von Neumann si cette arête venait à être perturbée. Le principal avantage de cette méthode est de pouvoir prendre en compte aisément les éventuelles pondérations du graphe. Cependant, quantifier la vulnérabilité de toutes les arêtes est coûteux en temps de calcul ce qui est problématique pour de grands graphes. Pour pallier cette contrainte, nous proposons l'utilisation de deux approximations de l'entropie de von Neumann permettant une réduction drastique du temps de calcul de l'algorithme précédent qui relève d'une complexité en $O(mn^3)$, le tout pour une erreur contenue. En effet, cette entropie étant une somme de fonction $f(\nu)$ où les ν sont les valeurs propres de la matrice de densité $\boldsymbol{\rho}$ du graphe, il est possible de l'approcher à l'aide d'un développement limité de la fonction f ou bien en utilisant une approximation des valeurs propres ν grâce à la théorie de perturbations matricielles.

✍ Dans le chapitre 4, nous introduisons une nouvelle matrice de représentation de graphes, notée $\mathbf{P}_{\alpha,k}$, possédant deux paramètres et permettant d'obtenir $\mathbf{A}$, $\mathbf{D}$, $\mathbf{L}$ et $\mathbf{Q}$. Cette matrice révèle de nombreux avantages. En effet, elle permet d'unifier, par définition, les théories inhérentes aux matrices précédentes et elle possède de nombreuses propriétés algébriques, notamment des domaines nécessaires de semi-définie positivé ainsi que la superposition des spectres selon le pa-

ramètre α. Bénéficier d'une seule matrice de représentation pour étudier l'évolution du contenu spectral entre les matrices traditionnelles constitue l'intérêt majeur quant à l'utilisation de $\mathbf{P}_{\alpha,k}$. De plus, pour la classification spectrale de graphes prenant en compte des éventuelles cospectralités, nous construisons un nouveau noyau de SVM basé sur une corrélation entre les spectres standardisés des matrices $\mathbf{P}_{\alpha,k}$ mis en lieu et place de la distance euclidienne d'un noyau RBF. Les performances obtenues par ce noyau pour des classifications de bases de données connues de la littérature surpassent les noyaux sur graphes standards, de surcroît avec un temps de calcul plus avantageux.

Perspectives

Bien que cette thèse ait apporté des éléments de réponse à la problématique générale relative au développement de méthodes permettant la classification, la caractérisation et l'analyse de graphes et de signaux vus comme des graphes, elle permet de dégager des perspectives pour la suite des travaux de recherches parmi les problèmes ouverts par ces derniers.

En effet, dans le chapitre 2, les signaux sont traduits en graphe à l'aide de l'algorithme de visibilité. Or, ce dernier possède de nombreuses variantes : le graphe de visibilité différentielle, signé, pénétrant, etc. Il serait intéressant de voir quelles informations ces variantes peuvent apporter dans le cadre d'une tâche de classification mais aussi pour estimer le coefficient de Hurst de processus stochastiques. En effet, nous avons mis en avant la pertinence des distributions des degrés tronquées des graphes de visibilité naturelle mais peut-être existe-t-il une variante permettant de s'affranchir de cette troncature. Par ailleurs, si cette troncature est inévitable, le choix du seuil ε fixé à 100 dans ce manuscrit est discutable. Il serait profitable de trouver une stratégie subjective pour fixer ε : un pourcentage du nombre d'échantillons de la série temporelle par exemple. L'introduction du plan Fisher-Shannon comme espace de représentation de ces distributions pour estimer le coefficient de Hurst est également un choix fort. D'autres métriques de la théorie de l'information subsistent à l'image de l'entropie de Tsallis, l'entropie de Havrda-Charvát ou encore l'entropie de Burg, qui permettent toutes de caractériser de manière globale une distribution de probabilités. Ainsi, la question de la plus-value d'utiliser un plan Fisher-Tsallis ou un plan Fisher-Burg se pose. La méthode d'estimation du coefficient de Hurst développée dans ce travail est essentiellement basée sur des fBm et, lorsqu'un fGn est mis en entrée, il est intégré pour être de type fBm et traité comme tel. Il serait avisé de développer une méthode prenant en compte les fGn natifs d'autant qu'à notre connaissance, il existe peu de stratégies abordant ce point. En outre, nous avons utilisé des séries temporelles provenant principalement du monde financier pour évaluer notre démarche. Il serait cohérent d'analyser des signaux issus de domaines variés où les phé-

nomènes qu'ils transcrivent admettent des modélisations caractéristiques de processus stochastiques. Enfin, nous avons, pour la classification et la caractérisation de séries temporelles, fait le choix d'utiliser l'attribut le plus simple d'un graphe, à savoir sa distribution des degrés. Bien qu'elle ait fait ses preuves dans ce domaine, d'autres éléments peuvent être extraits des graphes de visibilité (centralités, pondérations, etc.) et il serait judicieux d'étudier l'apport de ces derniers.

Dans le cadre de l'étude de la vulnérabilité de réseaux complexes faisant l'objet du chapitre 3, plusieurs pistes sont également à explorer. Premièrement, l'algorithme **EIVP** utilisé est basé sur des perturbations individuelles de toutes les arêtes. Il pourrait être intéressant de développer une stratégie adaptée où l'analyse de la vulnérabilité est restreinte à une sélection d'arêtes et/ou de sommets d'intérêt, permettant de modéliser une défaillance groupée et localisée dans le réseau. Par ailleurs, l'algorithme **EIVP** pourrait, bien entendu, être étendu aux sommets, donnant lieu à un algorithme **NIVP** (pour ***N**ode **I**nformational **V**ulnerability to **P**erturbation*), qui fera l'objet d'un travail prochain. Une question pratique demeure : la suppression d'un sommet est-elle comprise comme une réduction de la taille du graphe (suppression de la ligne et la colonne correspondante au sommet dans la matrice d'adjacence) ou comme une déconnexion du sommet au reste du graphe (apparition de 0 sur toute la ligne et la colonne correspondante au sommet dans la matrice d'adjacence) ? Cette interrogation n'est pas anodine car beaucoup de méthodes dépendent du nombre de sommets du graphe étudié. Enfin, une autre perspective serait de chercher un éventuel remplaçant à l'entropie de von Neumann. En effet, par construction, la matrice de densité $\boldsymbol{\rho}$ admet des valeurs propres toutes positives ayant une somme égale à 1. Ainsi, son spectre peut être interprété comme une distribution de probabilité et, dans ce cas, les distances statistiques introduites dans le chapitre 2 (Jensen-Shannon, Helligner, Bhattacharya, ...) ou encore la divergence de Kullback-Leibler peuvent être envisagées pour comparer le graphe avant et après perturbation.

Quant à l'étude de matrices d'adjacence généralisées proposée dans le chapitre 4, c'est un domaine nettement plus exploratoire. L'utilisation de ce type de matrices n'est plus à prouver surtout quand la communauté scientifique s'attache à trouver la meilleure matrice de représentation. Ainsi, pour se démarquer des autres matrices pouvant être introduites dans la littérature, il serait nécessaire d'appuyer la définition de $\mathbf{P}_{\alpha,k}$ en lui trouvant d'autres propriétés, spectrales notamment. Établir ces propriétés permettrait à cette matrice de jouer son rôle de trait d'union entre les matrices de représentation classiques telles que la matrice d'adjacence $\mathbf{A}$, la matrice Laplacienne $\mathbf{L}$ et Laplacienne sans-signe $\mathbf{Q}$. De plus, il est clair que la matrice $\mathbf{P}_{\alpha,k}$ est « pilotée par les données » (*data driven* en anglais). En effet, il existe, pour chaque base de données de graphes ayant des structures bien différentes, un couple (α, k) optimal au sens des résultats de classification obtenus grâce au noyau CorS développé dans cette thèse. Est-il possible de pressentir quel serait ce couple de paramètres optimaux à partir des propriétés spec-

trales précédentes et certains attributs structurels ? Apporter des éléments de réponse à cette question améliorerait de manière évidente la compréhension de l'évolution des spectres entre matrices de représentation et, *de facto*, des liens existants entre structure et spectre.

Pour conclure, l'analyse, la caractérisation et la classification de graphes, qu'ils proviennent de réseaux physiques réels ou bien qu'ils soient construits à partir de séries temporelles, a encore beaucoup à apporter dans un monde dans lequel la quantité de données ne fait que croître et où il est nécessaire de les structurer.

Annexe A

Attributs d'un signal discret

$$\mathbf{s} = (s_i)_{1 \leq i \leq n}$$

Les éléments qui suivent sont issus d'une recherche d'attributs pouvant être extraits d'un signal discret, tirés de la littérature ou encore de la documentation Matlab [286–290]. Bien entendu, cette liste ne se veut pas être exhaustive et est amenée à être complétée.

Attributs temporels

Nom	Formule
Énergie	$E = \sum_{i=1}^{n} \lvert s_i \rvert^2$
Puissance	$P = \frac{1}{n} \sum_{i=1}^{n} \lvert s_i \rvert^2$
Moyenne quadratique	$\mathrm{RMS} = \sqrt{P}$
Racine	$R = \left(\frac{1}{n} \sum_{i=1}^{n} \sqrt{\lvert s_i \rvert} \right)^2$
Longueur	n
Minimum	$\min_{1 \leq i \leq n} s_i$
Maximum	$\max_{1 \leq i \leq n} s_i$

suite à la page suivante

Nom	Formule
Moyenne (espérance)	$\mu = \frac{1}{n}\sum_{i=1}^{n} s_i$
Écart-type	$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (s_i - \mu)^2}$
Variance	σ^2
Asymétrie	$\mu_3 = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{s_i - \mu}{\sigma}\right)^3$
Kurtosis	$\mu_4 = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{s_i - \mu}{\sigma}\right)^4$
Facteur d'asymétrie	$\frac{\mu_3}{\mathrm{RMS}^3}$
Facteur de kurtosis	$\frac{\mu_4}{\mathrm{RMS}^4}$
Moyenne absolue	$\overline{s} = \frac{1}{n}\sum_{i=1}^{n} \lvert s_i \rvert$
Facteur de forme	$\frac{\mathrm{RMS}}{\overline{s}}$
Facteur de crête	$\frac{\max_{1 \leq i \leq n} \lvert s_i \rvert}{\mathrm{RMS}}$
Facteur d'impulsion	$\frac{\max_{1 \leq i \leq n} \lvert s_i \rvert}{\overline{s}}$
Facteur de dégagement	$\frac{\max_{1 \leq i \leq n} \lvert s_i \rvert}{R}$
Centroïde temporel	$\frac{\sum_{i=1}^{n} i s_i}{\sum_{i=1}^{n} s_i}$
Taux d'attaque	$\max_{1 \leq i \leq n} \left(\frac{s_i - s_{i-1}}{n}\right)$
Taux de décroissance	$\min_{1 \leq i \leq n} \left(\frac{s_i - s_{i+1}}{n}\right)$
Taux de passages par zéro	$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\mathbb{R}^-}(s_i s_{i-1})$
Taux de passages par la moyenne	$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\mathbb{R}^-}((s_i - \mu)(s_{i-1} - \mu))$
Entropie de Shannon	$-\sum_{i=1}^{n} \mathbb{P}(s_i) \log_2(\mathbb{P}(s_i))$
Entropie de Rényi	$\frac{1}{1-\alpha} \log_2 \left(\sum_{i=1}^{n} \mathbb{P}(s_i)^\alpha\right)$

Attributs spectraux

Soit $\mathbf{S} = (S(f_k))_{1 \leq k \leq N}$ le module de la transformée de Fourier discrète du signal $\mathbf{s}$ calculée sur N fréquences $(f_k)_{1 \leq k \leq N}$.

Nom	Formule
Fréquence moyenne	$\bar{f} = \frac{1}{N} \sum_{k=1}^{N} f_k$
Amplitude moyenne	$\bar{S} = \frac{1}{N} \sum_{k=1}^{N} S(f_k)$
Centroïde spectral	$\mu_1 = \frac{\sum_{k=1}^{N} f_k S(f_k)}{\sum_{k=1}^{N} S(f_k)}$
Étalement spectral	$\mu_2 = \sqrt{\frac{\sum_{k=1}^{N} (f_k - \mu_1)^2 S(f_k)}{\sum_{k=1}^{N} S(f_k)}}$
Asymétrie spectral	$\mu_3 = \sqrt{\frac{\sum_{k=1}^{N} (f_k - \mu_1)^3 S(f_k)}{\mu_2^3 \sum_{k=1}^{N} S(f_k)}}$
Kurtosis spectral	$\mu_4 = \sqrt{\frac{\sum_{k=1}^{N} (f_k - \mu_1)^4 S(f_k)}{\mu_2^4 \sum_{k=1}^{N} S(f_k)}}$
Largeur de bande	$\sqrt{\frac{\sum_{k=1}^{N} f_k^2 \lvert S(f_k)\rvert^2}{\sum_{k=1}^{N} \lvert S(f_k)\rvert^2}}$
Entropie spectrale	$\frac{-\sum_{k=1}^{N} \mathbb{P}(f_k) \log(\mathbb{P}(f_k))}{\log(N)}, \quad \mathbb{P}(f_k) = \frac{\lvert S(f_k)\rvert^2}{\sum_{j=1}^{N} \lvert S(f_j)\rvert^2}$
Aplatissement spectral	$\frac{\sqrt[N]{\prod_{k=1}^{N} S(f_k)}}{\frac{1}{N} \sum_{k=1}^{N} S(f_k)}$
Crête spectrale	$\frac{\max_{1 \leq k \leq N} (S(f_k))}{\frac{1}{N} \sum_{k=1}^{N} S(f_k)}$
Pente spectrale	$\frac{\sum_{k=1}^{N} (f_k - \bar{f})(S(f_k) - \bar{S})}{\sum_{k=1}^{N} (f_k - \bar{f})^2}$
Décroissance spectrale	$\frac{\sum_{k=2}^{N} \frac{S(f_k) - S(f_1)}{k-1}}{\sum_{k=2}^{N} S(f_k)}$
Point d'inflexion spectrale	Valeur i telle que $\sum_{k=1}^{i} \lvert S(f_k)\rvert = 0.95 \sum_{k=1}^{N} S(f_k)$

Annexe B

Algorithme de Davies et Harte

Cette annexe est extraite d'une traduction du rapport complet de Ton Dieker [196].

L'algorithme de génération de fBm de Davies et Harte [194] a été introduit en 1987 avant d'être généralisé simultanément par Dietrich et Newsam [291] et Wood et Chan [292]. Comme les autres méthodes de génération de fBm (méthode de Hosking ou méthode de Cholesky), cette dernière tente de trouver une « racine carrée » de la matrice de covariance $\mathbf{\Gamma}$, c'est-à-dire une matrice carrée $\mathbf{G}$ telle que $\mathbf{\Gamma} = \mathbf{G}\mathbf{G}^\top$. Supposons qu'une série temporelle à n échantillons soit recherchée et que la taille de la matrice de covariance $\mathbf{\Gamma}$ soit une puissance de deux, c'est-à-dire $n = 2^g$ avec $g \in \mathbb{N}$. Pour pouvoir obtenir une telle matrice $\mathbf{G}$ de manière efficiente, l'idée principale est de plonger $\mathbf{\Gamma}$ dans une matrice de covariance circulante $\mathbf{C}$ de taille $2n = 2^{g+1}$. Plus précisément, la matrice $\mathbf{C}$ est définie par

$$\begin{pmatrix}
\gamma(0) & \gamma(1) & \cdots & \gamma(n-1) & 0 & \gamma(n-1) & \gamma(n) & \cdots & \gamma(2) & \gamma(1) \\
\gamma(1) & \gamma(0) & \cdots & \gamma(n-2) & \gamma(n-1) & 0 & \gamma(n-1) & \cdots & \gamma(3) & \gamma(2) \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
\gamma(n-1) & \gamma(n-2) & \cdots & \gamma(0) & \gamma(1) & \gamma(2) & \gamma(3) & \cdots & \gamma(n-1) & 0 \\
0 & \gamma(n-1) & \cdots & \gamma(1) & \gamma(0) & \gamma(1) & \gamma(2) & \cdots & \gamma(n-2) & \gamma(n-1) \\
\gamma(n-1) & 0 & \cdots & \gamma(2) & \gamma(1) & \gamma(0) & \gamma(1) & \cdots & \gamma(n-3) & \gamma(n-2) \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
\gamma(1) & \gamma(2) & \cdots & 0 & \gamma(n-1) & \gamma(n-2) & \gamma(n-3) & \cdots & \gamma(1) & \gamma(0)
\end{pmatrix}. \tag{B.1}$$

Notons que la i^{e} ligne peut être construite en décalant la première ligne de $i-1$ « places » vers la droite et en remplissant les éléments enlevés à gauche. Notons également que la matrice est symétrique et que le coin supérieur gauche (la zone grisée) est en réalité la matrice de covariance $\mathbf{\Gamma}$. Lorsque $\gamma(\cdot)$ est la fonction de covariance du bruit Gaussien fractionnaire de coefficient de Hurst H

$$\gamma(k) = \frac{1}{2}\left[|k-1|^{2H} - 2|k|^{2H} + |k+1|^{2H}\right], \quad 1 \leq k \leq n \tag{B.2}$$

et que les zéros de la matrice sont remplacés par $\gamma(n)$, la matrice est définie positive [293]. Il convient de préciser que la matrice circulante n'est pas nécessairement définie positive pour les fonctions générales d'autocovariance. La manière de traiter cette situation est décrite dans [291, 292]. Pour certaines fonctions d'autocovariance, il est même impossible de plonger la matrice de covariance dans une matrice définie positive. Dans ce cas, les auteurs proposent de rendre la méthode approximative. Si le nombre d'échantillons requis n'est pas une puissance de 2, plus de zéros doivent être ajoutés sur la première ligne pour obtenir une matrice circulante. Cependant, cela ne change pas la philosophie de la méthode. En réalité, l'algorithme repose sur le théorème suivant.

Théorème B.1. *Toute matrice circulante* $\mathbf{C}$ *peut être décomposée comme suit :*

$$\mathbf{C} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{\mathsf{H}} \tag{B.3}$$

où $\mathbf{\Lambda}$ *est la matrice diagonale composée des valeurs propres de* $\mathbf{C}$ *et* $\mathbf{Q}$ *est la matrice unitaire*[1] *définie par*

$$[\mathbf{Q}]_{jk} = \frac{1}{\sqrt{2n}} \mathrm{e}^{-2\pi \mathrm{i} \frac{jk}{2n}}, \qquad j, k \in [\![0, 2n-1]\!]. \tag{B.4}$$

Les valeurs propres λ_k, qui constituent la matrice $\mathbf{\Lambda}$, sont données par

$$\lambda_k = \sum_{j=0}^{2n-1} r_j \mathrm{e}^{2\pi \mathrm{i} \frac{jk}{2n}}, \qquad k \in [\![0, 2n-1]\!] \tag{B.5}$$

avec r_j le $(j+1)^{\mathrm{e}}$ élément de la première ligne de $\mathbf{C}$. Or, la matrice $\mathbf{C}$ est définie positive et symétrique. Ainsi, les valeurs propres λ_k sont toutes positives et réelles. Par conséquent, la matrice ayant pour valeurs propres $(\sqrt{\lambda_k})_{1 \leq k \leq 2n-1}$ et les mêmes vecteurs propres que $\mathbf{C}$ est également définie positive et réelle. Notons que, comme $\mathbf{C} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{\mathsf{H}}$ et que $\mathbf{Q}$ est unitaire, la matrice

$$\mathbf{S} = \mathbf{Q}\mathbf{\Lambda}^{1/2}\mathbf{Q}^{\mathsf{H}} \tag{B.6}$$

satisfait $\mathbf{S}\mathbf{S}^{\mathsf{H}} = \mathbf{S}\mathbf{S}^{\top} = \mathbf{C}$. En conclusion, la matrice $\mathbf{S}$ vérifie exactement la propriété désirée. À présent, pour obtenir un tirage du processus, il faut trouver un moyen de simuler

$$\mathbf{Q}\mathbf{\Lambda}^{1/2}\mathbf{Q}^{\mathsf{H}}\mathbf{v} \tag{B.7}$$

où $\mathbf{v}$ est un vecteur dont les éléments sont i.i.d et suivent une loi normale. Cela peut être fait *via* trois étapes successives :

① Calculer les valeurs propres à l'aide de l'équation (B.5). Pour rendre cette étape plus efficiente, il est possible de se servir de la FFT (*Fast Fourier Transform*). En effet, pour une suite complexe $(\alpha_k)_{0 \leq k \leq j-1}$, la

1. La matrice $\mathbf{Q}$ est dite unitaire si $\mathbf{Q}\mathbf{Q}^{\mathsf{H}} = \mathbf{I}$

FFT permet de calculer efficacement la transformée de Fourier de cette suite, c'est-à-dire

$$\sum_{k=0}^{j-1} \alpha_k \mathrm{e}^{2\pi \mathrm{i} \frac{\ell k}{j}}, \quad 0 \leq \ell \leq j-1. \tag{B.8}$$

Quand j est une puissance de 2, le nombre de calculs requis par la FFT est de l'ordre de $j \log(j)$, ce qui est un gain considérable par rapport aux j^2 calculs requis par le calcul direct.

② Calculer $\mathbf{w} = \mathbf{Q}^{\mathsf{H}}\mathbf{v}$. Utilisant la structure de covariance de $\mathbf{w}$, le principe de simulation est le suivant :

— Générer deux variables aléatoires w_0 et w_n suivant toutes deux des lois normales ;

— Pour $1 \leq j < n$, générer deux variables aléatoires $v_j^{(1)}$ et $v_j^{(2)}$ indépendantes suivant toutes deux des lois normales et définir

$$w_j = \frac{1}{\sqrt{2}} \left(v_j^{(1)} + \mathrm{i} v_j^{(2)} \right) \tag{B.9}$$

$$w_{2n-j} = \frac{1}{\sqrt{2}} \left(v_j^{(1)} - \mathrm{i} v_j^{(2)} \right). \tag{B.10}$$

Le vecteur $\mathbf{w}$ résultant a la même distribution que $\mathbf{Q}^{\mathsf{H}}\mathbf{V}$.

③ Calculer $\mathbf{z} = \mathbf{Q}\mathbf{\Lambda}^{1/2}\mathbf{w}$:

$$z_k = \frac{1}{\sqrt{2n}} \sum_{j=0}^{2n-1} \sqrt{\lambda_j} w_j \mathrm{e}^{-2\pi \mathrm{i} \frac{jk}{2n}}. \tag{B.11}$$

Une fois de plus, ce calcul peut être fait grâce à la FFT. En effet, la suite $(z_k)_{0 \leq k \leq 2n-1}$ est la transformée de Fourier de

$$w_k := \begin{cases} \sqrt{\frac{\lambda_k}{2n}} v_k^{(1)}, & k = 0 \\ \sqrt{\frac{\lambda_k}{4n}} \left(v_k^{(1)} + \mathrm{i} v_k^{(2)} \right), & k = 1, \ldots, n-1 \\ \sqrt{\frac{\lambda_k}{2n}} v_k^{(1)}, & k = n \\ \sqrt{\frac{\lambda_k}{4n}} \left(v_{2n-k}^{(1)} - \mathrm{i} v_{2n-k}^{(2)} \right), & k = n+1, \ldots, 2n-1 \end{cases} \tag{B.12}$$

Un tirage d'un bruit Gaussien fractionnaire est obtenu en prenant les n premiers éléments de $\mathbf{z}$.

Il est aisé de constater, à l'aide de l'équation (B.1), que les n derniers éléments de $\mathbf{z}$ ont également la structure de covariance souhaitée. Il semble donc possible d'obtenir un deuxième tirage « gratuitement ». Cependant, ces deux tirages ne peuvent pas être agrégés pour obtenir un tirage double car la structure de corrélation entre les deux tirages n'est pas conforme au bruit Gaussien fractionnaire. De plus, les deux tirages ne sont pas indépendants. Ainsi, ce deuxième tirage est souvent mis de côté. Le principal avantage de cette méthode est sa rapidité. Plus précisément, le nombre de calculs est de l'ordre de $n \log(n)$ pour un tirage à n échantillons. Si plus de tirages sont requis, les valeurs propres n'ont besoin que d'être une unique fois. Toutefois, les calculs de l'étape ② doivent être effectués séparément pour chaque tirage. Ainsi, le nombre de calculs est toujours en $n \log(n)$.

Bibliographie

[1] Tristan Gaudiaut. Le Big Bang du Big Data. `https://fr.statista.com/infographie/17800/big-data-evolution-volume-donnees-numeriques-genere-dans-le-monde/`, 2021. Consulté le : 14/03/2025.

[2] Ljubisa Stankovic, Danilo Mandic, Milos Dakovic, Milos Brajovic, Bruno Scalzo et Anthony G. Constantinides. Graph signal processing–Part II : Processing and analyzing signals on graphs. *Preprint arXiv :1909.10325*, 2019.

[3] Réka Albert, István Albert et Gary L. Nakarado. Structural vulnerability of the North American power grid. *Physical Review E*, 69(2):1–5, 2004.

[4] Ryan Kinney, Paolo Crucitti, Réka Albert et Vito Latora. Modeling cascading failures in the North American power grid. *The European Physical Journal B-Condensed Matter and Complex Systems*, 46(1): 101–107, 2005.

[5] Rui Carvalho, Lubos Buzna, Flavio Bono, Eugenio Gutiérrez, Wolfram Just et David Arrowsmith. Robustness of trans-European gas networks. *Physical Review E*, 80(1):016106, 2009.

[6] Min Ouyang. Review on modeling and simulation of interdependent critical infrastructure systems. *Reliability engineering & System safety*, 121:43–60, 2014.

[7] Shuliang Wang, Jianhua Zhang et Na Duan. Multiple perspective vulnerability analysis of the power network. *Physica A : Statistical Mechanics and its Applications*, 492:1581–1590, 2018.

[8] Fan R. K. Chung. *Spectral graph theory*, volume 92. American Mathematical Society, 1997.

[9] Dragoš Cvetković, Peter Rowlinson et Slobodan K. Simić. Signless Laplacians of finite graphs. *Linear Algebra and its Applications*, 423(1):155–171, 2007.

[10] David I. Shuman, Sunil K. Narang, Pascal Frossard, Antonio Ortega et Pierre Vandergheynst. The Emerging Field of Signal Processing on Graphs : Extending High-Dimensional Data Analysis to Networks and Other Irregular Domains. *IEEE Signal Processing Magazine*, 30(3):83–98, 2013.

[11] Aliaksei Sandryhaila et José M. F. Moura. Discrete Signal Processing on Graphs : Frequency Analysis. *IEEE Transactions on Signal Processing*, 62(12):3042–3054, 2014.

[12] Benjamin Girault, Paulo Gonçalves et Éric Fleury. Translation on graphs : An isometric shift operator. *IEEE Signal Processing Letters*, 22(12):2416–2420, 2015.

[13] Guidong Zhang, Zhong Li, Bo Zhang et Wolfgang A. Halang. Understanding the cascading failures in Indian power grids with complex networks theory. *Physica A : Statistical Mechanics and its Applications*, 392(15):3273–3280, 2013.

[14] Moniek de Jong. Tracing the downfall of the Nord Stream 2 gas pipeline. *Wiley Interdisciplinary Reviews : Energy and Environment*, 13(1), 2024.

[15] Weiwang Wang, Xilin Yan, Shengtao Li, Lina Zhang, Jun Ouyang et Xianfeng Ni. Failure of submarine cables used in high-voltage power transmission : Characteristics, mechanisms, key issues and prospects. *IET Generation, Transmission & Distribution*, 15(9):1387–1402, 2021.

[16] Charline Vergne. Un câble sous-marin saboté en mer Baltique : la Russie, qui a laissé un vaste indice, dans le viseur de la Finlande. `https://www.geo.fr/geopolitique/un-cable-sous-marin-sabote-en-mer-baltique-la-russie-qui-a-laisse-un-vaste-indice-dans-le-viseur-de-la-finlande-223932`, 2024. Consulté le : 14/03/2025.

[17] Antonio Ortega, Pascal Frossard, Jelena Kovačević, José M. F. Moura et Pierre Vandergheynst. Graph signal processing : Overview, challenges, and applications. *Proceedings of the IEEE*, 106(5):808–828, 2018.

[18] Benjamin Girault. *Signal processing on graphs-contributions to an emerging field*. Thèse de doctorat, ENS Lyon, 2015.

[19] Benjamin Ricaud, Pierre Borgnat, Nicolas Tremblay, Paulo Gonçalves et Pierre Vandergheynst. Fourier could be a data scientist : From graph Fourier transform to signal processing on graphs. *Comptes Rendus Physique*, 20(5):474–488, 2019.

[20] Nicolas Tremblay. *Réseaux et signal : des outils de traitement du signal pour l'analyse des réseaux*. Thèse de doctorat, ENS Lyon, 2014.

[21] Lucas Lacasa, Bartolo Luque, Fernando Ballesteros, Jordi Luque et Juan Carlos Nuno. From time series to complex networks : The visibility graph. *Proceedings of the National Academy of Sciences*, 105(13): 4972–4975, 2008.

[22] Lucas Lacasa, Bartolo Luque, Jordi Luque et Juan Carlos Nuno. The visibility graph : A new method for estimating the Hurst exponent of fractional Brownian motion. *Europhysics Letters*, 86(3):30001, 2009.

[23] Bartolo Luque, Lucas Lacasa, Fernando Ballesteros et Jordi Luque. Horizontal visibility graphs : Exact results for random time series. *Physical Review E*, 80(4):046103, 2009.

[24] Bruna Amin Goncalves. *Análise de séries temporais via grafo de visibilidade horizontal e teoria da informação*. Thèse de doctorat, Universidade Federal de Minas Gerais, 2016.

[25] Bruna Amin Gonçalves, Laura Carpi, Osvaldo A. Rosso et Martín G. Ravetti. Time series characterization via horizontal visibility graph and Information Theory. *Physica A : Statistical Mechanics and its Applications*, 464:93–102, 2016.

[26] Vladimir N. Vapnik. Pattern recognition using generalized portrait method. *Automation and remote control*, 24(6):774–780, 1963.

[27] Mark A. Aizerman, Emmanuil M. Braverman et Lev I. Rozonoer. Theoretical foundations of potential function method in pattern recognition. *Automation and Remote Control*, 25(6):917–936, 1964.

[28] Karsten Michael Borgwardt. *Graph Kernels*. Thèse de doctorat, Ludwig-Maximilians-Universität München, 2007.

[29] S. V. N. Vishwanathan, Nicol N. Schraudolph, Risi Kondor et Karsten M. Borgwardt. Graph Kernels. *Journal of Machine Learning Research*, 11:1201–1242, 2010.

[30] Karsten M. Borgwardt et Hans-Peter Kriegel. Shortest-path kernels on graphs. In *Fifth IEEE international conference on data mining (ICDM'05)*, pages 1–8. IEEE, IEEE, 2005.

[31] Nataša Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2): 177–183, 2007.

[32] Hadj-Ahmed Bay-Ahmed, Abdel-Ouahab Boudraa, Delphine Daré-Emzivat et Yves Préaux. Classification des signaux sur graphes par mesures spectrales algébriques. In *GRETSI*, pages 1–4, 2017.

[33] Pierre Geurts. Pattern extraction for time series classification. In *European conference on principles of data mining and knowledge discovery*, pages 115–127. Springer, 2001.

[34] Pari Jahankhani, Vassilis Kodogiannis et Kenneth Revett. EEG signal classification using wavelet feature extraction and neural networks. In *IEEE John Vincent Atanasoff 2006 International Symposium on Modern Computing (JVA'06)*, pages 120–124. IEEE, 2006.

[35] Shiliang Sun et Changshui Zhang. Adaptive feature extraction for EEG signal classification. *Medical and Biological Engineering and Computing*, 44(10):931–935, 2006.

[36] Ahmet M. Elbir. DeepMUSIC : Multiple signal classification via deep learning. *IEEE Sensors Letters*, 4(4):1–4, 2020.

[37] Guohun Zhu, Yan Li et Peng Paul Wen. Epileptic seizure detection in EEGs signals using a fast weighted horizontal visibility algorithm. *Computer Methods and Programs in Biomedicine*, 115(2):64–75, 2014.

[38] Supriya Supriya, Siuly Siuly, Hua Wang, Jinli Cao et Yanchun Zhang. Weighted visibility graph with complex network features in the detection of epilepsy. *IEEE Access*, 4:6554–6566, 2016.

[39] T. Rajadurai et C Valliyammai. Epileptic Seizure Prediction Using Weighted Visibility Graph. In *International Conference on Soft Computing Systems (ICSCS)*, pages 453–461, 2018.

[40] Nino Shervashidze, Pascal Schweitzer, Erik Jan Van Leeuwen, Kurt Mehlhorn et Karsten M. Borgwardt. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12(9):2539–2561, 2011.

[41] Harold Edwin Hurst. Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, 116(1):770–799, 1951.

[42] Christophe Vignat et Jean-François Bercher. Analysis of signals in the Fisher–Shannon information plane. *Physics Letters A*, 312(1-2):27–33, 2003.

[43] Ivan Gutman. Graph-theoretical formulation of Forsman's equations. *The Journal of Chemical Physics*, 68(5):2523–2524, 1978.

[44] Ivan Gutman et Bo Zhou. Laplacian energy of a graph. *Linear Algebra and its Applications*, 414(1):29–37, 2006.

[45] Filippo Passerini et Simone Severini. Quantifying complexity in networks : the von Neumann entropy. *International Journal of Agent Technologies and Systems (IJATS)*, 1(4):58–67, 2009.

[46] Samuel L. Braunstein, Sibasish Ghosh et Simone Severini. The Laplacian of a graph as a density matrix : a basic combinatorial approach to separability of mixed states. *Annals of Combinatorics*, 10(3):291–317, 2006.

[47] Richard C. Wilson et Ping Zhu. A study of graph spectra for comparing graphs and trees. *Pattern Recognition*, 41(9):2833–2841, 2008.

[48] Edwin R. Van Dam et Willem H. Haemers. Which graphs are determined by their spectrum ? *Linear Algebra and its Applications*, 373:241–272, 2003.

[49] Edwin R. Van Dam, Willem H. Haemers et Jack H. Koolen. Cospectral graphs and the generalized adjacency matrix. *Linear Algebra and its Applications*, 423(1):33–41, 2007.

[50] Willem H. Haemers et Gholam Reza Omidi. Universal adjacency matrices with two eigenvalues. *Linear Algebra and its Applications*, 435(10):2520–2529, 2011.

[51] Dragoš Cvetkovič. Spectral recognition of graphs. *Yugoslav Journal of Operations Research*, 22(2):145–161, 2016.

[52] Ralucca Gera, Lázaro Alonso, Brian Crawford, Jeffrey House, J. A. Mendez-Bermudez, Thomas Knuth et Ryan Miller. Identifying network structure similarity using spectral graph theory. *Applied Network Science*, 3(1):1–15, 2018.

[53] Hadj-Ahmed Bay-Ahmed, Abdel-Ouahab Boudraa et Delphine Dare-Emzivat. A joint spectral similarity measure for graphs classification. *Pattern Recognition Letters*, 120:1–7, 2019.

[54] Lorenzo Dall'Amico, Romain Couillet et Nicolas Tremblay. Classification spectrale par la laplacienne déformée dans des graphes réalistes. In *GRETSI*, pages 1–4, 2019.

[55] Vladimir Nikiforov. Merging the $\mathbf{A}$- and $\mathbf{Q}$-spectral theories. *Applicable Analysis and Discrete Mathematics*, 11(1):81–107, 2017.

[56] Sai Wang, Dein Wong et Fenglei Tian. Bounds for the largest and the smallest $\mathbf{A}_\alpha$ eigenvalues of a graph in terms of vertex degrees. *Linear Algebra and its Applications*, 590:210–223, 2020.

[57] Petter Holme, Beom Jun Kim, Chang No Yoon et Seung Kee Han. Attack vulnerability of complex networks. *Physical Review E*, 65(5):1–15, 2002.

[58] Chen Chen, Hanghang Tong, B Aditya Prakash, Charalampos E Tsourakakis, Tina Eliassi-Rad, Christos Faloutsos et Duen Horng Chau. Node immunization on large graphs : Theory and algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):113–126, 2016.

[59] Vito Latora et Massimo Marchiori. Vulnerability and protection of infrastructure networks. *Physical Review E*, 71(1):015103, 2005.

[60] Luca Dall'Asta, Alain Barrat, Marc Barthélemy et Alessandro Vespignani. Vulnerability of weighted networks. *Journal of Statistical Mechanics : Theory and Experiment*, 2006(04):P04006, 2006.

[61] Giorgia Minello, Luca Rossi et Andrea Torsello. On the von Neumann entropy of graphs. *Journal of Complex Networks*, 7(4):491–514, 2019.

[62] Scott Freitas, Diyi Yang, Srijan Kumar, Hanghang Tong et Duen Horng Chau. Graph Vulnerability and Robustness : A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(6):5915–5934, 2022.

[63] Leonhard Euler. Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, pages 128–140, 1741.

[64] Brian Hopkins et Robin J. Wilson. The truth about Königsberg. *The College Mathematics Journal*, 35(3):198–207, 2004.

[65] Carl Hierholzer et Chr Wiener. Über die Möglichkeit, einen Linienzug ohne Wiederholung und ohne Unterbrechung zu umfahren. *Mathematische Annalen*, 6(1):30–32, 1873.

[66] Hassler Whitney. Congruent Graphs and the Connectivity of Graphs. *American Journal of Mathematics*, 54:61–79, 1932.

[67] Brendan D. McKay. Practical graph isomorphism. *Congressus Numerantium*, 30:45–87, 1981.

[68] Vladimir Nikiforov. Eigenvalues and degree deviation in graphs. *Linear Algebra and its Applications*, 414(1):347–360, 2006.

[69] Tom A. B. Snijders. The degree variance : an index of graph heterogeneity. *Social Networks*, 3(3):163–174, 1981.

[70] Ivan Gutman et Nenad Trinajstić. Graph theory and molecular orbitals. Total φ-electron energy of alternant hydrocarbons. *Chemical Physics Letters*, 17(4):535–538, 1972.

[71] Gholam Hassan Shirdel, Hassan Rezapour et AM Sayadi. The hyper-Zagreb index of graph operations. *Iranian Journal of Mathematical Chemistry*, 4(2):213–220, 2013.

[72] Milan Randic. Characterization of molecular branching. *Journal of the American Chemical Society*, 97(23):6609–6615, 1975.

[73] Siemion Fajtlowicz. On Conjectures of Graffiti. In J. Akiyama, Y. Egawa et H. Enomoto, éditeurs. *Graph Theory and Applications*, volume 38 de *Annals of Discrete Mathematics*, pages 113–118. Elsevier, 1988.

[74] Jelena Sedlar, Dragan Stevanović et Alexander Vasilyev. On the inverse sum indeg index. *Discrete Applied Mathematics*, 184:202–212, 2015.

[75] Harry Wiener. Structural determination of paraffin boiling points. *Journal of the American Chemical Society*, 69(1):17–20, 1947.

[76] Réka Albert et Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47, 2002.

[77] Stanley Milgram. The small world problem. *Psychology Today*, 2(1):60–67, 1967.

[78] Sergey Edunov, Carlos Diuk, Ismail Onur Filiz, Smriti Bhagat et Moira Burke. Three and a half degrees of separation. `https://research.facebook.com/blog/2016/2/three-and-a-half-degrees-of-separation/`, 2016. Consulté le : 14/03/2025.

[79] Raksha Ramakrishna et Anna Scaglione. Grid-graph signal processing (grid-GSP) : A graph signal processing framework for the power grid. *IEEE Transactions on Signal Processing*, 69:2725–2739, 2021.

[80] Alain Barrat, Marc Barthelemy, Romualdo Pastor-Satorras et Alessandro Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences*, 101(11):3747–3752, 2004.

[81] Antoniou Ioannis et Tsompa Eleni. Statistical Analysis of Weighted Networks. *Discrete Dynamics in Nature and Society*, pages 1–16, 2007.

[82] Hadj-Ahmed Bay-Ahmed. *Classification of signals and graphs by algebraic spectral approaches*. Thèse de doctorat, Université de Bretagne Occidentale, 2018.

[83] Paul Erdös et Alfréd Rényi. On random graphs, I. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

[84] Edgar N. Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

[85] Mark E. J. Newman, Steven H. Strogatz et Duncan J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2):026118, 2001.

[86] Duncan J. Watts et Steven H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393(6684):440–442, 1998.

[87] Bernard J. McClelland. Properties of the latent roots of a matrix : the estimation of π-electron energies. *The Journal of Chemical Physics*, 54(2):640–643, 1971.

[88] Lothar Von Collatz et Ulrich Sinogowitz. Spektren endlicher grafen. *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg*, 21:63–77, 1957.

[89] Allen J. Schwenk et Robin J. Wilson. On the eigenvalues of a graph. *Selected topics in graph theory*, 1978.

[90] Zoran Stanić. Graphs with small spectral gap. *The Electronic Journal of Linear Algebra*, 26:417–432, 2013.

[91] Norman Biggs. *Algebraic Graph Theory*. Cambridge University Press, 1993.

[92] Chris Godsil et Gordon F. Royle. *Algebraic Graph Theory*, volume 207. Springer Science & Business Media, 2001.

[93] Graovac, Ante and Gutman, Ivan and John, Peter E. and Vidović, Dušica and Vlah, Ivana. On Statistics of Graph Energy. *Zeitschrift für Naturforschung A*, 56:307–311, 2001.

[94] Ivan Gutman. The energy of a graph : old and new results. In *Algebraic Combinatorics and Applications (ALCOMA)*, pages 196–211. Springer, 2001.

[95] Dragos M. Cvetkovic, Michael Doob, Ivan Gutman et Aleksandar Torgašev. *Recent results in the theory of graph spectra*, volume 36. Elsevier, 1988.

[96] Bo Zhou, Ivan Gutman, José Antonio de la Pena, Juan Rada et Leonel Mendoza. On spectral moments and energy of graphs. *MATCH Communications in Mathematical and in Computer Chemistry*, 57:183–191, 2007.

[97] Xueliang Li, Yongtang Shi et Ivan Gutman. *Graph Energy*. Springer, 2012.

[98] Ivan Gutman et Jia-Yu Shao. The energy change of weighted graphs. *Linear Algebra and its Applications*, 435(10):2425–2431, 2011.

[99] Ivan Gutman, Tanja Soldatović et Dušica Vidović. The energy of a graph and its size dependence. A Monte Carlo approach. *Chemical Physics Letters*, 297:428–432, 1998.

[100] Gilles Caporossi, Dragoš Cvetković, Ivan Gutman et Pierre Hansen. Variable Neighborhood Search for Extremal Graphs. 2. Finding Graphs with Extremal Energy. *Journal of Chemical Information and Computer Science*, 39(6):984–996, 1999.

[101] Jack H. Koolen et Vincent Moulton. Maximal Energy Graphs. *Advances in Applied Mathematics*, 26(1): 47–52, 2001.

[102] Jack H. Koolen et Vincent Moulton. Maximal Energy Bipartite Graphs. *Graphs and Combinatorics*, 19(1):131–135, 2003.

[103] Vladimir Nikiforov. Graphs and matrices with maximal energy. *Journal of Mathematical Analysis and Applications*, 327(1):735–738, 2007.

[104] Diego O. Bravo, Florencia Cubría et Juan Rada. Energy of matrices. *Applied Mathematics and Computation*, 312:149–157, 2017.

[105] Ernesto Estrada. Characterization of 3D molecular structure. *Chemical Physics Letters*, 319:713–718, 2000.

[106] José Antonio De La Peña, Ivan Gutman et Juan Rada. Estimating the Estrada index. *Linear Algebra and its Applications*, 427(1):70–76, 2007.

[107] Wu Jun, Mauricio Barahona, Tan Yue-Jin et Deng Hong-Zhong. Natural connectivity of complex networks. *Chinese Physics Letters*, 27(7):078902, 2010.

[108] William C. Forsman. Graph theory and the statistics and dynamics of polymer chains. *The Journal of Chemical Physics*, 65(10):4111–4115, 1976.

[109] William N. Anderson et Thomas D. Morley. Eigenvalues of the Laplacian of a graph. *Linear and Multilinear Algebra*, 18(2):141–145, 1985.

[110] Robert Grone, Russell Merris et Viakalathur Shankar Sunder. The Laplacian Spectrum of a Graph. *SIAM Journal on Matrix Analysis and Applications*, 11(2):218–238, 1990.

[111] Bojan Mohar. Laplace eigenvalues of graphs - a survey. *Discrete Mathematics*, 109:171–183, 1992.

[112] Robert Grone et Russell Merris. The Laplacian Spectrum of a Graph II. *SIAM Journal on Discrete Mathematics*, 7(2):221–229, 1994.

[113] Daniel Spielman. *Spectral Graph Theory*. Combinatorial Scientific Computing, 2012.

[114] Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23:298–305, 1973.

[115] Miroslav Fiedler. Laplacian of graphs and algebraic connectivity. *Banach Center Publications*, 25:57–70, 1989.

[116] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[117] Johannes F. Lutzeyer et Andrew T. Walden. Comparing graph spectra of adjacency and laplacian matrices. *arXiv*, 2017.

[118] Kinkar Das. The Laplacian spectrum of a graph. *Computers & Mathematics with Applications*, 48:715–724, 2004.

[119] Alexander K. Kelmans. Comparison of graphs by their number of spanning trees. *Discrete Mathematics*, 16(3):241–261, 1976.

[120] Douglas J. Klein et Milan Randic. Resistance distance. *Journal of Mathematical Chemistry*, 12:81–95, 1993.

[121] Darko Babić, Douglas J Klein, István Lukovits, Sonja Nikolić et Nenad Trinajstić. Resistance-distance matrix : A computational algorithm and its application. *International Journal of Quantum Chemistry*, 90:166–176, 2002.

[122] Ivan Gutman et Bojan Mohar. The Quasi-Wiener and the Kirchhoff Indices Coincide. *Journal of Chemical Information and Computer Science*, 36(5):982–985, 1996.

[123] Istvan Lukovits, Sonja Nikolić et Nenad Trinajstić. Resistance distance in regular graphs. *International Journal of Quantum Chemistry*, 71(3):217–225, 1999.

[124] Yi-Zhe Song, Pablo Arbeláez, Peter M. Hall, Chuan Li et Anupriya Balikai. Finding Semantic Structures in Image Hierarchies Using Laplacian Graph Energy. In *European Conference on Computer Vision*, pages 694–707. Springer, 2010.

[125] Christoph Helmberg et Vilmar Trevisan. Spectral threshold dominance, Brouwer's conjecture and maximality of Laplacian energy. *Linear Algebra and its Applications*, 512:18–31, 2017.

[126] Jianpin Liu et Bolian Liu. A Laplacian-energy-like invariant of a graph. *MATCH Communications in Mathematical and in Computer Chemistry*, 59:355–372, 2008.

[127] Dragan Stevanovic. Laplacian-like energy of trees. *MATCH Communications in Mathematical and in Computer Chemistry*, 61(2):407, 2009.

[128] Ivan Gutman, Bo Zhou et Boris Furtula. The Laplacian-energy like invariant is an energy like invariant. *MATCH Communications in Mathematical and in Computer Chemistry*, 2010.

[129] Nair Abreu, Domingos M. Cardoso, Ivan Gutman, Enide A. Martins et al.. Bounds for the signless Laplacian energy. *Linear Algebra and its Applications*, 435(10):2365–2374, 2011.

[130] Michael Cavers, Shaun Fallat et Steve Kirkland. On the normalized Laplacian energy and general Randić index R_{-1} of graphs. *Linear Algebra and its Applications*, 433:172–190, 2010.

[131] Gopalapillai Indulal, Ivan Gutman et Ambat Vijayakumar. On distance energy of graphs. *MATCH Communications in Mathematical and in Computer Chemistry*, 60:461–472, 2008.

[132] Frank Harary, Clarence King, Abbe Mowshowitz et Ronald C. Read. Cospectral graphs and digraphs. *Bulletin of the London Mathematical Society*, 3:321–328, 1971.

[133] Charles R. Johnson et Morris Newman. A note on cospectral graphs. *Journal of Combinatorial Theory, Series B*, 28:96–103, 1980.

[134] Chris D. Godsil et Brendan D. McKay. Constructing cospectral graphs. *Aequationes Mathematicae*, 25:257–268, 1982.

[135] Willem H. Haemers et Edward Spence. Enumeration of cospectral graphs. *European Journal of Combinatorics*, 25(2):199–211, 2004.

[136] Allen J. Schwenk. Almost all trees are cospectral. *New directions in the theory of graphs*, pages 275–307, 1973.

[137] Edwin R. Van Dam et Willem H. Haemers. Developments on spectral characterizations of graphs. *Discrete Mathematics*, 309(3):576–586, 2009.

[138] Willem H. Haemers. Are almost all graphs determined by their spectrum? *Notices of the South African Mathematical Society*, 47:42–45, 2016.

[139] Andries E. Brouwer et Edward Spence. Cospectral Graphs on 12 Vertices. *Electronic Journal of Combinatorics*, 16, 2009.

[140] Steve Butler et Jason Grout. A Construction of Cospectral Graphs for the Normalized Laplacian. *Electronic Journal of Combinatorics*, 18, 2010.

[141] Ronan Hamon, Pierre Borgnat, Patrick Flandrin et Céline Robardet. Transformation from graphs to signals and back. *Vertex-Frequency Analysis of Graph Signals*, pages 111–139, 2019.

[142] Jie Zhang et Michael Small. Complex network from pseudoperiodic time series : Topology versus dynamics. *Physical Review Letters*, 96(23):238701, 2006.

[143] Fenglin Wang, Qingfang Meng, Weidong Zhou et Shanshan Chen. The Feature Extraction Method of EEG Signals Based on Degree Distribution of Complex Networks from Nonlinear Time Series. In *Intelligent Conference on Intelligence Computing (ICIC)*, pages 354–361. Springer, 2013.

[144] Norbert Marwan, Jonathan F. Donges, Yong Zou, Reik V. Donner et Jürgen Kurths. Complex network approach for recurrence analysis of time series. *Physics Letters A*, 373(46):4246–4254, 2009.

[145] Reik V. Donner, Yong Zou, Jonathan F. Donges, Norbert Marwan et Jürgen Kurths. Recurrence networks—A novel paradigm for nonlinear time series analysis. *New Journal of Physics*, 12(3):033025, 2010.

[146] Jean-Pierre Eckmann, Sylvie Oliffson Kamphorst, David Ruelle et al.. Recurrence plots of dynamical systems. *Europhysics Letters*, 4:973–977, 1987.

[147] Carlos Bergillos Varela. A study of visibility graphs for time series representations. B.S. thesis, Universitat Politècnica de Catalunya, 2020.

[148] Xin Lan, Hongming Mo, Shiyu Chen, Qi Liu et Yong Deng. Fast transformation from time series to visibility graphs. *Chaos : An Interdisciplinary Journal of Nonlinear Science*, 25(8):083105, 2015.

[149] Sheng-Li Zhu et Lu Gan. Specific emitter identification based on horizontal visibility graph. In *International Conference on Computer and Communications (ICCC)*, pages 1328–1332. IEEE, 2017.

[150] Chenyang Li, Lingfei Mo et Ruqiang Yan. Rolling bearing fault diagnosis based on horizontal visibility graph and graph neural networks. In *International Conference on Sensing, Measurement & Data Analytics in the era of Artificial Intelligence (ICSMD)*, pages 275–279. IEEE, 2020.

[151] Sayanjit Singha Roy et Soumya Chatterjee. Partial discharge detection framework employing spectral analysis of horizontal visibility graph. *IEEE Sensors Journal*, 21(4):4819–4826, 2020.

[152] Gregory Gutin, Toufik Mansour et Simone Severini. A characterization of horizontal visibility graphs and combinatorics on words. *Physica A : Statistical Mechanics and its Applications*, 390(12):2421–2428, 2011.

[153] Angel Nuñez, Lucas Lacasa, Eusebio Valero, Jose Patricio Gómez et Bartolo Luque. Detecting series periodicity with horizontal visibility graphs. *International Journal of Bifurcation and Chaos*, 22(07): 1250160, 2012.

[154] Guohun Zhu, Yan Li et Peng Paul Wen. Analysis and Classification of Sleep Stages Based on Difference Visibility Graphs From a Single-Channel EEG Signal. *IEEE Journal of Biomedical and Health Informatics*, 18(6):1813–1821, 2014.

[155] Gunjan Soni. Signed visibility graphs of time series and their application to brain networks. Mémoire de D.E.A., University of British Columbia, 2019.

[156] Jacopo Iacovacci et Lucas Lacasa. Visibility Graphs for Image Processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):974–987, 2019.

[157] J. K. Chung, P. L. Kannappan, Che Tat Ng et P. K. Sahoo. Measures of distance between probability distributions. *Journal of Mathematical Analysis and Applications*, 138:280–292, 1989.

[158] Mehran Ahmadlou, Hojjat Adeli et Anahita Adeli. New diagnostic EEG markers of the Alzheimer's disease using visibility graph. *Journal of Neural Transmission*, 117(9):1099–1109, 2010.

[159] Zeynab Mohammadpoory, Mahda Nasrolahzadeh et Javad Haddadnia. Epileptic seizure detection in EEGs signals based on the weighted visibility graph entropy. *Seizure*, 50:202–208, 2017.

[160] Supriya Supriya, Siuly Siuly, Hua Wang et Yanchun Zhang. Epilepsy Detection From EEG Using Complex Network Techniques : A Review. *IEEE Reviews in Biomedical Engineering*, 16:292–306, 2021.

[161] Xiao-Hui Ni, Zhi-Qiang Jiang et Wei-Xing Zhou. Degree distributions of the visibility graphs mapped from fractional Brownian motions and multifractal random walks. *Physics Letters A*, 373(42):3822–3826, 2009.

[162] Wen-Jie Xie et Wei-Xing Zhou. Horizontal visibility graphs transformed from fractional Brownian motions : Topological properties versus the Hurst index. *Physica A : Statistical Mechanics and its Applications*, 390(20):3592–3601, 2011.

[163] Corinna Cortes et Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[164] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc., 2022.

[165] James Mercer. Functions of positive and negative type, and their connection the theory of integral equations. *Philosophical Transactions of the Royal Society of London, Series A*, 209:415–446, 1909.

[166] Gebhard Kirchgässner, Jürgen Wolters et Uwe Hassler. *Introduction to modern time series analysis*. Springer, 2012.

[167] Denis V. Martynov, E. D. Hall, B. P. Abbott, R. Abbott, T. D. Abbott, C. Adams, R. X. Adhikari, R. A. Anderson, S. B. Anderson, K. Arai et al.. Sensitivity of the Advanced LIGO detectors at the beginning of gravitational wave astronomy. *Physical Review D*, 93(11):112004, 2016.

[168] Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large et Eamonn Keogh. The great time series classification bake off : a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 31:606–660, 2017.

[169] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar et Pierre-Alain Muller. Deep learning for time series classification : a review. *Data Mining and Knowledge Discovery*, 33(4):917–963, 2019.

[170] Thomas Gärtner, Peter Flach et Stefan Wrobel. On graph kernels : Hardness results and efficient alternatives. In *Learning Theory and Kernel Machines*, volume 2777, pages 129–143. Springer, 2003.

[171] Nino Shervashidze, S. V. N. Vishwanathan, Tobias Petri, Kurt Mehlhorn et Karsten Borgwardt. Efficient graphlet kernels for large graph comparison. *International Conference on Artificial Intelligence and Statistics*, pages 488–495, 2009.

[172] Alberto Sanfeliu et King-Sun Fu. A distance measure between attributed relational graphs for pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(3):353–362, 1983.

[173] Michel Neuhaus, Kaspar Riesen et Horst Bunke. Fast suboptimal algorithms for the computation of graph edit distance. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 163–172. Springer, 2006.

[174] Tristan Averty, Delphine Daré-Emzivat et Abdel-Ouahab Boudraa. Détection d'épilepsie dans les signaux EEG par graphe de visibilité et un noyau de SVM adapté. In *GRETSI*, pages 1–4, 2022.

[175] Vairavan Srinivasan, Chikkannan Eswaran et N. Sriraam. Artificial neural network based epileptic detection using time-domain and frequency-domain features. *Journal of Medical Systems*, 29(6):647–660, 2005.

[176] Ling Guo, Daniel Rivero, Julián Dorado, Cristian R Munteanu et Alejandro Pazos. Automatic feature extraction using genetic programming : An application to epileptic EEG classification. *Expert Systems with Applications*, 38(8):10425–10436, 2011.

[177] Muhammad U Abbasi, Anum Rashad, Anas Basalamah et Muhammad Tariq. Detection of epilepsy seizures in neo-natal EEG using LSTM architecture. *IEEE Access*, 7:179074–179085, 2019.

[178] Ralph G Andrzejak, Klaus Lehnertz, Florian Mormann, Christoph Rieke, Peter David et Christian E Elger. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity : Dependence on recording region and brain state. *Physical Review E*, 64(6):061907, 2001.

[179] Solomon Kullback et Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86, 1951.

[180] Dominik Maria Endres et Johannes E. Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860, 2003.

[181] Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[182] Mark E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[183] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte et Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics : Theory and Experiment*, 2008(10): P10008, 2008.

[184] James Lenz et Alan S. Edelstein. Magnetic sensors and their applications. *IEEE Sensors Journal*, 6(3):631–649, 2006.

[185] Timothée Roignant, Nicolas Le Josse, Abdel Boudraa, Jean-Jacques Szkolnik, Paul Penven et Hugues Henocq. Magnetic Anomaly Detection using Noise-Optimized Orthonormalized Functions on dual magnetometric sensor signals. In *European Signal Processing Conference (EUSIPCO)*, pages 2382–2386. IEEE, IEEE, 2024.

[186] E. S. Borovitskaya et Michael S. Shur. On theory of $1/f$ noise in semiconductors. *Solid-State Electronics*, 45(7):1067–1069, 2001.

[187] Stefan Rostek et Rainer Schöbel. A note on the use of fractional Brownian motion for financial modeling. *Economic Modelling*, 30:30–35, 2013.

[188] Marek Wolf. $1/f$ noise in the distribution of prime numbers. *Physica A : Statistical Mechanics and Its Applications*, 241:493–499, 1997.

[189] Benoit B. Mandelbrot et John W. Van Ness. Fractional Brownian motions, fractional noises and applications. *SIAM Review*, 10(4):422–437, 1968.

[190] Rachid Jennane, Rachid Harba et Gérard Jacquet. Méthodes d'analyse du mouvement brownien fractionnaire : théorie et résultats comparatifs. *Traitement Du Signal*, 18:419–436, 2001.

[191] Benoit B. Mandelbrot et James R. Wallis. Noah, Joseph, and operational hydrology. *Water Resources Research*, 4(5):909–918, 1968.

[192] Vivien Marmelat, Kjerstin Torre et Didier Delignières. Relative roughness : an index for testing the suitability of the monofractal model. *Frontiers in Physiology*, 3:208, 2012.

[193] Francesco Serinaldi. Use and misuse of some Hurst parameter estimators applied to stationary and non-stationary financial time series. *Physica A : Statistical Mechanics and its Applications*, 389(14):2770–2781, 2010.

[194] Robert B. Davies et David S. Harte. Tests for Hurst effect. *Biometrika*, 74:95–101, 1987.

[195] Patrice Abry et Fabrice Sellan. The wavelet-based synthesis for fractional Brownian motion proposed by F. Sellan and Y. Meyer : Remarks and fast implementation, 1996.

[196] Ton Dieker. Simulation of fractional Brownian motion. Mémoire de D.E.A., University of Twente, 2004.

[197] Vladas Pipiras. Wavelet-based simulation of fractional Brownian motion revisited. *Applied and Computational Harmonic Analysis*, 19:49–60, 2005.

[198] Jens Timmer et Michel Koenig. On generating power law noise. *Astronomy and Astrophysics*, 300:707, 1995.

[199] Hristo Zhivomirov. A method for colored noise generation. *Romanian Journal of Acoustics and Vibration*, 15:14–19, 2018.

[200] Mark E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323–351, 2005.

[201] Tiago A. Schieber, Laura Carpi, Alejandro C. Frery, Osvaldo A. Rosso, Panos M. Pardalos et Martín G. Ravetti. Information theory perspective on network robustness. *Physics Letters A*, 380(3):359–364, 2016.

[202] Tingyuan Nie, Zheng Guo, Kun Zhao et Zhe-Ming Lu. Using mapping entropy to identify node centrality in complex networks. *Physica A : Statistical Mechanics and its Applications*, 453:290–297, 2016.

[203] Ronald A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A*, 222:309–368, 1922.

[204] Pablo Sánchez-Moreno, R. J. Yánez et J. S. Dehesa. Discrete densities and Fisher information. In *International Conference on Difference Equations and Applications*, pages 291–298, 2009.

[205] Ingve Simonsen, Alex Hansen et Olav Magnar Nes. Determination of the Hurst exponent by use of wavelet transforms. *Physical Review E*, 58(3):2779, 1998.

[206] Chun-Feng Li. Rescaled-range and power spectrum analyses on well-logging data. *Geophysical Journal International*, 153:201–212, 2003.

[207] Chung-Kang Peng, Sergey V. Buldyrev, Shlomo Havlin, Michael Simons, H. Eugene Stanley et Ary L. Goldberger. Mosaic organization of DNA nucleotides. *Physical Review E*, 49(2):1685, 1994.

[208] Apostolos Serletis et Aryeh Adam Rosenberg. The Hurst exponent in energy futures prices. *Physica A : Statistical Mechanics and its Applications*, 380:325–332, 2007.

[209] Jose Alvarez-Ramirez et Rafael Escarela-Perez. Time-dependent correlations in electricity markets. *Energy Economics*, 32(2):269–277, 2010.

[210] Xin Yuan, Yanqing Hu, H Eugene Stanley et Shlomo Havlin. Eradicating catastrophic collapse in interdependent networks via reinforced nodes. *Proceedings of the National Academy of Sciences*, 114(13):3311–3315, 2017.

[211] Jianxi Gao, Baruch Barzel et Albert-László Barabási. Universal resilience patterns in complex networks. *Nature*, 530(7590):307–312, 2016.

[212] Santiago Segarra et Alejandro Ribeiro. Stability and continuity of centrality measures in weighted graphs. *IEEE Transactions on Signal Processing*, 64(3):543–555, 2015.

[213] Linton C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40:35–41, 1977.

[214] Michelle Girvan et Mark E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[215] Réka Albert, Hawoong Jeong et Albert Ĺaszló Barabási. Attack and error tolerance in complex networks. *Nature*, 406:378–382, 2000.

[216] Paolo Crucitti, Vito Latora, Massimo Marchiori et Andrea Rapisarda. Efficiency of scale-free networks : error and attack tolerance. *Physica A : Statistical Mechanics and its Applications*, 320:622–642, 2003.

[217] Paolo Crucitti, Vito Latora et Massimo Marchiori. Model for cascading failures in complex networks. *Physical Review E*, 69(4):1–4, 2004.

[218] Adilson E. Motter, Takashi Nishikawa et Ying-Cheng Lai. Range-based attack on links in scale-free networks : are long-range links responsible for the small-world phenomenon ? *Physical Review E*, 66(6): 065103, 2002.

[219] Kushal Kanwar, Harish Kumar et Sakshi Kaushal. A metric to compare vulnerability of the graphs of different sizes. *Electronic Notes in Discrete Mathematics*, 63:525–533, 2017.

[220] Kartik Anand, Ginestra Bianconi et Simone Severini. Shannon and von Neumann entropy of random networks with heterogeneous expected degree. *Physical Review E*, 83(3):036109, 2011.

[221] Jongkwang Kim et Thomas Wilhelm. What is a complex graph ? *Physica A : Statistical Mechanics and its Applications*, 387(11):2637–2652, 2008.

[222] Marcus Kaiser et Claus C. Hilgetag. Edge vulnerability in neural and metabolic networks. *Biological Cybernetics*, 90(5):311–317, 2004.

[223] Hadj-Ahmed Bay-Ahmed, Delphine Daré-Emzivat et Abdel-Ouahab Boudraa. Analyse de la vulnérabilité d'un réseau via la mesure de l'entropie de Von Neumann. In *GRETSI*, pages 1–4, 2019.

[224] Lin Han, Francisco Escolano, Edwin R. Hancock et Richard C. Wilson. Graph characterizations from von Neumann entropy. *Pattern Recognition Letters*, 33(15):1958–1967, 2012.

[225] Pin-Yu Chen, Lingfei Wu, Sijia Liu et Indika Rajapakse. Fast incremental von neumann graph entropy computation : Theory, algorithm, and applications. In *International Conference on Machine Learning*, pages 1091–1101. PMLR, 2019.

[226] Hayoung Choi, Jinglian He, Hang Hu et Yuanming Shi. Fast computation of von Neumann entropy for large-scale graphs via quadratic approximations. *Linear Algebra and its Applications*, 585:127–146, 2020.

[227] Xuecheng Liu, Luoyi Fu, Xinbing Wang et Chenghu Zhou. On the Similarity between von Neumann Graph Entropy and Structural Information : Interpretation, Computation, and Applications. *IEEE Transactions on Information Theory*, 68(4):2182–2202, 2022.

[228] Jian-Qiang Li, Xiu-Bo Chen et Yi-Xian Yang. Quantum state representation based on combinatorial Laplacian matrix of star-relevant graph. *Quantum Information Processing*, 14(12):4691–4713, 2015.

[229] Jianjia Wang, Richard C. Wilson et Edwin R. Hancock. Spin statistics, partition functions and network entropy. *Journal of Complex Networks*, 5(6):858–883, 2017.

[230] Ginestra Bianconi et Albert-László Barabási. Bose-Einstein condensation in complex networks. *Physical Review Letters*, 86(24):5632, 2001.

[231] John Von Neumann. *Mathematical foundations of quantum mechanics : New edition*. Princeton University Press, 2018.

[232] Michael A. Nielsen et Isaac L. Chuang. *Quantum computation and quantum information*. Cambridge University Press, 2010.

[233] Siddarth Srinivasan, Carlton Downey et Byron Boots. Learning and Inference in Hilbert Space with Quantum Graphical Models. In *Neural Information Processing Systems*, 2018.

[234] S. Perseguers, M. Lewenstein, A. Acin et J. I. Cirac. Quantum complex networks. *Preprint arXiv :0907.3283*, 2009.

[235] Dan Hu, Xueliang Li, Xiaogang Liu et Shenggui Zhang. The von Neumann entropy of random multipartite graphs. *Discrete Applied Mathematics*, 232:201–206, 2017.

[236] Jianjia Wang, Richard C. Wilson et Edwin R. Hancock. fMRI activation network analysis using bose-einstein entropy. In *International Workshop on Structural and Syntactic Pattern Recognition*. Springer, Springer, 2016.

[237] Michael Dairyko, Leslie Hogben, Jephian C.-H. Lin, Joshua Lockhart, David Roberson, Simone Severini et Michael Young. Note on von Neumann and Rényi entropies of a graph. *Linear Algebra and its Applications*, 521:240–253, 2017.

[238] Bing Wang, Huanwen Tang, Chonghui Guo et Zhilong Xiu. Entropy optimization of scale-free networks' robustness to random failures. *Physica A : Statistical Mechanics and its Applications*, 363(2):591–596, 2006.

[239] Wayne W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4):452–473, 1977.

[240] Daniel Baird et Robert E. Ulanowicz. The seasonal dynamics of the Chesapeake Bay ecosystem. *Ecological Monographs*, 59(4):329–364, 1989.

[241] Pablo M. Gleiser et Leon Danon. Community structure in jazz. *Advances in Complex Systems*, 6(04): 565–573, 2003.

[242] Carsten Matke, Wided Medjroubi, David Kleinhans et Sebastian Sager. Structure analysis of the German transmission network using the open source model SciGRID. In *Advances in Energy System Optimization : Proceedings of the first International Symposium on Energy System Optimization*, pages 177–188. Springer, 2017.

[243] Donald Ervin Knuth. *The Stanford GraphBase : a platform for combinatorial computing*. ACM, 1993.

[244] Nathanaël Perraudin, Johan Paratte, David Shuman, Lionel Martin, Vassilis Kalofolias, Pierre Vandergheynst et David K. Hammond. GSPBOX : A toolbox for signal processing on graphs. *Preprint arXiv :1408.5781*, 2014.

[245] Alexander Streltsov, Hermann Kampermann, Sabine Wölk, Manuel Gessner et Dagmar Bruß. Maximal coherence and the resource theory of purity. *New Journal of Physics*, 20(5):053058, 2018.

[246] Richard Von Mises et Hilda Pollaczek-Geiringer. Praktische Verfahren der Gleichungsauflösung. *ZAMM-Journal of Applied Mathematics and Mechanics*, 9:152–164, 1929.

[247] Roger A. Horn et Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[248] Tristan Averty, Delphine Daré-Emzivat, Abdel-Ouahab Boudraa et Yves Préaux. Approximation de l'entropie de von Neumann de graphes pour une analyse de vulnérabilité. In *GRETSI*, pages 1–4, 2022.

[249] Alexander Weinmann. *Uncertain models and robust control*. Springer, 1991.

[250] Gilbert W. Stewart et Ji-guang Sun. Matrix Perturbation Theory. *Academic Press*, 1990.

[251] Lucjan Piela. *Ideas of Quantum Chemistry*. Elsevier, 2006.

[252] Tommaso Coletta et Philippe Jacquod. Performance measures in electric power networks under line contingencies. *IEEE Transactions on Control of Network Systems*, 7:221–231, 2019.

[253] Shyam Boriah, Varun Chandola et Vipin Kumar. Similarity measures for categorical data : A comparative evaluation. In *International Conference on Data Mining*, pages 243–254. SIAM, Society for Industrial and Applied Mathematics (SIAM), 2008.

[254] Selim Aksoy et Robert M. Haralick. Feature normalization and likelihood-based similarity measures for image retrieval. *Pattern Recognition Letters*, 22(5):563–582, 2001.

[255] Simone Santini et Ramesh Jain. Similarity Measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):871–883, 1999.

[256] Bjørn Magnus Mathisen, Agnar Aamodt, Kerstin Bach et Helge Langseth. Learning similarity measures from data. *Progress in Artificial Intelligence*, 9(2):129–143, 2020.

[257] Maxat Kulmanov, Fatima Zohra Smaili, Xin Gao et Robert Hoehndorf. Semantic similarity and machine learning with ontologies. *Briefings in Bioinformatics*, 22(4):1–18, 2021.

[258] Karsten Roth, Biagio Brattoli et Bjorn Ommer. Mic : Mining interclass characteristics for improved metric learning. In *Proceedings IEEE/CVF International Conference on Computer Vision*, pages 8000–8009. IEEE, 2019.

[259] Miin-Shen Yang et Kuo-Lung Wu. A similarity-based robust clustering method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(4):434–448, 2004.

[260] Yung-Shen Lin, Jung-Yi Jiang et Shie-Jue Lee. A similarity measure for text classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 26(7):1575–1590, 2013.

[261] Michael M. Richter. Classification and learning of similarity measures. In *Information and Classification : Concepts, Methods and Applications*, pages 323–334. Springer, 1993.

[262] Mathieu Latourrette. Toward an explanatory similarity measure for nearest-neighbor classification. In *European Conference on Machine Learning*, pages 238–245. Springer, 2000.

[263] Alexander J. Stewart, Mohsen Mosleh, Marina Diakonova, Antonio A. Arechar, David G. Rand et Joshua B. Plotkin. Information gerrymandering and undemocratic decisions. *Nature*, 573(7772):117–121, 2019.

[264] William Cohen, Pradeep Ravikumar et Stephen Fienberg. A comparison of string metrics for matching names and records. In *KDD Workshop on Data Cleaning and Object Consolidation*, volume 3, pages 73–78, 2003.

[265] Pierre-Antoine Champin et Christine Solnon. Measuring the similarity of labeled graphs. In *Proceedings of the Conference on Case-Based Reasoning*, pages 80–95. Springer, 2003.

[266] Mladen Nikolić. Measuring similarity of graph nodes by neighbor matching. *Intelligent Data Analysis*, 16(6):865–878, 2012.

[267] Li Dengfeng et Cheng Chuntian. New similarity measures of intuitionistic fuzzy sets and application to pattern recognitions. *Pattern Recognition Letters*, 23:221–225, 2002.

[268] Guo-Dong Guo, Anil K. Jain, Wei-Ying Ma et Hong-Jiang Zhang. Learning similarity measure for natural image retrieval with relevance feedback. *IEEE Transactions on Neural Networks*, 13(4):811–820, 2002.

[269] Mehul P. Sampat, Zhou Wang, Shalini Gupta, Alan Conrad Bovik et Mia K. Markey. Complex wavelet structural similarity : A new image similarity index. *IEEE Transactions on Image Processing*, 18(11):2385–2401, 2009.

[270] Pierre Jean A. Colombo, Chloé Clavel et Pablo Piantanida. InfoLM : A New Metric to Evaluate Summarization & Data2Text Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10554–10562. AAAI Press, 2022.

[271] Maria Hatzigiorgaki et Athanassios N. Skodras. Compressed domain image retrieval : a comparative study of similarity metrics. In *Visual Communications and Image Processing*, volume 5150, pages 439–448. SPIE, SPIE, 2003.

[272] Xiuju Fu et Lipo Wang. Data dimensionality reduction with application to simplifying RBF network structure and improving classification performance. *IEEE Transactions on Systems, Man, and Cybernetics*, 33(3):399–409, 2003.

[273] Gustavo Camps-Valls et Lorenzo Bruzzone. Kernel-based methods for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 43(6):1351–1362, 2005.

[274] Vladimir Nikiforov et Oscar Rojo. A note on the positive semidefiniteness of $\mathbf{A}_\alpha(G)$. *Linear Algebra and its Applications*, 519:156–163, 2017.

[275] Haiyan Guo et Bo Zhou. On the α-spectral radius of graphs. *Applicable Analysis and Discrete Mathematics*, 14(2):431–458, 2020.

[276] Shariefuddin Pirzada, Bilal A. Rather, Hilal A. Ganie et Rezwan ul Shaban. On α-adjacency energy of graphs and Zagreb index. *AKCE International Journal of Graphs and Combinatorics*, 18:39–46, 2021.

[277] Yasuhiko Ikebe, Toshiyuki Inagaki et Sadaaki Miyamoto. The monotonicity theorem, Cauchy's interlace theorem, and the Courant-Fischer theorem. *The American Mathematical Monthly*, 94(4):352–354, 1987.

[278] Tinghuai Ma, Hongmei Wang, Lejun Zhang, Yuan Tian et Najla Al-Nabhan. Graph classification based on structural features of significant nodes and spatial convolutional neural networks. *Neurocomputing*, 423:639–650, 2021.

[279] Stijn Van Dongen et Anton J. Enright. Metric distances derived from cosine similarity and Pearson and Spearman correlations. *Preprint arXiv :1208.3145*, 2012.

[280] Asim Kumar Debnath, Rosa L. Lopez de Compadre, Gargi Debnath, Alan J. Shusterman et Corwin Hansch. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of Medicinal Chemistry*, 34(2): 786–797, 1991.

[281] Christoph Helma, Ross D. King, Stefan Kramer et Ashwin Srinivasan. The predictive toxicology challenge 2000–2001. *Bioinformatics*, 17:107–108, 2001.

[282] Karsten M. Borgwardt, Cheng Soon Ong, Stefan Schönauer, S. V. N. Vishwanathan, Alex J. Smola et Hans-Peter Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21:47–56, 2005.

[283] Pinar Yanardag et S. V. N. Vishwanathan. Deep Graph Kernels. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1365–1374, 2015.

[284] Eamonn Keogh, Li Wei, Xiaopeng Xi, Stefano Lonardi, Jin Shieh et Scott Sirowy. Intelligent icons : Integrating lite-weight data mining and visualization into GUI operating systems. In *Proceedings of the International Conference on Data Mining (ICDM)*, pages 912–916. IEEE, IEEE, 2006.

[285] Giannis Siglidis, Giannis Nikolentzos, Stratis Limnios, Christos Giatsidis, Konstantinos Skianis et Michalis Vazirgiannis. GraKel : A Graph Kernel Library in Python. *Journal of Machine Learning Research*, 21(54):1–5, 2020.

[286] James D. Johnston. Transform coding of audio signals using perceptual noise criteria. *IEEE Journal on Selected Areas in Communications*, 6(2):314–323, 1988.

[287] Eric Scheirer et Malcolm Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1331–1334. IEEE, 1997.

[288] Geoffroy Peeters. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. *CUIDADO First Project Report*, 54:1–25, 2004.

[289] Hemant Misra, Shajith Ikbal, Hervé Bourlard et Hynek Hermansky. Spectral entropy based feature for robust ASR. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages I–193. IEEE, 2004.

[290] Alexander Lerch. *An introduction to audio content analysis : Applications in signal processing and music informatics*. Wiley Online Library, 2012.

[291] Claude R. Dietrich et Garry N. Newsam. Fast and exact simulation of stationary gaussian processes through circulant embedding of the covariance matrix. *SIAM Journal on Scientific Computing*, 18(4): 1088–1107, 1997.

[292] Andrew T. A. Wood et Grace Chan. Simulation of stationary Gaussian processes in $[0, 1]^d$. *Journal of computational and graphical statistics*, 3(4):409–432, 1994.

[293] Matthew S. Crouse et Richard G. Baraniuk. Fast, exact synthesis of Gaussian and non-Gaussian long-range dependent processes. *IEEE Transactions on Information Theory*, pages 1–45, 1999.

COLLEGE SCIENCES
DOCTORAL DE LA MER
BRETAGNE ET DU LITTORAL

Université de Bretagne Occidentale

Titre : Matrices de représentation généralisées, mesures spectrales et distances statistiques pour l'analyse et la classification de graphes et de signaux

Mots-clés : Matrices de représentation, mesure de similarité, graphes de visibilité, entropies, vulnérabilité des réseaux

Résumé : Les graphes sont des objets mathématiques particulièrement adaptés pour représenter des réseaux complexes (infrastructures d'approvisionnement en énergie, routes terrestres, aériennes ou maritimes, réseau mondial de câbles sous-marins de communication, etc.), des données provenant de multiples capteurs (réseaux d'hydrophones ou réseaux de capteurs pour la surveillance environnementale) et enfin des séries temporelles vus comme des graphes à l'aide de l'algorithme dit de visibilité. Ces graphes sont souvent étudiés en utilisant les outils de l'algèbre linéaire tels que les matrices de représentation et leurs spectres associés. Différentes matrices existent chacune disposant d'avantages et d'inconvénients. Dans cette thèse, nous construisons une matrice de représentation généralisée bi-paramètre, notée $\mathbf{P}_{\alpha,k}$. L'introduction de $\mathbf{P}_{\alpha,k}$ permet d'unifier les théories et propriétés issues de celles traditionnellement utilisées comme la matrice d'adjacence et la matrice Laplacienne. Cette matrice nous permet également d'évaluer l'évolution des spectres entre ces matrices, essentielle pour une classification de graphes et de signaux vus comme des graphes, domaine d'application extrêmement porteur tant par la diversité des sources d'entrée que par les tâches à accomplir : détection d'anomalies magnétiques, détection d'épilepsie dans des signaux EEG, détermination de la cancérogénicité d'une protéine ou non. Pour assurer cette classification, il est nécessaire de disposer de mesures de similarité entre graphes qui comparent des éléments structurels et/ou spectraux. Ainsi, deux nouvelles mesures de similarité ont été développées : la première calcule la corrélation entre les spectres de la matrice de représentation généralisée précédemment évoquée et la deuxième compare, à l'aide de distances statistiques, les distributions de degrés des graphes de visibilité horizontale. En plus de la classification de séries temporelles, les graphes de visibilité constituent des outils précieux pour l'analyse de signaux tels que les processus stochastiques que ce sont les fBm ou les fGn. Ainsi, nous avons introduit une méthode d'estimation du coefficient de Hurst, paramètre caractérisant ces processus, basée sur l'extraction de deux grandeurs de la théorie de l'information des distributions de degrés des graphes de visibilité : une globale, soit l'entropie de Shannon, et la seconde locale, à savoir l'information de Fisher. Classer des graphes peut également signifier caractériser leurs éventuelles vulnérabilités. Nous introduisons pour cela une mesure de vulnérabilité d'une arête à partir de la variation relative de l'entropie de von Neumann basée sur la matrice de densité du graphe et traduisant le contenu informationnel du système physique sous-jacent si cette arête venait à être perturbée voire supprimée. Nous avons recours à deux approximations de l'entropie de von Neumann dans le but de réduire drastiquement le temps de calcul tout en conservant une erreur contenue : une basée sur un développement limité autour d'un point optimal, et la seconde sur une approximation des valeurs propres de la matrice de densité du graphe perturbé.

Title: Generalized representation matrices, spectral measures and statistical distances for graph and signal analysis and classification

Keywords: Representation matrices, similarity measures, visibility graphs, entropies, networks vulnerability

Abstract: Graphs are mathematical objects that are well-suited to represent complex networks (energy supply infrastructures, land, air or sea routes, network of underwater communication cables, etc.), data from multiple sensors (hydrophone networks or sensor networks for environmental monitoring) and finally time series seen as graphs using the so-called visibility algorithm. These graphs are often studied using linear algebra tools such as representation matrices and their associated spectra. Different matrices exist, each with its own advantages and disadvantages. In this thesis, we construct a generalized bi-parameter representation matrix, denoted $\mathbf{P}_{\alpha,k}$. The introduction of $\mathbf{P}_{\alpha,k}$ allows us to unify the theories and properties derived from those traditionally used, such as the adjacency matrix and the Laplacian one. This matrix also enables us to evaluate the evolution of spectra between these matrices, essential for the classification of graphs and signals seen as graphs, a field of application that is extremely promising in terms of both the diversity of input sources and the tasks to be accomplished: detection of magnetic anomalies, detection of epilepsy in EEG signals, determination of whether a protein is carcinogenic or not. To ensure this classification, it is necessary to have similarity measures between graphs that compare structural and/or spectral elements. Two new similarity measures have therefore been developed: the first calculates the correlation between the spectra of the generalized representation matrix mentioned above, and the second compares, using statistical distances, the degree distributions of horizontal visibility graphs. In addition to time series classification, visibility graphs are valuable tools for the analysis of signals such as stochastic processes like fBm or fGn. We have thus introduced a method for estimating the Hurst coefficient, a parameter characterizing these processes, based on the extraction of two information-theoretic quantities from the degree distributions of visibility graphs: one global, namely Shannon entropy, and the second local, namely Fisher information measure. Classifying graphs can also mean characterizing their potential vulnerabilities. To this end, we introduce a measure of edge vulnerability based on the relative variation of the von Neumann entropy (based on the density matrix of the graph and reflecting the information content of the underlying physical system) if this edge is disturbed or even removed. We use two approximations of the von Neumann entropy in order to drastically reduce computation time while keeping a reasonable error: one based on a Taylor series at an optimal point, and the second on an approximation of the eigenvalues of the density matrix of the perturbed graph.