$N°$ d'Ordre :
EDSPIC :

**Université Paris 13**

# THÈSE

réalisée par :

**Mouna RIFI**

pour obtenir le grade de :

DOCTEURE D'UNIVERSITÉ DÉLIVRÉ PAR L'UNIVERSITÉ SORBONNE PARIS CITÉ
ET PRÉPARÉ À L'UNIVERSITÉ PARIS 13
(Arrêté du 25 mai 2016)

École Doctorale Sciences, Technologie, Santé, Galilée (ED 146)

SPÉCIALITÉ : Informatique

# Modélisation et Analyse des Réseaux Complexes : Application à la sûreté nucléaire.

Soutenue le 03 mai 2019 devant le jury suivant :

| | |
|---|---|
| Directeur de thèse : | |
| Y. Bennani | Professeur des Universités, Université Paris 13 |
| Rapporteurs : | |
| R. Verde | Professeure des Universités, Università della Campania "Luigi Vanvitelli" |
| C. Bertelle | Professeur des Universités, Université du Havre Normandie |
| L. Tabourier | Maître de Conférences HDR, Sorbonne Université |
| Examinateurs : | |
| P. Leray | Professeur des Universités, Université de Nantes |
| M. Hibti | Chercheur expert Ph.D., EDF Lab Paris Saclay |
| R. Kanawati | Maître de Conférences, Université Paris 13 |
| A. Yahyaouy | Professeur des Universités, Université Sidi Mohamed Ben Abdellah |

# Résumé

Ce travail propose une modélisation adéquate en graphes pour les systèmes et séquences accidentelles de sûreté nucléaire. Ces systèmes et séquences proviennent des "Etudes Probabilistes de Sûreté" (EPS) qui consistent à analyser de façon exhaustive tous les scénarios accidentels envisageables, d'estimer leurs probabilités d'occurrence (en les regroupant par famille) et les conséquences associées.
Ensuite une analyse des réseaux complexes résultants est effectuée par des mesures de centralités.
Une première application consiste à la prédiction du Facteur d'Accroissement du Risque nucléaire en utilisant les algorithmes d'apprentissages supervisé : méthodes à base d'arbre de classification, régression logistique et méthodes ensemblistes, sur des données déséquilibrées.
Par ailleurs, un nouveau coefficient synthétique de centralité et une mesure de similarité sont conçus pour comparer les structures de réseaux, indépendamment de leurs caractéristiques topologiques, en se basant sur les interdépendances entre leurs vecteurs de centralités. Cette nouvelle approche utilise des techniques statistiques (échantillonnage, corrélation et homogénéité).
La pertinence et l'efficacité de cette nouvelle mesure de similarité sont validées sur le clustering de graphes théoriques classiques et la prédiction du type de graphes. Enfin, une application de cette approche est réalisée pour le clustering des réseaux complexes des systèmes de sûreté nucléaire.
**Mots clés :** Réseaux Complexes, Sûreté nucléaire, Centralités, Similarité de graphes, Facteur d'Accroissement du Risque, Corrélation de Kendall, Clustering de graphes.

# Abstract

This work aims to propose an adequate graph modeling approach for nuclear safety accident systems and sequences.
These systems and sequences come from "Probabilistic Safety Analysis" (PSA) which is an exhaustive analysis of all possible accident scenarios, to estimate their probabilities of occurrence (by grouping them by families) and the associated consequences.
Then, an analysis of the resulting networks is performed by network centrality measures.
A first application consists on predicting the nuclear Risk Increase Factor, which is a PSA importance factor, using supervised learning algorithms : classification tree methods, logistic regression and ensemble learning methods, on unbalanced data.
Furthermore, a new synthetic centrality coefficient and a similarity measure are developed to compare the networks structures and their topological characteristics, based on their centrality vectors interdependencies. This new approach uses statistical techniques (sampling, correlation and homogeneity).
The relevance and appreciation of this new measure of similarity are validated on the clustering of most popular theoretical graphs and on the prediction of the type of these graphs.
Finally, an application of this approach has been conducted for the clustering of nuclear safety systems networks.

**Keywords :** Complex Networks, Nuclear Safety, Centalities, Graph similarity, Risk Increase Factor, Kendall correlation, Graph clustering.

# Dédicace

*A mon défunt grand-père, à mes parents.*
*A toutes les femmes courageuses.*

# Table des matières

# Table des figures

# Liste des tableaux

# Liste des publications

## Revues internationales

- RIFI M., HIBTI M., BENNANI Y. et KANAWATI R. : A new similarity measure for directed or undirected graphs with an application to a real case of nuclear safety systems networks.[En préparation].

## Conférences internationales

- RIFI M., HIBTI M., VERMUSE S., BENNANI Y. et KANAWATI R. : A complex network analysis for balanced design verification, International Topical Meeting on Probabilistic Safety Assessment and Analysis (PSA ), 2019 [Soumis].
- RIFI M., HIBTI M., KANAWATI R. : A Complex Network Analysis Approach for Risk Increase Factor Prediction in Nuclear Power Plants, Conference on Complexity, Future Information Systems and Risk (Complexis), Madeira 2018.
  Nominé pour le prix du meilleur article.
- RIFI M., HIBTI M., KANAWATI R. : Exploring network metrics for accident scenarios : a Case of study of the Uncontrolled Level Drop, International Topical Meeting on Probabilistic Safety Assessment and Analysis (PSA ), Pittsburgh 2017.

## Conférences nationales

- RIFI M., HIBTI M., KANAWATI R. : Réseaux complexes pour la classification du Facteur d'Accroissement du Risque dans une centrale nucléaire, Conférence Internationale Francophone sur la Science des Données (CIFSD), Tanger 2018.
- RIFI M., HIBTI M.,KANAWATI R. : Applying complex networks analysis to risk estimation in Nuclear Power Plants, Modèles & Analyse des Réseaux : Approches Mathématiques & Informatiques (MARAMI), La Rochelle 2017.

# Chapitre 1

# Introduction

## Contexte et Problématique

Les réseaux sont omniprésents dans l'étude de nombreux systèmes complexes, tels que les réseaux sociaux, réseaux d'interaction protéine-protéine, l'Internet physique, le World Wide Web, entre autres [30]. En fait, un système complexe peut être modélisé par un réseau complexe, de telle sorte que le réseau complexe représente un modèle abstrait de la structure et des interactions des éléments dans le système complexe [70].
Par exemple, un réseau social peut être considéré comme un réseau (graphe) où les individus sont représentés par des sommets ; et les relations d'amitié entre individus sont des liens [86]. De même, le Web peut être modélisé sous forme de réseau, où les pages Web sont représentées sous forme de sommets reliés par un lien quand une page contient un hyperlien vers une autre [5, 61].
Plusieurs travaux définissent un réseau complexe comme étant un graphe de faible densité, de coefficient de clustering global plus important que cette dernière, et d'une distribution des degrés des sommets qui suit une loi de puissance. Une classe revient souvent dans les études, il s'agit de la classe des réseaux petit monde ("Small-World") qui tient son nom de la célèbre expérience du sociologue Milgram [83]. Ces réseaux ont une propriété en plus à savoir qu'ils ont un petit diamètre.
Il est souvent possible de prédire la fonctionnalité ou de comprendre le comportement d'un système complexe si nous arrivons à vérifier certaines « bonnes propriétés » en analysant le réseau sous-jacent. Par exemple, si nous détectons des groupes de sommets de même caractéristiques topologiques du réseau, nous pouvons obtenir des informations sur les rôles particuliers joués par chaque sommet (par exemple, hubs, outliers) ou comment des clusters entiers décrivent ou affectent le comportement général du système complexe [70].

Dans ces travaux, nous nous intéressons aux données d'études de la sûreté nucléaire. En effet, une centrale nucléaire est un système complexe dont la sûreté est un enjeux majeur

durant tout son cycle de vie. Effectivement, de la conception au démantèlement, en passant par la phase de demandes d'autorisations et celle de l'exploitation, la sûreté est une priorité au cœur de toutes les décisions.
Traditionnellement, les études de sûreté sont basées sur des approches déterministes. Cependant, les approches probabilistes commencent à gagner du terrain, et depuis quelques années, elles sont considérées comme outil essentiel capable de fournir un aperçu qualitatif et global en intégrant non seulement la gravité d'un évènement mais aussi sa probabilité d'occurrence obtenue grâce aux calculs probabilistes. L'approche utilisée dans le domaine de la sûreté nucléaire est appelée "Études Probabilistes de Sûreté" (EPS).
Malgré le succès connu par cette approche, depuis son apparition en 1975 dans le rapport nommé WASH-1400 [72], comme outil d'analyse, de décomposition, et de quantification du risque d'occurrence d'un évènement indésirable dans une centrale nucléaire, elle commence à atteindre certaines limites principalement calculatoires mais aussi par rapport au savoir-faire nécessaire pour réaliser rigoureusement de telles études.
Dans ces travaux, nous proposons d'utiliser les réseaux complexes pour explorer ces études de sûreté. L'objectif est en premier lieu de proposer une méthode de modélisation des systèmes de sauvegarde et des séquences accidentelles mises en place à la suite de l'occurrence d'un évènement initiateur et menant à des conséquences inacceptables sur l'installation.
Le deuxième objectif consiste à analyser les réseaux obtenus notamment en utilisant les différentes centralités réseaux introduites dans l'état de l'art des réseaux complexes. De telles centralités sont couramment exploitées pour caractériser les sommets dans l'analyse des réseaux sociaux. Ici, le but est de révéler certains patterns et identifier des composants qui ont été négligés par le calcul EPS comme conséquence des troncatures, bien qu'ils soient potentiellement important de point de vue sûreté.
Comme troisième objectif, on s'intéressera à prédire par les centralités le Facteur d'Accroissement du Risque (RIF) qui est la mesure d'importance la plus utilisée pour caractériser l'impact de la défaillance d'un composant sur la sûreté de l'installation nucléaire étudiée.
Le quatrième objectif est l'étude stratistique des centralités dans les réseaux : (i) l'homogénéité des distributions des valeurs d'une centralité dans deux ou plusieurs réseaux et (ii) la dépendance entre les centralités (deux à deux) dans un même réseau.
Le cinquième et dernier objectif de cette thèse est d'établir de nouvelles techniques et notions de similarité entre réseaux pour pouvoir les comparer structurellement : (i) proposition d'un coefficient de centralité synthétique, (ii) mesure de similarité entre réseaux (iii) validation de cette mesure pour le clustering graphes théoriques et application aux réseaux réels de systèmes de sûreté nucléaire.

## Organisation de la thèse

Ce manuscrit se compose de deux parties dont la 1ère compte 4 chapitres :

— Le chapitre 2 présente les principales notions des Etudes Probabilistes de Sûreté et quelques motivations de l'utilisation de l'approche réseau pour enrichir ces dernières.

— Le chapitre 3 est consacré aux définitions des différentes de la théorie des graphes ainsi que celles des mesures de centralités employées dans le cadre de ce travail.

— Le chapitre 4 décrit la méthode de construction de graphes de systèmes et de séquences accidentelles. Deux cas d'applications sont présentés : le premier est relatif aux séquences accidentelles de l'initiateur EPR "Baisse incontrôlée du niveau primaire" dans les états d'arrêts ; le second est relatif à la construction des réseaux correspondants aux systèmes de sauvegarde de l'EPR2.

— Le chapitre 5 présente une application de l' analyse des réseaux complexes pour la prédiction par les mesures de centralités du facteur d'importance Risk Increase Factor, et cela en utilisant les algorithmes de classification supervisée.

La seconde partie se compose de deux chapitres :

— Le chapitre 6 présente l'étude statistique des mesures de centralités dans un ou deux réseaux différents par les notions d'homogénéïté de distribution et la dépendance statistique,

— Le chapitre 7 introduit dans un premier lieu un coefficient synthétique de centralité pour capturer la structure du réseau grâce aux interdépendances entre ses vecteurs de centralité ; puis dans un second, une mesure de similarité entre réseaux est élaborée afin de les comparer structurellement.

Nous concluons ce manuscrit en exposant les points forts de nos contributions et les perspectives de ce travail.

## Première partie

# Réseaux complexes pour systèmes et séquences de sûreté, application à la classification du Risk Increase Factor

Plusieurs travaux de recherche ont été développés au cours des dernières années produisant une variété de techniques et de modèles pour aider à comprendre ou à prévoir le comportement de systèmes complexes tels que Internet, les réseaux sociaux ou les réseaux biologiques. Ils ont examiné les développements dans ce domaine, y compris des concepts tels que l'effet du petit monde, les distributions de degrés, les modèles de graphes aléatoires, les modèles de croissance du réseau et l'attachement préférentiel, ainsi que des processus dynamiques sur les réseaux.

Quelques récentes méthodes d'analyse de la sûreté des systèmes de processus industriels se basent aussi sur les réseaux complexes. Elles modélisent ces systèmes comme de réseaux complexes et examinent leurs propriétés topologiques. Ensuite, elles étudient les défaillances en cascade qui sont des processus dynamiques, et construisent des modèles simples intégrant les poids sur les sommets et l'efficacité des réseaux. Par exemple, Jiang et al. [55] constatent que le réseau d'un système industriel est en général un réseau non homogène et présente un coefficient de clustering élevé et une longueur de chemin moyenne réduite, ce qui diffère des réseaux aléatoires et des réseaux classiques. Ils remarquent également qu'une défaillance initiale d'un périphérique du système peut générer une défaillance en cascade, si la charge initiale (poids) des périphériques est extrêmement élevée. La charge d'un périphérique du réseau étant le flux maximal que peut supporter ce périphérique, par ex. formes d'énergie, d'informations, de matériaux et de données. Cette méthode peut aider à identifier les composants critiques du système et peut également être utilisée pour la prévention des pannes et la conception d'un système de contrôle de la sécurité.

Par ailleurs, les systèmes électriques et les technologies de l'information et de la communication (TIC) peuvent être considérés comme des réseaux ou infrastructures critiques. Ces infrastructures sont fortement interconnectées et interdépendantes. Une panne dans une infrastructure peut atteindre les autres infrastructures en cascades. Les menaces sur les TIC, telles que le manque de disponibilité, de confidentialité, d'authenticité et de traçabilité, peuvent affecter le comportement du système d'alimentation. De même, les pannes de courant des systèmes électriques et les défaillances des systèmes auxiliaires peuvent affecter les TIC qui contrôlent et surveillent le système d'alimentation au moyen de capteurs et de moyens de communication. Afin d'élaborer un modèle unique et intégré, Sanchez Torres [82] propose un modèle bidimensionnel, qui permet de comprendre et d'identifier les vulnérabilités inhérentes aux infrastructures couplées. Il décrit certaines des principales interactions et interdépendances entre ces infrastructures critiques. En particulier, il étudie les interdépendances entre les réseaux de distribution d'énergie, les infrastructures d'information et de communication et les différents niveaux hiérarchiques de contrôle et de supervision. Une analyse des principales propriétés de ces réseaux complexes lui permet d'identifier les éléments les plus critiques ou les plus centraux en ce qui concerne les analyses basées sur la topologie.

Dans le domaine du nucléaire, Ruiz-Martin et al. [75] utilisent les réseaux complexes pour

analyser la résilience organisationnelle (Business resilence) qui est la capacité d'une organisation à faire face à une épreuve (ou un incident) et à la surmonter dans des conditions qui peuvent être défavorables. Cette épreuve peut être une séquence d'incidents en cascades. Ils se basent sur le Plan d'Urgence Nucléaire externe issu du journal officiel espagnol 1985 et 2009 et étudient la résilience de son organisation contre la défaillance de différents canaux de communications. L'organisation est modélisée en réseau complexe dont les sommets représentent les membres du personnel et les liens les relations qui les lient. L'objectif est d'identifier les sommets ou liens cruciaux pour la connectivité du réseau et dont la suppression cause des dysfonctionnements.
Concernant les Etudes Probabilistes de Sûreté, des algorithmes basés sur la théorie des graphes sont utilisés dans [3] et [45] pour rechercher des coupes minimales conduisant à un événement indésirable à partir d'un initiateur. Les résultats sont représentés sous forme de réseaux. Un calcul direct des fréquences de séquences des réseaux construits sur des réseaux de systèmes et une représentation standard des scénarios sous forme de diagrammes de séquence d'événements permet d'obtenir les fréquences de séquences indésirables.
À partir d'un diagramme de séquence d'événements, représentant les scénarios suivant l'occurrence d'un initiateur, Hibti et al. [46] construisent un réseau qui représente les différentes relations entre les sommets du réseau et les différentes conséquences du diagramme des séquences d'événements. Une séquence indésirable est alors représentée par un réseau de flux "s-t" (source et destination) qui est analysé à l'aide de différentes métriques. Ils mettent en évidence l'existence d'une forte relation entre une intermédiarité (betweeness) élevée et la fréquence d'occurrence dans la liste des ensembles de coupes de la séquence correspondante. Ainsi, ils proposent une expérimentation de métriques de la théorie des graphes pour analyser des modèles probabilistes de sûreté des centrales nucléaires.
Dans cette partie, nous nous intéressons aux réseaux complexes pour modéliser les systèmes de sûreté nucléaire et séquences accidentelles mises en place pour mitiguer les effets d'un évènements initiateur et analysons leurs applications nottamment à la classification du Risk Increase Factor (Facteur d'Accroissement du Risque). Notons que le Facteur d'Accroissement du Risque tel que défini dans [84] évalue des liens de corrélation qui ne sont pas nécessairement causaux. Les méthodes statistiques sont fréquemment utilisées pour évaluer la force de ces associations et pour fournir des preuves de causalité.

Cette partie est structurée comme suit :
Le deuxième chapitre comprend une présentation des principales notions des Études Probabilistes de Sûreté, leurs objectifs, étapes de réalisation, données d'entrées et mesures d'importance. Ensuite, il comporte quelques motivations de l'utilisation des réseaux complexes comme outil de modélisation et d'analyse pour ces études.
Le troisième chapitre regroupe des définitions sur les principales mesures de la théorie des graphes celles qui caractérisent un graphe entier et celles qui mesurent l'importance de ses

sommets ou liens nommées centralités.
Le quatrième chapitre propose une méthode pour construire des réseaux complexes à partir de systèmes de sûreté et séquences accidentelles. Cette méthode est appliquée pour une étude réelle qui concerne l'évènement initiateur de "la baisse incontrôlée du niveau primaire" pour une installation de type EPR (European Pressurized Reactor), le réseau dirigé ainsi obtenu est enfin analysé par les mesures les plus célèbres de centralités.

Le cinquième chapitre de cette partie propose une application des réseaux complexes aux études de sûreté qui est la classification du Facteur d'accroissement du Risque nucléaire, métrique très utilisée dans les Études Probabilistes de sûreté, et ce en utilisant les centralités les plus célèbres dans l'analyse des réseaux complexes.

# Chapitre 2

# Etudes Probabilistes de Sûreté

Dans ce chapitre nous présentons les plus importants concepts des Études Probabilistes de Sûreté (EPS) nécessaires pour la compréhension de ce document. Introduisons dans un premier temps les principaux objectifs des EPS, leurs niveaux, les étapes de réalisation, les données d'entrées nécessaires pour la réalisation d'une EPS, les résultats produits par ces études et enfin les mesures d'importance les plus utilisées. Nous décrivons ensuite quelques raisons qui justifient l'intérêt d'utiliser les réseaux et la théorie des graphes pour enrichir les EPS.

## 2.1 Les Études Probabilistes de Sûreté

Comme mentionné précédemment, les EPS sont apparues dans le rapport de Rasmussen "Reactor Safety Study" célèbre sous le nom de WASH-1400 [72]. En France, à EDF, l'utilisation de cette approche n'a commencé qu'en début des années 90. Elle permet de procéder à une investigation systématique et un découpage de l'évènement indésirable (aussi appelé évènement redouté) en plusieurs évènements initiateurs (pouvant causer cet évènement indésirable). Ensuite, les EPS établissent pour chacun de ces initiateurs les séquences accidentelles qui passent par l'échec ou le succès des missions de sauvegarde mises en place pour mitiguer le risque et ramener l'installation à un état sûr appelé conséquence acceptable (CA). En étudiant l'installation nucléaire comme étant un système intégré, incluant à la fois les aspects techniques et humains, les EPS contribuent à la gestion du risque, identifient les séquences accidentelles, déterminent à quelles fréquences ces dernières peuvent avoir lieu et étudient pour chaque scénario l'ensemble des conséquences potentielles.

Ainsi, les EPS sont considérées comme un outil d'aide à la décision à la fois pour les actions de maintenance, de modification de l'installation, mais aussi pour la démonstration de sûreté vis à vis des Autorités de Sûreté.

Les Études Probabilistes de Sûreté sont de trois niveaux, nous nous intéressons au niveau le plus bas qui a pour objectif d'estimer le risque de fusion du cœur d'une installation nucléaire. L'EPS de niveau 2 utilise les résultats produits par celle de niveau 1 et s'intéresse

à évaluer la nature, l'importance et les fréquences des rejets potentiels précoces de matières radioactives dans l'environnement. Quant au niveau 3, il permet de quantifier les fréquences calculées en termes de contamination voire de maladies telles que le cancer, jusqu'aux décès. Pour la suite, nous utiliserons le terme EPS pour désigner une EPS de niveau 1.

### 2.1.1 Étapes de construction d'une EPS

Une EPS commence par une identification de la liste des évènements initiateurs qui peuvent causer l'évènement redouté qui est la fusion du cœur.

Un initiateur peut être un aléas interne (explosion, feu,..) ou externe (inondation, tornade, ...) affectant le site de l'installation.

La deuxième étape d'une EPS consiste à effectuer une Analyse Qualitative des Séquences (AQS), qui permet de modéliser toutes les séquences de scénarios possibles en fonction de l'évolution de la réponse de l'installation aux effets de l'initiateur.

Chaque scénario commence par l'initiateur étudié et parcourt les succès et échecs des systèmes de sûreté et actions humaines mis en place pour mitiger les effets de l'initiateur. Nous décrirons plus en détail cette étape plus loin dans ce document.

Ces scénarios accidentels sont ensuite traduits en arbres d'évènements (un arbre pour chaque initiateur, avec potentiellement des renvois vers d'autres initiateurs). Les échecs de missions systèmes sont étudiés par l'analyse d'arbres de défaillance, tandis que ceux des actions humaines sont des valeurs points injectées dans le modèle EPS mais qui proviennent des résultats d'analyses de fiabilité humaine. Concernant les séquences qui mènent à des conséquences inacceptables, leurs fréquences sont calculées et les listes de coupes sont établies à l'aide du logiciel RiskSpectrum.
Notons que l'analyse par arbre de défaillance [33] permet de modéliser les causes de l'échec d'un système en détail. Ce type d'analyse est utilisé pour identifier les causes potentielles d'échec d'un système grâce à une cascade de portes logiques (OU, ET,..) de défaillances de composants.

Elle permet aussi de calculer la probabilité d'échec de chaque mission car un arbre de défaillance peut être vu comme une expression booléenne reliant l'évènement indésirable aux évènements de base. Il a été démontré que toute expression booléenne admet une représentation unique (forme canonique) sous forme d'union de coupes minimales. Les coupes minimales étant la conjonction de plusieurs évènements de base. Ainsi, la probabilité de chaque coupe minimale est calculée, ensuite, la probabilité de l'évènement indésirable est estimée à partir des probabilités obtenues par l'application de la formule de Poincaré selon la formule suivante :

$$P(EI) = \sum_{i=1}^{n} P(C_i) - \sum_{i<j} P(C_i.C_j) + \sum_{i<j<k} P(C_i.C_j.C_k) - ... + (-1)^n P(C_1.C_2...C_n)$$

tel que EI est l'évènement indésirable, et les $C_i$ sont les coupes minimales.

En pratique, les probabilités de base sont souvent faibles, on se contente donc du premier terme de la formule de Poincaré :

$$P(EI) = \sum_{i=1}^{n} P(C_i)$$

### 2.1.2 Données d'entrées d'une Étude Probabiliste de Sûreté

La réalisation d'une EPS nécessite plusieurs types de données d'entrée telles que :

— Les probabilités des (différentes) défaillances des composants l'installation étudiée, ces paramètres sont soit obtenus par l'analyse du retour d'expérience, ou ont des valeurs forfaitaires établies par des méthodes statistiques.

— Les fréquences des évènements initiateurs, qui correspondent aux fréquences d'occurrence de la défaillance ou de l'aléa qui peut conduite à l'évènement indésirable.

— Les défaillances de Causes Communes, qui sont las défaillances multiples de composants appartenant à un même système par une même source de défaillance.

— Probabilités d'erreurs humaines, elles représentent l'échec de conduite d'une action manuelle par un opérateur. Ces probabilités sont estimées par les méthodes d'analyse de fiabilité humaine.

### 2.1.3 Mesures d'importances des Études Probabilistes de Sûreté

Les Études Probabilistes de Sûreté disposent de métriques dites mesures d'importance, qui permettent comme leur nom l'indique d'identifier le rôle d'un composant, et être capable de comparer les composants par exemple par rapport à leur contribution au risque d'une installation. Si la mesure de Fussel-Vesely (FV) [40] est souvent utilisée quand l'objectif est de quantifier l'importance d'un composant par rapport au risque global étudié, Le Risk Increase Factor (RIF), ou coefficient d'accroissement de risque est plutôt utilisé pour mesurer l'importance d'un composant de point de vue sûreté [84].

Dans le tableau 2.1, nous présentons les définitions de ces mesures d'importance pour un composant donné $x_i$.

Nous considérons que $R(x_i = 0)$ est l'estimation du risque en supposant que le composant $x_i$ est parfaitement fiable. Tandis que dans $R(x_i = 1)$ $x_i$ est supposé défaillant ou absent. $R_{base}$ correspond au risque de base sans aucune condition sur l'état du composant.

| Measure | Abbreviation | Pinciple |
|---|---|---|
| Risk Decrease | $RD(x_i)$ | $R_{base} - R(x_i = 0)$ |
| Fussell-Vesely | $FV(x_i)$ | $\frac{R_{base} - R(x_i=0)}{R_{base}}$ |
| Risk Decrease Factor | $RDF(x_i)$ | $\frac{R_{base}}{R(x_i=0)}$ |
| Criticality Importance | $CR(x_i)$ | $\frac{R(x_i=1) - R(x_i=0)}{R_{base}}$ |
| Risk Increase | $RI(x_i)$ | $R(x_i = 1) - R_{base}$ |
| Risk Increase Factor | $RIF(x_i)$ | $\frac{R(x_i=1)}{R_{base}}$ |

TABLE 2.1 – Mesures d'importances des EPS

## 2.2 Motivations

Bien qu'étant un puissant outil de calcul et d'analyse de risque les EPS présentent plusieurs limitations. Nous citerons quelques unes : Les calculs EPS sont basés sur des formules booléennes, à réduire en forme normale disjonctive. Ce type de calculs sont connus d'être NP-durs [9] [39]. Pour palier à ces limites des troncatures sont réalisées pour pouvoir avoir des résultats. Une troncature peut établir un seuil de probabilité ou bien un seuil sur la longueur de coupe. Ensuite, elle élimine les coupes qui sont en dessous du seuil choisi. Ces troncatures peuvent négliger certaines coupes, ou composants car supposés très-fiables, mais peuvent s'avérer importants de point de vue sûreté. Certaines coupes peuvent être graves, mais éliminées car considérées très rares, peuvent se produire soudainement à la suite de l'occurrence d'un aléa.

Les positions des différents composants dans une séquence accidentelle peuvent révéler certaines fragilités et donc compléter le champ d'analyse.

De plus, même avec les approximations effectuées, les moyens de calculs d'EPS commencent à atteindre leurs limites, une simple modification peut transformer le temps de calcul de polynomial à exponentiel.

A titre d'information, un calcul d'EPS peut durer de quelques minutes à quelques jours, ce qui n'est pas convenable dans un contexte industriel.

D'autre part, la réalisation d'une analyse EPS demande un très grand savoir-faire et des connaissances approfondies des différents systèmes, et fait appel à différents types d' analyses (comme précisé dans le paragraphe consacré aux données d'entrées des Études Probabilistes de Sûreté (voir paragraphe 2.2.2), provenant de bases de données différentes, tandis que le réseau d'une séquence accidentelle encapsule toutes sorte de données souhaitées et facile d'accès.

Chaque modification ou mise à jour de l'étude nécessite de modifier toutes les études où apparaissait (ou doit apparaître) un composant et de régénérer les arbres de défaillances et donc de tout recalculer, ce qui est fastidieux, comparé à l'effort d'ajout ou de suppression

d'un sommet, arc ou modification de la valeur d'un attribut dans un réseau.
L'utilisation des réseaux comme outil de modélisation, d'analyse et de visualisation fournit un éclairage supplémentaire et complémentaires aux études de sûreté réalisées par les EPS. En effet on peut voir un réseau comme un modèle compacte encapsulant toutes les spécificités d'un système ou d'une séquence accidentelles grâce aux attributs des sommets et des liens.

## 2.3 Conclusion

Ce chapitre décrit quelques concepts liés aux Études Probabilistes de Sûreté utilisés dans cette thèse. Quelques motivations de l'intérêt d'utiliser la modélisation et l'analyse de réseaux complexes pour enrichir les EPS y sont présentées.

# Chapitre 3

# Graphes et centralités

Ce chapitre présente les principales notions de la théorie des graphes et quelques familles de graphes de la littérature. Ces graphes seront utilisés dans les chapitres 7 et 8 pour valider les approches présentées. Ensuite, il s'intéresse aux principales mesures de centralités décrites dans l'état de l'art pour caractériser l'importance d'un sommet ou d'un lien dans un réseau (graphe).
Le domaine qui a défini très tôt des mesures d'importances c'est celui de la sociologie pour caractériser l'importance d'un individu dans un réseau social auquel il appartient, qui peut être sa famille, sa ville, son entreprise, son club de Karaté ou autres.
Plusieurs mesures de centralité sont utilisées dans la pratique. Le choix de la mesure dépendra du type d'importance que l'on souhaite quantifier. Bien entendu, il est possible d'en créer d'autres selon le besoin ou en utiliser plusieurs simultanément pour avoir plusieurs "angle de vue" et donc plusieurs classements d'importance.
On rappelle dans ce chapitre les définitions des centralités les plus utilisées dans l'analyse des réseaux sociaux et d'autres considérées comme sorte de méta-centralités usuellement employées pour la recherche d'information dans le Web.

## 3.1 Notions de la théorie des graphes

Dans cette section nous rappelons les principales notions de la théorie des graphes et plus précisément celles usuellement utilisées pour caractériser un graphe.
Soit $G$ un graphe, $G = (V, E \subseteq V \times V)$ où $V$ est l'ensemble de sommets et $E$ désigne l'ensemble de liens (orientés ou non).

### 3.1.1 Chemin et distance

Un chemin entre deux sommets d'un réseau est la suite des arêtes consécutifs reliants ces deux sommets. La longueur d'un chemin correspond au nombre de liens reliant ces deux sommets pour les réseaux binaires et la somme des poids des liens reliant ces deux sommets pour les réseaux pondérés. Pour les réseaux dirigés un chemin prends compte de l'orientation

des liens. Le plus court chemin entre deux sommets correspond au chemin dont la longueur est minimale.

### 3.1.2 Diamètre

Le diamètre d'un graphe est la plus longue des distances (minimale) entre deux sommets du graphe.

### 3.1.3 Coefficient de Clustering

Le coefficient de clustering (global) d'un graphe (aussi appelé coefficient d'agglomération, de connexion, de regroupement, d'agrégation ou de transitivité) est une mesure du regroupement des sommets dans un réseau. Il mesure à quel point le voisinage d'un sommet est connecté. C'est une propriété des graphes très utilisée dans l'analyse des réseaux sociaux. Une transitivité parfaite implique que, si $x$ est connecté (par une arête) à $y$, et que $y$ est connecté à $z$, alors $x$ est également connecté à $z$. C'est un cas rare dans les réseaux réels, car cela implique que chaque composante est une clique, c'est-à-dire que chaque paire de nœuds accessibles dans le graphique serait connectée par une arête. Cependant, la transitivité partielle est utile. Dans de nombreux réseaux, en particulier sociaux, le fait que $x$ connaisse $y$ et que $y$ connaisse $z$ ne garantit pas que $x$ connaisse aussi $z$, mais le rend beaucoup plus probable. L'ami de mon ami n'est pas nécessairement mon ami, mais il est beaucoup plus susceptible de l'être qu'un membre de la population choisi au hasard. Nous pouvons quantifier la transitivité dans un réseau non dirigé comme suit : Un cycle de longueur trois est une suite de sommets $x$, $y$, $z$, $x$ tels que $(x, y)$, $(y, z)$ et $(z, x)$ sont des arêtes du graphe. Notons qu'il s'agit de le plus petit cycle possible dans un réseau non dirigé. Un chemin de longueur deux est une suite de sommets $x$, $y$, $z$ tels que $(x, y)$ et $(y, z)$ sont des arêtes (l'arête $(z, x)$ peut être présente ou non). Remarquons que les chemins (et les cycles) sont des séquences, d'où le cycle $x$, $y$, $z$, $x$ est différente de $y$, $z$, $x$, $y$. Le coefficient de clustering (ou transitivité) d'un réseau est le rapport entre le nombre de cycles de longueur trois et le nombre de chemins de longueur deux. Il s'agit donc de la fréquence des cycles de longueur trois dans le réseau. Le coefficient de clustering global $C$ est définit comme suit :

$$C = \frac{\text{nombre de triplets fermés}}{\text{nombre de tous les triplets (ouverts et fermés)}} \tag{3.1}$$

Le nombre de triplets fermés correspond aussi dans la littérature à $3 \times$ triangles d'où :

$$C = \frac{3 \times \text{nombre de triangles}}{\text{nombre de tous les triplets (ouverts et fermés)}} \tag{3.2}$$

### 3.1.4 Connexité et composante connexe

Pour un graphe donné, une composante connexe est un sous-ensemble maximal de sommets tels que deux quelconques d'entre eux soient reliés par une chaîne (chemin) c'est à dire une suite de sommets et liens. Un graphe est dit connexe s'il existe un chemin entre toutes ses paires de sommets.

## 3.2 Quelques familles classiques de graphes

Plusieurs modèles de graphes artificiels existent dans la littérature, ces derniers ont été développés depuis la fin des années 50 pour tenter de représenter des graphes qui ressemblent à des réseaux réels. Chacun de ces modèles propose un sous-ensemble des caractéristiques observées dans les réseaux réels dits "réseaux complexes" à savoir une faible densité, un coefficient de clustering plus élevé que la densité, un faible diamètre et une distribution des degrés en loi de puissance, etc.

### 3.2.1 Modèle de réseaux aléatoires d'Erdös-Rényi

Ces modèles de graphes aléatoires ont été introduits par Erdös et Rényi en 1959 [31]. Il s'agit du cas le plus simple à imaginer. Considérons un ensemble de $N$ sommets complétement déconnectés.
L'idée est de commencer à sélectionner aléatoirement une paire de sommets et de les connecter par un lien avec une certaine probabilité $p, 0 \leq p \leq 1$ ensuite continuer ainsi et de manière uniforme.
Le résultat obtenu est un réseau avec plusieurs composantes séparées. Rappelons qu'une composante désigne un sous-ensemble de sommets connectés (aucun sommet n'est isolé). Ce processus de connexion aléatoire des paires de sommets révèle une des propriétés très remarquable et prévisible propre au réseau aléatoire.
Soit $p$ la proportion de liens connectés dans le graphe aléatoire. En fonction de la valeur du paramètre $p$, trois comportements très différents sont observés : Dans le cas où $p$ est suffisamment petit, le graphe a plusieurs composantes connexes, et donc toutes de petites tailles (Nombres faibles de sommets). Ensuite à partir d'une certaine valeur de $p$, le graphe a une unique composante "géante", de grande taille (qui tend vers $N$), et les autres composantes connexes sont très petites. Ce seuil est nommé "seuil critique de connectivité" et il est atteint quand $p$ s'approche de $\frac{log(N)}{N}$ [32]. Ainsi, le graphe tend à devenir connexe.
On compte deux variantes différentes mais étroitement liées [2] de modèles de graphes aléatoires d'Erdös-Rényi :

— La première variante $G(N, M)$ est un modèle de génération de graphes à $N$ sommets et $M$ liens. Un graphe est choisi de manière uniforme et aléatoirement à partir de cette collection de graphes possibles. Le nombre des graphes possible est de $\binom{\binom{N}{2}}{M}$.

Par exemple, dans un modèle $G(3,2)$, combinaison de $\binom{3}{2} = 3$ paires possibles donc 3 graphes non-orientés sont possibles. Ces graphes ont $N = 3$ sommets, donc 3 paires possibles et par conséquent $M = 2$ liens introduits avec une probabilité de $2/3$ chacun. Si on étend cela aux graphes orientés, à $N = 3$ sommets et $M = 2$ arcs (liens orientés), il y aura 6 paires d'arcs possibles, et par conséquent, $\binom{6}{2}$ soit 15 graphes orientés envisageables avec 3 sommets et 2 arcs orientés chacun.

— La deuxième variante $G(N,p)$ est un modèle de génération de graphes à $N$ sommets, mais cette fois-ci les sommets sont connectés aléatoirement et chaque lien est présent indépendamment des autres et avec une probabilité $p$, $0 \leq p \leq 1$.
Ainsi, tout graphe généré de la sorte, a une probabilité d'être choisi égale à :

$$p^M (1-p)^{\binom{N}{2}-M}$$

En moyenne, un graphe de $G(N,p)$ à $p^M(1-p)^{\binom{N}{2}-M}$ liens. La distribution des degrés de tout sommet particulier $s$ est binômiale [71].

$$P(deg(s) = k) = \binom{N-1}{k} p^k (1-p)^{N-1-k} \underset{N\to\infty}{\sim} \frac{z^k \exp(-z)}{k!}.$$

Avec $z = p(N-1) \underset{N\to\infty}{\sim} pN$ correspond au degré moyen de chaque sommet. Ainsi, quand le nombre de sommets $N$ est très grand ($N \to \infty$), la distribution des degrés suit une loi de Poisson de paramètre $z = pN$.

### 3.2.2 Modèle de réseaux "petit-monde" de Watts et Strogatz

Selon Watts [87], un graphe est dit aléatoire s'il obéit au moins aux conditions suivantes :

1. Réseau de grande taille : le nombre de sommets grand, ce qui permet de ne pas retrouver cas où tous les sommets sont liés, d'un graphe de petite taille.
2. Le réseau est de faible densité : c'est à dire que le nombre de liens présents dans le réseau est très faible par rapport au nombre de liens possibles (reliant toutes les paires de sommets), si $z$ est le nombre de liens qu'un sommet choisi aléatoirement possède alors $z << N$. Cela revient à dire qu'en moyenne, les sommets ne sont liés qu'à $z$ autres sommets, et donc chaque sommet n'a donc qu'un très petit voisinage directe (liens directs) comparé à la taille du réseau.
3. Le réseau est décentralisé : cette condition permet de s'assurer qu'aucun sommet n'est dominant, c'est-à-dire auquel tous les autres sommets sont connectés. Ainsi, le nombre de liens qu'un sommet peut avoir au maximum $z_{max}$ doit être très faible devant le nombre total de sommets ce qui réduit les distances géodésiques.

4. Le réseau a une structure communautaire sous-jacente : ce qui correspond à un coefficient de clustering élevé.

Le modèle de réseaux dit “petit-monde” (ou "Small-World") de Watts et Strogatz [88] s’inspire du phénomène petit-monde qui est également connu sous le nom des “six degrés de séparation”. Ce phénomène tient son nom de la célèbre expérience du sociologue Milgram réalisée aux États-Unis dans les années 60 [83]. Cette expérience avait pour objectif de faire arriver un message à un individus cible (personnalité célèbre par exemple) en formant une chaîne optimale d’intermédiaires. Chaque individu qui reçoit le message doit le transmettre à la personne qu’il connaît et qu’il considère la plus proche de la cible. L’idée est d’atteindre la cible avec le minimum d’intermédiaires possibles. 20% des chaînes de personnes initiées ont atteint les cibles avec une moyenne de longueurs de 6.5 personnes intermédiaires.
En plus des propriétés dénombrées par Watts dans l’article [87] qui ont été présentées dans le paragraphe (Modèle de réseaux "petit-monde" de Watts et Strogatz), les réseaux “petit-monde” se caractérisent par une faible distance géodésique moyenne. Il s’agit d’un graphe qui a une forme intermédiaire entre un graphe régulier de type graphe de Caveman connecté et le graphe aléatoire comme l’illurstre la figure 3.1.

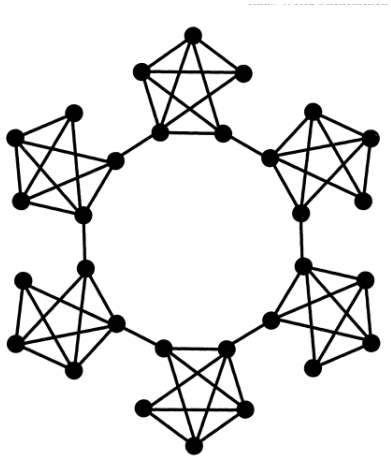


FIGURE 3.1 – Structure de construction du graphe de Caveman selon Watts [87]

Ils sont petit-monde car ils contiennent des liens particuliers (des sortes de raccourcis), s’ils sont supprimés cela induit une croissance de la distance entre deux sommets à une distance supérieure à 2. Un nombre très faible de ces liens particuliers (aléatoires) permet de rendre le monde petit [88].

Afin de caractériser les valeurs observées de la distance géodésique moyenne $L_{obs}$ et du coefficient de clustering $C_{obs}$ pour ces réseaux, Watts [87] s’est basé sur un graphe connu qui est celui de Caveman connecté connu par son coefficient de clustering $C_{caveman}$ élevé.
Les formules de calcul de $C_{caveman}$ et de la distance géodésique moyenne de ce graphe

$L_{caveman}$ sont les suivantes :

$$C_{caveman} \simeq 1 - \frac{6}{z^2 - 1};\ L_{caveman} \simeq \frac{N}{2(z+1)}.$$

Rappelons que $z$ représente la taille moyenne du voisinage d'un sommet, qui correspond à la centralité de degré moyen des sommets.
Pour les réseaux avec une distance géodésique moyenne $L$ faible est le graphe pour lequel $C_{random}$ et $L_{random}$ ont les formules suivantes :

$$C_{random} = \frac{z}{N}; L_{random} = \frac{ln(N)}{ln(z)}.$$

Ainsi, un graphe petit-monde est une forme intermédiaire entre les deux graphes extrèmes qui sont le graphe régulier de Caveman connecté et le graphe aléatoire.
Une quantification du caractère "petit-monde" $SW$ a été proposée par Watts. Elle compare la structure observée du réseau à celle d'un réseau aléatoire ayant la même taille et se calcule par la quotient suivant :

$$SW = \frac{C_{obs}}{L_{obs}} \times \frac{L_{random}}{C_{random}}.$$

Par conséquent, pour considérer qu'un réseau comme "petit-monde", il faut que son quotient soit très supérieur à 1, $SW >> 1$. Ce qui revient à dire que son coefficient de clustering est élevé $C_{obs} >> C_{random}$ et sa distance géodésique moyenne est faible $L_{obs} < L_{random}$.

### 3.2.3 Modèle de réseaux sans-échelles de Barabási-Albert

La croissance et l'attachement préférentiel (cf. (1) et (2) ci-dessous) sont deux propriétés importantes des réseaux réels. Barabási et Albert [7] se sont inspirés de ces deux propriétés pour créer un modèle minimal appelé modèle (BA), pouvant générer des réseaux sans échelle [1] comme suit :
Commençons avec les sommets $m_0$ ($m \leq m_0$), les liens entre eux sont choisis arbitrairement, tels que chaque sommet a au moins un lien. Le réseau se développe en deux étapes :

1. Un nouveau sommet est rajouté avec $m$ liens, $m \leq m_0$ qui connectent le nouveau sommet à $m$ sommets déjà dans le réseau.
2. L'attachement préférentiel est respecté dans la mesure où le nouveau sommet se connecte à un sommet $i$ avec probabilité $P(k_i)$ dépendant du degré $k_i$ comme suit :

1. Un réseau est dit sans échelle si la distribution des degrés suit une loi de puissance ce qui signifie qu'un petit nombre de sommets sont massivement liés et que la très grande majorité ne le sont jamais ou presque.

$$P(k_i) = \frac{k_i}{\sum_j (k_j)} \tag{3.3}$$

L'attachement préférentiel permet à un nouveau sommet de se connecter à n'importe quel sommet du réseau, qu'il s'agisse d'un hub [2] ou sommet à lien unique $degré = 1$, cependant l'équation (3.3) impose, par exemple, que si un nouveau sommet a le choix entre un sommet de degré 2 et un autre sommet de degré 4, il est deux fois plus probable qu'il se connecte au sommet de degré 4.

Après $t$ étapes, le modèle BA génère un réseau avec $N = t + m_0$ sommets et $m_0 + mt$ liens. Le réseau obtenu a une distribution de degré en loi de puissance.

La figure 3.3 et surtout la matrice d'adjacence d'un réseau de type BA, montre que la plupart des sommets n'a que quelques liens. Cependant, quelques-uns se transforment progressivement en hubs. Ces hubs sont le résultat l'attachement préférentiel qui génère un phénomène de riche-enrichissant : Les nouveaux sommets sont susceptibles de se connecter plus aux sommets les plus connectés qu'aux sommets faiblement connectés. Par conséquent, les sommets les plus importants acquièrent des liens au dépend des plus petits sommets, devenant finalement des hubs.

Bref, le modèle de BA indique que la croissance et l'attachement préférentiel, sont responsables de l'émergence de réseaux sans échelle.

La définition du modèle Barabási-Albert telle que présentée dans [7] laisse ouverts de nombreux détails dont les principaux sont les suivants :

— Elle ne spécifie de façon précise la configuration initiale des premiers $m_0$ sommets.

— Elle ne spécifie pas si les $m$ liens affectés à un nouveau sommets sont ajoutés un par un ou simultanément.

Cela conduit à des conflits logiques : si les liens sont vraiment indépendants, ils pourraient se connecter au même sommet $i$, ce qui donnerait un multi-liens. Bollobás et al. [10] ont proposé le diagramme de couplage linéarisé (LCD) pour remédier à ces problèmes. Selon le LCD, pour $m = 1$, on construit un graphe $G_1^{(t)}$ comme suit : (voir figure 3.2)

1. Le processus commence par $G_1^{(0)}$, correspondant à un graphe vide sans sommets,
2. Étant donné $G_1^{(t-1)}$ générer $G_1^{(t)}$ en ajoutant un sommet $v_t$ et un seul lien entre $v_t$ et $v_i$ un sommet de $G_1^{(t-1)}$, où $v_i$ est choisi avec la probabilité $p$ dépendant du degré $k_i$ de $v_i$ comme suit :

$$p = \begin{cases} \frac{k_i}{(2t-1)} & \text{si } 1 < i \leq t-1 \\ \frac{1}{(2t-1)} & \text{si } i = t \end{cases} \tag{3.4}$$

2. Un hub est un sommet dont un nombre de liens largement supérieur à la moyenne.

où le nouveau lien contribue déjà au degré de $v_t$. En conséquence, le sommet $v_t$ peut également être relié à lui-même avec la probabilité $\frac{1}{(2t-1)}$, le deuxième terme de l'équation (3.4). Notons également que le modèle permet les auto-cycles et multi-liens. Pourtant, leur nombre devient négligeable lorsque $t \to \infty$.
Pour $m > 1$ on construit $G_m(t)$ en ajoutant $m$ liens à partir du nouveau sommet $v_t$ un par un, à chaque étape, permettant au nouveau lien ajouté de contribuer aux degrés.
La figure 3.2 représente l'évolution de la construction d'un réseau Barabasi-Albert.

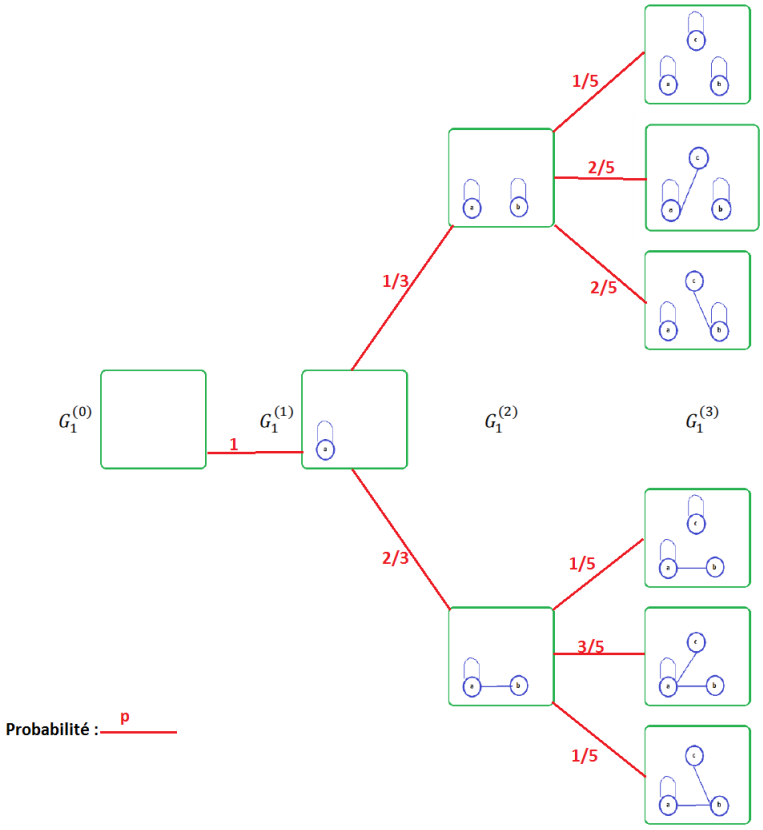



FIGURE 3.2 – Evolution de la construction d'un réseau Barabasi-Albert

### 3.2.4 Limites de chaque modèle

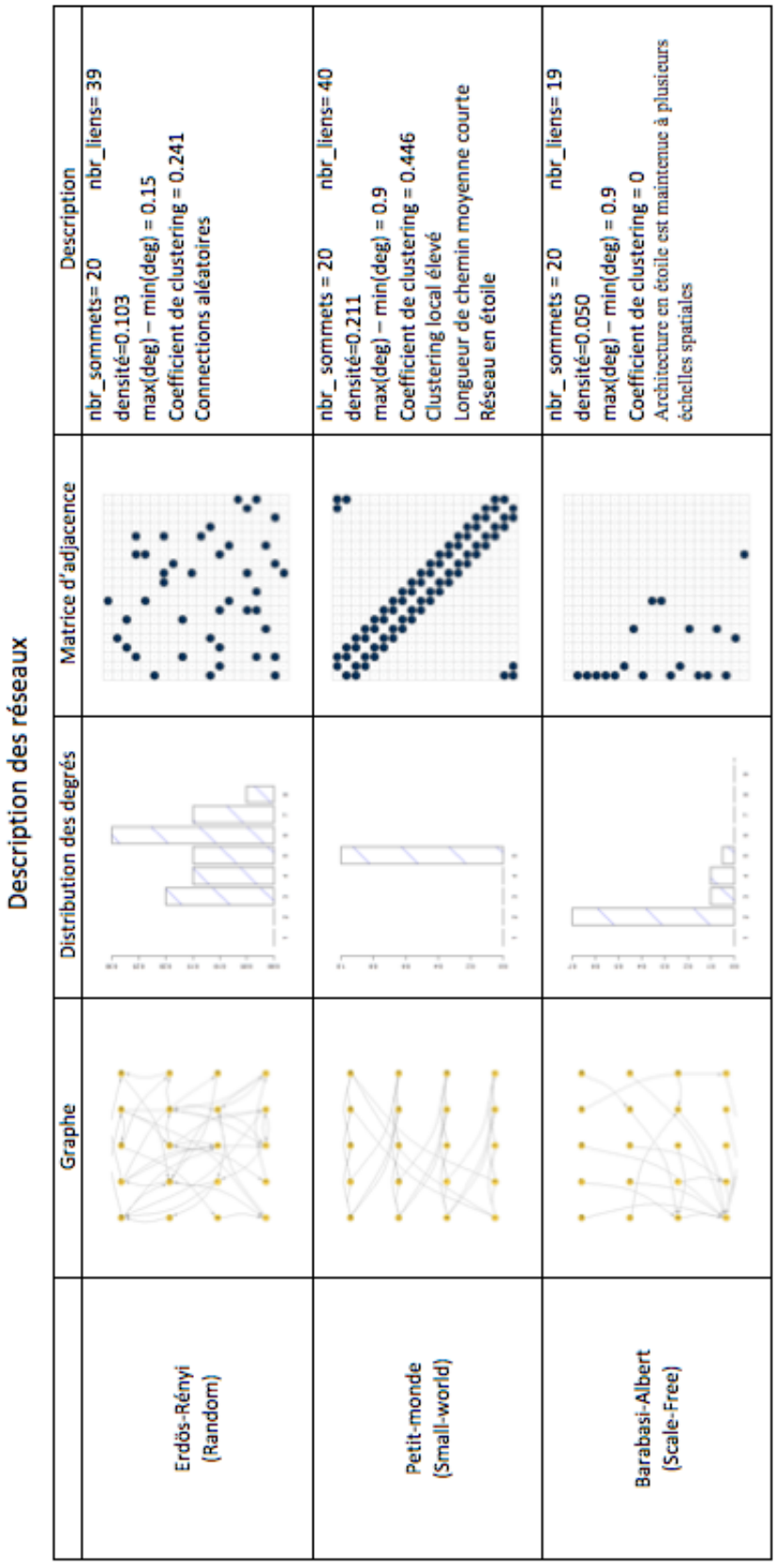
Comme évoqué précédemment, chaque modèle de génération de graphes synthétiques capture un sous-ensemble des propriétés souvent observés dans les graphes réels.
Les réseaux d'Erdös- Rényi par exemple n'ont pas deux propriétés importantes observées dans ces graphes du monde réels à savoir : d'une part, la non-génération de structure com-

munautaire locale [3] et de fermeture triadique [4].Puisque dans ces modèles deux sommets ont une probabilité d'être connectés constante, aléatoire et surtout indépendante. Cela génère des graphes avec un faible coefficient de clustering.
D'autre part, la distribution des degrés des graphes générés par le modèle d'ER converge vers une loi de poisson plutôt qu'une loi de puissance qui est le cas de plusieurs réseaux du monde réels sans-échelle (scale-free) tels que ceux de Barabasi-Albert, ce qui rend les modèles de type ER moins représentatifs de la réalité puisqu'ils ne tiennent pas compte de la formation des hubs.
Quant aux limites du modèle "Small-World" de Watts & Strogatz, elles sont essentiellement dûent à la distribution des degrés qui tout comme le modèle ER ne suit pas une loi de puissance.
Le modèle "sans-échelle" de Barabsi-Albert a comme principale limite l'absence de structure communautaire, en effet, ce type de graphe assure la loi en puissance de la distribution de degrés par l'attachement préférentiel, mais produit un coefficient de clustering nul.
Plusieurs autres modèles de génération de graphes existent dans la littérature. Cependant, nous nous restreignons à ces trois modèles qui serviront d'exemples de validation dans les chapitres suivant. La figure 3.3 représente un réseau de chaque famille, sa matrice d'adjacence et un résumé de ses caractéristiques topologiques.

3. Structure communautaire : c'est une partition des nœuds telle que les nœuds de chaque communauté sont plus connectés entre eux qu'avec l'extérieur

4. Fermeture triadique : la présence des liens entres les sommets A et B et entre les sommets A et C suggère la présence du lien entre B et C.

FIGURE 3.3 – Tableau de bord récapitulatif de ces modèles



La section suivante présente les plus importantes mesures de centralités utilisées dans la théorie des graphes.

## 3.3 Centralités

Dans cette sections nous présentons les définitions des centralités les plus utilisées dans l'analyse des réseaux sociaux et d'autres considérées comme méta-centralités usuellement employées pour la recherche d'information dans le Web.

### 3.3.1 Centralités issues de l'Analyse des Réseaux Sociaux

#### 3.3.1.1 Centralité de degré

Cette centralité comme la plupart des centralités issues de l'Analyse des Réseaux Sociaux fut introduite par Freeman [35]. Elle est considérée comme la centralité la plus intuitive car elle mesure l'importance d'un individu au sein d'un environnement (groupe) par le nombre de relations qu'il a. Cela inclue par exemple les autres individus du groupe qu'il connaît, avec qui il intéragit ou collabore dans son environnement professionnel à titre d'exemple. Un individu peut représenter une personne physique, morale telle qu'une entreprise ou tout autre type d'objet modélisé par un sommet dans le graphe étudié. Ainsi, le degré d'un sommet correspond au nombre de sommets auxquels il est directement lié, qui représentent aussi son voisinage immédiat, ce qui est aussi équivalent au nombre de liens qui lui sont adjacents. Cela correspond à l'appelation propre à la théorie des graphe qui est le "degré d'un sommet".
Soit un graphe $G(V, E)$ avec $N$ sommets. Ce graphe est représenté par sa matrice d'adjacence $A = (a_{(ij)}), 1 \leq i, j \leq N$.

— Si $G$ est un graphe non-orienté, $A$ est symétrique et la centralité de degré d'un sommet $s_i \in V$ est définie par la formule suivante :

$$deg(s_i) = \frac{\sum_{j=1}^{N} a_{ij}}{N-1} \tag{3.5}$$

Le vecteur de centralité de degré s'écrit sous la forme : $deg = \frac{A \times 1}{N-1}$ ; où 1 est un vecteur colonne de taille $N$ avec des composantes toutes égales à 1. La figure 3.4 représente un exemple de graphe non-orienté où les couleurs des sommets représentent les valeurs de leurs centralité de degré.

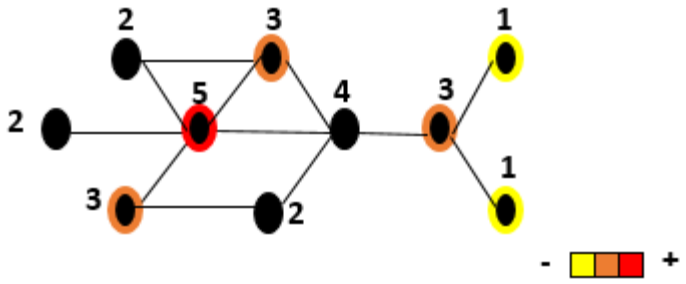



FIGURE 3.4 – Centralités de degré dans un graphe non-orienté.

— Si $G$ est un graphe orienté, sa matrice d'adjacence $A$ n'est plus forcément symétrique. Dans ce cas la centralité de degré peut être dissociée en deux mesures. L'une entrante et l'autre sortante ; la première prend en compte uniquement les liens qui arrivent vers le sommet étudié, tandis que la seconde considère ceux qui en sortent. Elles sont calculées respectivement par les formules suivantes :

$$\begin{cases} degIn(s_i) & = \frac{\sum_{j=1}^{N} a_{ij}}{N-1} \\ degOut(s_i) & = \frac{\sum_{j=1}^{N} a_{ji}}{N-1} \end{cases} \tag{3.6}$$

Les vecteurs de centralité respectifs s'écrivent comme suit :

$$\begin{cases} degIn & = \frac{A \times 1}{N-1} \\ degOut & = \frac{A^T \times 1}{N-1} \end{cases} \tag{3.7}$$

Remarques :

— Dans le cas d'un graphe orienté : $deg(s_i) = degIn(s_i) + degOut(s_i)$.

— Dans le cas d'un graphe non-orienté : $deg(s_i) = degIn(s_i) = degOut(s_i)$.

#### 3.3.1.2 Centralité de proximité

Cette centralité est une mesure globale également introduite par Freeman [35]. Elle s'intéresse à l'étude de la position qu'occupe un sommet dans un graphe et mesure l'intensité de proximité qu'il a avec les autres éléments du graphe. Dans un réseau informatique par exemple, elle correspond à l'idée qu'un équipement puisse atteindre rapidement et facilement l'ensemble des autres équipements. Ce qui correspond concrétement au calcul de sa proximité moyenne par rapport aux autres sommets du graphe.
Comme pour la centralité de degré, son calcul dépend du type de graphe étudié (orienté ou

non).

— Dans le cas d'un graphe non-orienté $G(V, E)$, ayant $N$ sommets et représenté par la matrice d'adjacence $A$, la centralité de proximité d'un sommet $s_i \in V$ s'écrit comme suit :

$$c(s_i) = \frac{N-1}{\sum_{j=1}^{N} d(s_i, s_j)},$$

où $d(s_i, s_j)$ représente la distance entre les sommets $s_i$ et $s_j$.
Il existe plusieurs métrique pour calculer la distance entre deux sommets. Freeman suggère l'utilisation de la distance dite géodésique qui correspond à la longueur du plus court chemin entre deux sommets.
La figure 3.5 représente un graphe où les couleurs des sommets correpondent à leur classement selon la mesure de centralité de proximité.

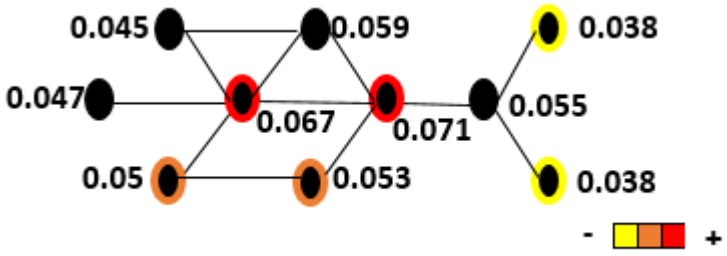



FIGURE 3.5 – Centralités de proximité dans un graphe non-orienté.

— Si $G(V, E)$ est un graphe orienté, deux mesures de centralité de proximité sont distinguées : l'une entrante $cIn$ et l'autre sortante $cOut$.

$$\begin{cases} cIn(s_i) & = \frac{N-1}{\sum_{j=1}^{N} d(s_i, s_j)} \\ cOut(s_i) & = \frac{N-1}{\sum_{j=1}^{N} d(s_j, s_i)} \end{cases} \tag{3.8}$$

telle que $d(s_i, s_j)$ est la distance orientée entre les sommets $s_i$ et $s_j$.
La distance orientée prend en compte l'orientation des liens dans le calcul, du plus court chemin, si la distance utilisée est la distance géodésique.

#### 3.3.1.3 Centralité d'intermédiarité

Cette mesure de centralité globale fut également introduite par Freeman [35]. Elle a comme idée de base qu'un sommet est d'autant plus important, s'il est indispensable de passer par lui pour se déplacer d'un sommet à un autre. En d'autres termes, un sommet est important par cette centralité s'il est parcouru par plusieurs chemins géodésiques dans

le graphe.
Selon Borgatti et Everett [13], un individu appartenant à un réseau social avec une valeur élevée de la centralité d'intermédiarité correspond à un sommet dont dépendent beaucoup d'interactions entres sommets ne lui étant pas adjacent dans le graphe.
Soit $G = (V, E)$ un graphe (orienté ou non) avec $N$ sommets. La centralité d'intermédiarité d'un sommet $s_i$ se calcule par la formule suivante :

$$betw(s_i) = \sum_{j=1}^{N} \sum_{k=1}^{N} \frac{g_{jk}(s_i)}{g_{jk}},$$

tel que $g_{jk}(s_i)$ représente le nombre de chemin géodésiques entre les sommets $s_j$ et $s_k$ passant par le sommet $s_i$ et $g_{jk}$ le nombre total de chemins géodésiques entre $s_j$ et $s_k$.
Notons que la prise en compte de l'orientation des liens se fait dans le calcul de la distance géodésique. La mesure de la centralité d'intermédiarité se base sur le fait que les sommets n'intéragissent entre eux que via les plus courts chemins (chemins géodésiques). D'autres mesures ne se basant pas sur cette même hypothèse ont été proposées dans la litterature telle que la centralité du flux d'intermédiarité [34]. La figure 3.6 représente un graphe où les sommets sont colorés selon leurs intermédiarités.

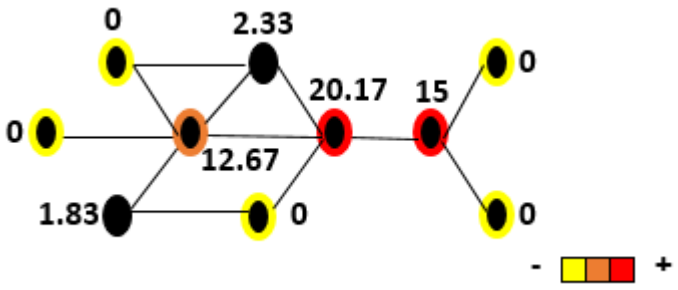



FIGURE 3.6 – Centralités d'intermédiarité dans un graphe non-orienté.

#### 3.3.1.4 Limites de certaines centralités issues de l'Analyse des Réseaux Sociaux

Etant donné que la centralité de degré est une mesure de centralité locale [77], elle ne considère que le voisinage immédiat du sommet concerné et donc n'inclut pas la structure globale du réseau, ce qui la rend peu informative quand l'objectif est d'étudier un graphe dans sa globalité.
La centralité de proximité n'est applicable que lorsque le graphe étudié est connexe, car dans le cas contraire, les distances géodésiques entre certains sommets peuvent être indéfinies s'il n'existe aucun chemin entre eux.
D'autre part, pour que la centralité d'intermédiarité d'un sommet ne soit pas nulle, il faut que ce dernier ait au moins deux liens, l'un entrant et l'autre sortant. Typiquement, dans

un graphe ayant la majorité de ses sommets qui sont soit des feuilles (un degré sortant nul) ou des sources qui sont la configuration inverse (un degré entrant nul), la centralité d'intermédiarité n'est pas pertinente pour étudier ses sommets.

### 3.3.2 Centralités issues de la recherche d'information dans le Web

Cette partie s'intéresse à des centralités utilisées pour mesurer l'importance d'une page dans le recherche d'information dans le web. Pour cela, on peut définir le web de manière simplifié comme étant une collection de $N$ pages avec $N$ un nombre entier très grand.
Selon le site internet [1], ce nombre s'éléverait à au moins 47 milliards de pages référencées dans le WWW et estime de 47.4 milliards le nombre de pages référencées par Google (valeurs relevée le 17 septembre 2018).

La majorité de ces pages web contiennent des liens hypertextes faisant référence à d'autres pages (internes ou externes). Le terme employé est "pointer vers".

Le principe de base considéré par les moteurs de recherche pour classer les pages selon leurs pertinences (ordre décroissant) est le fait de considérer que plus une page est pointée par des liens venant d'autres pages, plus elle est a de chance d'être intéressante et fiable en terme de contenu pour l'utilisateur ; et réciproquement. Ainsi, l'objectif est de quantifier cette pertinence pour chacune des pages.

Il s'agit de la logique adoptée par l'algorithme de référencement de Google qui est le PageRank présenté dans le paragraphe PageRank.
Tandis que son ancêtre, l'algorithme HITS suggère non pas une seule mesure d'importance d'une page mais deux centralités la centralité de Hubité (ou hubscore) et celle de l'autorité. Ces dernières sont présentées dans la partie qui suit.
Les autres mesures de centralités également utilisées dans la recherche d'information dans le web telles que les autres centralités de feedback (l'indice de Katz [57] qui est une généralisation du degré d'un sommet, la centralité de vecteurs propres de Bonacich[11, 12] et l'indice de Hubbell [48]) ne sont pas abordées ici.

#### 3.3.2.1 Hubs et autorités

L'algorithme HITS pour "Hypertext Induced Topics Search" (également connu sous le nom Hubs and Authority) est un algorithme d'analyse de liens qui classe les pages Web, développé par Jon Kleinberg [59, 60]. Il utilise la structure du Web afin d'améliorer la qualité et la pertinence de la recherche.

Il détermine deux valeurs pour une page : son autorité et son degré d'hubité, qui sont des mesures de centralités relatives respectivement aux liens entrants et sortants. L'autorité d'une page estime la valeur de son contenu, tandis que son hubscore évalue les liens qu'elle a vers d'autres pages. Une autorité est une page référencée (liens entrants) par plusieurs

hubs, et donc un hub est une page qui pointe (liens sortants) vers plusieurs autorités.
HITS n'opère que sur un petit sous-graphe (défini grâce à la graine relative à la requête de recherche $Q$) à partir du graphe Web. Ce sous-graphe dépend de la requête. A chaque fois que l'utilisateur cherche avec une phrase de requête différente, la graine change également. HITS classe les sommets de départ en fonction de leurs valeurs d'autorité et de leurs hubscores. Les pages ayant les valeurs les plus élevées sont affichées à l'utilisateur par le moteur de requête.
Pour bien comprendre les concepts de hubs et autorités, prenons l'exemple des sites webs tels que "TripAdvisor", "La Fourchette", "Booking". Ces sites web listent des endroits touristiques, des restaurants ou des hotels et hébergements. Ces derniers sont notés et commentés par des utilisateurs. Une autorité est par exemple un restaurant qui apparait dans plusieurs sites et qui est bien noté. Un hub est une sorte de répertoire (un site web) qui regroupe plusieurs autorités qui sont des restaurants (ou hotels, ...) bien notés. Le sous-graphe modélisant les liens entres les hubs et autorités peut être vu comme un graphe biparti où la relation est le référencement.
Ainsi, les valeurs d'autorités de sommets peuvent être déduits à partir de celles des hubs. Voici les étapes de l'algorithme HITS :

1. À partir de la requête $Q$ fournie par l'utilisateur, HITS assemble le jeu initial $S$ de pages :
   L'ensemble initial de pages est appelé ensemble racine. Ces pages sont ensuite étendues à un ensemble racine plus grand $T$ en ajoutant les pages liées à ou provenant de n'importe quelle page du jeu initial $S$.
2. associe ensuite à chaque page $p$ un poid de hubité $hub(p)$ et un poids d'autorité $aut(p)$, tous initialisés à 1.
3. met à jour itérativement les valeurs du hubscore et de l'autorité de chaque page. $p \to q$ dénote le fait d'avoir un hyperlien de la page $p$ vers la page $q$. HITS met à jour les hubscores et les autorités de la manière suivante :

$$\begin{cases} aut(p) = & \sum_{q \to p} hub(q). \\ hub(p) = & \sum_{q \to p} aut(q). \end{cases} \tag{3.9}$$

Prenons l'exemple de $G_{web}(V, E)$ un graphe web contenant 4 sommets $V(G_{web}) = \{A, B, C, D\}$. Ce graphe est dirigé et représenté dans la figure 3.7.

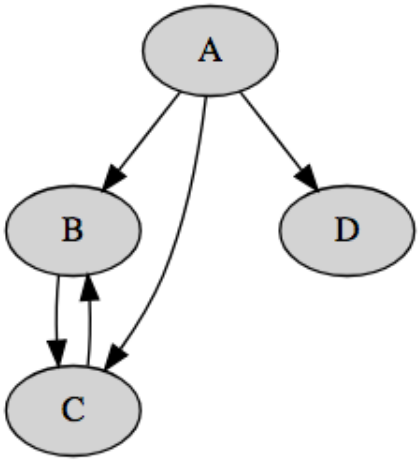



FIGURE 3.7 – Exemple de graphe Web

La représentation en graphe biparti avec les deux sous-ensembles hubs et autorités qui représentent respectivement les origines et destinations des liens dirigés de $G_{web}$ est dans la figure 3.8.

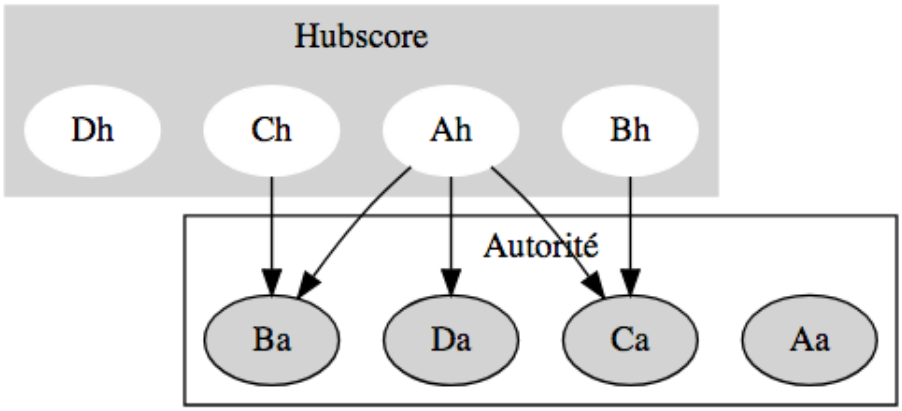



FIGURE 3.8 – Hubs et autorités du graphe web $G_{web}$

La calcul du degré de hubité (hub) et d'autorité (aut) pour les sommets de $G_{web}$ est résumé dans le tableau 3.1. Les formules utilisées proviennt de (3.9).

| Sommets | Hub-Score | Autorité-Score |
|---|---|---|
| A | 0.7887 | 0 |
| B | 0.5774 | 0.4597 |
| C | 0.2113 | 0.6280 |
| D | 0 | 0.6280 |

TABLE 3.1 – Calcul des valeurs du hubscore et d'autorité des sommets du graphe $G_{web}$

Le sommet $A$ a un fort hubscore tandis que son autorité est nulle cela est dû au fait qu'il n'a pas de lien entrant et qu'il pointe vers les sommets $B$ et $C$ qui ont des valeurs d'autorité élevées. Le cas inverse est celui du sommet $D$ qui a un hubscore nul (pas de lien sortant) et une autorité élevée (plusieurs liens venant de hubs).

#### 3.3.2.2 PageRank

L'algorithme PageRank a été développé par Larry Page le cofondateur de Google et Sergey Brin à l'Université de Stanford. Contrairement à l'algorithme HITS qui propose à chaque page deux mesures d'importance (centralités), le PageRank ne produit qu'un seul classement.
Son objectif est donc d'attribuer à chaque page un score proportionnel au nombre de fois qu'un utilisateur (surfeur) explorant de manière aléatoire le réseau du Web passerait par cette page. L'exploration aléatoire consiste à cliquer aléatoirement sur un des liens apparaissant sur chaque page. Ainsi, une page a un PageRank d'autant plus important qu'est grande la somme des PageRanks des pages qui pointent vers elle (elle, y compris dans le cas où elle contient des liens internes). Le PageRank est une mesure de centralité sur le réseau du web qui donc va du principe suivant "si les voisins d'un sommets sont nombreux et importants, il est important. Et plus un sommet est important, plus il rayonne son importance sur ses voisins."
Ainsi, on pourrait voir le PageRank comme une sorte de vote effectué par toutes les autres pages pour déterminer l'importance d'une page. Typiquement, un lien vers une page $p$ est considéré comme un vote. Cependant, l'absence de lien est considérée comme une abstention de vote et il n'existe pas de notion "vote contre". Selon l'article de Google [74], le PageRank est défini comme suit :
On suppose qu'une page $p$ est pointée par les pages $T_1, ..., T_n$ (sous forme de citations). Soit $d$ le facteur d'amortissement qui peut prendre des valeurs réelles allant de 0 à 1. Il est généralement fixé à 0.85. $C(p)$ correspond au nombre de liens sortants de la page $p$ (son degré sortant). Le PageRank de la page $p$ est donné par la formule récursive suivante :

$$pRank(p) = (1-d) + d\left(\frac{pRank(T_1)}{C(T_1)} + ... + \frac{pRank(T_n)}{C(T_n)}\right)$$

Selon l'article de Google le $pRank(p)$ peut être calculé grâce à un simple algorithme itératif et il correspond à la principale valeur propre de la matrice normalisée des liens du Web. $pRank(T_n)$ représente l'importance de la $n$ ième et dernière page ($pRank(T_1)$ l'importance de la 1ère page), puisque chaque page a une importance propre. $C(T_n)$ le nombre de vote accordés par la $n$ième page, ainsi $C(T_1)$ représente le nombre de vote de la 1ère page qui n'est autre que son degré sortant. L'introduction du facteur $d$ permet, comme son nom l'indique, de stoper les pages d'avoir une grande influence. Quant au terme $(1 - d)$ est plus un terme qui permet d'avoir un PageRank sous forme de probabilité et il a aussi la signification qu'une page même sans lien entrant aura une petite valeur du pageRank de 0.15 ( 1 – 0.85, car les fractions sont toutes nulles).
Etant donné que le PageRank de chaque page dépend de ceux des pages pointant vers elle, l'impression est de devoir calculer tous leurs PageRanks pour pouvoir obtenir celui de la page étudiée, ce qui peut former un cycle et qui renderait le calcul impossible. Cependant, il n'est pas nécessaire de devoir calculer le PageRank de toutes les pages pointant vers la page étudiée car dans l'article de Google, il est mentionné que le PageRank d'une page peut être calculé par un simple algorithme itératif et correspond au vecteur propre principal de matrice des liens normalisée du Web. Ainsi, à chaque fois que les calculs sont lancés, la valeur obtenue est d'autant plus proche de l'estimation de la valeur finale. Cela revient à répéter les calculs plusieurs fois jusqu'à ce que les valeurs se stabilient.

## 3.4 Conclusion

Ce chapitre regroupe quelques notions de la théorie des graphes nécessaires à la lecture de ce document. Il décrit aussi les principales familles de graphes théoriques présentes dans la littérature et utilisées dans la partie 2 pour la validation des mesures de similarité entre graphes monovariables (homogénéité et dépendance) et puis notre nouvelle mesure de similarité entre graphes multivariables. Ce chapitre présente également les principales mesures de centralité issues des domaines de l'Analyse des Réseaux Sociaux et de la recherche de contenu dans le Web. Ces centralités ont été définies pour les sommets d'un graphe néanmoins elles peuvent s'appliquer également aux arrêtes/arcs d'un graphe exactement de la même manière.

# Chapitre 4

# Construction de réseaux de systèmes et de séquences accidentelles

Ce chapitre décrit la méthode que nous proposons pour la construction de réseau à partir d'un système ou d'une séquence accidentelle. Ensuite sont présentées un exemple réel de séquences accidentelles relatives à l'initiateur de "baisse incontrolée du niveau (du circuit) primaire" dans les états d'arrêt du réacteur pour une installation de type EPR. Enfin, nous présentons les exemples réels de constructions des systèmes de sûreté du modèle EPR 2.

## 4.1 Méthode de construction de graphe d'une étude de sûreté

Dans ce paragraphe nous décrivons la méthode adoptée pour construire les réseaux d'une étude de sûreté. On s'intéresse à la modélisation d'une séquence accidentelle en réseau. Les séquences accidentelles sont décrites dans des notes EDF nommée Analyses Qualitatives de Séquences.

Une note AQS a pour objectif de présenter l'identification et la description des séquences accidentelles initiées par une famille d'initiateurs. Il est donc question de déterminer de façon précise les différentes missions systèmes (frontaux voire support, contrôle-commande ou facteur humain) intervenant dans la conduite de tels scénarios.

Une famille d'initiateurs englobe plusieurs initiateurs qui ont des effets similaires sur l'installation même s'ils peuvent être dans des contextes (états) différents, par conséquent, les missions systèmes à mettre en place pour les mitiguer sont les mêmes (ou presque). Une séquence accidentelle est établie pour un état de l'installation donné, certains états peuvent être regroupés. On peut distinguer 5 grands états du réacteur :

1. réacteur en puissance (état A1),
2. réacteur en attente à chaud (état A2),

3. réacteur en arrêt normal sur le générateur de vapeur (GV) (états A3, B),
4. réacteur en arrêt normal sur le système de refroidissement à l'arrêt (RRA) (état C),
5. réacteur en arrêt pour intervention (états D11, D12, D13, D21 et D22).

Les conséquences d'une séquence accidentelle sont soit acceptables (CA) soit inacceptables (CI) et il est possible d'avoir un renvoi vers d'autres familles d'initiateurs.

Chaque séquence accidentelle est résumée dans un diagramme appelé "Diagramme de requis fonctionnels" comme illustré dans la figure 4.1. Les blocs rectangulaires représentent les missions de sauvegarde, une transition verticale (flèche verte) représente le succès de la mission amont, tandis qu'une horizontale (flèche rouge) signifie son échec. Cette étape est faite manuellement [1]

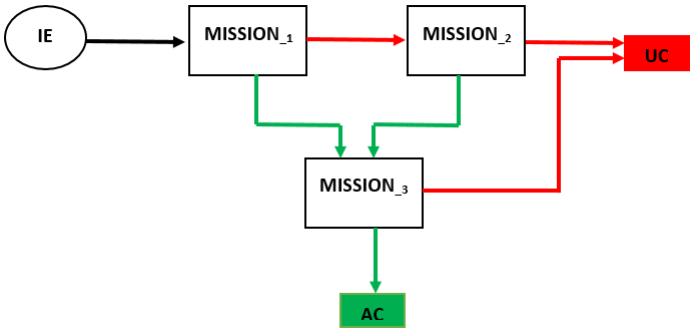



FIGURE 4.1 – Principe d'un diagramme de requis fonctionnels pour un évènement initiateur (IE) et deux conséquences inacceptable (UC) et acceptable (AC)

A partir d'une séquence accidentelle on obtient un "macro-réseau" avec des "macro-sommets" qui ne sont que les missions systèmes, et les liens sont des arcs qui représentent les transitions du diagramme de requis fonctionnels. Ce qui signifie qu'à partir de n'importe quelle séquence accidentelle, on peut construire le réseau correspondant.
L'étape suivante est de détailler les "macro-sommets" en réseau. Il faut donc récupérer les systèmes correspondants aux missions.

EDF dispose d'un outil nommé KB3 [14] qui permet de générer automatiquement les arbres de défaillances correspondant à des missions, ou études. Ces missions sont exportables en instanciant le "FIGARO.0" qui est un langage qui décrit l'ensemble des composants intervenant dans la mission, leurs caractéristiques, leurs interfaces qui sont les autres composants auxquels ils sont connectés (amont/aval). Dans le fichier obtenu, il existe aussi une partie "Interaction" qui décrit les conditions de succès de missions.

1. Cette étape a été automatisée récemment et une nouvelle fonctionnalité a été incluse dans l'outil Andromeda [36] d'EDF qui permet de générer un arbre d'évènement.

La construction des réseaux de missions inclut un travail d'analyse qui consiste à sélectionner les types de composants à conserver, car certains objets sont modélisés dans KB3 mais ne sont utilisés que pour des besoins de modélisation. L'analyse consiste aussi à se documenter sur chaque type des composants considéré pour identifier les attributs pertinents à modéliser, les types d'alimentations existantes, les missions d'alimentation correspondantes, les types de liens à modéliser.

Les listes des types et attributs sélectionnées sont susceptibles d'évoluer selon le palier étudié.

Pour la modélisation des AQS du palier EPR2 ces listes ont été finalisées. On modélise 55 types de composants, 3 types de liens. Une fois les réseaux des missions obtenus, ils sont intégrés dans le "macro-réseau".

Ainsi, à une séquence accidentelle correspond un réseau dirigé : ses sommets sont les différents composants (pompes, vannes, clapets, capteurs, tableaux électriques, et bien d'autres), les arcs sont les différents liens présents entre ses composants en plus des liens (échecs et succès de missions) déjà modélisés dans le "macro-réseau" à partir de diagramme de requis fonctionnel. Les sommets sont attribués, et les arcs aussi avec le type de liens modélisés appartenant au domaine de valeurs [ `lien_fluide`, `lien_electrique`, `lien_signal`, `lien_fictif` ].

## 4.2 Cas d'étude : Étude de la baisse incontrôlée du niveau primaire

Nous appliquons cette méthode pour construire le réseau correspondant l'initiateur de "la baisse incontrôlée du niveau primaire", pour une installation du palier EPR et dans les états d'arrêts. Afin de mitiguer les effets de cet initiateur, il faut assurer trois fonctions de sûreté qui sont :

— le contrôle de l'inventaire en eau primaire

— l'évacuation de la puissance résiduelle

— le maintien du confinement enceinte

Nous obtenons le réseau illustré dans la figure 4.2.

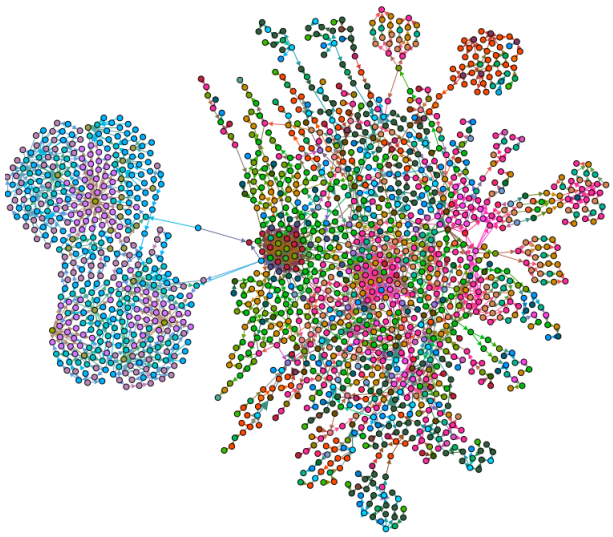


FIGURE 4.2 – Réseau de l'EPS EPR "baisse incontrôlée du niveau primaire" dans les états d'arrêts

Les couleurs des sommets représentent les différents types de composants modélisés (`pompes`, `vannes`, `tableau_HT`, `disjoncteur`, `capteur`, ...). Les arcs sont les liens entre ces composants. Ils sont de différents types (`lien_fluide`, `lien_electrique`, `lien_signal`).

Nous remarquons que le réseau obtenu est assez petit (environ 1700 sommets et 2700 arcs) comparé aux réseaux présents dans la littérature, ce réseau est dirigé, et attribué (sommets et arcs). Ce réseau est peu dense (*densité* = 0.0009), donc peut être considéré comme réseau complexe. De plus, il a un petit diamètre (*diamètre* = 33), ce qui fait de lui un réseau de type "petit monde", où les centralités se calculent rapidement et peuvent certainement révéler la structure sous-jacente. Le coefficient de clustering de ce réseau est de 0.012.

Rappelons que le coefficient de clustering d'un graphe mesure la probabilité que des sommets adjacents d'un sommet soient connectés.

D'autre part, nous appliquons les plus célèbres mesures de centralités sur le réseau obtenu, nous remarquons que les valeurs élevées de certaines sont spécifiques soit à des types de composants, soit à des composants de systèmes spécifiques. Par exemple, nous obtenons que les composants ayant les *proximités sortantes* les plus élevées sont tous de **types électriques**. D'autre part, les composants du **contrôle commande** sont très présents pour *l'autorité*. Les composants les plus *intermédiaires* sont tous **hydrauliques**. Nous remarquons aussi que les composants du système de **l'injection de sûreté** sont très *centraux* par rapport à toutes les centralités, sauf par rapport à *l'autorité*, ce type d'information nous permet de pouvoir déduire approximativement le rôle du système en question.

## 4.3 Réseaux réels des systèmes de sûreté nucléaire

Dans cette partie nous avons appliqué la méthode de construction de graphe à partir d'étude de sûreté pour modéliser les graphes correspondant aux systèmes du modèle EPR 2 initialement modélisées par l'outil Kb3 d'EDF. Nous disposons ainsi de 23 réseaux représentant les systèmes de sauvegarde utilisés en sûreté nucléaire. Pour EPR 2, certain systèmes ont été modélisés ensemble tels que CFI, CRF, SEC, SEN et SRU, ils sont donc modélisés en un seul réseau nommé CFI_ CRF_ SEC_ SEN_ SRU. Les réseaux GCT_ VDAVVP et de AAD_APA_ARE correspondent respectivement à celui des systèmes GCT, VDA et VVP pour le premier, et AAD, APA et ARE pour le second. Notons que nous avons 23 réseaux systèmes, les systèmes supports font partie des différents réseaux. Ainsi, nous avons formé un jeu de données que nous avons nommé $P_{sys-EPR2}$ composé des réseaux systèmes du modèle EPR2, il est noté pour la suite $P_{sys-EPR2}$ et il contient $K = card(P_{sys-EPR2}) = 23$ éléments.

Ces réseaux ont les propriétés structurelles résumées dans le tableau 4.1 suivant.

| Graphe | n | m | Diamètre | Coeff-clustering | Densité |
|---|---|---|---|---|---|
| AAD_APA_ARE | 1378 | 2049 | 13 | 0.0052 | 0.0011 |
| ASG | 1222 | 1776 | 13 | 0.0047 | 0.0012 |
| CEX | 1135 | 1659 | 13 | 0.0030 | 0.0013 |
| CFI_CRF_SEC_SEN_SRU | 1484 | 2149 | 13 | 0.0069 | 0.0010 |
| SRI | 1133 | 1635 | 13 | 0.0026 | 0.0013 |
| DEL | 1174 | 1715 | 13 | 0.0062 | 0.0012 |
| DER | 1104 | 1592 | 13 | 0.0027 | 0.0013 |
| DVD | 1376 | 2072 | 16 | 0.0040 | 0.0011 |
| DVL | 1260 | 1785 | 13 | 0.0027 | 0.0011 |
| DVP | 1064 | 1545 | 13 | 0.0029 | 0.0014 |
| EVU_EVi | 1193 | 1736 | 13 | 0.0029 | 0.0012 |
| GCT_VDAVVP | 1421 | 2144 | 13 | 0.0073 | 0.0011 |
| PTR | 1335 | 2006 | 13 | 0.0078 | 0.0011 |
| RBS | 1111 | 1600 | 14 | 0.0029 | 0.0013 |
| RCP | 1302 | 1974 | 16 | 0.0049 | 0.0012 |
| RCV | 1075 | 1553 | 13 | 0.0026 | 0.0013 |
| REA | 1054 | 1531 | 13 | 0.0027 | 0.0014 |
| RGL | 1052 | 1529 | 13 | 0.0027 | 0.0014 |
| **RIS** | 1577 | 2628 | 13 | **0.0187** | 0.0011 |
| **RRI** | 1793 | 2767 | 17 | **0.0108** | 0.0009 |
| SAR | 1123 | 1613 | 13 | 0.0026 | 0.0013 |
| SED | 1053 | 1530 | 13 | 0.0027 | 0.0014 |
| SER | 1068 | 1546 | 13 | 0.0027 | 0.0014 |

TABLE 4.1 – Profil topologique des réseaux des systèmes de sauvegarde

Nous remarquons que ces réseaux ont presque les mêmes propriétés topologiques : une densité $\in [0.0009, 0.0014]$, un coefficient de clustering $\in [0.0026, 0.0187]$, une distribution des degré en loi de puissance et un diamètre presque toujours égale à 13.
Un des objectifs de ce travail de thèse est de proposer une méthode de comparaison entre réseaux quelles que soient leurs tailles (nombres de sommets, nombres de liens, diamètres), en se basant uniquement sur leurs vecteurs de centralités. Cette étude sera présentée dans les chapitres 6 Approche statistique pour la comparaison monovariable de graphes et 7 Nouvelle mesure de similarité entre graphes multivariables et validée sur des jeux de données composés de graphes théoriques et sur $P_{sys-EPR2}$ dans le paragraphe Application au cas des réseaux réels des systèmes nucléaires du modèle EPR 2 $P_{sys-EPR2}$.

## 4.4 Conclusion

Ce chapitre présente comment construire à partir d'un système complexe un réseau correspondant modélisant les intéractions existantes. Nous avons pris le cas particulier qui est celui des systèmes de sûreté nucléaire et les séquences accidentelles. Nous exposons un exemple réel de séquences accidentelles de l'initiateur de "baisse incontrolée du niveau (du circuit) primaire" dans les états d'arrêt du réacteur pour une installation de type EPR. Les réseaux relatives aux systèmes de sûreté du modèle EPR 2 sont aussi présentés et caractérisés par les principales mesures topologiques.

# Chapitre 5

# Classification du Risk Increase Factor par les centralités des réseaux dirigés

Pour des besoins divers, l'exploitant nucléaire a besoin d'effectuer des calculs d'importance sur les différents composants de ses installations. le Facteur d'Accroissement du Risque (RIF) est un facteur d'importance le plus souvent utilisé pour mesurer l'aspect sûreté. Un calcul exact de ce facteur d'importance implique un calcul de probabilité d'occurence de l'évènement indésirable (risque de fusion du cœur pour l'EPS de niveau 1) pour chaque composant étudié (en supposant son indisponibilité) comme expliqué dans le paragraphe Mesures d'importances des Études Probabilistes de Sûreté. Ce calcul est très coûteux et les temps de calculs sont très variables. En effet, comme évoqué précédemment, les moyens de calcul EPS commençent à atteindre leurs limites. Une simple modification du modèle EPS peut transformer le temps de calcul de polynomial à exponentiel. Pour palier à cela, ce calcul se fait uniquement pour les composants qui apparaissent dans les coupes minimales qui ne constituent qu'environ 10% des composants.
Nous proposons d'utiliser les mesures de centralités des réseaux dirigés présentées dans la section 3.3 du chapitre 3 Graphes et centralités pour prédire le RIF. Ainsi le RIF est considéré comme une variable cible à prédire et les variables potentiellement prédictives sont les différentes centralités.
Nous prenons comme cas d'étude l'EPS EPR pour l'initiateur "Baisse incontrôlée du niveau primaire" déjà modélisée en réseau dans le paragraphe Cas d'étude : Étude de la baisse incontrôlée du niveau primaire.

Ce chapitre se structure comme suit :
Dans un premier temps nous définissons la problématique, les données étudiées et la variable à prédire. Ensuite, de présenter un choix judicieux des variables prédictives est effectué, pour enfin décrire les principaux résultats obtenus en appliquant les méthodes de classification

présentées dans l'annexe B. Une conclusion et des perspectives de ce travail clôturent ce chapitre.

## 5.1 Problématique

L'objectif est de prédire le RIF d'un composant appartenant à une séquence accidentelle. Les composants ayant des valeurs RIF élevées sont des composants critiques de point de vue de la sûreté, en d'autre terme, la défaillance d'un de ces composants a un grand impacte sur le risque.
Etant donné les redondances et diversifications prévues dès la conception des différents systèmes de sauvegardes, en plus des parades mises en place pour réduire l'impact de la défaillance d'un composant, le nombre de composants ayant des RIF élevés est très petit. On s'intéresse à prédire non pas la valeur du RIF d'un composant, mais plutôt si cette valeur est élevée.
Ainsi, il est question de prédire une classe rare, et donc moins présente dans notre échantillon d'étude.
Pour la prédiction de classe sur un échantillon étiqueté, on parle de classification ou de discrimination qui est une branche de l'apprentissage supervisé.
La classification a pour but de prédire la classe d'appartenance (l'étiquette) d'un individus (observation) et donc une fonction qualitative catégoriale. A partir d'un échantillon d'apprentissage, un classifieur doit être capable de prédire à quelle classe appartient une nouvelle observation. L'objectif est en premier lieu de pouvoir construire un modèle basé sur un jeu d'apprentissage et des valeurs (nom des catégories) et utiliser ce modèle pour classer (prédire la catégorie) des données nouvelles.
Nous utilisons pour la classification du RIF les méthodes d'apprentissage supervisé présentées dans l'annexe B. Rappelons que nous utilisons comme variables pour la prédiction les centralités des réseaux dirigés présentées dans la section 3.3 du chapitre 3 Graphes et centralités.

## 5.2 Données étudiées

Nous reprenons le réseau dirigé modélisé pour l'EPS EPR pour l'initiateur "Baisse incontrôlée du niveau primaire" et présenté dans le paragraphe Cas d'étude : Étude de la baisse incontrôlée du niveau primaire.
Nous ne pouvons pas traiter tous les sommets du réseau, car le calcul exact du RIF est coûteux, comme précisé anterieurement, puisqu'il implique pour chaque valeur, le calcul du Risque de Fusion du cœur en supposant que le composant étudié $x_i$ est défaillant (ou absent) par la formule présentée dans le tableau 2.1, ce qui est fastidieux.
Nous utilisons dans ce cas un échantillon formé par les 20 composants les mieux classés par chacune des 8 centralités réseau. L'échantillon est composé de 156 observations, par rapport

à un réseau d'environ 1700 sommets.
Pour chaque observation (composant), nous disposons de ses valeurs de centralités et de son RIF, et nous souhaitons utiliser ces valeurs de centralités pour prédire ce dernier. Dans la partie suivante, nous rappelons la définition de la variable à prédire (le RIF), les variables prédictives (les centralités) sont présentées dans la section 3.3 du chapitre Graphes et centralités.

## 5.3 Variable à prédire : le Risk Increase Factor

Le Risk Increase Factor (RIF), ou le facteur d'accroissement de risque est souvent considéré comme l'indicateur d'importance de point de vue sûreté [84] d'un composant. La formule de calcul du RIF est présentée dans le tableau 2.1 pour un composant $x_i$.
Nous prélevons un échantillon aléatoire sur lequel sera construit le modèle d'apprentissage supervisé. Cet échantillon est composé de 65% des données initiales. Le rest sera consacré au test du modèle.
Les valeurs de la variable RIF calculées sur l'échantillon d'apprentissage sont représentées dans la figure 5.1 où nous traçons la valeur prise par le RIF sur les observations de l'échantillon d'apprentissage.

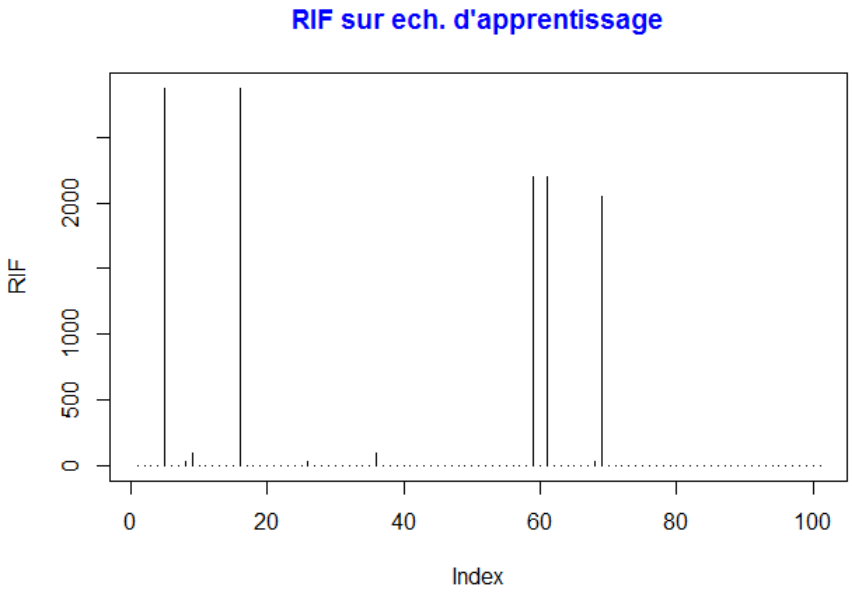



FIGURE 5.1 – Valeurs du RIF sur l'échantillon

Nous remarquons à partir de la figure 5.1 que la variable RIF prend de "petites" valeurs à quelques exceptions près. Ces exceptions (valeurs extrêmes) correspondent à des composants

dont la défaillance intrinsèque a un très grand impact sur le risque de fusion du cœur.
De plus, l'écart type de la variable RIF sur cet échantillon est 4 fois sa moyenne, ce qui indique que cette variable est dispersée donc difficile à ajuster.
Nous pouvons aussi, en se basant sur la figure 5.1 discrétiser la variable RIF en ne gardant que deux modalités comme suit :

— RIF élevé ($RIF > 2$) représenté par la catégorie **RIF=1** ;

— RIF faible ($RIF < 2$) représenté par la catégorie **RIF=0**.

Ainsi, le problème devient un problème de classification binaire puisque la variable à prédire est catégorielle et l'objectif est de prédire la classe "1" qui correspond aux composants dont la défaillance contribue fortement à l'augmentation du risque de fusion du cœur, et donc qui sont sensibles pour la sûreté de l'installation nucléaire. Comme évoqué précedemment, cette classe est minoritaire, nous sommes en présence d'un fort déséquilibre de classes car la classe à prédire ne constitue que 5% de l'échantillon. Nous réalisons une étape de sélection de variables dans la section suivante pour ne garder que celles pertinentes, ce choix se fait uniquement sur les données d'apprentissage sélectionnées précedemment de manière aléatoire. Nous utiliserons comme méthode de validation la méthode Hold-Out décrite dans l'annexe C.

## 5.4 Choix des variables de prédiction du Risk Increase Factor

Dans ce paragraphe, nous souhaitons affiner les variables de la classification. Pour cela, nous proposons de garder les variables qui sont fortement corrélées avec le RIF, ensuite nous éléminons celles qui sont fortement corrélées entre elles.
La figure 5.2 représente les distributions des valeurs des centralités degIn, degOut, cIn, cOut, betw, hub, aut et pRank ainsi que celle de la variable à prédire RIF prises sur l'échantillon d'apprentissage.

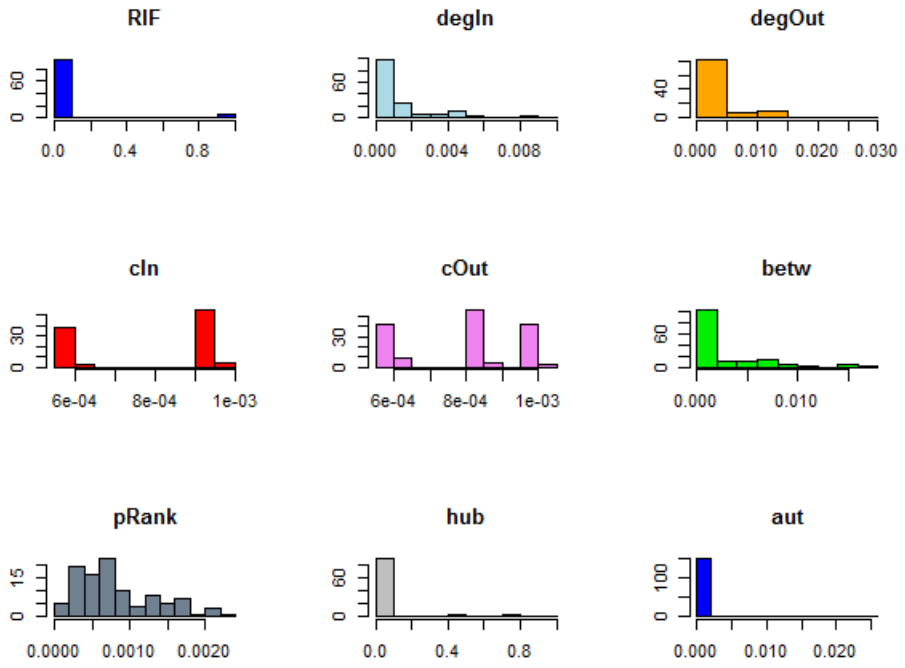



FIGURE 5.2 – Distributions des valeurs des centralités et du RIF sur l'échantillon d'apprentissage

Nous remarquons qu'aucune de ces variables n'est normale. Une étude de la normalité de ces variables a été réalisée par le test de normalité de Lilliefors-Van Soest présenté dans le paragraphe 6.2.3 Test de normalité de Lilliefors-Van Soest confirme bien la non-normalité observée grâce aux histogrammes de la figure 5.2.
La corrélation linéaire de Pearson est donc peu pertinente car suppose la normalité des variables étudiées. Pour cela, nous utilisons le coefficient de corrélation non-paramétrique de Spearman présenté dans le paragraphe 7.3.4.

La figure 5.3 représente les corrélogrammes de Pearson et de Spearman entre variables sur l'échantillon d'apprentissage. L'intensité (l'aire) des disques indique le degré de corrélation et sa couleur est bleue ou rouge selon que la corrélation est positive ou négative.
Dans la figure 5.3, nous représentons les valeurs absolues des corrélations car on ne s'intéresse qu'à l'intensité de la dépendance. Par conséquent, tous les disques sont bleus.


FIGURE 5.3 – Corrélogrammes de Pearson et de Spearman sur l'échantillon d'apprentissage

Les calculs de corrélations des rangs de Spearman révèlent que la variable cible RIF est très peu corrélée aux variables cOut et aut. De plus, elle est faiblement corrélée avec degIn et betw. Les tests d'indépendance non-paramétrique présentés dans la deuxième ligne du tableau 5.1 confirment les résultats des corrélations de Spearman et donc permettent d'éliminer degIn, cOut, et aut. Ceux de Pearson bien que peu pertients suggèrent d'après la première ligne du tableau 5.1 de garder comme variables explicatives du RIF les centralités degOut, cOut et pRank.

| Centralités | degIn | degOut | cIn | cOut | betw | pRank | hub | aut |
|---|---|---|---|---|---|---|---|---|
| Pearson | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 |
| Spearman | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 |

TABLE 5.1 – Tests de Pearson et Spearman entre les centralités et le RIF sur l'échantillon d'apprentissage

Pour résumer, le test de corrélation linéaire de Pearson suggère de prendre comme variables explicatives : le Degré sortant (degOut), la Proximité entrante (cIn) et le Page Rank (pRank). Celui de Spearman suggère de considérer les suivantes : le Degré sortant (degOut), la proximité entrante (cIn), le Page Rank (pRank), l'intermédiarité (betw) et le degré de hubité (hub). Puisque le maximum de correlation de Spearman avec RIF est

atteint pour la variable hub et le fait que l'intermédiarité (betw) est fortement corrélée aux autres variables, nous proposons d'inclure hub et éliminer betw des variables explicatives. Ainsi, les variables retenues sont : Le degré sortant (degOut), la proximité entrante (cIn), le Page Rank (pRank) et le degré de hubité (hub).

## 5.5 Prédiction du Risk Increase Factor par les centralités réseaux

Dans cette section, nous utilisons comme variables prédictives celles retenues dans la section précédente à savoir : le degré sorant (degOut), la proximité entrante (cIn), le Page Rank (pRank) et le degré de hubité (hub). Nous considérons comme données d'apprentissage celles prélevées aléatoirement avant l'étape de choix de variables (65% des données initiales) et comme données de test les 35% restants. Nous utilisons des méthodes d'apprentissage supervisé [44], plus précisément celle de l'arbre de classification puis celle de la régression logistique. Notons que ces méthodes sont présentées dans l'annexe B.

### 5.5.1 Arbre de classification

La méthode de classification à base d'arbre est exposée dans l'annexe B et les techniques de validation (d'une classification) sont décrites dans l'annexe C. Dans ce paragraphe, nous utilisons la méthode Hold-Out avec 65% des données pour l'apprentissage et les 35% des données restantes pour le test.

#### 5.5.1.1 Méthode Classification And Regression Tree (CART)

Nous appliquons la méthode de l'arbre de classification (CART) [19]. Dans un premier temps, nous construisons un modèle (arbre) sur les données d'apprentissage. L'arbre de classification obtenu sur l'ensemble d'apprentissage est présenté dans la figure 5.4.

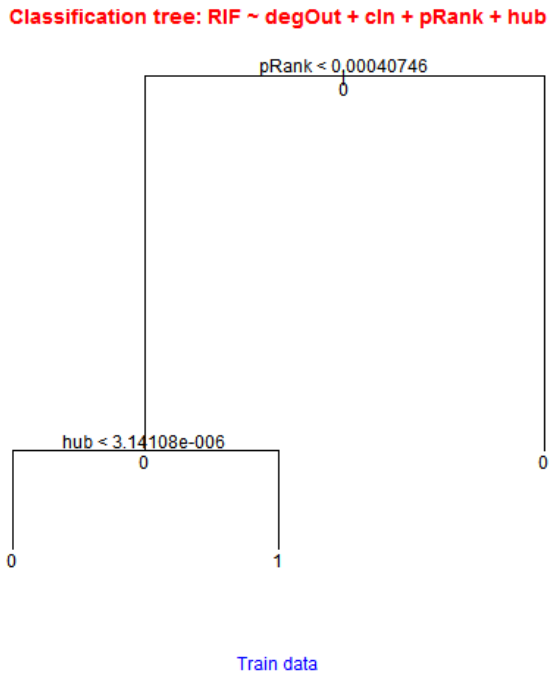



FIGURE 5.4 – Arbre de Classification sur l'ensemble d'apprentissage.

Nous remarquons d'après la figure 5.4 que l'arbre de classification n'utilise que 2 variables sur 4 pour discriminer les classes du RIF. En effet, cette méthode utilise en premier lieu de Page Rank (pRank), pour isoler un ensemble de sommets tous de $RIF = 0$ puis ensuite le degré de hubité (hub) qui permet de touver un ensemble de sommets tous de $RIF = 1$ et le reste de $RIF = 0$. Cela permet d'avoir une règle de décision à base des deux variables pRank et hub.
Pour évaluer ce modèle sur les données d'apprentissage nous calculons les différents indicateurs de performances d'un classifieur présentés dans l'annexe D.
Les prédictions faites par ce modèle sur les données d'apprentissage permettent d'avoir les valeurs présentées dans la première ligne du tableau 5.2

| sample | specificity | sensitivity | precision | F-meas | AUC |
|---|---|---|---|---|---|
| train | 0.979 | 0.400 | 0.500 | 0.444 | 0.690 |
| test | 0.981 | 0.333 | 0.500 | 0.400 | 0.657 |

TABLE 5.2 – Indicateurs de performances de l'arbre de classification

Ce classifieur est faiblement performant de plus il est sensible à l'échantillonnage : les résultats dépendent du choix aléatoire de l'échantillon d'apprentissage.

Pour y remédier, nous utilisons les méthodes ensemblistes (Ensemble Learning methods), décrites dans l'annexe B3, qui sont des méta-algorithmes qui combinent plusieurs techniques d'apprentissage automatique dans un même modèle prédictif afin de réduire la variance (Bagging), les biais (Boosting) ou d'améliorer les prédictions (Stacking).

#### 5.5.1.2 Méthode Random Forest

Nous utilisons ici une méthode de classification ensembliste très connue sous le nom de forêts décisionnelles aléatoires (Random Forest) présentée dans l'annexe B.
Le tableau 5.3 présente les valeurs prises par chacun des indicateurs de performance pour ce classifieur.

| sample | specificity | sensitivity | precision | F-meas | AUC |
|---|---|---|---|---|---|
| train | 1 | 0.600 | 1 | 0.750 | 0.800 |
| test | 1 | 0.333 | 1 | 0.500 | 0.667 |

TABLE 5.3 – Indicateurs de performances de la méthode Random Forest

Nous remarquons une amélioration notable de tous les paramètres. En particulier, la précision passe de 0.5 à 1. Cependant, le rappel (sensitivity) reste faible à 1/3 pour l'échantillon de test. Nous proposons d'explorer la méthode Bagging afin de continuer à améliorer ces paramètres.

#### 5.5.1.3 Méthode Bagging

Nous utilisons dans ce paragraphe la méthode Bagging (Boosted Aggregation) présentée dans l'annexe B Bagging. Pour cela, nous proposons de réaliser 100 itérations.
Les indicateurs de performances obtenus sont présentés dans le tableau 5.4.

| sample | specificity | sensitivity | precision | F-meas | AUC |
|---|---|---|---|---|---|
| train | 1 | 0.800 | 1 | 0.889 | 0.900 |
| test | 1 | 0.33 | 1 | 0.500 | 0.667 |

TABLE 5.4 – Indicateurs de performances obtenus par un Bagging avec 100 itérations

Nous remarquons que le Bagging améliore les indicateurs de performance sur l'échantillon d'apprentissage. Cependant, il réalise exactement les mêmes performances que la méthode Random Forest sur l'échantillon de test et donc un compromis entre la précision et le rappel (F-Mesure) d'environ $F - meas = 50\%$ ce qui est un bon résultat étant donné le fort déséquilibre de classe et la très faible représentativité de la classe à prédire $RIF = 1$

dans le jeu de données utilisé.

#### 5.5.1.4 Méthode Gradient Boosted Machine

Comme décrit dans l'annexe B, cette méthode permet encore d'améliorer les indicateurs surtout sur l'échantillon de test. Le tableau 5.5 présentent ces indicateurs.

| sample | specificity | sensitivity | precision | F-meas | AUC |
|---|---|---|---|---|---|
| train | 1 | 0.600 | 1 | 0.750 | 0.800 |
| test | 1 | 0.667 | 1 | 0.800 | 0.833 |

TABLE 5.5 – Indicateurs de performances obtenus par le Gradient Boosted Machine

Cette méthode donne la meilleure aire sous la courbe de ROC (AUC) (cf. annexe D). La précision reste parfaite comme dans les deux précédentes méthodes, le rappel (sensitivity) est doublé et devient égale à 0.667 ce qui est intéressant vu le fort déséquilible de classes. Dans la section suivante, nous utilisons l'arbre de classification sur des échantillons stratifiés.

#### 5.5.1.5 Arbre de classification sur échantillons stratifiés

Rappelons que la classe $RIF = 1$ ne représente que 5% du jeu de données. Une des techniques utilisées, notamment lorsque l'objectif est de mettre en évidence un sous-groupe spécifique au sein de la population, est l'échantillonnage aléatoire stratifié. Cette technique est utile pour de telles applications car garantit la présence du sous-groupe clé dans l'échantillon. Elle consiste à prélever aléatoirement l'échantillon d'apprentissage sous la contrainte de proportionnalité de chaque classe. Il s'agit de l'échantillonnage aléatoire stratifié proportionnel. Nous appliquons cette technique pour prélever l'échantillon d'apprentissage (le reste constitue celui de test). Il s'agit de subdiviser les données initiales en 2 parties selon que le $RIF = 0$ ou $RIF = 1$, puis, prélever aléatoirement 65% de chaque partie pour constituer l'échantillon d'apprentissage. Les 35% restant constituent l'échantillon test.
Ainsi, les échantillons d'apprentissage et de test ont les mêmes proportions de chaque classe que le jeu de données étudié. Le tableau 5.6 présente les résultats de performances de cette méthode d'échantillonnage proportionnel.

| sample | specificity | sensitivity | precision | F-meas | AUC |
|---|---|---|---|---|---|
| **Echantillonnage aléatoire** | | | | | |
| train-imb | 0.979 | 0.400 | 0.500 | 0.444 | 0.690 |
| test-imb | 0.981 | 0.333 | 0.500 | 0.400 | 0.657 |
| **Echantillonnage aléatoire proportionnel** | | | | | |
| train-proportionnel | 0.969 | 0.800 | 0.571 | 0.667 | 0.884 |
| test-proportionnel | 0.981 | 0.667 | 0.667 | 0.667 | 0.824 |

TABLE 5.6 – Comparaison des performances de l'arbre de classication avec et sans stratification

D'après le tableau 5.6, tous les indicateurs de performance sont remarquablement meilleurs en utilisant l'échantillonnage stratifié. Par ailleurs, ces résultats sont comparables à ceux obtenus par la méthode Gradient Boosted Machine mais restent néanmoins moins bons.

### 5.5.2 Arbre de classification sans déséquilibre des classes

Dans ce paragraphe, nous proposons d'utiliser les méthodes de rééquilibrage de classes sur le jeu de données et comparer les résultats obtenus par l'arbre de classification sur ces échantillons et ceux présentés précédemment obtenus sur les données brutes.
Plusieurs méthodes ont été proposées dans la littérature, pour améliorer les performances d'un classifieur dans un problème de déséquilibre de classes, parmi lesquelles les méthodes suivantes :

- Le `sur-échantillonnage` : qui consiste à augmenter aléatoirement (avec remise) la classe minoritaire, cette méthode présente un risque de sur-apprentissage.
- Le `sous-échantillonnage` : qui consiste à réduire aléatoirement (sans remise) la classe majoritaire. Elle ne semble pas intéressante dans notre cas car on a très peu de données.
- Méthodes `hybrides` : combinant un sous-échantillonnage et la génération de données synthétiques.

L'une des plus populaires est la méthode Random Over-Sampling Examples `ROSE` [65]. La méthode ROSE génère des échantillons synthétiques équilibrés et permet ainsi de renforcer l'estimation ultérieure de tout classifieur binaire. ROSE (Random Over-Sampling Examples) est une technique basée sur le bootstrap qui facilite la tâche de la classification binaire en présence de classes rares. Cela gère les données continues et catégorielles en générant des exemples synthétiques à partir de l'estimation de densité conditionnelle des deux classes. Le tableau 5.7 présente les résultats et performances de ces différentes méthodes d'échantillonnage.

| sample | specificity | sensitivity | precision | F-meas | AUC |
|---|---|---|---|---|---|
| **Données brutes** | | | | | |
| train-imb | 0.979 | 0.400 | 0.500 | 0.444 | 0.690 |
| test-imb | 0.981 | 0.333 | 0.500 | 0.400 | 0.657 |
| **ROSE** | | | | | |
| train-rose | 0.635 | 1 | 0.125 | 0.222 | 0.818 |
| test-rose | 0.596 | 1 | 0.125 | 0.222 | 0.798 |
| **Over-sampling** | | | | | |
| train-over | 0.927 | 1 | 0.417 | 0.588 | 0.964 |
| test-over | 0.904 | 1 | 0.375 | 0.545 | 0.952 |
| **Under-sampling** | | | | | |
| train-under | 0 | 1 | 0.050 | 0.094 | 0.500 |
| test-under | 0 | 1 | 0.055 | 0.103 | 0.500 |
| **Both (unber+over)** | | | | | |
| train-both | 0.885 | 1 | 0.312 | 0.476 | 0.943 |
| test-both | 846 | 0.667 | 0.200 | 0.308 | 0.756 |

TABLE 5.7 – Indicateurs de performances obtenus par l'arbre de classification après rééquilibre de classes

D'après le tableau 5.7, nous remarquons que la méthode de sous-échantillonnage (under-sampling) donne de très mauvais résultats et un modèle équivalent à un modèle aléatoire (AUC=0.5) ce qui est cohérent avec la faible taille du jeu de données.
Les méthodes both et ROSE dégradent la précision, la F-mesure est faible, cependant le rappel (sensitivity) est remarquablement amélioré. Enfin, la méthode de sur et sous échantillonnage (both) ne permet pas d'améliorer les indicateurs à part le rappel (sensitivity).
La méthode de sur-échantillonnage (over-sampling) rend le rappel parfait mais dégrade la précision et donne une F-mesure meilleure que celle obtenue sur les données brutes, l'AUC est améliorée.
Ainsi, la méthode sur-échantillonnage (over-sampling) est la meilleure méthode de ré-équilibrage de classes et améliore presque tous les indicateurs de performances pour l'arbre de classification.
Notons que l'utilisation de la méthode Bagging n'améliore pas les performances de l'arbre de classification sur les échantillons après ré-équilibre de classes.

### 5.5.3 Régression logistique

Rappelons que l'objectif est de prédire l'importance d'un composant de point de vue sûreté (i.e la catégorie de $RIF = 1$) et cela en utilisant les variables sélectionnées à savoir le degré sortant (degOut), la proximité entrante (cIn), le page Rank (pRank) et le degré de hubité (hub). Puisque notre problème de classification est à deux classes, la régression logistique est une méthode convenable pour la classification binômiale. La fonction "glm" (Generalized Linear Model) est utilisée pour la régression logistique. Il s'agit d'une version généralisée de la régression linéaire qui permet de relier le modèle linéaire à la variable de réponse (variable à prédire) via une fonction de lien et garantit la dépendance de l'intensité de la variance de chaque prédicteur à la valeur prédite.
Dans la suite, nous construisons plusieurs modèles logistiques à une, deux, trois et quatres variables prédictives.
Il faut également définir un critère pour déterminer la qualité d'un modèle. L'un des critères les plus utilisés est le Critère d'Information d'Akaike (AIC) [4]. Ce critère a pour objectif de trouver un compromis entre la qualité de l'ajustement et la complexité du modèle en terme de nombre de variables, ce qui permet d'éviter le sur-apprentissage (sur-ajustement). Ainsi, plus l'AIC sera faible, meilleur sera le modèle.

#### 5.5.3.1 Régression logistique sur échantillons aléatoires

Nous reprenons les mêmes échantillons d'apprentissage et de test précedemment utilisés pour l'arbre de classification dans le paragraphe 5.5.1. les 65% des données prélevées aléatoirement serviront pour l'apprentissage et le reste (35%) sera utilisé pour le test.

Le tableau 5.8 présente les valeurs du critère d'AIC prises par chacun des modèles construits.

| **Logistic model** | |
|---|---|
| **RIF $\sim$** | **AIC** |
| degOut | 40.70 |
| cIn | 31.83 |
| pRank | 31.92 |
| hub | 43.80 |
| deg0ut+cIn | 30.50 |
| deg0ut+pRank | 33.36 |
| degOut+hub | 42.45 |
| cIn+pRank | 32.41 |
| cIn+hub | 31.69 |
| pRank+hub | 33.83 |
| degOut+cIn+pRank | 32.29 |
| **degOut+cIn+hub** | **30.00** |
| degOut+pRank+hub | 35.34 |
| cIn+pRank+hub | 32.77 |
| degOut+cIn+pRank+hub | 31.90 |

TABLE 5.8 – AIC de chaque modèle logistique sur l'échantillon d'apprentissage

Nous remarquons d'après le tableau 5.8 que le modèle minimisant le critère d'AIC est : `RIF` $\sim$ `degOut+cIn+hub` composé des trois variables prédictives le degré sortant (degOut), la proximité entrante (cIn) et le degré de hubité (hub). Dans la suite, nous étudions ce modèle et présentons ses performances.
Ce choix de modèle a été aussi confirmé par le bootstrap : Nous avons tiré 1000 fois aléatoirement 65% des données de l'échantillon d'apprentissage sur lesquelles nous avons calculé le critère d'AIC pour chacun des 15 modèles précédents, ensuite nous avons choisi le meilleur modèle de chaque itération. Les résultats obtenus sont présentés dans le figure 5.5.

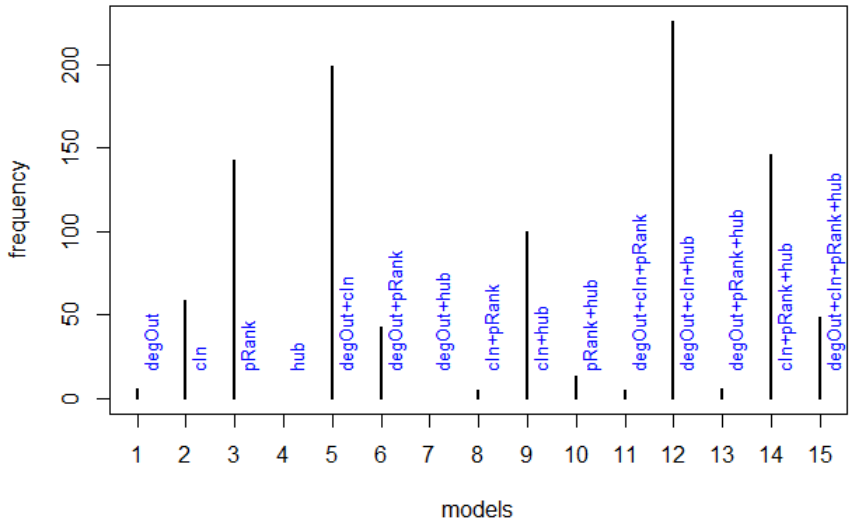



FIGURE 5.5 – Choix du meilleur modèle logistique par 1000 itérations de Boostrap

Après le choix du meilleur modèle logistique : `RIF` $\sim$ `degOut+cIn+hub`, nous évalons ses performances.

La F-mesure ou F1-score est un indicateur de performance d'un classifieur couramment utilisé car elle représente un compromis entre la précision et le rappel. Souvent elle correspond à la moyenne harmonique (non pondérée). Ici nous souhaitons accorder plus d'importance au rappel (sensitivity), car cela permet de détecter le plus grand nombre de composants critiques de point de vue de la sûreté nucléaire, quitte à avoir quelques fausses alertes. Pour cela, nous utilisons la formule générale qui est le $F_\beta$ définie par :

$$F_\beta = (1 + \beta^2)\frac{precision.sensitivity}{\beta^2.precision + sensitivity}$$

La figure 5.6 montre l'influence du choix de $\beta$.

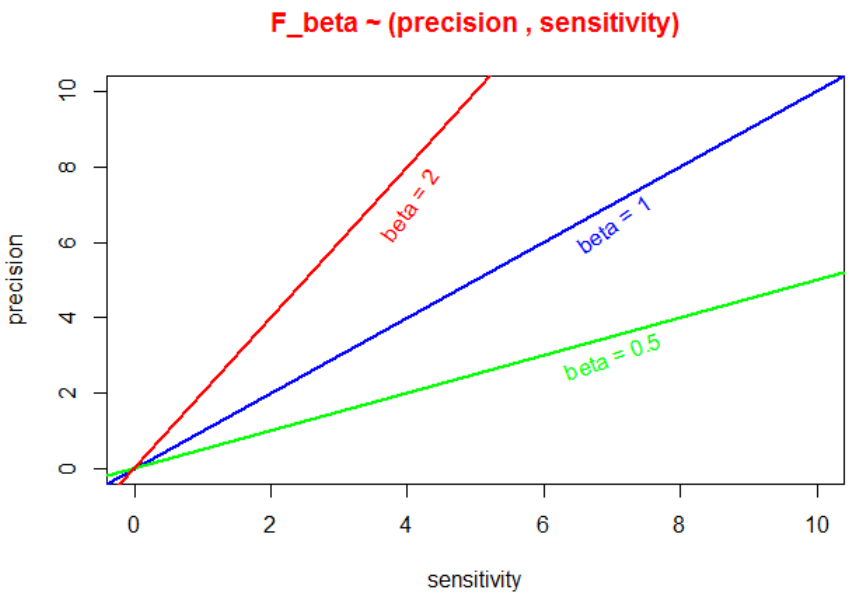



FIGURE 5.6 – Influence du choix de $\beta$ sur la moyenne harmonique $F_\beta$

— $\beta > 1$ donne plus de poids au rappel (sensitivity).

— $\beta < 1$ donne plus de poids à la précision (precision).

— $\beta = 1$ donne les mêmes poids à la précision et au rappel (cas particulier de la F-Mesure).

Nous proposons de fixer $\beta = 2$ ce qui accorde un poids de 80% au rappel (sensitivity), et de 20% à la précision.

Dans un premier temps, nous cherchons un bon seuil de prédiction pour calibrer le modèle logistique sélectionné sur les données d'apprentissage. Pour se faire, nous souhaitons que ce seuil donne de bons résultats pour les différents indicateurs de performances à savoir rappel (sensitivity), précision (precision), $F_\beta$ et spécificité (specificity).

Les courbe suivantes (figure 5.7) présentent l'évolution de ces différents indicateurs en fonction du seuil de prédiction.

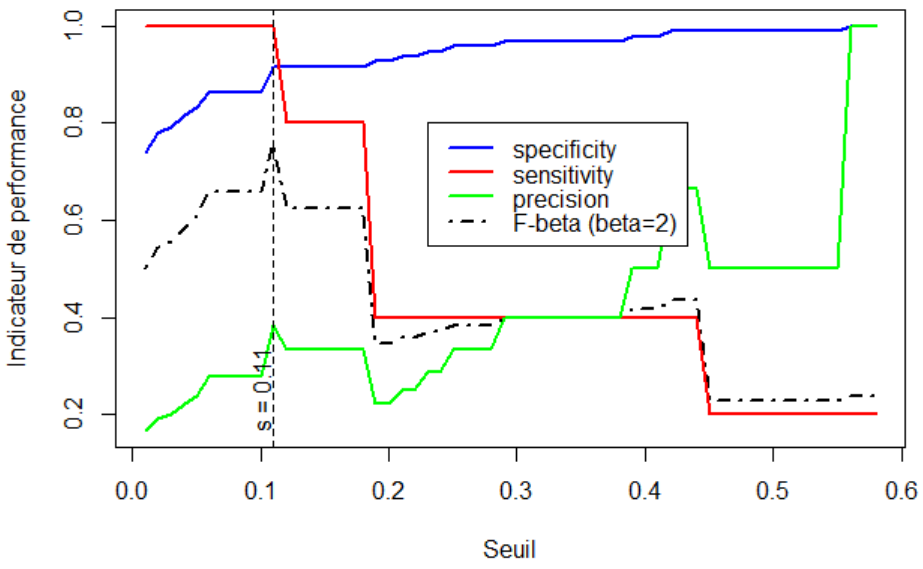



FIGURE 5.7 – Choix du seuil de prédiction pour le modèle `RIF ~ degOut+cIn+hub`

Le meilleur seuil compromis entre les différents indicateurs de performance est $s = 0.11$. Ce seuil maximise le rappel (sensitivity) et la $F_\beta$.
Le tableau 5.9 résume les indicateurs obtenus par ce modèle sur les échantillons d'apprentissage et de test où le seuil est fixé à la valeur optimale $s = 0.11$.

| `RIF ~ degOut+cIn+hub` | | | | | |
|---|---|---|---|---|---|
| **Sample** | **specificity** | **sensitivity** | **precision** | $F_\beta$ | **AUC** |
| train | 0.92 | 1 | 0.38 | 0.76 | 0.94 |
| test | 0.88 | 1 | 0.33 | 0.71 | 0.94 |

TABLE 5.9 – Indicateurs de performance pour le modèle logistique `RIF ~ degOut+cIn+hub`

Les indicateurs de performances ont légèrement dimunué sur l'échantillon de test. Cependant ils restent raisonablement acceptables étant donné le fort déséquilibre de classes.
Dans le paragraphe suivant, nous explorons l'apport de la méthode d'échantillonnage stratifié telle que décrite dans le paragraphe 5.5.1.5 pour l'amélioration des performances de la régression logistique.

#### 5.5.3.2 Régression logistique sur échantillons stratifiés

Nous reprenons les échantillons utilisés pour l'apprentissage et le test dans le paragraphe Arbre de classification sur échantillons stratifiés. Rappelons que l'échantillon d'apprentissage est prélevé aléatoirement à partir des données initiales sous la contrainte de contenir 65% des observations appartenant à la classe $RIF = 0$ et 65% appartenant à la classe $RIF = 1$. Ainsi, l'échantillon de test comprend les 35% restant de chaque classe.
Le tableau 5.10 présente les valeurs prises par le critère AIC pour chaque modèle sur l'échantillon d'apprentissage stratifié.

| **Logistic model** | |
|---|---|
| **RIF $\sim$** | **AIC** |
| degOut | 42.09 |
| cIn | 33.23 |
| pRank | 35.45 |
| hub | 41.07 |
| deg0ut+cIn | 33.83 |
| deg0ut+pRank | 37.08 |
| degOut+hub | 42.55 |
| cIn+pRank | 35.08 |
| cIn+hub | 23.16 |
| pRank+hub | 34.93 |
| degOut+cIn+pRank | 35.82 |
| degOut+cIn+hub | 23.62 |
| degOut+pRank+hub | 36.92 |
| **cIn+pRank+hub** | **16.81** |
| degOut+cIn+pRank+hub | 18.72 |

TABLE 5.10 – Valeurs de AIC pour chacun des modèles logistiques sur l'échantillon stratifié d'apprentissage

Nous remarquons d'après le tableau 5.10 que le modèle minimisant le critère d'AIC est `RIF` $\sim$ `cIn+pRank+hub` composé de 3 variables prédictives la proximité entrante (cIn) le Page Rank (pRank) et le degré de hubité (hub). Ce modèle est différent de celui retenu sans stratification.
Dans la suite, nous étudions ce modèle (`RIF` $\sim$ `cIn+pRank+hub`) et présentons ses performances.
Dans un premier temps, comme dans le paragraphe précédent, nous cherchons un bon seuil de prédiction qui soit un compromis entre le rappel (sensitivity) et la $F_\beta$.

La figure 5.8 présente l'évolution de ces différents indicateurs en fonction du seuil de prédiction.

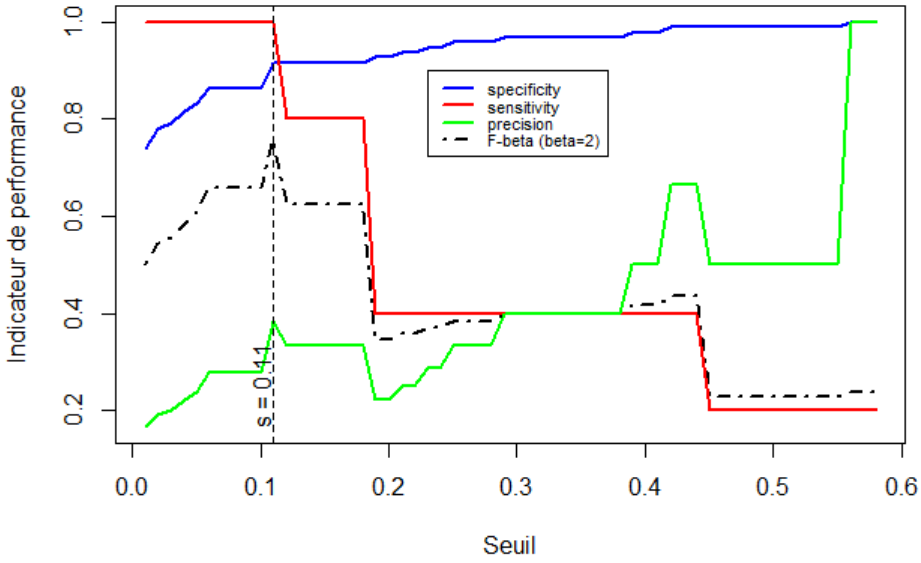



FIGURE 5.8 – Choix du seuil de prédiction pour le modèle `RIF ~ cIn+pRank+hub`

Le meilleur seuil compromis entre ces différents indicateurs de performance est $s = 0.13$. Ce seuil maximise le rappel (sensitivity), la précision (precision) et la $F_\beta$.
Le tableau 5.11 résume les indicateurs obtenus par ce modèle sur les échantillons d'apprentissage et de test au seuil $s = 0.13$.

| `RIF ~ cIn+pRank+hub` | | | | | |
|---|---|---|---|---|---|
| **Sample** | **specificity** | **sensitivity** | **precision** | $F_\beta$ | **AUC** |
| train | 0.99 | 0.80 | 0.80 | 0.80 | 0.89 |
| test | 0.98 | 0.33 | 0.50 | 0.36 | 0.66 |

TABLE 5.11 – Indicateurs de performance pour le modèle logistique `RIF ~ cIn+pRank+hub` avec stratification

Les indicateurs de performances ont régressé sur l'échantillon de test.
Nous proposons d'étudier le deuxième meilleur modèle retenu par le critère d'AIC.
Ce modèle `RIF ~ degOut+cIn+pRank+hub` est composé des 4 variables prédictives. Le tableau 5.12 résume les indicateurs obtenus par ce modèle sur les échantillons d'apprentissage et de test.

| RIF ∼ degOut+cIn+pRank+hub | | | | | |
|---|---|---|---|---|---|
| **Sample** | **specificity** | **sensitivity** | **precision** | $F_\beta$ | **AUC** |
| train | 0.99 | 0.80 | 0.80 | 0.80 | 0.89 |
| test | 0.98 | 0.33 | 0.50 | 0.36 | 0.66 |

TABLE 5.12 – Indicateurs de performance pour le modèle logistique `RIF` ∼ `degOut+cIn+pRank+hub` avec stratification

Nous remarquons que les performances sont exactement les mêmes que le premier modèle. D'autres part, contrairement à ce qui est attendu, l'échantillonnage stratifié n'apporte aucune amélioration aux performances de la régression logistique sur ce jeu de données fortement déséquilibré.

## 5.6 Conclusion

Les centralités sont assez performantes pour la prédiction du RIF. En effet, malgré le fort déséquilibre de classes les résultats de prédictions sont suffisamment performants. Cependant, le choix du modèle dépend de son utilisation. On recommandera celui avec une précision maximale si l'utilisation implique un investissement financier, par exemple dans le cadre d'une modification d'une installation nucléaire. On recommandera plutôt celui qui maximise le rappel (sensitivity) si l'objectif est d'être exhaustif par rapport aux composants qui sont critique ($RIF = 1$), par exemple fournir à l'analyste une checklist des composants potentiellement critiques pour vérification plus fine.

## Deuxième partie

# Similarité entre réseaux : Validation sur des graphes théoriques et applications aux réseaux des systèmes de sûreté nucléaire

Plusieurs métriques ont été introduites pour mesurer la l'importance d'un sommet ou lien dans un réseaux. Par exemple la centralité du degré (locale), la centralité (globale) de proximité (closeness) [70], la centralité d'intermédiarité (betweeness) [15], et la centralité de Bridgeness [54] caractérisent cette importance.
La similarité entre sommets ou liens d'un réseau est définie en terme de voisinage : deux sommets sont « proches » s'il y a un fort recouvrement entre leurs voisinages. Il existe des mesures de similarité structurelles locales entre les sommets à l'instar de l'indice de Jaccard [52], la similarité Cosinus ou l'indice de Sorensen-Dice [78, 28] qui sont basés sur le modèle de connectivité des sommets dans leur voisinage immédiat [23].
En outre, il existe des mesures plus sophistiquées comme PageRank qui classe les sommets d'un réseau à l'aide de modèles de chaîne de markov [20], ou SimRank qui calcule la similarité entre deux sommets sur des graphes dirigés en utilisant une définition de similarité récursive [53].
Les mesures de similarité structurelles, entre sommets ou liens, mentionnées ci-dessus et d'autres similarités ont été utilisées efficacement dans la détection de communautés dans des graphes [23, 22, 24]. Cependant, ces similarités présentent un inconvénient majeur car qu'ils se limitent au voisinage immédiat des sommets du réseau. Cette limitation conduit à un manque de discernement vis-à-vis de propriétés structurelles importantes qui peuvent améliorer la qualité de la similarité structurelle et donc de ses applications.
Bien que certains réseaux peuvent avoir des valeurs proches des indicateurs usuels (ordre, taille, densité, coefficient de clustering, diamètre,...,etc), ils peuvent avoir d'autres propriétés qui les rendent différents.
Notre objectif est de concevoir une mesure de similarité entre réseaux qui utilise les vecteurs de centralités (en totalité) pour avoir un meilleur outil de comparaison globale.

Nous validons dans un premier temps l'approche proposée sur des graphes théoriques classiques décrit dans le paragraphe 3.2 du chapitre 3. Ensuite, nous appliquons cette méthode aux réseaux réels des systèmes de sûreté de l'EPR 2 dont les caractéristiques topologiques sont présentées dans le paragraphe 4.3 du chapitre 4. Ainsi cette partie se structure comme suit :
Le chapitre 6 exploite des outils statistiques classiques pour deux finalités :

1. effectuer des tests d'homogénéïté qui permettent des comparaisons des distributions d'une même centralité (vecteurs entiers) dans deux ou plusieurs réseaux afin d'avoir une vue globale sur le comportement de celle-ci dans différents réseaux.
2. réaliser des statistiques d'indépendance entre deux ou plusieurs centralités dans un même réseau dans le but d'avoir une idée précise sur les relations potentielles qui lient et coordonnent ces centralités entre elles.

Le chapitre 7 définit un coefficient synthétique de centralité (vectoriel) susceptible de réunir et de résumer l'information concerant les variations communes des distributions des différentes centralités dans un même réseau. Ensuite, une mesure de similarité est introduite

afin de pouvoir comparer et évaluer la proximité entre les structures topologiques globales de différents réseaux. Enfin ; nous validons dans un premier temps l'approche proposée sur des graphes théoriques classiques décrits dans la partie 1. Et on cloture cette partie, en appliquant cette méthode aux réseaux réels des systèmes de sauvegarde de l'EPR2 dont les caractéristiques topologiques ont été présentées dans la partie 1.

# Chapitre 6

# Approche statistique pour la comparaison monovariable de graphes

La première section de ce chapitre est consacrée à la comparaison deux réseaux $P_1$ et $P_2$ par rapport à une même mesure de centralité $X$. Nous commençons par présenter les fondements théoriques du coefficient d'écart entre deux réseaux. Ensuite nous présentons le test d'homogénéité de Mann-Whitney qui permet de comparer la position de deux populations, puis nous proposons un exemple de calcul et une application aux graphes classiques théoriques de types Erdös-Rényi, Small-World et Barabasi-Albert. Enfin, nous utilisons cette méthode pour comparer à titre d'exemples deux des réseaux réels de $P_{sys-EPR2}$ présentés dans le paragraphe 4.3 à savoir $P_1 =$ "$AAD\_APA\_ARE$" et $P_2 =$ "$ASG$" : à titre d'exemples dans un premier temps par rapport à la centralité du degré entrant (degIn) puis par rapport à la centralité de proximité sortante (cOut).

La seconde section compare deux vecteurs de centralités $X$ et $Y$ dans un même réseau $P$. Nous commençons par présenter une mesure de l'intensité de dépendance entre deux vecteurs de centralités $X$ et $Y$ dans un même réseau. Ensuite, les tests de Spearman et de Kendall sont utilisés pour étudier cette dépendance. Ils seront appliqués dans un premier temps aux graphes classiques, et enfin à un exemple du réseau réel $P_1 =$ "$AAD\_APA\_ARE$".

## 6.1 Étude d'une centralité dans deux ou plusieurs réseaux

Rappelons que l'objectif ici est de comparer les distributions d'une centralité sur deux ou plusieurs réseaux différents.

### 6.1.1 Notations

— Centralité : $X$

— Réseaux : $P_k$ $(k = 1, 2)$

— Distribution de $X$ sur le réseau $P_k$ $(k = 1, 2)$
$P^{(k)}(X <= x) = F^{(k)}(x)$ : fonction de distribution en $x$ de $X$ dans $P_k$ $(k = 1, 2)$.

— Observations de $X$ dans $P_k$ $(k = 1, 2)$ : $(x_1^{(k)}, x_2^{(k)}, ..., x_{n^{(k)}}^{(k)})$
où $n^{(k)}$ est le nombre de sommets du réseau $P_k$ : $(k = 1, 2)$.

### 6.1.2 Homogénéité

L'homogénéité consiste à comparer les distributions d'une centralité $X$ dans deux réseaux $P_1$ et $P_2$, du point de vue de leurs positions. On évalue la différence de position entre ces distributions de $X$ dans les deux réseaux au moyen d'une mesure $d$ appelée coefficient d'écart entre les deux réseaux et qui s'écrit sous la forme suivante :

$$d = 2P(X^{(1)} < X^{(2)}) - 1,\ (-1 \leq d \leq 1),$$

où

- $X^{(1)}$ et $X^{(2)}$ sont deux vecteurs indépendants définis comme suit : $X^{(k)}$ admet la même distribution que la centralité $X$ dans le réseau $P_k$ et de ce fait la même fonction de répartition $F^{(k)}(x)$, $(k = 1, 2)$.
- $P(X^{(1)} < X^{(2)})$, représente la probabilité qu'un sommet prélevé au hasard dans le réseau $P_1$ ait une valeur de la centralité $X$ inférieure à celle d'un autre sommet prélevé au hasard dans le réseau $P_2$.

**Remarques**

— Si les distributions sont identiques, alors
$P(X^{(1)} < X^{(2)}) = P(X^{(2)} < X^{(1)}) = 1/2 \iff d = 0$

— $P(X^{(1)} < X^{(2)}) = 1 \iff d = 1$ signifie que presque toutes les valeurs prises par la centralité $X$ dans $P_1$ sont inférieures à presque toutes celles prises dans $P_2$.

— $P(X^{(1)} < X^{(2)}) = 0 \iff d = -1$ signifie que presque toutes les valeurs prises par la centralité $X$ dans $P_1$ sont supérieures à presque toutes celles prises dans $P_2$.

### 6.1.3 Estimateur du coefficient d'écart $d$

**Définition de l'estimateur de $d$**

$$\bar{d} = 2 * \frac{U}{n^{(1)} \times n^{(2)}} - 1 \tag{6.1}$$

où $U$ est la statistique de Mann-Whitney définie par l'une des deux façons suivantes :

$$U = \sum_{i=1}^{n^{(2)}} U_i \tag{6.2}$$

où $U_i$ est le nombre de sommets du réseau $P_1$ dont les valeurs de la centralité $X$ sont inférieures à celle du $i^{eme}$ sommet de $P_2$ $(i = 1, ..., n^{(2)})$.

$U$ peut aussi être calculé d'une autre manière comme suit :

$$U = S_2 - \frac{n^{(2)}(n^{(2)}+1)}{2} = \sum_{i=1}^{n^{(2)}} R^*(x_i^{(2)}) - \frac{n^{(2)}(n^{(2)}+1)}{2} \tag{6.3}$$

où $R^*(x_i^{(2)})$ est le rang de la centralité $x_i^{(2)}$ du sommet $i$ du réseau $P_2$ dans l'échantillon global (constitué de toutes les valeurs de cette centralité dans les deux réseaux) ordonné.

**Exemple**

Soit $x^{(1)}$ et $x^{(2)}$ les valeurs observées de la centralité $X$ sur les réseaux $P_1$ et $P_2$.

$x^{(1)} = (19, 17, 18, 14)$
$x^{(2)} = (12, 16, 13, 15)$
$n^{(1)} = n^{(2)} = 4$

La figure (6.1) représente les distributions de $X^{(1)}$ et $X^{(2)}$.

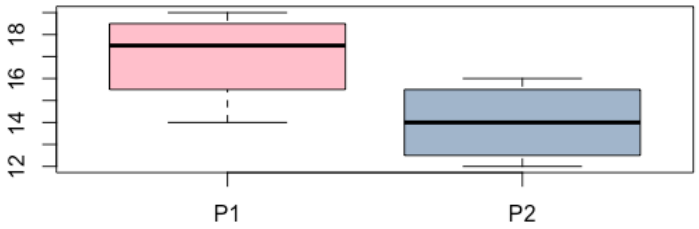



FIGURE 6.1 – Distributions des valeurs de $X$ dans $P_1$ et $P_2$

On constate qu'il y'a une différence de position de la centralité $X$ entre les réseaux $P_1$ et $P_2$. En utilisant la formule (6.2) :

$$U = 0 + 1 + 0 + 1 = 2.$$

2ème méthode de calcul de $U$ : On commence par calculer les rangs dans le tableau suivant :

| **réseau** | **x** | **rangs** |
|---|---|---|
| P1 | 19 | 8 |
| P1 | 17 | 6 |
| P1 | 18 | 7 |
| P1 | 14 | 3 |
| P2 | 12 | 1 |
| P2 | 16 | 5 |
| P2 | 13 | 2 |
| P2 | 15 | 4 |

TABLE 6.1 – Calcul des rangs dans l'échantillon global

Calcul de la statistique de Mann-Whitney utilisant la formule (6.3) :

$$U = (1 + 5 + 2 + 4) - (4 * 5)/2 = 2.$$

L'estimation de la distance de Mann-Whitney $\bar{d}$ par la formule 6.1 donne $d = 2 * \frac{2}{4*4} - 1 = -0.75$.

**Interprétation**

$\bar{d} = -0.75$ (proche de $-1$) signifie que les distributions sont décalées à gauche i.e. presque toutes les valeurs de $X$ sur $P_1$ sont supérieures à presque toutes celles sur $P_2$ ce qui est confirmé grâce à la figure (6.1).

**Remarque**

Dans le cas où il y'a des exe-aequos, on attribue à ces coïncidences le rang moyen et on utilise la formule (6.3).

### 6.1.4 Test de Mann-Whitney

Ce test a pour objectif de comparer les distributions d'une variable quantitative (centralité) dans deux populations différentes (réseaux dans ce cas).
Il s'agit d'un test non-paramétrique basé sur les rangs des valeurs de la centralité dans les deux réseaux et non sur les valeurs $x$ comme les tests $z$ et $t$ de Student.

#### 6.1.4.1 Hypothèses

Les hypothèses à tester sont les suivantes :

$$\begin{cases} H_0 : & F^{(1)}(x) = F^{(2)}(x) \text{ (distributions homogènes)} ; \\ H_1 : & d \neq 0; (P(X^{(1)} < X^{(2)}) \neq 1/2) \text{ (distributions décalées)} \end{cases} \tag{6.4}$$

Ou bien

$$\begin{cases} H_0^{'} : & F^{(1)}(x) = F^{(2)}(x) \text{ (distributions homogènes)} ; \\ H_1^{'} : & d < 0; (P(X^{(1)} < X^{(2)}) < 1/2) \end{cases} \tag{6.5}$$

Dans $H_1^{'}$ les distributions sont décalées à gauche i.e : presque tous les sommets du réseau $P_1$ ont des valeurs de la centralité $X$ supérieures à celles des sommets du réseau $P_2$.

Ou bien

$$\begin{cases} H_0^{''} : & F^{(1)}(x) = F^{(2)}(x) \text{ (distributions homogènes)} ; \\ H_1^{''} : & d > 0; (P(X^{(1)} < X^{(2)}) > 1/2) \end{cases} \tag{6.6}$$

Dans ce cas ($H_1^{''}$), les distributions sont décalées à droite c'est à dire que presque tous les sommets du réseau $P_1$ ont des valeurs de la centralité $X$ inférieures à celles des sommets du réseau $P_2$.

**Interprétation**

Si les deux distributions sont identiques (homogénéité), alors $d = 0$. Intuitivement, on rejette :

— $H_0$ ou $H_0^{'}$ si $U$ est trop petit par rapport à $n^{(1)}n^{(2)}/2$.

— $H_0$ ou $H_0^{''}$ si $U$ est trop grand par rapport à $n^{(1)}n^{(2)}/2$.

#### 6.1.4.2 Distribution de la statistique de Mann-Whitney $U$ sous $H_0$

— Distribution exacte : voir Conover [25] pages 231-236. On peut aussi calculer une approximation de la p-value par simulations.

— Distribution asymptotique :
Normale de moyenne $n^{(1)}n^{(2)}/2$ et de variance $\frac{n^{(1)}n^{(2)}(n^{(1)}+n^{(2)}+1)}{12}$.
Cette approximation est valable si $n^{(1)}$ ou $n^{(2)} > 20$.

**Remarque**

Pour que ce test soit applicable il faut que le nombre de sommets de chacun des deux réseaux soit supérieur à 10 et qu'il n'y ait pas trop de valeurs exe-aequos, par contre la normalité des distributions n'est pas exigée même pour les réseaux de petites tailles.

### 6.1.5 Homogénéité des réseaux artificiels classiques

Cette partie s'intéresse à étudier l'homogénéité d'une centralité dans deux réseaux artificiels du même type avec deux valeurs différentes de $p$ pour Erdös-Rényi et Small-World et *power* pour Barabasi-Albert.
Dans un premier temps, cela est illustré par la méthode descriptive en présentant rangs des valeurs des deux populations. Ensuite, le test d'homogénéité de Mann-Whitney est effectué afin de confirmer ou d'infirmer statistiquement cette homogénéité observée.
Nous présentons uniquement les cas des centralités degré entrant (degIn), intermédiarité (betw) et hubité (hub). Pour les réseaux Small-World, qui sont non-dirigés, le degré (deg) remplacera le degré entrant (degIn).

#### 6.1.5.1 Homogénéité du degré entrant (degIn)

Pour étudier l'homogénéité du degré entrant dans les graphes artificiels, nous comparons les rangs ses valeurs dans deux populations différentes.
Dans la figure de gauche, les deux populations sont deux graphes de type Erdös-Rényi, le premier avec $p1 = 0.06$ et le deuxième avec $p2 = 0.10$.
Dans la figure centrale les deux populations sont deux réseaux de type Small-World avec p1 et p2 égales à celles du premier cas. La variable étudiée dans ce cas est le degré car les réseaux Small-World sont non-dirigés.
La figure de droite représente cette fois-ci deux populations correspondantes à des réseaux Barabasi-Albert avec des puissances d'attachement préférentiel respectives de $p1 = 1$ et $p2 = 2$.

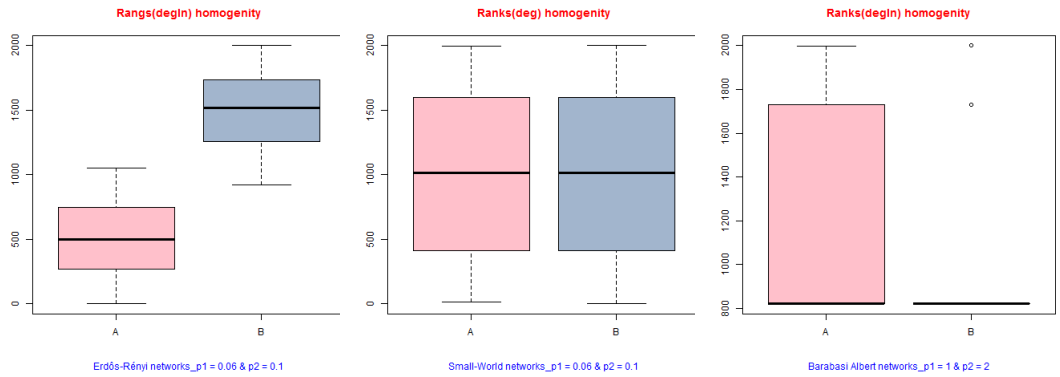



FIGURE 6.2 – Comparaison des rangs du degré entrant entre paires de type Erdös-Rényi, Small-World et Barabasi-Albert

D'après la figure 6.2, nous remarquons les points suivants.
Les distributions du degré entrant sur les deux réseaux de type Erdös-Rényi sont très décalées. En effet, presque tous les sommets du second réseau ($p2 = 0.10$) ont des degrés entrants supérieurs à presque tous ceux des sommets du premier réseau ($p1 = 0.06$). Ainsi on retrouve la propriété caractéristique des réseaux de type Erdös-Rényi à savoir le paramètre $p$ représente la probabilité d'existence d'un lien entre deux sommets quelconques du réseau.
Le degré entrant est réparti de la même manière dans les réseaux Small-World $p1 = 0.06$ et $p2 = 0.10$, contrairement aux deux autres cas.
Pour les réseaux de type Barabasi-Albert, nous remarquons que le degré entrant est dénéneré pour $power = 2$ . En fait, à partir de $power = 2$, le degré entrant est dégénéré pour Barabasi-Albert.
Nous effectuons à présent le test de Mann-Whitney pour étudier statistiquement l'homogénéité du degré entrant. Pour se faire, nous l'appliquons pour chaque paire de réseaux de même type avec des paramètres $p$ différents.
Les tableaux 6.2, 6.3 et 6.4 présentent les résultats de ces tests.
Avec la convention :

— Si 1, il y a homogénéité.

— Si 0, il n'y a homogénéité.

| ER degIn Homogenity | ER_0.02 | ER_0.06 | ER_0.1 |
|---|---|---|---|
| ER_0.02 | | 0 | 0 |
| ER_0.06 | | | 0 |
| ER_0.1 | | | |

TABLE 6.2 – Test d'homogénéité de Mann-Whitney du degré entrant sur Erdös-Rényi

Le test de Mann-Whitney confirme la non-homogénéité du degré entrant pour tous les réseaux d'Erdös-Rényi considérés.

| SW deg Homogenity | SW_0.02 | SW_0.06 | SW_0.1 |
|---|---|---|---|
| SW_0.02 | | 1 | 1 |
| SW_0.06 | | | 1 |
| SW_0.1 | | | |

TABLE 6.3 – Test d'homogénéité de Mann Whitney pour le degré sur Small-World

Le test de Mann-Whitney confirme l'homogénéité du degré entrant pour tous les réseaux Small-World considérés.

| BA degIn Homogenity | BA_1 | BA_1.5 | BA_2 | BA_2.5 | BA_3 |
|---|---|---|---|---|---|
| BA_1 | | 0 | 0 | 0 | 0 |
| BA_1.5 | | | 0 | 0 | 0 |
| BA_2 | | | | 1 | 1 |
| BA_2.5 | | | | | 1 |
| BA_3 | | | | | |

TABLE 6.4 – Test d'homogénéité de Mann Whitney pour le degré entrant sur Barabasi-Albert

La notion d'homogénéité pour le degré entrant sur type de réseau, n'apparaît qu'à partir d'une puissance d'attachement préférentiel $power = 2$. Cela est dû au fait que cette variable devient dégénérée comme illustré dans la figure de droite de 6.2.

#### 6.1.5.2 Homogénéité de l'intermédiarité (betw)

Nous reprenons les mêmes paires de réseaux Erdös-Rényi, Small-World et Barabasi-Albert utilisées pour l'étude de l'homogénéité du degré entrant.

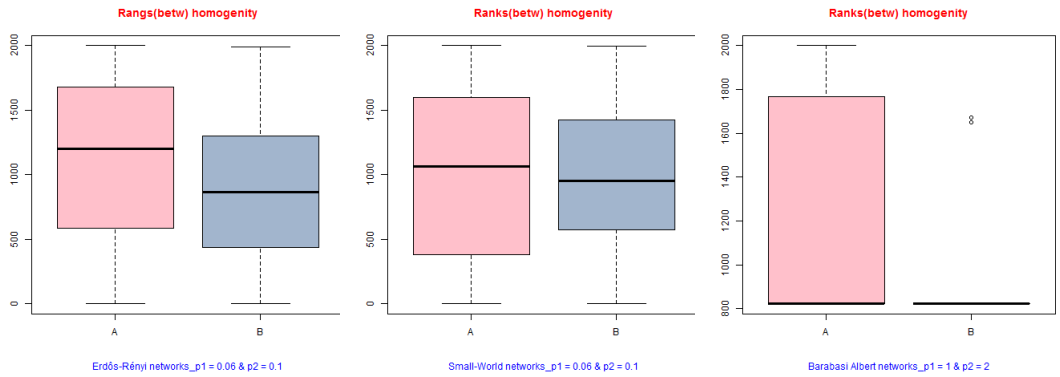



FIGURE 6.3 – Comparaison des rangs de l'intermédiarité entre paires de type Erdös-Rényi, Small-World et Barabasi-Albert

D'après la figure 6.3, nous notons que :
Les distributions de l'intermédiarité sur les deux réseaux de type Erdös-Rényi sont décalées, cependant, contrairement au degré entrant, plusieurs sommets du second réseau ($p2 = 0.10$) ont des intermédiarités inférieures à celles des sommets du premier réseau ($p1 = 0.06$). Cependant, on ne peut ni confirmer ni infirmer l'homogénéité dans ce cas.
La boîte à moustache de l'intermédiarité du second réseau Small-World ($p2 = 0.10$) est complétement incluse dans celle du premier ($p2 = 0.06$).
L'intermédiarité est dégénérée pour Barabasi-Albert à partir de $power = 2$.
A présent, nous réalisons le test de Mann-Whitney pour étudier statistiquement l'homogénéité de l'intermédiarité pour chaque paire de réseaux de même type avec des paramètres $p$ différents.

Les tableaux 6.5, 6.6 et 6.7 présentent les résultats de ces tests.

| ER betw Homogenity | ER_0.02 | ER_0.06 | ER_0.1 |
|---|---|---|---|
| ER_0.02 | | 0 | 0 |
| ER_0.06 | | | 0 |
| ER_0.1 | | | |

TABLE 6.5 – Test d'homogénéité de Mann Whitney de l'intermédiarité sur Erdös-Rényi

Comme précédemment, le fait que la probabilité d'existence d'un lien dépend du paramètre p d'un réseau d'Erdös-Rényi en plus en caractère aléatoire de ces réseaux rend l'intermédiarité non-homogène entre réseaux qui n'ont pas le même $p$.

| SW betw Homogenity | SW_0.02 | SW_0.06 | SW_0.1 |
|---|---|---|---|
| SW_0.02 | | 1 | 1 |
| SW_0.06 | | | 1 |
| SW_0.1 | | | |

TABLE 6.6 – Test d'homogénéité de Mann Whitney de l'intermédiarité sur Small-World

Les réseaux Small-World ont une intermédiarité homogène pour tout $p$ comme pour le degré entrant.

| BA betw Homogenity | BA_1 | BA_1.5 | BA_2 | BA_2.5 | BA_3 |
|---|---|---|---|---|---|
| BA_1 | | 0 | 0 | 0 | 0 |
| BA_1.5 | | | 0 | 0 | 0 |
| BA_2 | | | | 1 | 1 |
| BA_2.5 | | | | | 1 |
| BA_3 | | | | | |

TABLE 6.7 – Test d'homogénéité de Mann Whitney de l'intermédiarité sur Barabasi-Albert

Le test d'homogénéité pour l'intermédiarité sur les réseaux Barabasi-Albert sont identiques à ceux obtenus pour le degré entrant.

#### 6.1.5.3 Homogénéité de la hubité (hub)

Les mêmes paires de réseaux Erdös-Rényi, Small-World et Barabasi-Albert précédemment considérées pour l'étude de l'homogénéité du degré entrant et de l'intermédiarité sont également utilisées pour la hubité.

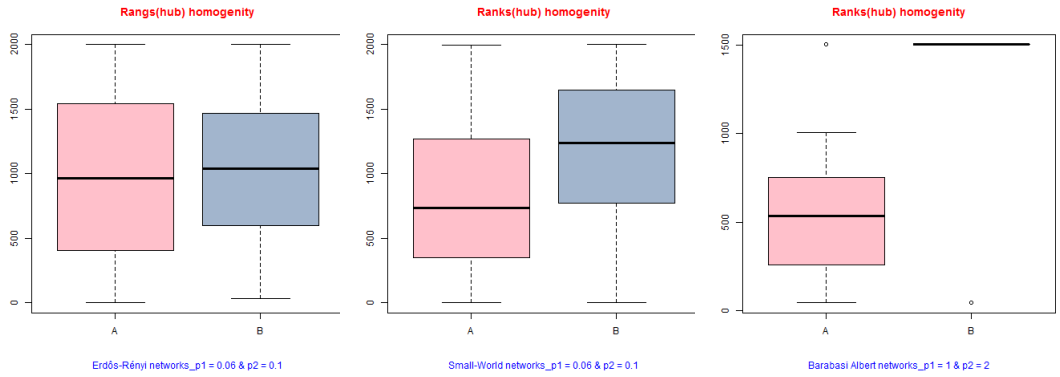



FIGURE 6.4 – Comparaison des rangs de la hubité entre paires de type Erdös-Rényi, Small-World et Barabasi-Albert

D'après la figure 6.4, nous remarquons que :
La boîte à moustache de la hubité du second réseau Edrös-Rényi ($p2 = 0.10$) est complétement incluse dans celle du premier ($p2 = 0.06$), contrairement aux cas de degIn et de betw où il y'a décalage entre les distributions.
Les distributions de la hubité sur les deux réseaux de type Small-World sont décalées.
Cependant, on ne peut ni confirmer ni infirmer l'homogénéité d'aucun des deux cas précédents, seuls les tests permettent de conclure.

Les distributions de toutes les centralités pour Barabasi-Albert sont dégénérées à partir de $power = 2$.
La réalialisation du test de Mann-Whitney permet d'étudier statistiquement l'homogénéité de la hubité pour chaque paire de réseaux de même type avec des paramètres $p$ différents.
Les tableaux 6.8, 6.9 et6.10 présentent les résultats de ces tests.

| ER hub Homogenity | ER_0.02 | ER_0.06 | ER_0.1 |
|---|---|---|---|
| ER_0.02 | | 0 | 0 |
| ER_0.06 | | | 0 |
| ER_0.1 | | | |

TABLE 6.8 – Test d'homogénéité de Mann Whitney de la hubité sur Erdös-Rényi

Comme attendu, la hubité est non-homogène partout sur les paires de réseaux d'Erdös-Rényi.

| SW hub Homogenity | SW_0.02 | SW_0.06 | SW_0.1 |
|---|---|---|---|
| SW_0.02 | | 0 | 0 |
| SW_0.06 | | | 0 |
| SW_0.1 | | | |

TABLE 6.9 – Test d'homogénéité de Mann Whitney de la hubité sur Small-World

Contrairement au degré entrant et à l'intermédiarité, la hubité est non-homogène sur toutes les paires de réseaux Small-World.

| BA hub Homogenity | BA_1 | BA_1.5 | BA_2 | BA_2.5 | BA_3 |
|---|---|---|---|---|---|
| BA_1 | | 0 | 0 | 0 | 0 |
| BA_1.5 | | | 0 | 0 | 0 |
| BA_2 | | | | 1 | 1 |
| BA_2.5 | | | | | 1 |
| BA_3 | | | | | |

TABLE 6.10 – Test d'homogénéité de Mann Whitney de la hubité sur Barabasi-Albert

Les résultats des test pour la hubité sur les réseaux Barabasi-Albert sont exactement les mêmes que dans les cas du degré entrant et l'intermédiarité. Cela est dû, rappelons le, au caractère dégénéré des centralités à partir de $power = 2$.

### 6.1.6 Application au cas de paires de réseaux réels

Dans ce paragraphe, nous reprenons le jeu de données des réseaux de systèmes de sûreté nucléaire $P_{sys-EPR2}$ présentés dans la paragraphe Réseaux réels des systèmes de sûreté nucléaire.

#### 6.1.6.1 Homogénéité de la centralité de degré entrant

Pour étudier l'homogénéité du degré entrant (degIn), nous considérons à titre d'exemple, le cas des deux réseaux reels "AAD_APA_ARE" et "ASG" de $P_{sys-EPR2}$. En premier lieu, représentons graphiquement les rangs des sommets en fonction du réseau de provenance.

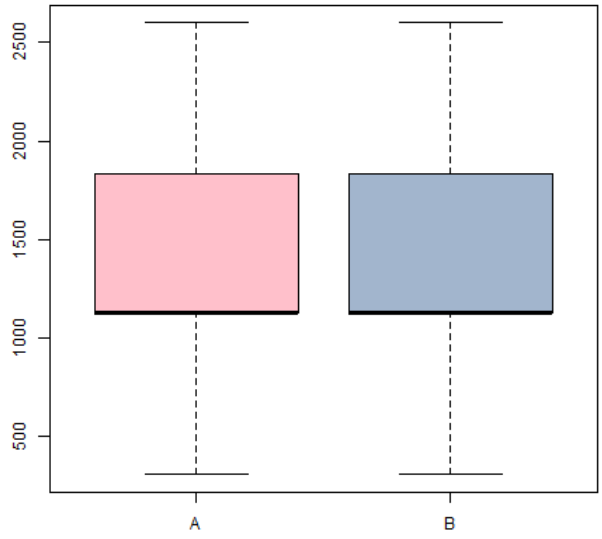



FIGURE 6.5 – Rangs du degré entrant sur les réseaux "AAD_APA_ARE" et "ASG"

D'après la figure 6.5, nous remarquons qu'il n'y a pas de différence de positions, c'est à dire qu'il y a homogénéité de la centralité "degIn" sur les réseaux "AAD_APA_ARE" et "ASG", le test de Mann-Whitney permettera d'étudier statistiquement cette homogénéité.

**Test de Mann-Whitney (Wilcoxon rank sum test)**

Le calcul de la statistique de Mann-Whitney donne $U = 861740$.

La distribution asymptotique de $U$ définie dans le paragraphe Distribution de la statistique de Mann-Whitney $U$ sous $H_0$ est gaussienne de moyenne $n^{(1)}n^{(2)}/2$ de variance $\frac{n^{(1)}n^{(2)}(n^{(1)}+n^2+1)}{12}$.

Ainsi $z$ s'écrit sous la forme suivante :

$$z = \frac{U - \frac{n^{(1)}n^{(2)}}{2}}{\sqrt{\frac{n^{(1)}n^{(2)}(n^{(1)}+n^{(2)}+1)}{12}}} \tag{6.7}$$

Ce qui donne un $z = 1.035$. Au seuil de probabilité $\alpha = 0.05$, sous $H_0$, $z \in [-1.96, 1.96]$.

Ici, $z = 1.035 \in [-1.96, 1.96]$, donc rien ne permet de rejeter l'hypothèse $H_0$, donc il y'a homogénéité de la centralité degIn sur les deux réseaux réels étudiés.

On peut aussi utiliser la $p - value = 0.2804 > 1.96$ pour obtenir la même conclusion.

La centralité degré entrant "degIn" admet ainsi la même distribution sur les deux réseaux "AAD_APA_ARE" et "ASG". Cela confirme les résultats observés sur la figure 6.5.

#### 6.1.6.2 Homogénéité de la centralité de proximité sortante

Dans cette partie, nous nous proposons d'étudier la proximité sortante (cOut) sur les mêmes réseaux réels précédemment étudiés à savoir "AAD_APA_ARE" et "ASG".

La figure 6.6 représente graphiquement les rangs des sommets en fonction du réseau de provenance.

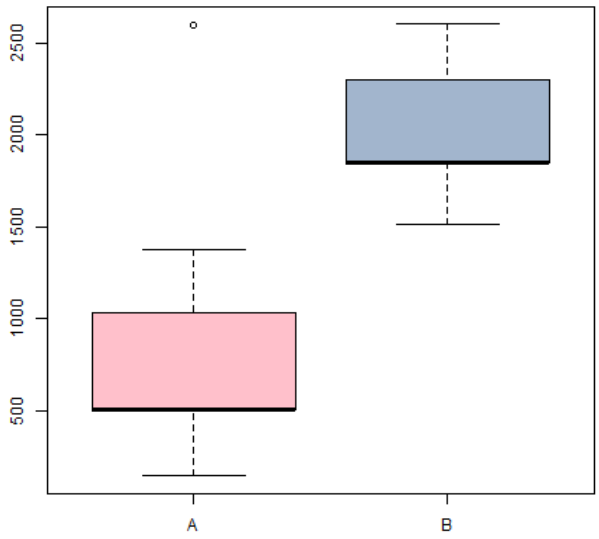



FIGURE 6.6 – Rangs de la proximité sortante sur les réseaux "AAD_APA_ARE" et "ASG"

La figure 6.6 montre que la proximité sortante prend des valeurs plus petites sur le réseau "AAD_APA_ARE" que celles prises sur le réseau "ASG", c'est à dire qu'il n'y a visiblement pas d'homogénéité.

**Test de Mann-Whitney (Wilcoxon rank sum test)**

Le calcul de la statistique de Mann-Whitney donne $U = 1221$.

La distribution asymptotique de $U$ définie dans le paragraphe Distribution de la statistique de Mann-Whitney $U$ sous $H_0$ est gaussienne de moyenne $n^{(1)}n^{(2)}/2$ et de variance $\frac{n^{(1)}n^{(2)}(n^{(1)}+n^{(2)}+1)}{12}$.

$$z = \frac{U - \frac{n^{(1)}n^{(2)}}{2}}{\sqrt{\frac{n^{(1)}n^{(2)}(n^{(1)}+n^{(2)}+1)}{12}}}. \tag{6.8}$$

ce qui correspond à un $z = -44.007$, ainsi :

$$z \notin [-1,96, 1,96]\,.$$

L'hypothèse d'homogénéité est rejetée, au seuil $\alpha = 0.05$. Il est aussi possible d'utiliser la $p - value = 2.2e - 16 < 1.96$ pour obtenir la même conclusion. La centralité de proximité sortante cOut n'admet pas la même distribution sur les réseaux "AAD_APA_ARE" et "ASG".

Plus précisément,

$$\bar{d} = 2\frac{U}{n^{(1)} * n^{(2)}} - 1 \tag{6.9}$$

selon la formule 6.1, $\bar{d} = -0.9985498 \approx -1$. Cela signifie que presque toutes les valeurs de la centralité proximité sortante cOut sur le réseau "AAD_APA_ARE" sont inférieures à celles sur réseau "ASG". Ceci est confirmé par la représentation graphique 6.6.

## 6.2 Étude de deux centralités d'un même réseau

L'objectif de cette partie est d'étudier la dépendance entre deux centralités dans un même réseau $P$. Il s'agit de quantifier le lien pouvant exister entre ces deux centralités.

### 6.2.1 Notations

- Variables (centralités) : $X$ et $Y$.
- Valeurs de $X$ : $x \in D(X)$ (domaine des valeurs de $X$).
- Valeurs de $Y$ : $y \in D(Y)$.
- Population (réseau) : $P$.
- Distribution de $(X, Y)$ dans $P$.
  $P(X \leq x \text{ et } Y \leq y) = F(x, y)$ : fonction de distribution en $(x, y)$ de $(X, Y)$ dans $P$.
  - distribution marginale de $X$ : $P(X \leq x) = F_X(x)$.
  - distribution marginale de $Y$ : $P(Y \leq y) = F_Y(y)$.
- Observations de $(X, Y)$ : $(x_1, y_1)$, $(x_2, y_2)$,..., $(x_n, y_n)$.

### 6.2.2 Mesure de la dépendance

On veut étudier l'intensité de la dépendance entre les centralités $X$ et $Y$ dans $P$. Celle-ci est mesurée (sauf dans Test de Spearman) par :

$$\tau = 2P_c - 1 = 2P([(X^{(1)} > Y^{(1)}) \text{ et } (X^{(2)} > Y^{(2)})] \text{ ou } [(X^{(1)} < Y^{(1)}) \text{ et } (X^{(2)} < Y^{(2)})]) - 1$$

$$(-1 \leq \tau \leq 1)$$

où $(X^{(1)}, Y^{(1)})$ et $(X^{(2)}, Y^{(2)})$ sont deux réalisations aléatoires indépendantes du couple de centralités $(X, Y)$.
$(X^{(1)}, Y^{(1)})$ et $(X^{(2)}, Y^{(2)})$ peuvent donc être considérés comme une paire d'observations indépendantes prélevées sur le réseau $P$. Cette paire est dite :

- **concordante** si $X^{(1)} > X^{(2)}$ et $Y^{(1)} > Y^{(2)}$ ou si
  $X^{(1)} < X^{(2)}$ et $Y^{(1)} < Y^{(2)}$ (probabilité : $P_c$)
- **discordance** si $X^{(1)} > X^{(2)}$ et $Y^{(1)} < Y^{(2)}$ ou si
  $X^{(1)} < X^{(2)}$ et $Y^{(1)} > Y^{(2)}$ (probabilité : $P_d = 1 - P_c$)
- **coïncidence** si $X^{(1)} = X^{(2)}$ ou $Y^{(1)} = Y^{(2)}$ (probabilité nulle).

On voit donc que $\tau = P_c - P_d$ est la différence entre la probabilité pour que deux observations prises au hasard soient concordantes et la probabilité pour que deux observations prises au hasard soient discordantes.
Si les variables $X$ et $Y$ sont indépendantes, c'est-à-dire si
$F(x, y) = F_X(x)F_Y(y)$ alors $\tau = 0$ ; la réciproque n'est pas vraie en général.

Afin de décider du type de test à appliquer, l'étude de la normalité des variables (centralités) s'impose. Le test de Kolmogorov-Smirnov et plus précisément sa variante qui est celui de Lilliefors-Van Soest permet d'étudier cette normalité.

### 6.2.3 Test de normalité de Lilliefors-Van Soest

La normalité des variables est au coeur de l'utilisation de tous les tests paramétriques tels que le test du Khi-deux et le test (basé sur les corrélations) de Pearson. Dans notre cas non-seulement il faut s'assurer que les variables (centralités) suivent une loi normale sur chacun des réseaux mais aussi vérifier que les interdépendances entre ces centralités suivent aussi des lois normales. La plus part des méthodes utilisées pour tester cette hypothèse se basent sur une variante du test de Kolmogorov-Smirnov qui est un test d'hypothèse non-paramétrique étudiant la qualité d'ajustement qui est le test de normalité de Lilliefors-Van Soest. Ce test a été construit indépendamment par Lilliefors[64] et par Van Soest (1967). L'hypothèse nulle de ce test correspond à une erreur normalement distribuée (i.e., pas de différence entre la distribution observée de l'erreur et la distribution normale). Quant à l'hypothèse alternative, elle suppose que l'erreur n'est pas normalement distribué. Comme la plupart des tests statistiques ce test de normalité défini une statistique et donne sa distribution d'échantillonage.

Lorsque la probabilité associée à la statistique est inférieure à un niveau donné $\alpha$ . Lilliefors calcule une approximation de la distribution d'échantillonage utilisant la méthode de Monte-Carlo car l'approche classique qui se base sur des techniques analytiques pour définir une statistique ne marche pas. Pour se faire, la procédure se base sur l'extraction d'un grand nombre d'échantillon à partir d'une population obéïssant à une loi normale et calculer la valeur de la statistique pour chacun de ces échantillons. La distribution observée des valeurs de la statistique donne une approximation de la distribution de l'échantillon de la statistique sous l'hypothèse nulle.

En particulier, Lilliefors et Van Soest utilisent tous les deux, pour chaque échantillon choisi, 1000 échantillon aléatoires dérivés d'une distribution normale centrée et réduite pour approximer la distribution de l'échantillon du critère d'ajustement de Kolmogorov-Smirnov. Les valeurs critiques données par Lilliefors et Van Soest sont presque similaires et l'erreur relative était de l'ordre de $10^2$.

D'après Lilliefors [64] ce test de normalité est plus puissant que les autres méthodes pour plusieurs conditions non-normales.

L'application du test de Lilliefors-Van Soest permet d'attester de la normalité de nos variables afin de juger de la pertinence de l'utilisation le test de corrélation de Pearson.

#### 6.2.3.1 Hypothèses

$$\begin{cases} H_0 : & \forall x : F(x) = F_0(x) \\ H_1 : & \exists x : F(x) \neq F_0(x) \end{cases} \tag{6.10}$$

où $F_0(x)$ : fonction de répartition normale ($m$ et $\sigma$ non spécifiés).

Dans $H_0$ La distribution de la variable suit une loi normale.

#### 6.2.3.2 Statistique du test de Lilliefors-Van Soest

$$D = \sup_{-\infty<X<+\infty} \mid \hat{F}(x) - F_0^*(x) \mid$$

Telle que $\hat{F}(x)$ est la fonction de répartition observée ; et $F_0^*(x)$ est la fonction de répartition d'une loi normale $N(\bar{x}; S)$ :
$F_0^*(x) = F_\nu(\frac{x-\bar{x}}{S})$ où $F_\nu(y)$ est la fonction de répartition normale réduite.
La formule pour le calcul de la statistique $D$ s'écrit comme suit :

$$D = max(D^+; D^-) \tag{6.11}$$

où :

$$\begin{cases} D^+ = & max_{1\leq i\leq n}\left[\hat{F}(x_{(i)}) - F_0(x_{(i)})\right] \\ D^- = & max_{1\leq i\leq n}\left[F_0(x_{(i)}) - \hat{F}(x_{(i-1)})\right] \end{cases} \tag{6.12}$$

#### 6.2.3.3 Distribution de la statistique de Lilliefors-Van Soest sous $H_0$

- Pour $n \leq 30$ : les quantiles $D_{1-\alpha}$
- Pour $n > 30$ : une approximation des quantiles $D_{1-\alpha}$ d'ordre $1-\alpha$ de la distribution de $D$ sous $H_0$

Ainsi, on rejette $H_0$ au niveau de probabilité $\alpha$ si $D > D_{1-\alpha}$.

#### 6.2.3.4 Application au jeu de données $P_{sys-EPR2}$

Nous avons effectué ce test aux réseaux du jeu de données $P_{sys-EPR2}$, nous présentons uniquement les figures relatives au réseau AAD_APA_ARE.

Dans un premier temps, nous représentons les distributions de chaque centralité dans la figure 6.7.

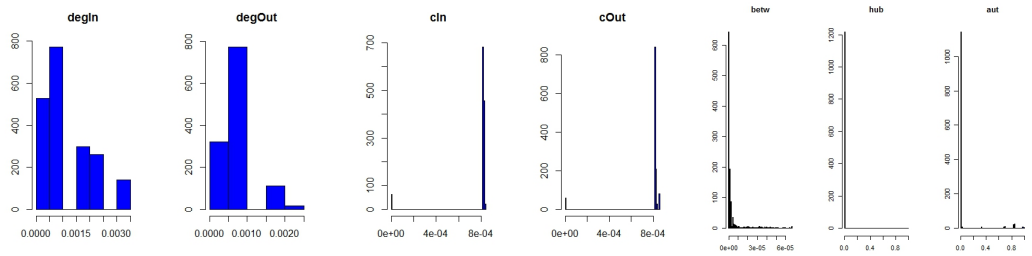



FIGURE 6.7 – Distributions des 7 centralités sur le réseau AAD_APA_ARE.

La figure 6.7 représente les valeurs prises par les centralités sur le réseau AAD_APA_ARE. Nous remarquons qu'aucune de ces distributions n'est normale. Les mêmes résultats ont été observés sur les 22 autres réseaux du jeu de données $P_{sys-EPR2}$.

Le test Lilliefors-Van Soest confirme ces résultats de non-normalité des centralités pour tous les réseaux de $P_{sys-EPR2}$.
En effet, on observe une $p-value$ preque nulle pour chacune des centralités sur chaque réseau du jeu de données. Ainsi, aucune centralité ne suit la loi normale sur tout le jeu de données $P_{sys-EPR2}$. Les conditions d'utilisation de la corrélation paramétrique de Pearson ne sont pas réalisées. Par conséquent, seuls les tests non-paramétriques de corrélation de Spearman ou de Kendall conviennent.

### 6.2.4 Test de Spearman

#### 6.2.4.1 Hypothèses

$$\begin{cases} H_0 : & X \text{ et } Y \text{ sont indépendantes} ; \\ H_1 : & X \text{ et } Y \text{ ne sont pas indépendantes} \end{cases} \tag{6.13}$$

ou bien

$$\begin{cases} H_0' : & X \text{ et } Y \text{ sont indépendantes} ; \\ H_1' : & \text{il y a tendance pour les petites valeurs de } X \text{ à être liées} \\ & \text{aux grandes valeurs de } Y \text{ et inversement (dépendance inverse).} \end{cases} \tag{6.14}$$

ou bien

$$\begin{cases} H_0^{''} : & X \text{ et } Y \text{ sont indépendantes}\,; \\ H_1^{''} : & \text{il y a tendance pour les grandes valeurs de X à être liées} \\ & \text{aux grandes valeurs de } Y \text{ et invesement (dépendance directe).} \end{cases} \tag{6.15}$$

#### 6.2.4.2 Statistique de Spearman

Le coefficient de corrélation de rang de Spearman :

$$r_S = \frac{\sum_{i=1}^{n} \left[R(x_i) - \bar{R}_x\right] \left[R(y_i) - \bar{R}_y\right]}{\sqrt{\sum_{i=1}^{n} \left[R(x_i) - \bar{R}_x\right]^2 \sum_{i=1}^{n} \left[R(y_i) - \bar{R}_y\right]^2}} \tag{6.16}$$

où :

- $R(x_i)$ est la rang de $x_i$ dans $(x_1, x_2, \ldots, x_n)$.
- $R(y_i)$ est la rang de $y_i$ dans $(y_1, y_2, \ldots, y_n)$.
- $\bar{R}_x = \frac{1}{n}\sum_{i=1}^{n} R(x_i) = \frac{n+1}{2}$.
- $\bar{R}_y = \frac{1}{n}\sum_{i=1}^{n} R(y_i) = \frac{n+1}{2}$.

$r_S$ est donc le coefficient de corrélation portant sur :

$$\{(R(x_i), R(y_i)); i = 1, 2, ...., n\} \Rightarrow -1 \leq r_S \leq 1 \tag{6.17}$$

Le signe de $r_S$ indique le sens de dépendance entre $X$ et $Y$ dans le réseau $P$.

S'il n'y a pas de coïncidence on peut utiliser la formule suivante :

$$r_S = 1 - \frac{6\sum_{i=1}^{n} \left[R(x_i) - R(y_i)\right]^2}{n(n^2-1)} \tag{6.18}$$

Intuitivement, on rejette $H_0$ ou $H_0^{'}$ si $r_S$ est suffisamment proche de $-1$ ; on rejette $H_0$ et $H_0^{''}$ si $r_S$ est suffisamment proche de 1.

#### 6.2.4.3 Distribution de $r_S$ sous $H_0$

— distribution exacte : symétrique par rapport à 0 [41], pages 228-232

— distribution asymptotique : $r_S \approx N(0; \frac{1}{\sqrt{n-1}})$ (approximation par la loi normale de moyenne nulle et d'écart type $\frac{1}{\sqrt{n-1}}$, valable si $n > 30$).

#### 6.2.4.4 Traitement des coïncidences

Si plusieurs $x_i$ sont confondus ou si plusieurs $y_i$ sont confondus, il faut utiliser la formule 6.16 pour le calcul de $r_S$ et non 6.18. on procédera alors au calcul des rangs moyens pour déterminer les rangs $R(x_i)$ et $R_{y_i}$.
Notons que la distribution normale asymptotique reste inchangée.
Étant donné que le nombre de sommets dans notre cas est toujours $> 30$, on se basera sur la distribution asymptotique et donc on utilisera la méthode générale relative à une statistique de distribution normale.

#### 6.2.4.5 Exemple d'application

Considérons les observations du couple de variables $(X, Y)$ :

$$(7, 22); (12, 25); (19, 8); (25, 13); (28, 26)$$

Représentons ces données graphiquement, la couleur rose correspond aux valeurs de la variable $X$ et bleu celles de $Y$ :

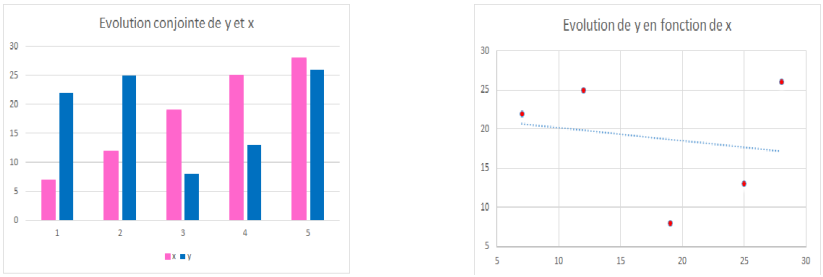



FIGURE 6.8 – Evolution conjointe de x et y

On observe ici une dépendance inverse i.e. les petites valeurs de $x$ correspondent aux grandes valeurs de $y$ et inversement.

**Corrélation non-paramétrique de Spearman**

Rappelons que le coefficient de corrélation de Spearman tel que défini dans la formule 6.18 s'écrit sous la forme suivante :

$$r_S = 1 - \frac{6 \sum_{i=1}^{n} [R(x_i) - R(y_i)]^2}{n(n^2 - 1)} \tag{6.19}$$

Le tableau suivant présente les rangs des observations $x_i$ et $y_i$ des centralités $X$ et $Y$ au sommet $i$.

| i | $x_i$ | $y_i$ | $R(x_i)$ | $R(y_i)$ | $R(x_i) - R(y_i)$ |
|---|---|---|---|---|---|
| 1 | 7 | 22 | 1 | 3 | (1-3) |
| 2 | 12 | 25 | 2 | 4 | (2-4) |
| 3 | 19 | 8 | 3 | 1 | (3-1) |
| 4 | 25 | 13 | 4 | 2 | (4-2) |
| 5 | 28 | 26 | 5 | 5 | (5-5) |

D'après la formule 6.18

$$r_{S(X,Y)} = 1 - \frac{(6 * ([1-3]^2 + [2-4]^2 + [3-2]^2 + [4-2]^2 + [5-5]^2)}{5 * 24} = -0.2$$

$r_{S(X,Y)} = -0.2$ est négatif ce qui confirme la dépendance inverse mise en évidence dans le graphique 6.8.

#### 6.2.4.6 Etude des réseaux artificiels classiques par Spearman

Cette partie s'intéresse à l'indépendance entre les différentes centralités d'un même réseau dans un premier temps par la méthode descriptive en présentant les corrélogrammes de Spearman. Ensuite, le test d'indépendance de Spearman est effectué afin de confirmer statistiquement cette dépendance.

**A- Réseaux de type Erdös-Rényi**

Les corrélogrammes représentés dans la figure 6.9 montrent que dans un réseau Erdös-Rényi les centralités degIn, cIn et aut sont fortement corrélées entre elles. Les centralités degOut, cOut et hub le sont aussi. Tandis que betw est moyennement corrélées avec toutes les centralités.

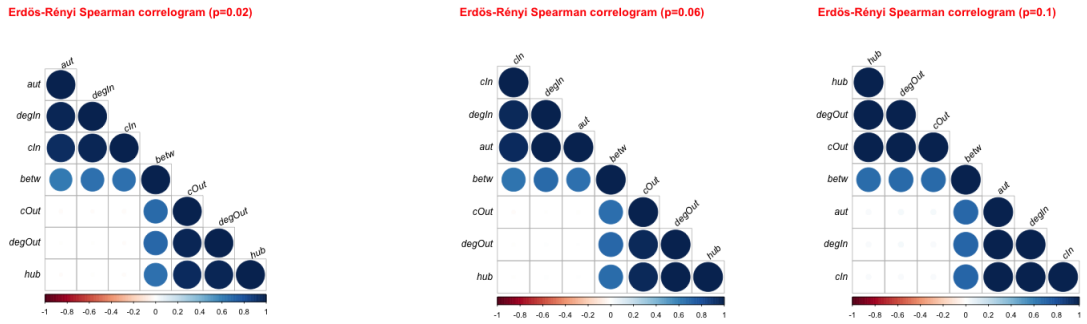



FIGURE 6.9 – Corrélogrammes de Spearman pour Erdös-Rényi : $p = 0.02$, $p = 0.06$, $p = 0.10$

### B- Réseaux de type Small-World

La figure 6.10 montre que dans les trois corrélogrammes : la proximité (c) est moyennement corrélée à l'intermédiarité (betw), l'intermédiarité (betw) un peu moins corrélée au degré (deg), de plus, la corrélation entre le degré (deg) et la hubité (hub) est plus faible.
Ces corrélations augmentent proportionnellement à $p$.

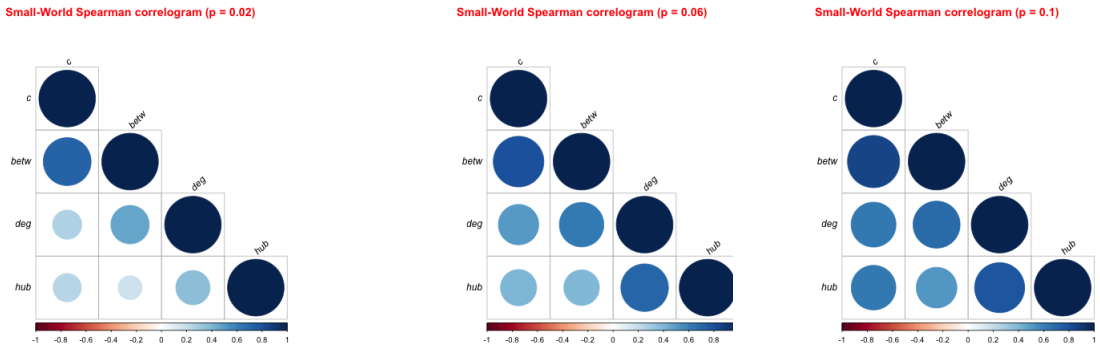



FIGURE 6.10 – Corrélogrammes de Spearman pour Small-World : $p = 0.02$, $p = 0.06$, $p = 0.10$

### C-Réseaux de type Barabasi-Albert

La figure 6.11 montre une forte dépendance entre degIn, cIn, betw et aut et une corrélation moyenne négative entre cOut et hub pour $power = 1$. Pour $power = 2$, la forte corrélation entre degIn, cIn et betw se confirme, cependant, avec aut leur corrélation devient négative (passe en rouge). La corrélation entre cOut et hub s'intensifie mais reste négative. Pour $power = 3$ toutes les corrélations s'intensifient, mais restent positives (bleues) ou négatives(rouges).
Notons que les différentes centralités à partir de $power = 2$ sont dégénérées (ne prennent que très peu de valeurs).

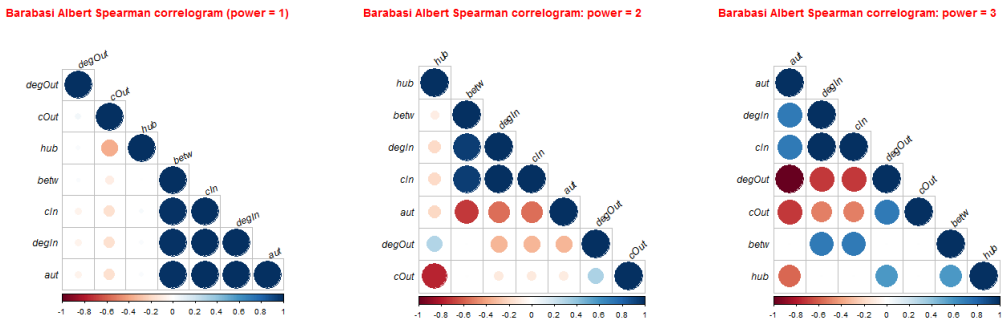



FIGURE 6.11 – Corrélogrammes de Spearman pour Barabasi-Albert : $power = 1$, $power = 2$, $power = 3$

**D- Test non paramétrique des rangs de Spearman**

Avec la convention :

— Si 1, on rejette l'hypothèse d'indépendance (donc il y'a dépendance).

— Si 0, on ne rejette pas l'hypothèse d'indépendance .

Etudions l'indépendance entre les centralités dans chacun des réseaux artificiels.

— Réseaux de type Erdös-Rényi

| ER Spearman test | degIn | degOut | cIn | cOut | betw | hub | aut |
|---|---|---|---|---|---|---|---|
| degIn | | 0 | 1 | 0 | 1 | 0 | 1 |
| degOut | | | 0 | 1 | 1 | 1 | 0 |
| cIn | | | | 0 | 1 | 0 | 1 |
| cOut | | | | | 1 | 1 | 0 |
| betw | | | | | | 1 | 1 |
| hub | | | | | | | 0 |
| aut | | | | | | | |

TABLE 6.11 – Test d'indépendance de Spearman pour Erdös-Rényi avec $p = 0.02$

Ce test confirme les résultats observés sur les corrélogrammes de la figure 6.9 (voir interprétation). De plus, betw est corrélée à toutes les autres centralités.

— Réseaux de type Small-World

| SW Spearman test | deg | c | betw | hub |
|---|---|---|---|---|
| deg | | 1 | 1 | 1 |
| c | | | 1 | 1 |
| betw | | | | 1 |
| hub | | | | |

TABLE 6.12 – Test d'indépendance de Spearman pour Small-World avec $p = 0.02$

Nous remarquons que toutes les centralités sont dépendantes dans le réseau Small-World $p = 0.02$ tel que interprété dans le corrélogramme.

— Réseaux de type Barabasi-Albert

| BA Spearman test | degIn | degOut | cIn | cOut | betw | hub | aut |
|---|---|---|---|---|---|---|---|
| degIn | | 1 | 1 | 0 | 1 | 1 | 1 |
| degOut | | | 1 | 1 | 0 | 1 | 1 |
| cIn | | | | 0 | 1 | 1 | 1 |
| cOut | | | | | 1 | 1 | 1 |
| betw | | | | | | 1 | 0 |
| hub | | | | | | | 1 |
| aut | | | | | | | |

TABLE 6.13 – Test d'indépendance de Spearman pour Barabasi-Albert avec *power* = 2

Ces tests confirment les résultats obtenus par le corrélogramme du milieu de la figure 6.11.

#### 6.2.4.7 Étude du réseau du système "AAD_APA_ARE" par Spearman

Rappelons que pour chaque sommet du réseau, nous avons calculé les centralités de base à savoir : degré entrant (degIn), degré sortant (degOut), proximité entrante (cIn), proximité sortante (cOut), intermédiarité (betw), hubité (hub) et autorité (aut).

Le corrélogramme suivant permet de visualiser les coefficients de corrélations de Spearman entre ces centralités.

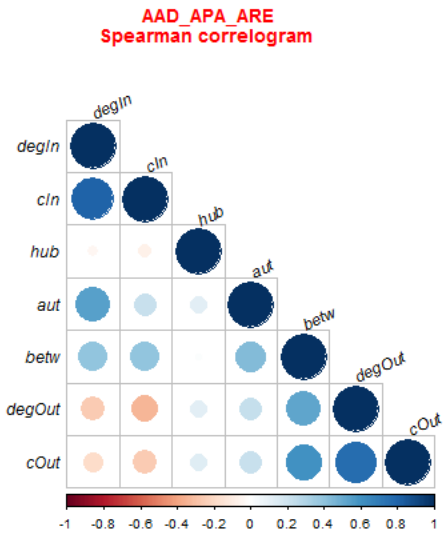



FIGURE 6.12 – Corrélogramme de Spearman entre les centralités pour le réseau "AAD_APA_ARE"

Nous remarquons qu'il y a une forte corrélation positive (selon Spearman) entre les centralités (degIn, cIn) et (degOut, cOut). Tandis que (degIn, cOut) et (cIn, cOut) sont faiblement négativement corrélées. D'autre part, la centralité betw est moyennement corrélée à toutes les autres, à part hub. De plus, degIn et hub présentent une très faible corrélation.

**Test non paramétrique des rangs de Spearman**

Avec la convention :

— Si 1, le test est significatif donc on rejette l'hypothèse d'indépendance (donc il y'a dépendance).

— Si 0, on ne rejette pas l'hypothèse d'indépendance (Rien ne permet d'assurer la dépendance).

Le tableau suivant donne les résultats de ce test par rapport aux paires de centralités sur le réseau "AAD_APA_ARE".

| AAD_APA_ARE Spreaman test | degIn | degOut | cIn | cOut | betw | hub | aut |
|---|---|---|---|---|---|---|---|
| degIn | | 1 | 1 | 1 | 1 | 0 | 1 |
| degOut | | | 1 | 1 | 1 | 1 | 1 |
| cIn | | | | 1 | 1 | 1 | 1 |
| cOut | | | | | 1 | 1 | 1 |
| betw | | | | | | 0 | 1 |
| hub | | | | | | | 1 |
| aut | | | | | | | |

TABLE 6.14 – Test d'indépendance de Spearman entre les centralités pour le réseau "AAD_APA_ARE"

Ces tests de Spearman confirment l'interprétation du corrélogramme 6.12, ils confirment de plus certaines dépendances qui étaient peu visibles dans le corrélogramme.

### 6.2.5 Test de Kendall

#### 6.2.5.1 Hypothèses

$$\begin{cases} H_0 : & X \text{ et } Y \text{ sont indépendantes} ; \\ H_1 : & \tau \neq 0 \text{ selon le paragraphe 6.2.2.} \end{cases} \tag{6.20}$$

Ou bien

$$\begin{cases} H_0^{'} : & X \text{ et } Y \text{ sont indépendantes} ; \\ H_1^{'} : & \tau < 0 \end{cases} \tag{6.21}$$

Ou bien

$$\begin{cases} H_0^{''} : & X \text{ et } Y \text{ sont indépendantes} ; \\ H_1^{'} : & \tau > 0 \end{cases} \tag{6.22}$$

Où $\tau$ est défini dans la paragraphe Mesure de la dépendance.

#### 6.2.5.2 Statistique de Kendall

La statistique de kendall est définie comme suit :

$$T = N_c - N_d; (-\frac{n(n-1)}{2} \leq T \leq \frac{n(n-1)}{2}) \tag{6.23}$$

où $N_c$ est le nombre de paires concordantes, $N_d$ est le nombre de paires discordantes parmi les $\frac{n(n-1)}{2}$ paires d'observations $(x_i, y_i)$ et $(x_j, y_j)$ avec $(i < j)$.

En pratique, pour faciliter les comparaisons, on range les observations par ordre croissant (par exemple) des $x_i$.

**Interprétation**

Si $X$ et $Y$ sont des variables indépendantes, alors $\tau = 0$ et on peut s'attendre à ce que $N_c$ soit proche de $N_d$.

Intuitivement, on rejette $H_0$ ou $H_0^{'}$ si $N_c$ est trop petit par rapport à $N_d$ ($T < 0$ et $T$ trop grand).

Par contre, on rejette $H_0$ ou $H_0^{''}$ si $N_c$ est trop grand par rapport à $N_d$ ($T > 0$ et trop grand).

#### 6.2.5.3 Exemple d'application

Reprenons l'exemple précédent présenté dans le paragraphe Exemple d'application où les $x_i$ sont rangés par ordre croissant. Ensuite, on compte les nombres de concordances et de discordances avec les observations suivantes.

| $x_i$ | $y_i$ | **concordances** | **discordances** |
|---|---|---|---|
| 7 | 22 | 2 | 2 |
| 12 | 25 | 1 | 2 |
| 19 | 8 | 1 | 1 |
| 25 | 13 | 0 | 1 |
| 28 | 26 | 0 | 0 |
| Total | | 4 | 6 |

TABLE 6.15 – Concordances et discordances dans l' Exemple d'application

Calculons la statistique de Kendall en utilisant la formule 6.23.

$T = N_c - N_d = 4 - 6 = -2\,;\ (-\frac{n(n-1)}{2} = -10 \leq T = -2 \leq \frac{n(n-1)}{2} = 10$

$T = -2$ négatif, ce qui confirme la dépendance inverse obtenue dans la représentation graphique 6.8.

#### 6.2.5.4 Distribution de la statistique de Kendall $T$ sous $H_0$

— Distribution exacte : symétrique par rapport à [41] pages 215-218.

— Distribution asymptotique :

$$T \approx N(0; \sqrt{(\frac{n(n-1)(2n+5)}{18}}) \tag{6.24}$$

(cette approximation est valable si $n > 40$).

**Remarque :** Comme généralement nos réseaux ont un nombre de sommets $n > 40$, on peut se baser sur la distribution asymptotique Distribution de la statistique de Kendall $T$ sous $H_0$ et utiliser la méthode générale relative à une statistique de distribution normale.

#### 6.2.5.5 Etude des réseaux artificiels classiques par Kendall

Cette partie s'intéresse à l'indépendance entre les différentes centralités d'un même réseau dans un premier temps par la méthode descriptive en présentant les corrélogrammes de Kendall. Ensuite, le test d'indépendance de Kendall est effectué afin de confirmer ou d'infirmer statistiquement cette dépendance.
Nous remarquerons, au passage que les résultats ici, sont identiques à ceux obtenus par Spearman.

**A- Réseaux de type Erdös-Rényi**

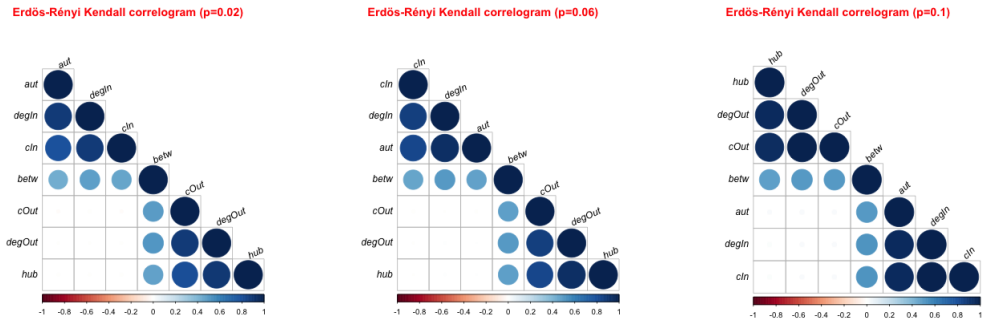



FIGURE 6.13 – Corrélogrammes de Kendall pour Erdös-Rényi : $p = 0.02$, $p = 0.06$, $p = 0.01$

**B- Réseaux de type Small-World**

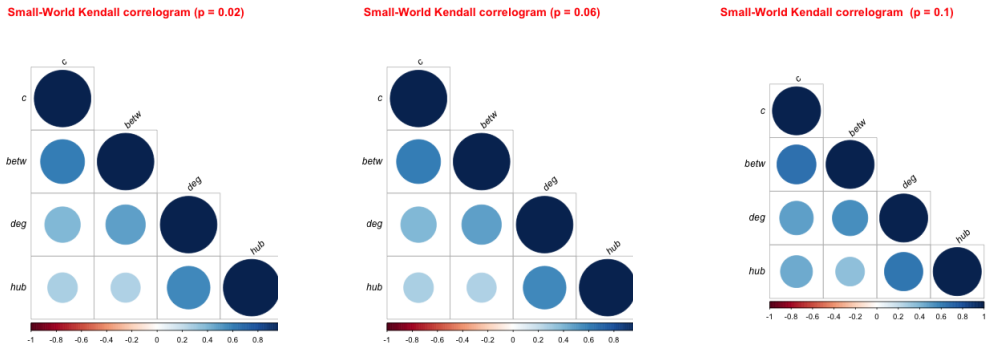



FIGURE 6.14 – Corrélogrammes de Kendall pour Small-World : $p = 0.02$, $p = 0.06$, $p = 0.1$

### C-Réseaux de type Barabasi-Albert

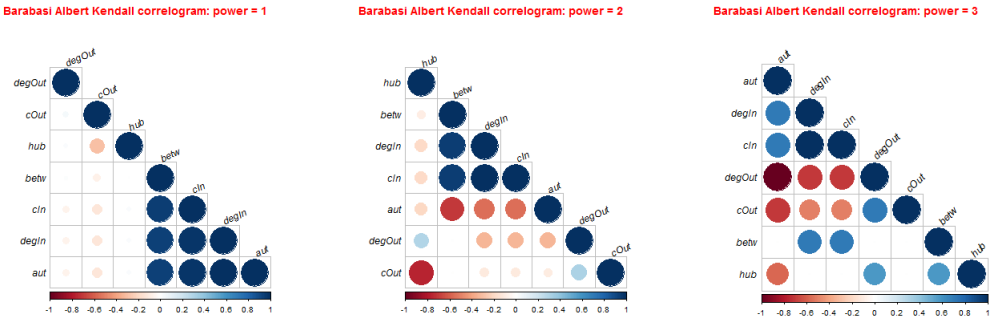



FIGURE 6.15 – Corrélogrammes de Kendall pour Barabasi-Albert : $power = 1$, $power = 2$, $power = 3$

### D- Test non paramétrique des rangs de Kendall

Avec la convention :

— Si 1, le test est significatif donc on rejette l'hypothèse d'indépendance (donc il y'a dépendance).

— Si 0, on ne rejette pas l'hypothèse d'indépendance (Rien ne permet d'assurer la dépendance).

Les tableaux suivants donnent les résultats des tests de Kendall entre les centralités dans chacun des réseaux artificiels. Ces résultats sont identiques à ceux obtenus par Spearman.

— Réseaux de type Erdös-Rényi

| ER Kendall test | degIn | degOut | cIn | cOut | betw | hub | aut |
|---|---|---|---|---|---|---|---|
| degIn | | 0 | 1 | 0 | 1 | 0 | 1 |
| degOut | | | 0 | 1 | 1 | 1 | 0 |
| cIn | | | | 0 | 1 | 0 | 1 |
| cOut | | | | | 1 | 1 | 0 |
| betw | | | | | | 1 | 1 |
| hub | | | | | | | 0 |
| aut | | | | | | | |

TABLE 6.16 – Test d'indépendance de Kendall pour Erdös-Rényi avec $p = 0.02$

— Réseaux de type Small-World

| SW Kendall test | deg | c | betw | hub |
|---|---|---|---|---|
| deg | | 1 | 1 | 1 |
| c | | | 1 | 1 |
| betw | | | | 1 |
| hub | | | | |

TABLE 6.17 – Test d'indépendance de Kendall pour Small-World avec $p = 0.02$

— Réseaux de type Barabasi-Albert

| BA Kendall test | degIn | degOut | cIn | cOut | betw | hub | aut |
|---|---|---|---|---|---|---|---|
| degIn | | 1 | 1 | 0 | 1 | 1 | 1 |
| degOut | | | 1 | 1 | 0 | 1 | 1 |
| cIn | | | | 0 | 1 | 1 | 1 |
| cOut | | | | | 1 | 1 | 1 |
| betw | | | | | | 1 | 0 |
| hub | | | | | | | 1 |
| aut | | | | | | | |

TABLE 6.18 – Test d'indépendance de Kendall pour Barabasi-Albert avec $power = 2$

#### 6.2.5.6 Étude du réseau du système "AAD_APA_ARE" par Kendall

Reprenons l'étude du réseau "AAD_APA_ARE" et rappelons que pour chaque sommet, nous avons calculé les centralités : degré entrant (degIn), degré sortant (degOut), proximité entrante (cIn), proximité sortante (cOut) et intermédiarité (betw).
Le corrélogramme suivant permet de visualiser les coefficients de corrélations de Kendall entre ces centralités.

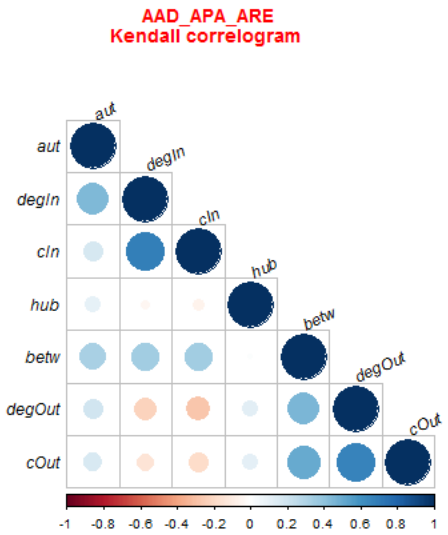



FIGURE 6.16 – Corrélogramme de Kendall entre les centralités du réseau "AAD_APA_ARE"

**Interprétation**

Nous remarquons qu'il une forte corrélation positive (selon Kendall) entre les centralités (degIn, cIn) et (degOut, cOut). Tandis que (degIn, cOut) et (cIn, cOut) sont faiblement négativement corrélées. D'autre part, la centralité betw est moyennement corrélée avec toutes les centralités. Notons que nous obtenons les mêmes relations avec Spearman.

**Test non paramétrique des rangs de Kendall**

Avec la convention :

- Si 1, le test est significatif donc on rejette l'hypothèse d'indépendance (donc il y'a dépendance).
- Si 0, on ne rejette pas l'hypothèse d'indépendance (Rien ne permet d'assurer la dépendance).

Le tableau suivant donne les résultats de ce test.

| AAD_APA_ARE Kendall test | degIn | degOut | cIn | cOut | betw | hub | aut |
|---|---|---|---|---|---|---|---|
| degIn | | 1 | 1 | 1 | 1 | 0 | 1 |
| degOut | | | 1 | 1 | 1 | 1 | 1 |
| cIn | | | | 1 | 1 | 1 | 1 |
| cOut | | | | | 1 | 1 | 1 |
| betw | | | | | | 0 | 1 |
| hub | | | | | | | 1 |
| aut | | | | | | | |

TABLE 6.19 – Test d'indépendance de Kendall entre les centralités du réseau "AAD_APA_ARE"

Ces tests de Kendall conduisent exactement aux mêmes résultats que ceux de Spearman et confirment encore l'interprétation du corrélogramme représenté dans la figure 6.16.

## 6.3 Conclusion

Dans ce chapitre, nous avons étudié une centralité sur plusieurs réseaux, ce sont des propriétés qui caractérisent les sommets du réseau. Nous avons montré que certaines de ces centralités ont des comportements similaires sur différents réseaux, elles sont homogènes. L'homogénéité est une similarité partielle entre réseaux qui ne considère qu'une seule centralité à la fois.
Par ailleurs nous avons étudié deux centralités à la fois dans un même réseau pour détecter les relations de dépendances qui peuvent lier deux centralités caractérisant un même réseau. Nous avons prouvé, par exemple, que presque toutes les centralités dans le réseau AAD_APA_ARE sont liées. Encore ici, nous essayons de détecter des similarités mais cette fois ci-entre centralités dans un même réseau.
Le chapitre suivant propose une nouvelle mesure de similarité qui permet de comparer globalement les structures des réseaux : Une nouvelle représentation d'un réseau est proposée basée sur les corrélations de Kendall entre ses vecteurs de centralités. Ensuite une mesure de similarité basée sur cette représentation sera introduite pour permettre de comparer des réseaux de différentes tailles, indépendamment de leurs densités, diamètres,..., etc.

# Chapitre 7

# Nouvelle mesure de similarité entre graphes multivariables

L'objectif de ce chapitre est de proposer une nouvelle mesure de dissimilarité pour pouvoir comparer deux ou plusieurs réseaux. Cette mesure aura pour mission de comparer des réseaux indépendamment de leurs tailles, nombres de liens, densité ou autre mesure de caractérisation de réseaux, mais uniquement en se basant sur les classements des sommets fournis par chacune des centralités présentées dans la section 3.3 du chapitre 3 Graphes et centralités. Chacune des centralités étant une représentation différente du réseau et le fait de combiner toutes ces représentations permet d'avoir une plus riche représentation de la structure globale du réseau étudié, qui n'est nullement assurée ni par la densité, ni par le diamètre, le coefficient de clustering ou la distribution des degrés.
A ce propos, nous proposons d'étudier les interdépendances entre les différents vecteurs de centralités pour en capturer la structure. Ensuite, nous définissons une nouvelle mesure "synthétique de centralité" efficace pour collecter et rassembler l'information concernant cette structure globale et les caractéristiques topologiques du réseau. Enfin, dans le but de comparer un réseau à un ou plusieurs autres réseaux, nous définissons une mesure de similarité entre réseaux basée sur cette mesure synthétique de centralité.
Plusieurs illustrations de cette mesure de similarité (semi-distance), seront effectuées à l'aide de clustering sur des collections de graphes générés artificiellement tels que les graphes d'Erdös-Rényi, graphes « petit-monde » et graphes de Barabasi-Albert, ainsi que des familles de graphes obtenus par percolation.

## 7.1 Coefficient synthétique de centralité et notion de similarité

Soit $G$ un graphe, $G = (V, E \subseteq V \times V)$ où $V$ est l'ensemble de sommets et $E$ désigne l'ensemble de liens (orientés ou non).

Considérons les vecteurs de centralités calculées sur le graphe entier :

(i) Le degré : $deg = degree(G, V(G))$

(ii) La proximité : $c = closeness(G, V(G))$

(iii) L'intermédiarité : $betw = betweenness(G, V(G))$

(iv) Le degré de hubité : $hub = hub.score(G)$.

Rappelons que nous avons prouvé dans le paragraphe Test de normalité de Lilliefors-Van Soest la non-normalité des centralités en réalisant le test de "Lilliefors-Van Soest", il n'est donc pas convenable d'utiliser la corrélation linéaire de Pearson, ainsi le choix est fait sur la corrélation non-paramétrique.
Nous définissons le coefficient synthétique de centralité $Kcor(G)$ à partir des corrélations entre les vecteurs de centralités ci-dessus pris par paires.
Nous considérons (de préférence) la corrélation de Kendall pour ces qualités décrites dans le paragraphe Test de Kendall et surtout pour son indépendance vis-à-vis des lois sous-jacentes et pour son interprétabilité en tant que probabilité. Notons que la corrélation de Spearman est équivalente à celle de Kendall [27] et donne les mêmes résultats. La méthode présentée ci-après reste inchangée quant au type de corrélation utilisé.

Ainsi $Kcor(G)$ est un vecteur à $\binom{4}{2} = 6$ composantes dont la première composante, par exemple, est $Corr(deg, c)$ tel que $Corr$ est la corrélation de Kendall.
Remarquons que $\|Kcor(G)\| = 0$ signifie que toutes ces centralités sont non-corrélées, c'est-à-dire il n'y a pas de relation (de dépendance) entres les centralités des sommets de $G$.
Au contraire, $\|Kcor(G)\| = \|(1, \ldots, 1)\|$ signifie que toutes ces centralités sont parfaitement liées deux à deux. C'est à dire, il existe de forte dépendance entres les centralités.
Autrement dit, un sommet important pour une centralité l'est aussi pour les autres et inversement. Dans ce cas les sommets « importants » pour toutes les centralités sont mis en évidence.

Définissons la similarité entre deux graphes $G1$ et $G2$ par

$$Sim(G1, G2) = dist(Kcor(G1), Kcor(G2)),$$

où $dist()$ est une distance sur $\mathbb{R}^2$.
Notons d'abord que $Sim(G, G) = 0$, pour tout graphe G. $Sim()$ est symétrique et vérifie l'inégalité triangulaire. $Sim$ est donc une semi-distance entre graphes.
Dans le cas de graphe orientés, le coefficient synthétique de centralité $Kcor(G)$ peut être défini en distinguant les liens entrants des liens sortants dans les calculs des centralités, ainsi les centralités calculées sur le réseau deviennent :

(i) Le degré entrant, noté : $degIn = degree(G, V(G), mode = in)$

(ii) Le degré sortant : $degout = degree(G, V(G), mode = out)$

(iii) La proximité entrante : $cIn = closeness(G, V(G), mode = in)$

(iv) la proximité sortante : $cOut = closeness(G, V(G), mode = out)$

(v) L'intermédiarité : $betw = betweenness(G, V(G))$

(vi) Le degré de hubité : $hub = hub.score(G)$,

(vii) Le degré d'autorité : $aut = authority.score(G)$

Ainsi le coefficient synthétique de centralité $Kcor(G)$ devient un vecteur à $\binom{7}{2} = 21$ composantes dont la première composante, par exemple, est $Corr(degIn, cIn)$, $Corr$ étant la corrélation de Kendall entre les deux variables.
La similarité entre deux graphes $G1$ et $G2$ reste toujours définie par :

$$Sim(G1, G2) = dist(Kcor(G1), Kcor(G2)),$$

où $dist()$ est une distance sur $\mathbb{R}^2$.
La figure 7.1 représente graphiquement les distances usuelles. Les couleurs noire, bleue et rouge représentent respectivement la distance de Manhattan, l'euclidienne et l'infinie entre deux élements $M1$ et $M2$.
Il s'agit d'une représentation à deux dimensions $X$ et $Y$.

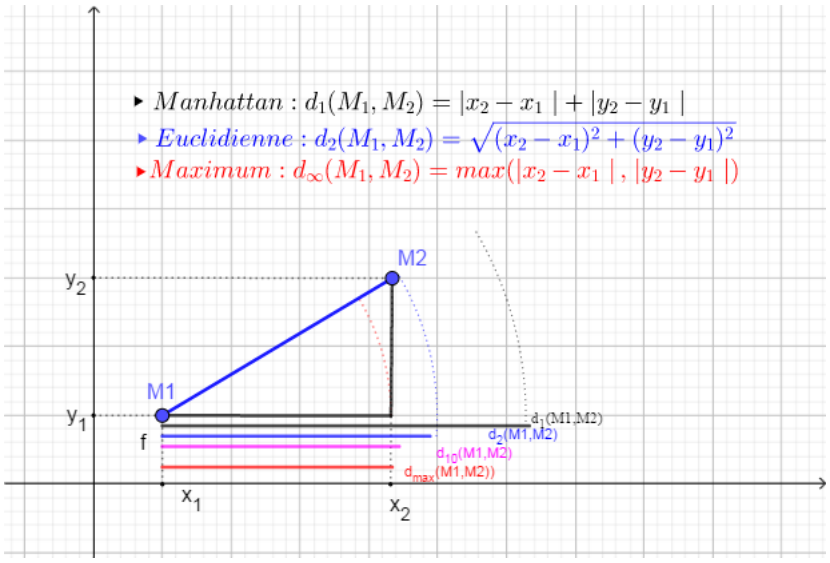



FIGURE 7.1 – Distances de Manhattan $d_1$, Euclidienne $d_2$ et Infinie $d_\infty$

## 7.2 Mesure de similarité entre réseaux

1. Soit $P$ un ensemble de graphes de $card(P) = K$ graphes notés $P_k$, $1 \leq k \leq K$.
2. Pour chaque graphe $P_k$, $1 \leq k \leq K$ de $P$ , on procède comme suit :

(a) Extraire la plus grande composante connexe du graphe $P_k$ et l'utiliser pour toute la suite.

(b) Calculer les $N_c$ vecteurs des centralités choisies. Rappelons que pour un réseau non-dirigé, il n'y a pas de distinction entre centralité entrante et centralité sortante, ainsi le nombre de vecteurs des centralités sera réduit.

(c) Sélectionner les sommets les mieux classés par chaque centralité : $top_c^{(j)}$ est l'ensemble des sommets ayant les $c$ meilleures valeurs par la centralité $X^{(j)}$, $1 \leq j \leq N_c$.

$$Top = \underset{1 \leq j \leq N_c}{} top_c^{(j)}.$$

$Top$ est donc la réunion disjointe de ces $top_c^{(j)}$, qui correspondent aux sommets les mieux classés ($c$ meilleures valeurs) par au moins l'une des $N_c$ centralités $X^{(j)}$, $1 \leq j \leq N_c$ considérées.

Notons qu'il est possible de conserver les vecteurs de centralités (pour tous les sommets), mais cela donne des résultats moins précis, étant donné que seuls les sommets les mieux classés ont une réelle signification. Par exemple, seules les premières images qui apparaissent suite à une requête de recherche sur google sont pertinantes [51].

**Construction du coefficient synthétique de centralité $Kcor[k]$ de $P_k$**

Bien entendu, les vecteurs de centralités sont réduits aux représentants $Top$ choisis dans l'étape précédente.

Notons que $Kcor[k]$, $1 \leq k \leq K$ est un vecteur à 21 composantes dans le cas de graphe dirigé, dont chaque composante désigne la corrélation de Kendall entre une paire de vecteurs de centralités $X^{(i)}$ et $X^{(j)}$ $1 \leq j < i \leq N_c$ parmi par exemple ("degré entrant", "degré sortant", "proximité entrante", "proximité sortante", "intermédiarité", "hub score" et "autorité").

Ainsi le vecteur de corrélation de Kendall $Kcor[k]$ relatif au réseau $P_k$ est de taille $\frac{N_c \times (N_c - 1)}{2}$. Ainsi à chaque réseau $P_k$; $(1 \leq k \leq K)$ correspond un vecteur ligne $Kcor[k]$ à $\frac{N_c \times (N_c - 1)}{2}$ coordonnées (21, si on considère les $N_c = 7$ centralités, dans le cas de graphes orientés).

Nous obtenons ainsi, un recodage des données des graphes de $P$ sous forme de vecteurs $Kcor[k]$, $1 \leq k \leq K$ tous à $\frac{N_c \times (N_c - 1)}{2}$ composantes qui sont les coefficients de corrélations de Kendall entre $N_c$ centralités.

Dans le cas de graphes non dirigés, le nombre de centralités à considérer sera réduit à $N_c = 4$ et les vecteurs $Kcor[k]$ auront 6 composantes seulement.

(d) Former la matrice $Kcor$ de l'ensemble $P$ où chaque ligne représente les corrélations obtenues pour un même réseau et chaque colonne les valeurs des corrélations

entre deux centralités obtenues sur les différents réseaux. Cette matrice se compose de $\frac{N_c \times (N_c - 1)}{2}$ colonnes et de $K$ lignes qui correspond au nombre de réseaux de $P$.
Ainsi $Kcor$ est une nouvelle représentation de l'ensemble de $K$ réseaux de l'étude.

$$Kcor = \begin{pmatrix} Kcor[P_1] \\ Kcor[P_2] \\ \vdots \\ \\ \vdots \\ Kcor[P_K] \end{pmatrix}$$

$Kcor$ est une matrice dont les colones sont les 21 corrélations de Kendall entre les vecteurs de centralités (réduits aux représentants $Top$) nommées à l'aide de la liste des 21 croisements ("degOut x degIn", "cIn x degIn","cOut x degIn", "betw x degIn", "hub x degIn", "aut x degIn", "cIn x degOut", "cOut x degOut" , "betw x degOut", "hub x degOut", "aut x degOut", "cOut x cIn", "betw x cIn", "hub x cIn", "aut x cIn", "betw x cOut", "hub x cOut", "aut x cOut", "hub x betw", "aut x betw", "aut x hub").

Rappelons que nous avons déjà défini une mesure de similarité $Sim$ entre deux réseaux $P_1$ et $P_2$ par :

$$Sim(P_1, P_2) = dist(Kcor[P_1], Kcor[P_2]),\ P_1, P_2 \in P.$$

Et $dist(Kcor)_{K \times K}$ permet d'évaluer toutes les mesures de similarité entre tous les réseaux de $P$ pris par paires et peut prendre la forme d'une matrice symétrique de diagonale nulle et dont les éléments sont : $Sim(P_k, P_l),\ 1 \leq k \neq l \leq K$.

## 7.3 Validation : Clustering des familles de graphes artificiels

Afin de tester les performances et valider l'efficacité cette nouvelle mesure de similarité, nous proposons de la mettre en oeuvre pour effectuer un clustering sur plusieurs ensemble de graphes différents. Une étape importante de ce travail est celle du recodage des données qui consiste à construire le coefficient synthétique de centralité pour chaque réseau.

### 7.3.1 Cas de plusieurs types de graphes de mêmes ordres

#### 7.3.1.1 Description du jeu de données

Ce cas d'étude s'intéresse à un jeu de données $P_{ER-SW-BA}$ qui est un ensemble de graphes tous d'ordre $N = 1000$ sommets, générés de $cl = 3$ manières différentes. $P_{ER-SW-BA}$ est composé de $m = 40$ graphes de chaque type.
Les 3 types de réseaux générés sont :

- des graphes Erdös- Rényi générés avec un même paramètre $p_{ER}$ qui représente la probabilité d'existence d'un lien entre deux sommets que nous fixons à $p_{ER} = 0.01$
- des graphes de type Small-World de Watts et Strogatz générés avec les mêmes paramètres une dimension de 1, une constante de voisinage $nei_{SW} = 5$ et une probabilité de reconnexion $p_{SW} = 0.05$.
- des graphes de type invariants d'échelle de Barabasi-Albert générés avec le même paramètre : une puissance d'attachement préférentiel égale $pwr_{BA} = 0.1$.

#### 7.3.1.2 Recodage des données

Les mesures de centralités utilisées sont les $N_c = 7$ centralités présentées dans Centralités à savoir le "degré entrant", "degré sortant", "proximité entrante", "proximité sortante", "intermédiarité", "hubité" et l'"autorité".
Pour les 120 graphes $P_k$ $(1 \leq k \leq 120)$ de ce jeu de données, nous appliquons la méthode de construction des vecteurs de corrélations (cf. Mesure de similarité entre réseaux) $Kcor[k]_{1 \leq k \leq 120}$ de Kendall composés chacun de 21 composantes qui sont les corrélations de Kendall entre les diffrents vecteurs de centralités :

$$Kcor^{(l)}[k] = cor(X^{(i)}[k], X^{(j)}[k]) \ (1 \leq k \leq 120;\ 1 \leq j < i \leq 7;\ 1 \leq l \leq 21).$$

Notons que pour construire le vecteur de corrélations $Kcor$, nous choisissons un nombre de meilleures valeurs à retenir pour chaque centralité $X^{(j)}$ égale à 7 $(c = 7)$

et donc $top_7^{(j)}$ si le nombre de sommets concernés par ces valeurs est inférieur ou égale à 10 (i.e $card(top_7^{(j)}) \leq 10$), sinon on se restreint uniquement aux sommets ayant les 5 meilleures valeurs ($c = 5$ et donc $top_5^{(j)}$ pour la centralité $X^{(j)}_{1 \leq j \leq N_c=7)}$. Ces sommets (constituant l'ensemble $Top$) sont considérés comme les représentants du réseau $P_k$ étudié.
Ainsi nous obtenons la représentation de l'ensemble de 120 réseaux sous forme de la matrice $Kcor$ telle que décrit ci-dessus.
L'étape suivante dans la mise en œuvre d'un clustering consiste à déterminer le nombre optimal de clusters à considérer. Plusieurs méthodes peuvent effectuer cette tâche. Le paragraphe suivant présente quelques-unes.

#### 7.3.1.3 Exploration du nombre de clusters du jeu de données $P_{ER-SW-BA}$

**Méthode "Elbow"[79]**
La figure 7.2 présente le nombre optimal de clusters pour le jeu de données $P_{ER-SW-BA}$.

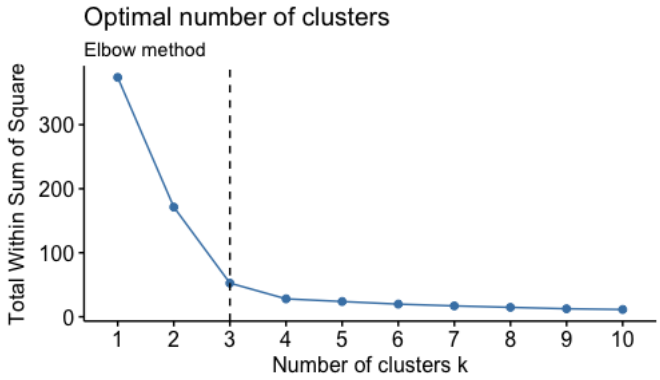



FIGURE 7.2 – Nombre optimal de clusters par la méthode Elbow pour $P_{ER-SW-BA}$.

Cette méthode propose un nombre optimal de clusters $K = 3$ qui correspond aux 3 types de données générés, reste à vérifier qu'il s'agit des mêmes clusters pour chaque type.

**Méthode "Silhouette" [58]**
La méthode Silhouette propose aussi un nombre optimal de clusters tel que représenté dans la figure 7.3.

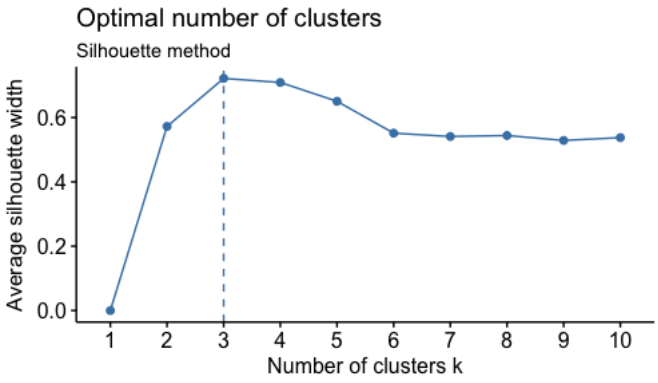



FIGURE 7.3 – Nombre optimal de clusters par la méthode Silhouette pour $P_{ER-SW-BA}$.

**Méthode "Choix du nombre optimal en faisant voter 20 indices"**
Selon le vote majoritaire, 12 sur 20 votants qui sont différents indices, le meilleur nombre de clusters est également $K = 3$ (voir figure 7.4). La liste de ces 12 indices est la suivante : "Hartigan" [43],"Scott" [76] ,"Marriott" [66], "TrCovW" [69], "TraceW" [69], "Friedman" [37], "Cindex" [49], "DB" [26], "Silhouette" [58], "Ratkowsky" [73], "Ball" [6] et "SDindex" [42].

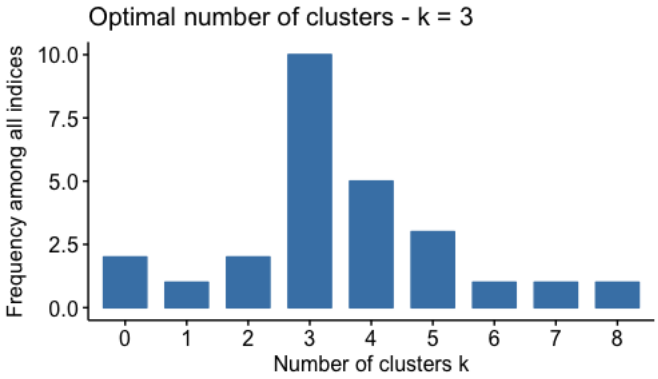



FIGURE 7.4 – Vote majoritaire pour nombre optimal de clusters pour $P_{ER-SW-BA}$.

Les deux figures 7.5 correspondent à l'indice de Hubert [50], et l'indice D [63] où le nombre optimal de clusters est celui qui permet de former un coude.

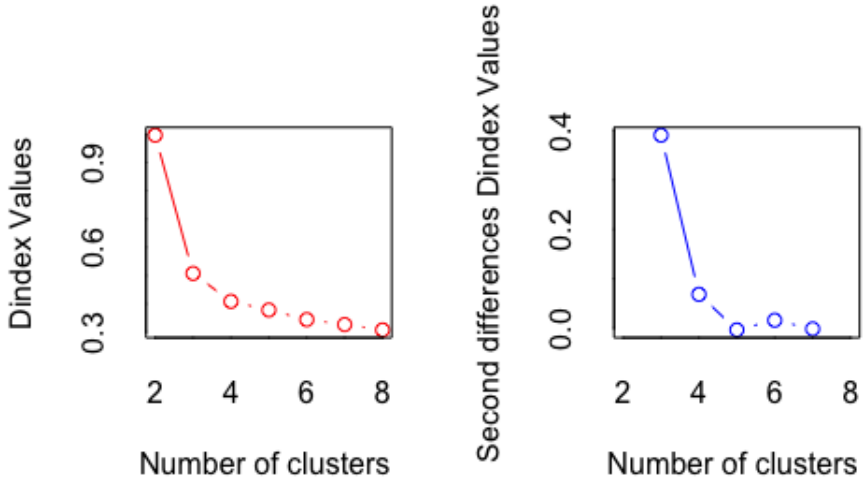



FIGURE 7.5 – Indice de Hubert et indice D pour déterminer du nombre optimal de clusters pour $P_{ER-SW-BA}$.

En conclusion, ces différentes méthodes permettent de retrouver le nombre de clusters de départ. Dans le paragraphe suivant, nous présentons une méthode qui permet de visualiser les données (multidimensionnelles) initiales, sélectionné le nombre optimal de clusters et de les déterminer entièrement.

#### 7.3.1.4 Clustering du jeu de données $P_{ER-SW-BA}$

Le paragraphe suivant montre qu'il est parfois possible d'obtenir des résultats satisfaisants en effectuant une CAH sans passer par l'étape ACP. Cependant, étant donné le caractère multidimensionnel des données, aucune représentation graphique n'est possible dans ce cas.

**Clustering Ascendant Hiérarchique (CAH) sur les Composantes principales de $P_{ER-SW-BA}$**

Nous effectuons une Analyse en Composantes Principales (ACP) pour mieux comprendre le jeu de données et pour réduire tout potentiel bruit (variables de faibles influences).

La figure 7.6 représente la projection des différents réseaux du jeu de données $P_{ER-SW-BA}$ sur les deux composantes (dimensions) les plus importantes constituant ainsi le premier plan factoriel.

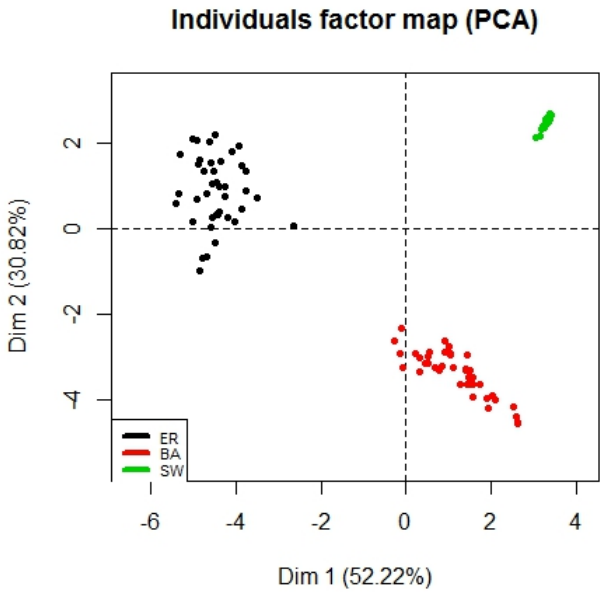



FIGURE 7.6 – Projection des différents réseaux du jeu de données $P_{ER-SW-BA}$

Nous remarquons que les réseaux de type Small-World (SW) représentés en vert sont bien compacts, ceux de type Barabasi-Albert (BA) en rouge sont moyennement dispersés tandis que les réseaux Erdös- Rényi (ER) en noir sont beaucoup plus dispersés. Néanmoins, les 3 types de réseaux sont bien différenciés avec ces deux dimensions et vraiment séparés. Nous effectuons une CAH sur les composantes principales issues de l'ACP. La figure 7.7 représente le dendrogramme obtenu, en utilisant la méthode de Ward.

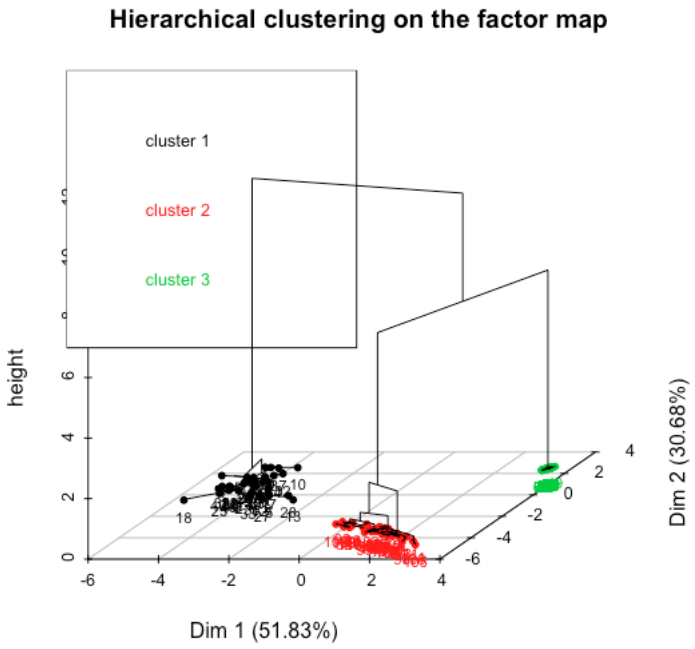



FIGURE 7.7 – Dendrogramme CAH+ACP sur $P_{ER-SW-BA}$

Confrontons à présent ce clustering avec les classes initialement générées dans $P_{ER-SW-BA}$. Le tableau 7.1 présente en ligne les classes générées dans $P_{ER-SW-BA}$ et en colones les clusters obtenus par CAH+ACP sur $P_{ER-SW-BA}$.

| clusters.acp-cah | 1 | 2 | 3 |
|---|---|---|---|
| ER | 40 | 0 | 0 |
| BA | 0 | 39 | 0 |
| SW | 0 | 0 | 40 |

TABLE 7.1 – Confrontation du clustering avec CAH+ACP avec classes de $P_{ER-SW-BA}$

Nous remarquons que les classes initialement générées dans le jeu de données sont parfaitement retrouvées par ce clustering sur les composantes principales retenues par l'ACP.

**Clustering Ascendant Hiérarchique sur** $P_{ER-SW-BA}$

Nous effectuons une CAH sur toutes les variables de $P_{ER-SW-BA}$ en utilisant le critère de Ward comme méthode d'aggrégation et nous obtenons un dendrogramme avec une multitude de branches mais montrant clairement que le nombre de classes qui maximise le saut d'intertie est $K = 3$, nous optons ainsi pour ce clustering. Confrontons à présent les groupes obtenus par ce clustering avec les classes générées dans le jeu de données. Le tableau 7.2 illustre cette confrontation, où les lignes sont

les classes du jeu de données et les colonnes les clusters obtenus par la CAH sur toutes les variables du jeu de données $P_{ER-SW-BA}$.

| clusters.cah | 1 | 2 | 3 |
|---|---|---|---|
| ER | 40 | 0 | 0 |
| BA | 0 | 39 | 0 |
| SW | 0 | 0 | 40 |

TABLE 7.2 – Confrontation du clustering avec CAH et les classes de $P_{ER-SW-BA}$.

Nous remaquons que les résultats obtenus par cette confrontation sont identique que le cas précédent où la CAH a été réalisée sur les résultats de l'ACP. En effet, on constate une "erreur d'affectation" nulle. Ainsi, nous considérons l'utilisation l'ACP avant le clustering pour $P_{ER-SW-BA}$ très facultative. Dans ce cas, néanmoins elle permet de représenter différents graphes dans le plan.

#### 7.3.1.5 Classification du jeu de données $P_{ER-SW-BA}$

Dans ce paragraphe, nous proposons de réaliser un apprentissage supervisé sur le jeu de données $P_{ER-SW-BA}$, cette fois-ci en l'étiquettant avec les classes relatives au modèle de génération de départ. Ainsi, $P_{ER-SW-BA}$ contient 3 classes à retrouver par un modèle de classification.
Pour se faire, nous utilisons plusieurs méthodes à base d'arbre de décision. Le modèle est bâti sur l'échantillon d'apprentissage (2/3 formé aléatoirement à partir des données initiales) et validé sur l'échantillon de test (1/3 des données qui constituent les données restantes non utilisées pour l'apprentissage). Le tableau 7.3 présente les différentes proportions classes dans les échantillons d'apprentissage et de test.

| Echantillon | ER | SW | BA |
|---|---|---|---|
| Apprentissage | 28 | 26 | 26 |
| Test | 12 | 14 | 14 |

TABLE 7.3 – Représentation des classes dans les échantillons d'apprentissage et de test de $P_{ER-SW-BA}$

**Classification par la méthode CART**

La figure 7.8 représente l'arbre obtenu sur les données d'apprentissage.

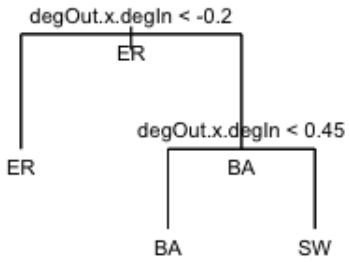



FIGURE 7.8 – Arbre de classification CART sur l'ensemble d'apprentissage

Nous remarquons que cette méthode ne prend en compte qu'une seule variable (corrélation entre le degré entrant et le degré sortant) : seules la loi des degrés entrants et celle des degrés sortants sur les 7 meilleures valeurs permettent de différencier et séparer les 3 types de graphes. L'erreur d'apprentissage obtenue dans ce cas est nulle. En effectuant la prédiction sur l'échantillon de test, l'erreur de prédiction est aussi nulle comme illustrée dans le tableau 7.4 de prédiction.

| Prédictions | ER | SW | BA |
|---|---|---|---|
| ER | 12 | 0 | 0 |
| SW | 0 | 14 | 0 |
| BA | 0 | 0 | 14 |

TABLE 7.4 – Prédictions obtenues par CART sur le jeu de données $P_{ER-SW-BA}$

Les prédictions obtenues sont parfaites (erreur de prédiction nulle), bien que l'arbre de classification est connu par sa sensiblibité aux données utilisées pour l'apprentissage. Dans ce cas étant donné que les classes sont équitablement représentées (aléatoirement) dans la phase d'apprentissage cela ne constitue aucun problème. De plus comme nous l'avons remarqué lors de l'intepretation de la figure 7.6, la projection des différents réseaux sur le plan factoriel met en évidence le fait que les différentes classes sont vraiment séparées ce qui permet d'obtenir un résultat de classification parfait. Les méthodes Bagging, Gradient Boosted Machine et Random Forest donnent les mêmes résultats et ne sont pas nécessaires pour la classification de $P_{ER-SW-BA}$.

### 7.3.2 Cas de plusieurs types de graphes de tailles différentes

#### 7.3.2.1 Description du jeu de données

Ce cas d'étude s'intéresse à un jeu de données $P'_{ER-SW-BA}$ qui est un ensemble de graphes générés presque comme $P_{ER-SW-BA}$, avec $cl = 3$ types de générateurs. Les 3 types de réseaux générés sont :

— des graphes Erdös- Rényi générés avec un même paramètre $p_{ER}$ qui représente la probabilité d'existance d'un sommet entre deux sommets que nous fixons cette expérience à $p_{ER} = 0.01$

— des graphes de type Small-World de Watts et Strogatz générés avec une même dimension de 1, une constante de voisinage $nei_{SW} = 5$ et une probabilité de reconnexion $p_{SW} = 0.05$.

— des graphes de type sans-échelle de Barabasi-Albert générés avec une meême puissance d'attachement préférentiel égale $pwr_{BA} = 2$.

Cependant la spécificité de $P'_{ER-SW-BA}$ est le fait que chaque classe contient des réseaux de tailles $N$ (nombre de sommets) différentes. En effet, chaque classe contient $m = 30$ réseaux de tailles $N^{(1)} = 1000$, $N^{(2)} = 2000$ et $N^{(3)} = 3000$, 10 de chaque tailles, pour chacune des 3 classes ("ER", "SW" et "BA") qui sont "hétérogènes" dans ce cas.

#### 7.3.2.2 Recodage des données

Le recodage des données de $P'_{ER-SW-BA}$ est fait exactement de la même manière que dans le cas d'étude sur $P_{ER-SW-BA}$ présenté dans le paragraphe Recodage des données. La matrice $Kcor = t(Kcor[1], ..., Kcor[K])$ qui regroupe tous les coefficients synthétiques de centralités a maintenant 21 colonnes comme précédemment mais $K = 90$ lignes, le nombre de graphes étudiés.

#### 7.3.2.3 Exploration du nombre optimal de clusters de $P'_{ER-SW-BA}$

**Méthode "Elbow" [79]**

La figure 7.9 représente le nombre optimal de clusters pour le jeu de données $P'_{ER-SW-BA}$.

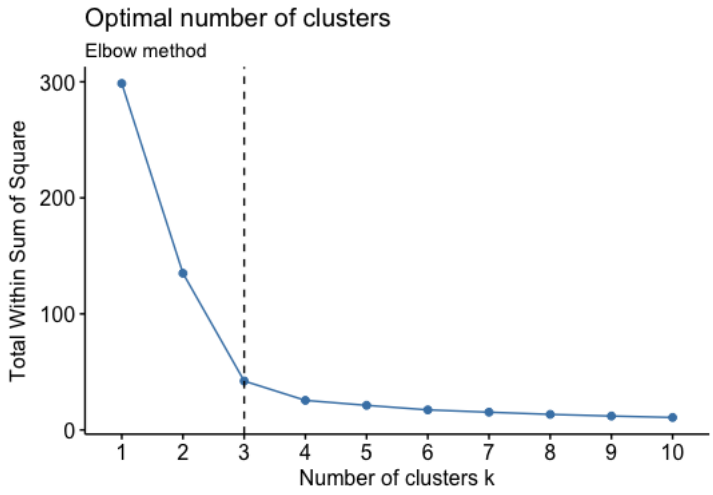



FIGURE 7.9 – Nombre optimal de clusters par la méthode Elbow $P'_{ER-SW-BA}$.

Cette méthode propose un nombre optimal de clusters $K = 3$ qui correspond au types de données générés au départ, reste à vérifier s'il s'agit des mêmes clusters pour chaque type.

**Méthode "Silhouette"**

La méthode Silhouette propose aussi un nombre optimal de clusters tel que représenté dans la figure 7.10.

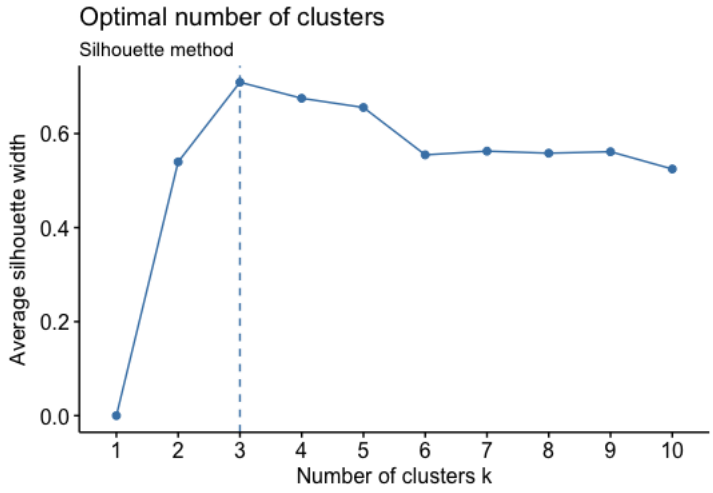



FIGURE 7.10 – Nombre optimal de clusters par la méthode Silhouette $P'_{ER-SW-BA}$.

#### 7.3.2.4 Clustering du jeu de données $P'_{ER-SW-BA}$

**Clustering Ascendant Hiérarchique sur les Composantes principales de $P'_{ER-SW-BA}$**

Nous effectuons une Analyse en Composantes Principales pour ne garder que les composantes à fortes influences dans la discrimination du jeu de données $P'_{ER-SW-BA}$.
La figure 7.11 représente la projection des différents réseaux du jeu de données $P'_{ER-SW-BA}$ sur les deux composantes (dimensions) les plus importantes constituant ainsi le premier plan factoriel.

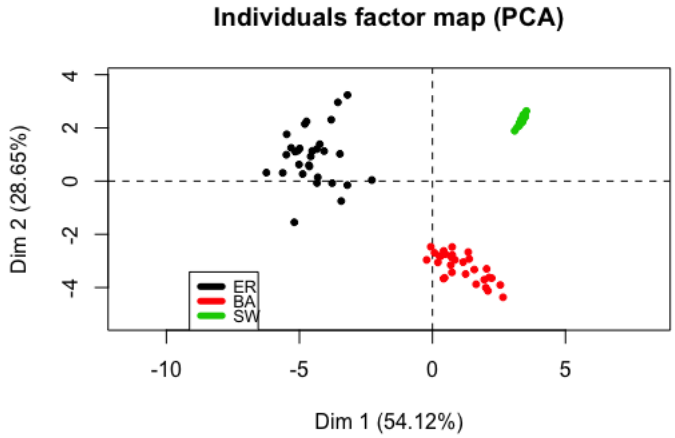



FIGURE 7.11 – Projection des différents réseaux du jeu de données $P'_{ER-SW-BA}$

Ici aussi, nous remarquons que les réseaux type Small-World (SW) en vert restent compactes et ceux de Barabasi et Albert (BA) représentés en rouge sont moyennement dispersés, tandis que les réseaux Erdös- Rényi (ER) en noir sont toujours beaucoup plus dispersés. Néanmoins, les 3 types de réseaux sont bien différenciés avec ces deux dimensions et vraiment séparés. Notons que la première dimension représente environ 54% de l'information et la seconde y ajoute presque 29% ce qui fait environ 83% de l'ensemble de l'information reprise par deux composantes.

Nous effectuons une CAH sur les composantes principales issues de l'ACP. La figure 7.12 représente le dendrogramme obtenu, en utilisant la méthode de Ward.


FIGURE 7.12 – Dendrogramme CAH+ACP sur $P'_{ER-SW-BA}$ et Sauts d'inerties

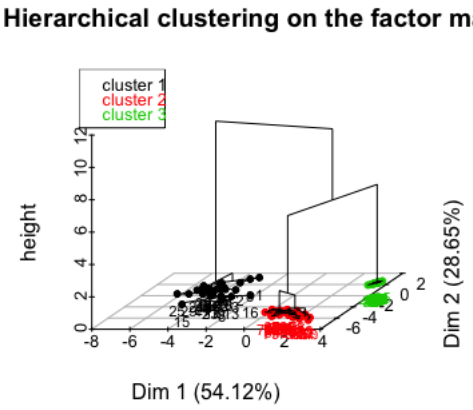



FIGURE 7.13 – Dendrogramme CAH+ACP sur $P'_{ER-SW-BA}$

Ce dendrogramme propose une coupe optimale à 3 classes qui correspond au saut le plus important dans la courbe d'inertie.

Confrontons à présent ce clustering avec les classes générées dans $P'_{ER-SW-BA}$. Le tableau 7.5 représente en ligne les classes générées dans $P'_{ER-SW-BA}$ et en colonnes les clusters obtenus par CAH+ACP sur $P'_{ER-SW-BA}$.

Nous remarquons que ce clustering matérialise parfaitement les classes qui ont été générées dans le jeu de données $P'_{ER-SW-BA}$.

**Clustering Ascendante Hiérarchique sur $P'_{ER-SW-BA}$**

Les clusters obtenus par une CAH sur toutes les variables de $P'_{ER-SW-BA}$ en utilisant

| clusters.acp-cah | 1 | 2 | 3 |
|---|---|---|---|
| BA | 0 | 30 | 0 |
| ER | 30 | 0 | 0 |
| SW | 0 | 0 | 30 |

TABLE 7.5 – Confrontation du clustering avec CAH+ACP avec classes de $P'_{ER-SW-BA}$

le critère de Ward sont confrontés avec les classes générées au départ. Le tableau 7.6 résume cette confrontation.

| clusters.cah | 1 | 2 | 3 |
|---|---|---|---|
| BA | 0 | 0 | 30 |
| SW | 0 | 30 | 0 |
| ER | 30 | 0 | 0 |

TABLE 7.6 – Confrontation du clustering avec CAH et les classes de $P'_{ER-SW-BA}$.

Nous remaquons que les résultats avec et sans ACP donnent les mêmes performances. En effet, on constate une " erreur d'affectation" nulle pour les deux cas.
Ainsi, nous retenons l'utilisation de l'ACP avant le clustering pour $P'_{ER-SW-BA}$ n'est pas particulièrement pertinente.

#### 7.3.2.5 Classification du jeu de données $P'_{ER-SW-BA}$

Ce paragraphe présente l'apprentissage supervisé réalisé sur le jeu de données $P'_{ER-SW-BA}$ qui contient également, tout comme $P_{ER-SW-BA}$, 3 classes à prédire par les méthodes à base d'arbre de décision.
La construction du modèle se fait encore sur 2/3 des données sélectionnés aléatoirement, et la validation sur le 1/3 restant. Le tableau 7.7 résume les effectifs de chaque échantillon.

| Echantillon | ER | SW | BA |
|---|---|---|---|
| Apprentissage | 22 | 19 | 19 |
| Test | 8 | 11 | 11 |
| $P'_{ER-SW-BA}$ | 30 | 30 | 30 |

TABLE 7.7 – Représentation des classes dans les échantillons d'apprentissage et de test de $P'_{ER-SW-BA}$

**Classification par la méthode CART**

La figure 7.14 représente l'arbre obtenu sur les données d'apprentissage.

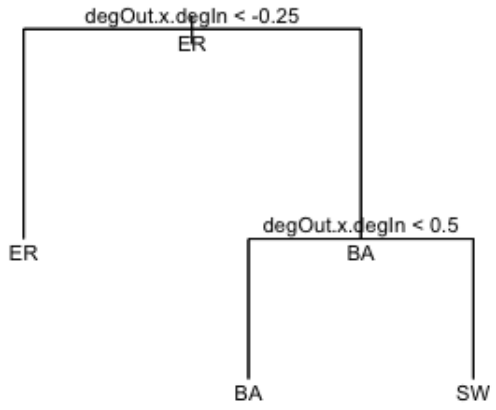



FIGURE 7.14 – Arbre de classification CART sur l'ensemble d'apprentissage de $P'_{ER-SW-BA}$

Nous remarquons que tout comme sur le jeu de données précédent, cette méthode ne prend en compte qu'une seule variable (corrélation entre le degré entrant et le degré sortant) pour séparer les 3 classes. En premier lieu, elle isole la classe Erdös- Rényi, puis sépare les deux autres. L'erreur d'apprentissage est comme dans le cas précédent nulle. Nous effectuons la prédiction sur l'échantillon de test. Le tableau 7.8 illustre les prédictions obtenues.

| Prédictions | ER | SW | BA |
|---|---|---|---|
| ER | 7 | 0 | 1 |
| SW | 0 | 11 | 0 |
| BA | 0 | 0 | 11 |

TABLE 7.8 – Prédictions obtenues par CART sur le jeu de données $P'_{ER-SW-BA}$

Contrairement aux résultats sur $P_{ER-SW-BA}$, les prédictions sur $P'_{ER-SW-BA}$ ne sont pas parfaites . En effet, nous observons d'après le tableau 7.8 une erreur de prédiction de 3.33% qui reste une valeur faible dans la mesure où l'arbre se trompe uniquement

pour classifier un des réseaux de type Erdös- Rényi qu'il considère de type Barabasi-Albert. Les méthodes Bagging et Random Forest donnent la même valeur d'erreur. Cependant, la méthode Gradient Boosted Machine élimine cette erreur de prédiction. Les 3 sections suivantes traitent le clustering de réseaux de même type ayant des paramètres différents. Plus précisément, nous vérifions que la mesure de similarité définie dans la section Mesure de similarité entre réseaux permet de différencier entre réseaux ayant des paramètres différents, même s'ils sont de même type.

### 7.3.3 Cas de réseaux de type "Small-World"

Pour tester notre approche sur des réseaux de type "Small-World", nous générons aléatoirement un ensemble de réseaux $P_{SW}$ tous d'ordre $N = 1000$ (nombre de sommets), de $cl = 3$ manières différentes.

#### 7.3.3.1 Description du jeu de données

Pour tester notre approche sur des réseaux de type "Small-World", nous générons dans un premier temps 3 réseaux "Small-World" ayant le même ordre $N = 1000$ (nombre de sommets), une même valeur de dimension $dim_{SW} = 1$, une même constante de voisinage $nei_{SW} = 5$ mais, des probabilités de recablage $p_{SW}$ différentes.

- Le 1er réseau autour duquel sera générée la classe $cl1$ à $p1 = 0.0075$
- Le 2ème réseau permet de générer la classe $cl2$ à $p2 = 0.0125$
- Le 3ème réseau produit la classe $cl3$ à $p3 = 0.0175$

Ensuite, les réseaux de chaque classe sont obtenus par percolation du réseau graine, en le pertubant chaque fois et supprimant aléatoirement $perturb = 2\%$ de ses liens. Cela est réalisé $m = 40$ fois pour générer les 40 réseaux de chaque classe.

Ainsi $P_{SW}$ est composé de $m = 40$ graphes chacun généré avec un même $p_{SW}$, et perturbé de 2%. Ce tau de perturbation *perturb* est bien entendu fixe et peut être changé pour former un autre jeu de données. La perturbation peut aller jusqu'à 5%. Au-dela, les clusters commencent à se chevaucher et le nombres de clusters devient difficile à déterminer. Cependant, les prédictions par les différentes méthodes restent de bonne qualité si on connait le nombre de classes.

#### 7.3.3.2 Recodage des données

Le recodage des données de $P_{SW}$ se fait exactement de la même manière que dans le cas d'étude sur $P_{ER-SW-BA}$ présentée dans le paragraphe Recodage des données.

#### 7.3.3.3 Exploration du nombre de clusters du jeu de données $P_{SW}$

**Méthode "Elbow" (coude) [79]**

La figure 7.15 présente le nombre optimal de clusters pour le jeu de données $P_{SW}$.

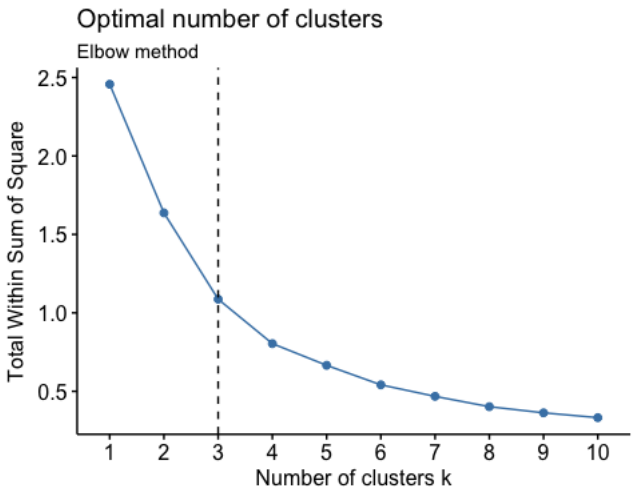



FIGURE 7.15 – Nombre optimal de clusters par la méthode Elbow $P_{SW}$

Nous remarquons que cette méthode donne un nombre optimal de clusters $K = 3$, malgré que le coude n'est pas vraiment clair.

**Méthode "Silhouette"**

La figure 7.16 représente le nombre optimal de clusters pour le jeu de données $P_{SW}$ selon la méthode "Silhouette" décrite dans le paragraphe Méthodes de détermination du nombre de clusters.

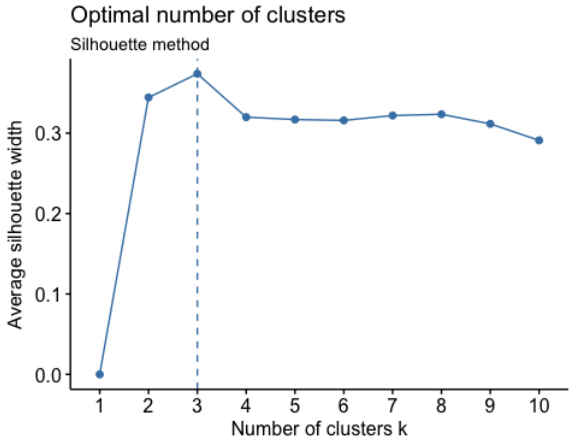



FIGURE 7.16 – Nombre optimal de clusters par la méthode Silhouette $P_{SW}$.

Nous remarquons que cette méthode donne aussi un nombre optimal de clusters $K = 3$.

Nous avons également fait voter un ensemble de 20 indices présents dans la littérature pour choisir le nombre optimal de clusters et le résultat par vote majoritaire donne aussi $K = 3$ avec 7 votants sur 20 qui sont : "Hartigan" [43], "TrCovW" [69], "TraceW" [69], "Silhouette" [58], "Ball" [6], "PtBiserial" [68, 67] et "SDindex" [42]. Reste à confirmer si les clusters correspondent vraiment aux classes de réseaux générées.

#### 7.3.3.4 Clustering du jeu de données $P_{SW}$

**Clustering Ascendant Hiérarchique sur les Composantes principales de $P_{SW}$**

Nous effectuons une Analyse en Composantes Principales pour mieux comprendre le jeu de données et pour réduire tout potentiel bruit (variables de faibles influences).
La figure 7.17 représente la projection des différents réseaux du jeu de données $P_{SW}$ sur les deux composantes (dimensions) les plus importantes constituant ainsi le premier plan factoriel. Ce plan est formé par une première dimension représantant environ 73% des données, puis une seconde qui y contribue avec environ 18%, par conséquent ce premier plan factoriel reprend 91% des informations.

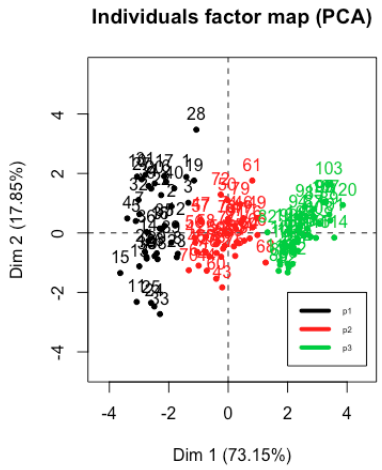



FIGURE 7.17 – Projection des différents réseaux du jeu de données $P_{SW}$

Les réseaux représentés en noir correspondent à ceux générés avec $p1 = 0.0075$, en rouge ceux avec $p2 = 0.0125$ et ceux en vert sont ceux avec $p3 = 0.0175$. Nous observons que chaque groupe occupe une région verticalement séparée du plan factoriel. Les réseaux en noir sont plus dispersés que les autres.

Nous effectuons une CAH sur les composantes principales issues de l'ACP. La figure 7.18 représente le dendrogramme obtenu en utilisant la méthode de Ward.

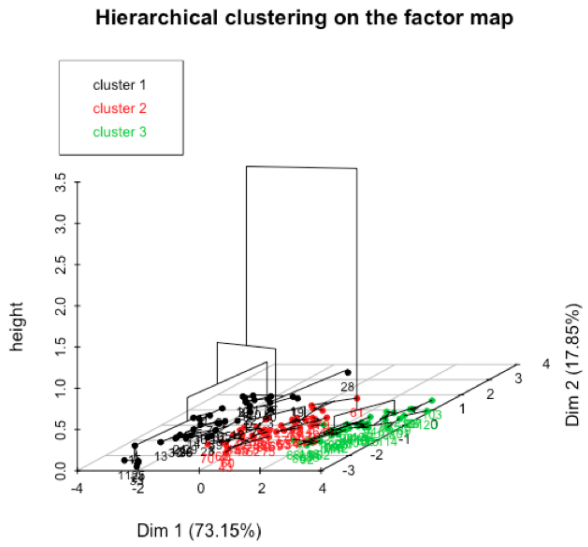



FIGURE 7.18 – Dendrogramme CAH+ACP sur $P_{SW}$

Le tableau 7.9 confronte les clusters obtenus par la CAH sur les composantes principales retenues par l'ACP. Nous remarquons que sur les classes générées avec les probabilités $p1 = 0.0075$ et $p3 = 0.0175$ sont parfaitement retrouvées, quant à la classe $p2 = 0.0125$, 1 individu est considéré appartenant à la 3ème. Ainsi, l'erreur est de 0.83%.

| clusters.acp-cah | 1 | 2 | 3 |
|---|---|---|---|
| Cl1 | 40 | 0 | 0 |
| Cl2 | 0 | 39 | 1 |
| Cl3 | 0 | 0 | 40 |

TABLE 7.9 – Confrontation des clusters obtenus par CAH après ACP avec classes générées dans $P_{SW}$

Nous remarquons que ce clustering sur ACP donne des résultats presque parfait.

**Clustering Ascendante Hiérarchique sur données brutes $P_{SW}$**

Nous effectuons une CAH sur toutes les variables de $P_{SW}$ en utilisant le critère de Ward comme méthode d'agrégation .Le dendrogramme obtenu est difficilement visualisable à cause des multiples branches néamoins, il permet de voir que le meilleur saut d'intertie est réalisé avec $K = 3$ clusters.

Confrontons à présent les clusters obtenus aux classes générées dans le jeu de données initial ($p1 = 0.0075$, $p2 = 0.0125$ et $p3 = 0.0175$). Le tableau 7.10 présente en colonnes les clusters obtenus et en ligne les classes initialement générées dans le jeux de données $P_{SW}$.

| clusters.cah | 1 | 2 | 3 |
|---|---|---|---|
| Cl1 | 26 | 14 | 0 |
| Cl2 | 33 | 5 | 2 |
| Cl3 | 1 | 0 | 39 |

TABLE 7.10 – Confrontation clusters avec classes générées dans $P_{SW}$

D'après le tableau 7.10, une "erreur de clustering" est observée et représente 28.33% majoritairement faite sur la classe $Cl1$ ($p1 = 0.0075$) ensuite sur $Cl2$ ($p2 = 0.0125$).

En confrontons les deux clustering, nous notons que les résultats avec ACP sont nettement meilleurs, ce qui confirme la pertinence d'utiliser l'ACP pour ce jeu de données.

#### 7.3.3.5 Classification du jeu de données $P_{SW}$

Dans ce paragraphe, nous proposons de réaliser un apprentissage supervisé sur le jeu de données $P_{SW}$ cette fois-ci en l'étiquettant avec les classes relatives à la probabilité $p_{SW}$ utilisées pour générer le réseau graine. Ainsi, $P_{SW}$ contient 3 classes à retrouver par un modèle de classification.
Pour se faire, nous utilisons plusieurs méthodes à base d'arbre de décision. Le modèle est bâti sur un échantillon d'apprentissage (2/3 formé aléatoirement à partir des données initiales) et validé sur l'échantillon de test (1/3 des données qui constituent les données restantes non utilisées pour l'apprentissage).
Le tableau 7.11 présente les différentes classes dans les échantillons d'apprentissage et de test.

| Echantillon | Cl1 | Cl2 | Cl3 |
|---|---|---|---|
| Apprentissage | 28 | 26 | 26 |
| Test | 12 | 14 | 14 |
| $P_{Sw}$ | 40 | 40 | 40 |

TABLE 7.11 – Représentation des classes dans les échantillons d'apprentissage et de test de $P_{SW}$

Nous remarquons que les classes sont presque également représentées dans les échantillons d'apprentissage et de test.

**Classification par la méthode CART**

A présent réalisons une classification en utilisons la méthode d'arbre de décision. La figure 7.19 représente l'arbre obtenu sur les données d'apprentissage.

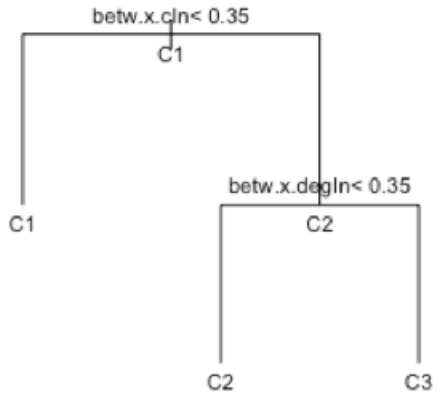



FIGURE 7.19 – Arbre de classification CART sur l'ensemble d'apprentissage de $P_{SW}$

Nous remarquons dans la figure 7.19 que cette méthode de classification utilise une variable $betw \times cIn$ (corrélation entre l'intermédiarité et la proximité entrante) pour séparer la classe $Cl1$ des deux autres puis la variable $betw \times degIn$ (corrélation entre l'intermédiarité et le degré entrant) pour séparer les deux classes $Cl2$ et $Cl3$. L'erreur d'apprentissage dans ce cas est faible est égale à 0.025, soit environ 2.5%.
Nous effectuons la prédiction sur l'échantillon de test, l'erreur de prédiction est aussi nulle comme illustrée dans le tableau 7.12 de prédiction.

| Prédictions | Cl1 | Cl2 | Cl3 |
|---|---|---|---|
| Cl1 | 12 | 0 | 0 |
| Cl2 | 0 | 14 | 0 |
| Cl3 | 0 | 0 | 13 |

TABLE 7.12 – Prédictions obtenues par CART sur le jeu de données $P_{SW}$

Les prédictions obtenues sont parfaites sur toutes les classes et l'erreur de prédiction est nulle. Bien que l'arbre de classification est connu par sa sensibilité aux données utilisées pour l'apprentissage. Dans notre cas étant donné que les classes sont équi-représentées (aléatoirement) dans la phase d'apprentissage cette sensibilité ne consti-

tue pas de problème. Les méthodes Bagging, Gradient Boosted Machine et Random Forest donnent les mêmes résultats et ne sont pas nécessaires pour la classification de $P_{SW}$.
L'approche a également été testé avec un jeu de données de même type, mais contenant 4 classes, les résultats obtenus sont aussi satisfaisants.

### 7.3.4 Cas de réseaux de type "Barabasi-Albert"

Pour tester notre approche sur des réseaux de type invariants d'échelle de Barabasi-Albert, nous générons un ensemble de réseaux $P_{BA}$ tous d'ordre $N = 1000$ (nombre de sommets), de $cl = 3$ manières différentes.

#### 7.3.4.1 Description du jeu de données

Dans un premier temps, nous générons 3 réseaux Barabasi-Albert dirigés en utilisant le même algorithme (psumtree), mais avec des puissances d'attachement préférentiel $power_{BA}$ différentes

- Le 1er réseau autour duquel sera construite la classe $cl1$ a $power1 = 1$
- Le 2ème réseau permet de construire la classe $cl2$ a $power2 = 2$
- Le 3ème réseau produit la classe $cl3$ a $power3 = 3$

Ensuite, les réseaux de chaque classe sont obtenus par percolation du réseau graine, en le pertubant chaque fois et supprimant aléatoirement $perturb = 2\%$ de ses liens. Cela est réalisé $m = 40$ fois pour obtenir les 40 réseaux de chaque classe.
Ainsi $P_{BA}$ est composé de 120 graphes répartis en 3 classes, chacune est formée $m = 40$ graphes ayant un même $power_{BA}$ et obtenus par perturbation de 2% de l'un des réseaux graines tel que décrits ci-dessus.
Ce tau de perturbation $perturb$ est bien entendu fixe et peut être changé pour former un autre jeu de données.
La perturbation peut aller jusqu'à 5%. Au-delà, les clusters commencent à se chevaucher et le nombres de clusters devient difficile à déterminer. Cependant, les prédictions par les différentes méthodes restent de bonne qualité si on connait le nombre de classes.

#### 7.3.4.2 Recodage des données

Le recodage des données de $P_{BA}$ se fait exactement de la même manière que dans les cas précédents.

#### 7.3.4.3 Exploration du nombre de clusters du jeu de données $P_{BA}$

**Méthode "Choix du nombre optimal en faisant voter 20 indices"**
Selon le vote majoritaire, 6 sur 20 votants qui sont différents indices, le meilleur nombre de clusters est également $K = 3$ (voir figure 7.20). La liste des 6 indices est la suivante : "Hartigan" [43], "Marriott" [66], "TrCovW" [69], "TraceW" [69], "Ball" [6] et "PtBiserial" [68, 67].

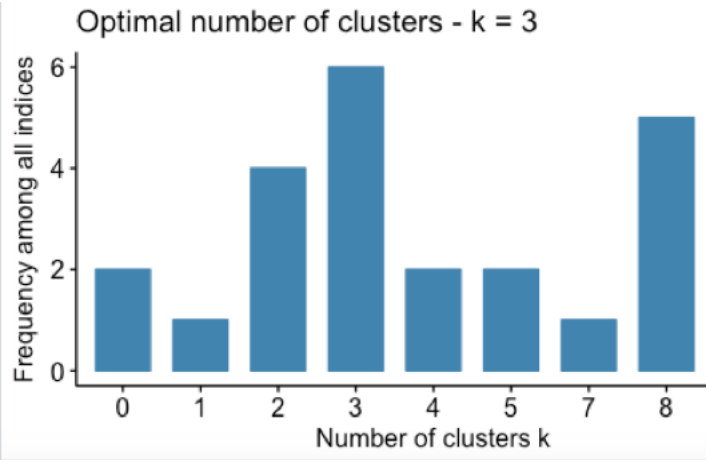



FIGURE 7.20 – Vote majoritaire pour nombre optimal de clusters de $P_{BA}$

#### 7.3.4.4 Clustering du jeu de données $P_{BA}$

**Clustering Ascendant Hiérarchique sur les Composantes principales de $P_{BA}$**
Nous effectuons une Analyse en Composantes Principales pour mieux comprendre le jeu de données et pour réduire tout potentiel bruit (variables de faibles influences).
La figure 7.21 représente la projection des différents réseaux du jeu de données $P_{BA}$ sur les deux composantes (dimensions) les plus importantes constituant ainsi le premier plan factoriel.

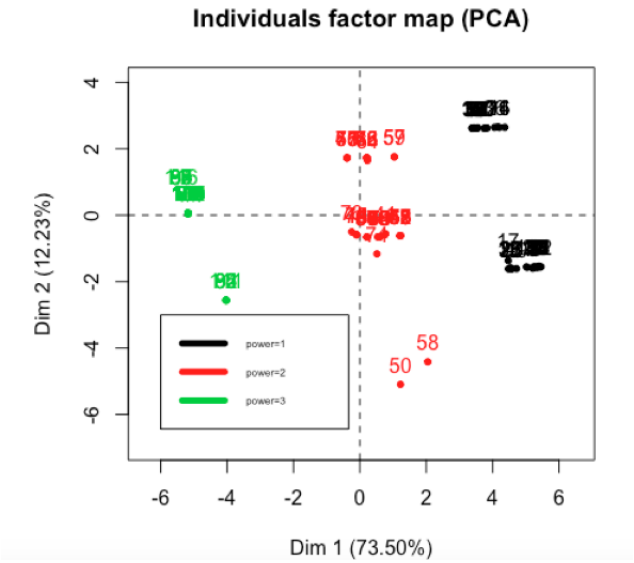



FIGURE 7.21 – Projection des différents réseaux du jeu de données $P_{BA}$

Nous remarquons une aggloémation des réseaux de même type et la séparation par rapport aux autres types.

Nous effectuons une CAH sur les composantes principales issues de l'ACP. La figure 7.22 représente le dendrogramme obtenu en utilisant la méthode de Ward.

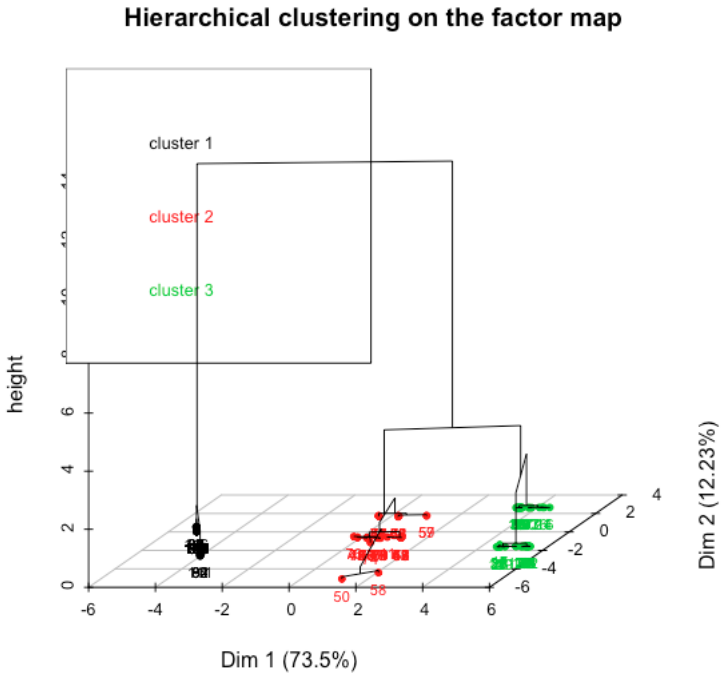



FIGURE 7.22 – Dendrogramme CAH+ACP sur $P_{BA}$

Le tableau 7.13 confronte les clusters obtenus par la CAH sur les composantes principales retenues par l'ACP. Nous remarquons que sur toutes les classes générées avec les puissances $power1 = 1$, $power2 = 2$, et $power3 = 3$, le clustering est parfait et retrouve parfaitement ces classes-ci. Ainsi, l'erreur est de 0%.

| clusters.acp-cah | 1 | 2 | 3 |
|---|---|---|---|
| Cl1 | 0 | 0 | 40 |
| Cl2 | 0 | 39 | 0 |
| Cl3 | 40 | 0 | 0 |

TABLE 7.13 – Confrontation des clusters obtenus par CAH après ACP avec classes générées dans $P_{BA}$

**Clustering Ascendant Hiérarchique sur $P_{BA}$**

Nous effectuons une CAH sur toutes les variables de $P_{BA}$ en utilisant le critère de Ward comme méthode d'aggrégation et nous obtenons le dendrogramme présenté dans la figure 7.23.

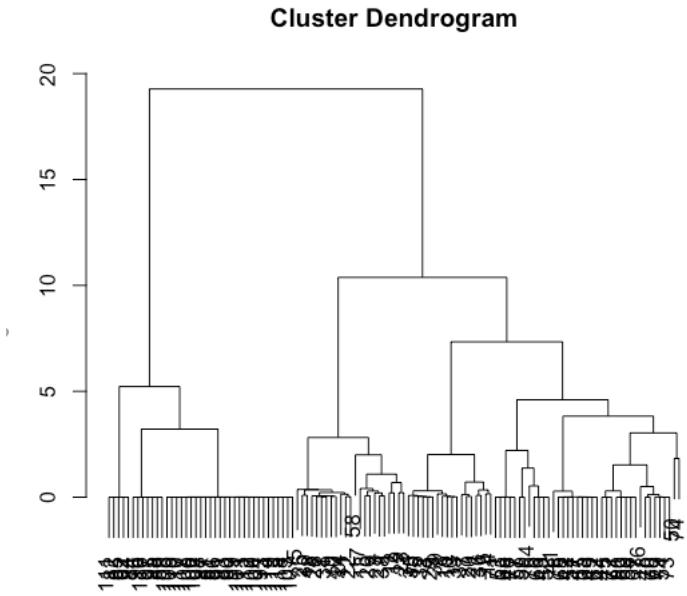



FIGURE 7.23 – Dendrogramme CAH directement sur $P_{BA}$

La figure 7.23 ne permet pas de distinguer les individus néamoins, elle met en évidence la présence de 3 clusters .

Confrontons à présent les clusters obtenus aux classes générées dans le jeu de données ($power1 = 1$ $power2 = 2$ et $power3 = 3$). Le tableau 7.14 présente en colonnes les clusters obtenus et en ligne les classes initiales générées dans le jeux de données $P_{BA}$.

| clusters.cah | 1 | 2 | 3 |
|---|---|---|---|
| Cl1 | 22 | 18 | 0 |
| Cl2 | 1 | 39 | 0 |
| Cl3 | 0 | 0 | 39 |

TABLE 7.14 – Confrontation clusters avec classes générées dans $P_{BA}$

Ce qui représente une erreur de 19/120 soit 15.8% qui est faite majoritairement sur la classe $Cl1$ ($power1 = 1$) de 18/40 confondus avec $Cl2$ ($power2 = 2$) et 1/40 de la classe $Cl2$ ($power2 = 2$) confondu pour cet individu avec $Cl1$ ($power1 = 1$).

#### 7.3.4.5 Classification du jeu de données $P_{BA}$

Dans cette partie, nous proposons de réaliser un apprentissage supervisé sur le jeu de données $P_{BA}$ cette fois-ci en l'étiquettant avec les classes relatives à la puissance d'attachement préférentiel $power_{BA}$ utilisées pour générer le réseau graine. Ainsi, $P_{BA}$ contient 3 classes à retrouver par un modèle de classification.
Pour se faire, nous utilisons plusieurs méthodes à base d'arbre de décision. Comme précédemment, le modèle est bâti sur un échantillon d'apprentissage (2/3 formé aléatoirement à partir des données initiales) et validé sur l'échantillon de test (1/3 des données qui constituent les données restantes non utilisées pour l'apprentissage).
Le tableau 7.15 présente les différentes proportions classes dans les échantillons d'apprentissage et de test.

| Echantillon | Cl1 | Cl2 | Cl3 |
|---|---|---|---|
| Apprentissage | 28 | 26 | 26 |
| Test | 12 | 14 | 13 |
| $P_{BA}$ | 40 | 40 | 39 |

TABLE 7.15 – Représentation des classes dans les échantillons d'apprentissage et de test de $P_{BA}$

**Classification par la méthode CART**
La figure 7.24 représente l'arbre obtenu sur les données d'apprentissage.

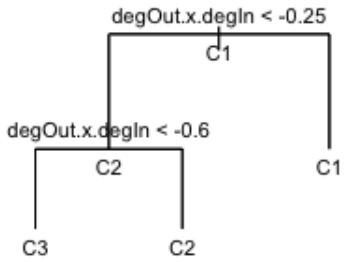



FIGURE 7.24 – Arbre de classification CART sur l'ensemble d'apprentissage de $P_{BA}$

Nous remarquons dans la figure 7.24 que cette méthode de classification utilise une seule variable (corrélation entre le degré entrant et sortant) pour séparer les classes. On pourrait en extraire la règle de décision :
Si $corr(degOut, degIn) \geq -0.25 \rightarrow$ le réseau de $P_{BA}$ est de type $C1$ ($power_{BA} = 1$),
sinon si $corr(degOut, degIn) \geq -0.6 \rightarrow$ le réseau de $P_{BA}$ est de type $C2$,
sinon le réseau de $P_{BA}$ est de type $C3$.

L'erreur d'apprentissage obtenue sur $P_{BA}$ est nulle.
Nous effectuons la prédiction sur l'échantillon de test l'erreur de prédiction est aussi nulle comme illustrée dans le tableau 7.16 de prédiction.

| Prédictions | 1 | 2 | 3 |
|---|---|---|---|
| Cl1 | 12 | 0 | 0 |
| Cl2 | 0 | 14 | 0 |
| Cl3 | 0 | 0 | 13 |

TABLE 7.16 – Prédictions obtenues par CART sur le jeu de données $P_{BA}$

Les prédictions obtenues sont parfaites sur toutes les classes.

### 7.3.5 Cas de réseaux de type "Erdös-Rényi"

#### 7.3.5.1 Description du jeu de données

Afin tester notre approche sur des réseaux de type "Erdös-Rényi", nous générons aléatoirement un ensemble de réseaux $P_{ER}$ tous d'ordre $N = 1000$ (nombre de sommets), générés de $cl = 3$ manières différentes. Pour se faire, nous générons 3 réseaux "Erdös-Rényi", avec des probabilités $p_{ER}$ différentes.

— Le 1er réseau autour duquel sera générée la classe $cl1$ à $p1 = 0.01$

— Le 2ème réseau permet de générer la classe $cl2$ à $p2 = 0.03$

— Le 3ème réseau produit la classe $cl3$ à $p3 = 0.05$

Ensuite, les réseaux de chaque classe sont obtenus par percolation du réseau graine, en le perturbant chaque fois et supprimant aléatoirement $perturb = 1\%$ de ses liens. Cela est réalisé $m = 40$ fois pour obtenir les 40 réseaux de chaque classe.
Ainsi $P_{ER}$ est composé de 120 graphes répartis en 3 classes, chacune est formée $m = 40$ graphes ayant un même $P_{ER}$ et obtenus par perturbation de 1% de l'un des réseaux graines définis ci-dessus.
Ce tau de perturbation *perturb* est bien entendu fixe et peut être changé pour former un autre jeu de données. Notons que pour ce type de réseaux, avec une valeur de perturbation supérieure ou égale à 2% ($perturb \geq 2\%$) les clusters commencent à se chevaucher et le nombres de clusters devient difficile à déterminer. Cependant, les prédictions par les différentes méthodes restent de bonne qualité si on connait le nombre de classes.

#### 7.3.5.2 Recodage des données

Le recodage des données de $P_{ER}$ est fait exactement de la même manière que dans le cas d'étude sur $P_{ER-SW-BA}$ présenté dans le paragraphe Recodage des données.

#### 7.3.5.3 Exploration du nombre optimal de clusters du jeu de données $P_{ER}$

**Méthode "Elbow" (coude) [79]**
La figure 7.25 représente le nombre optimal de clusters pour le jeu de données $P_{ER}$.

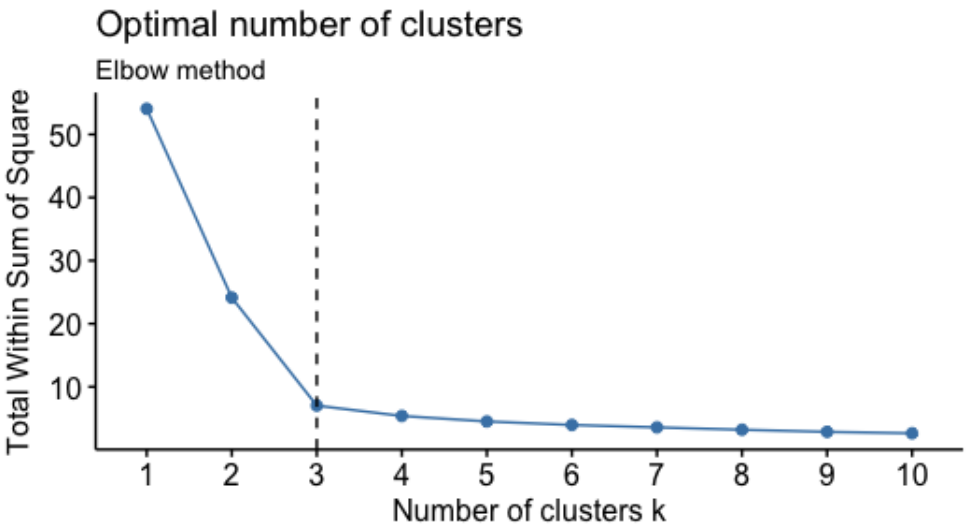



FIGURE 7.25 – Nombre optimal de clusters selon la méthode "Elbow" pour $P_{ER}$.

Nous remarquons que cette méthode met clairement en évidence un nombre optimal de clusters $K = 3$.

**Méthode "Slihouette"**

La figure 7.26 représente le nombre optimal de clusters pour le jeu de données $P_{ER}$ selon la méthode "Silhouette" décrite dans la paragraphe Méthodes de détermination du nombre de clusters.

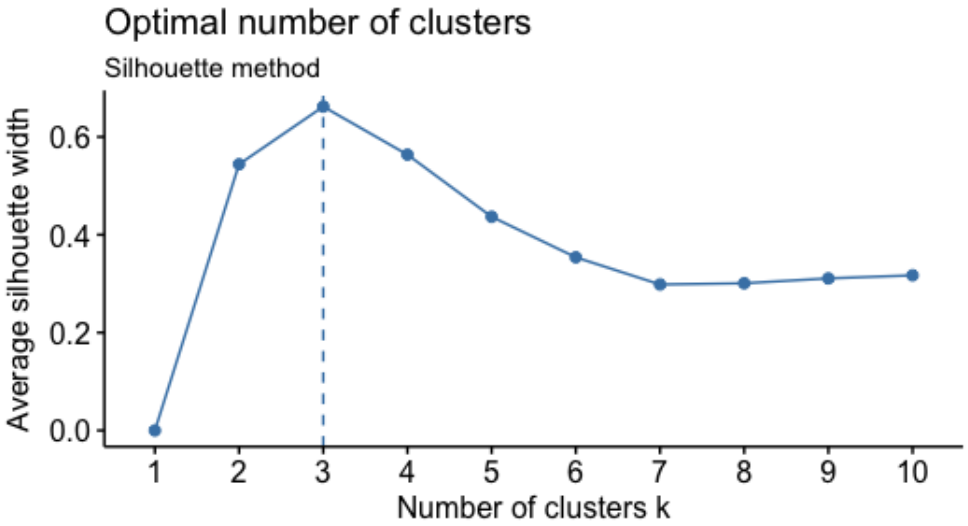



FIGURE 7.26 – Nombre optimal de clusters selon la méthode "Silhouette" pour $P_{ER}$.

Nous remarquons d'après la figure 7.26 que cette méthode donne aussi un nombre optimal de clusters $K = 3$.

Nous avons également fait voter un ensemble de 20 indices présents dans la littérature pour choisir le nombre optimal de clusters pour $P_{ER}$ et le résultat par vote majoritaire donne aussi $K = 3$ avec 16 votants sur 20 qui sont : "KL" [62],"CH" [21], "Hartigan" [43], "Scott" [76], "Marriot" [66], "TrCovW"[69], "TraceW" [69], "Friedman" [37], "Rubin" [37], "Cindex" [49], "DB" [26], "Silhouette" [58], "Ratkowsky" [73], "Ball" [6] , "PtBiserial" [68, 67] et "SDindex" [42].
La figure 7.27 représente ce vote-ci.

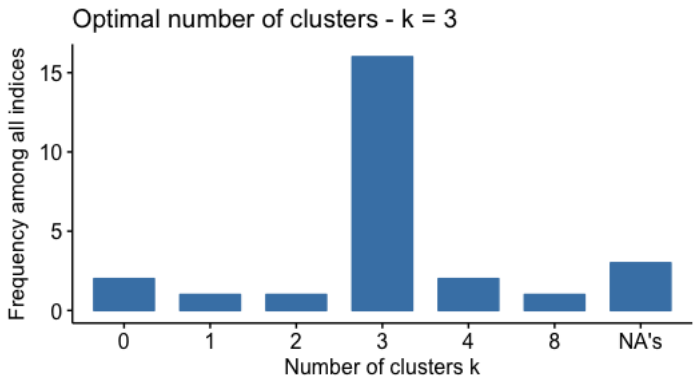



FIGURE 7.27 – Vote majoritaire de 20 indices pour choisir le nombre optimal de clusters de $P_{ER}$

Reste à confirmer si les clusters correspondent vraiment aux classes de réseaux générées.

#### 7.3.5.4 Clustering du jeu de données $P_{ER}$

**Clustering Ascendant Hiérarchique sur les Composantes Principales de $P_{ER}$**

Une Analyse en Composantes Principales nous permet de mieux comprendre le jeu de données et retenir les composantes à forte influence.
La représentation des données sur le premier plan factoriel est illustrée par la figure 7.28. La première dimension représente environ 67% des données. Quant à la deuxième, elle contribue avec environ 25%. Ce qui donne un premier plan factoriel avec 92% des informations.

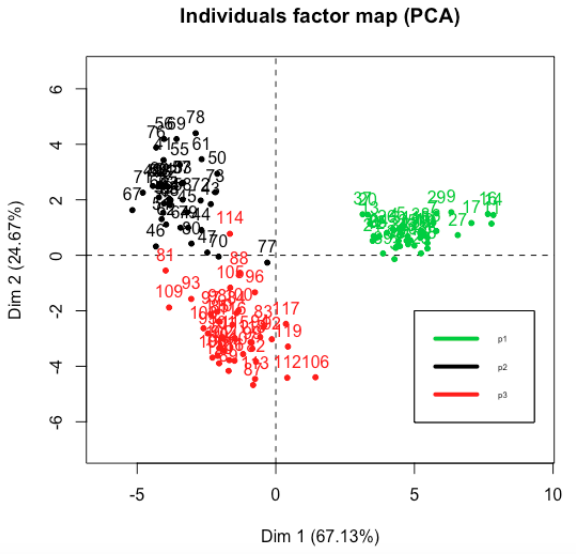



FIGURE 7.28 – Représentation du jeu de données $P_{ER}$ sur le premier plan factoriel de l'ACP

Les réseaux représentés en vert correspondent à ceux générés avec $p1 = 0.01$, en noir ceux avec $p2 = 0.03$ et ceux en rouge sont ceux avec $p3 = 0.05$. Nous remarquons que les clusters sont assez bien séparés dans le plan factoriel. Nous effectuons une CAH sur les composantes principales issues de l'ACP. La figure 7.29 représente le dendrogramme obtenu en utilisant la méthode de Ward.

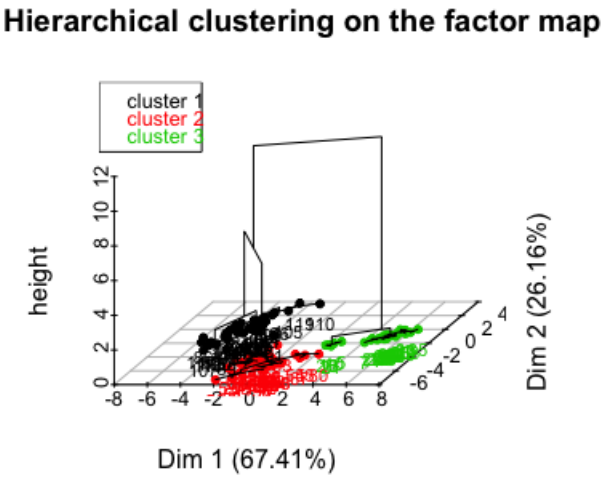



FIGURE 7.29 – Dendrogramme CAH+ACP sur $P_{ER}$

Le tableau 7.17 confronte les clusters obtenus par la CAH sur les composantes principales retenues par l'ACP. Nous remarquons que sur les classes générées avec les probabilités $p1 = 0.01$ et $p3 = 0.05$ sont retrouvées parfaitement par ce clustering, quant à celle $p2 = 0.03$, on remarque une seule erreur. L'erreur d'affectation sur le jeu de données $P_{ER}$ est de 1.6%.

| clusters.acp-cah | 1 | 2 | 3 |
|---|---|---|---|
| Cl1 | 40 | 0 | 0 |
| Cl2 | 0 | 39 | 1 |
| Cl3 | 0 | 0 | 40 |

TABLE 7.17 – Confrontation des clusters obtenus par CAH après ACP avec classes générées dans $P_{ER}$

**Clustering Ascendante Hiérarchique sur données brutes sans ACP de $P_{ER}$**
Nous effectuons une CAH directement sur toutes les variables de $P_{ER}$ en utilisant le critère de Ward comme méthode d'aggrégation. La coupe qui maximise le saut d'intertie est réalisée avec $K = 3$ clusters.

Confrontons à présent les clusters obtenus par cette coupe avec les classes générées dans le jeu de données initial ($p1 = 0.01$, $p2 = 0.03$ et $p3 = 0.05$). Le tableau 7.18 présente en colonnes les clusters obtenus et en ligne les classes initialement générées dans le jeux de données $P_{ER}$.

| clusters.cah | 1 | 2 | 3 |
|---|---|---|---|
| Cl1 | 40 | 0 | 0 |
| Cl2 | 0 | 40 | 0 |
| Cl3 | 0 | 4 | 36 |

TABLE 7.18 – Confrontation clusters avec classes générées dans $P_{ER}$

D'après 7.18, une faible "erreur de clustering" est observée et représente 3.33% faite sur la classe $Cl3$ ($p2 = 0.05$)
En comparant les deux clustering, nous remarquons que l'ACP améliore légèrement les résultats du clustering.

#### 7.3.5.5 Classification du jeu de données $P_{ER}$

Dans cette partie, nous rélisons une classification sur le jeu de données $P_{ER}$ en gardant les étiquettes des classes qui correspondent à la probabilité $p_{ER}$ utilisées pour

générer le réseau graine. Ainsi, $P_{ER}$ contient 3 classes à retrouver par un modèle de classification.
Pour se faire, nous utilisons plusieurs méthodes à base d'arbre de décision. Le modèle est bâti sur un échantillon d'apprentissage (2/3 formé aléatoirement à partir des données initiales) et validé sur l'échantillon de test (1/3 des données qui constituent les données restantes non utilisées pour l'apprentissage).
Le tableau 7.19 présente les différentes classes des échantillons d'apprentissage et de test.

| Echantillon | Cl1 | Cl2 | Cl3 |
|---|---|---|---|
| Apprentissage | 28 | 26 | 26 |
| Test | 12 | 14 | 14 |
| $P_{ER}$ | 40 | 40 | 40 |

TABLE 7.19 – Représentation des classes dans les échantillons d'apprentissage et de test de $P_{ER}$

Nous remarquons que les classes sont équi-représentées dans les échantillons d'apprentissage et de test.

**Classification par la méthode CART**

A présent réalisons une classification en utilisons la méthode d'arbre de décision. La figure 7.30 représente l'arbre obtenu sur les données d'apprentissage.

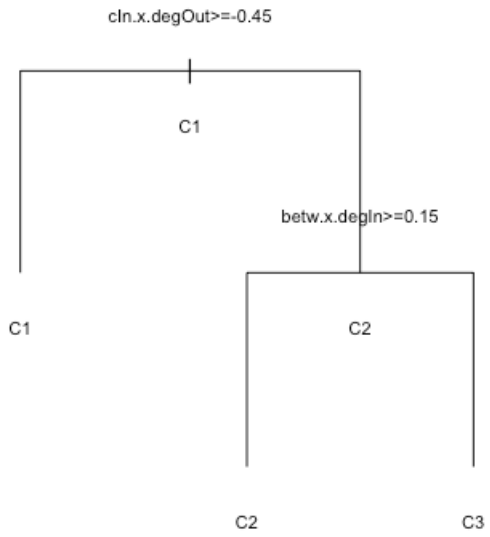



FIGURE 7.30 – Arbre de classification CART sur l'ensemble d'apprentissage de $P_{ER}$

Nous remarquons dans la figure 7.30 que cette méthode de classification (comme dans le cas du jeu de données "Small-World" $P_{SW}$) utilise une variable $cIn \times degOut$ (corrélation entre l'intermédiarité et la proximité entrante) pour séparer la classe $Cl1$ des deux autres puis la variable $betw \times degIn$ (corrélation entre l'intermédiarité et le degré entrant) pour séparer les deux classes $Cl2$ et $Cl3$. L'erreur d'apprentissage dans ce cas est faible est égale à .0.0125 soit environ 1.25%.
Nous effectuons la prédiction sur l'échantillon de test, l'erreur de prédiction est aussi nulle comme illustrée dans le tableau 7.20 de prédiction.

| Prédictions | Cl1 | Cl2 | Cl3 |
|---|---|---|---|
| Cl1 | 12 | 0 | 0 |
| Cl2 | 0 | 14 | 0 |
| Cl3 | 0 | 0 | 14 |

TABLE 7.20 – Prédictions obtenues par CART sur l'échantillon de test de $P_{ER}$

Les prédictions obtenues sont parfaites et l'erreur de prédiction est nulle pour $P_{ER}$.

## 7.4 Application au cas des réseaux réels des systèmes nucléaires du modèle EPR 2 $P_{sys-EPR2}$

### 7.4.1 Description du jeu de données

Dans cette partie, nous utilisons le jeu de données $P_{sys-EPR2}$ précédemment présenté dans le paragraphe Réseaux réels des systèmes de sûreté nucléaire. Rappelons qu'il s'agit de réseaux de systèmes de sûreté nucléaire. Ils sont 23 réseaux systèmes et donc $K = card(P_{sys-EPR2}) = 23$.

### 7.4.2 Recodage des données

Le recodage des données de $P_{sys-EPR2}$ est fait exactement de la même manière que celui des cas d'études précédents qui appliquent tous la méthode décrite dans le paragraphe (cf : Mesure de similarité entre réseaux).
Ainsi, à chaque réseau $P_k$ de $P_{sys-EPR2}$ correspond un vecteur ligne $Kcor[k]$, $1 \leq k \leq card(P_{sys-EPR2}) = 23$, tel que chaque composante $Kcor^{(l)}[k]$; $1 \leq k \leq 23$ s'écrit comme suit :

$$Kcor^{(l)}[k] = cor(X^{(i)}[k], X^{(j)}[k]);\ 1 \leq j < i \leq 7;\ 1 \leq l \leq 21).$$

La matrice $Kcor = t(Kcor[1], ..., Kcor[K])$ qui regroupe tous les coefficients synthétiques de centralités a maintenant 21 colonnes comme précédemment mais $K = 23$ lignes, le nombre de graphes étudiés.

### 7.4.3 Exploration du nombre optimal de clusters de $P_{sys-EPR2}$

**Méthode "Elbow" (coude) [79]**
La figure 7.31 représente le nombre optimal de clusters pour le jeu de données $P_{sys-EPR2}$.

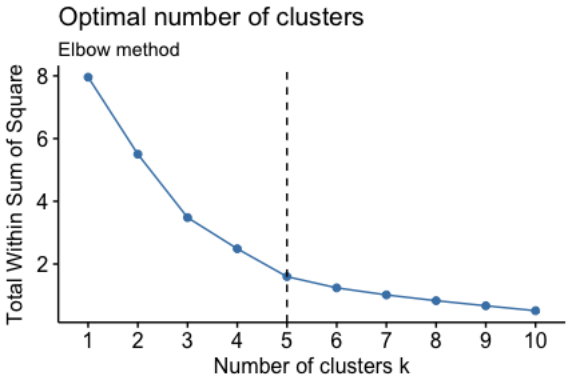



FIGURE 7.31 – Nombre optimal de clusters par la méthode Elbow

Nous remarquons que cette méthode donne un nombre optimal de clusters $K = 5$, malgré que le "coude" n'est pas clairement identifié ici.

**Méthode "Silhouette"**

La figure 7.32 représente le nombre optimal de clusters pour le jeu de données $P_{sys-EPR2}$ selon la méthode "Silhouette" décrite dans la partie Méthodes de détermination du nombre de clusters.

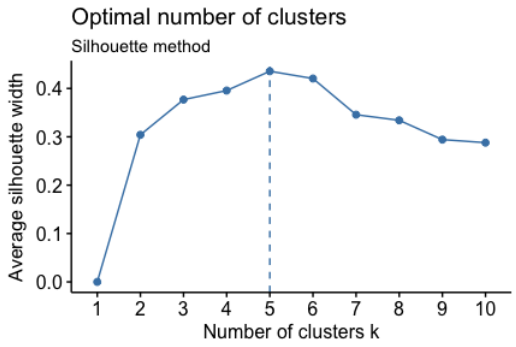



FIGURE 7.32 – Nombre optimal de clusters par la méthode Silhouette

Nous remarquons d'après la figure 7.32 que cette méthode donne aussi un nombre optimal de clusters $K = 5$.

Ces deux méthodes nous suggèrent un nombre optimal de clusters de 5, néamoins, étant donné que la CAH propose une hierarchie de clustering grâce au dendrogramme, il est possible de reconsidérer ce clustering et plus précisément fusionner deux clusters ou plus, ou à l'inverse partitionner un cluster de ceux que nous allons présenté dans le partitionnement à $K = 5$ clusters.

### 7.4.4 Clustering du jeu de données $P_{sys-EPR2}$

Nous effectuons, comme dans les différents cas d'études précédents, un clustering sur les résultats de l'ACP, puis un autre avec uniquement la CAH et dans ce cas sur les données brutes.

**Clustering Ascendant Hiérarchique sur ACP et sans ACP de $P_{sys-EPR2}$**

La figure 7.33 représente la projection des données de $P_{sys-EPR2}$ sur le premier plan factoriel retenu par l'ACP.

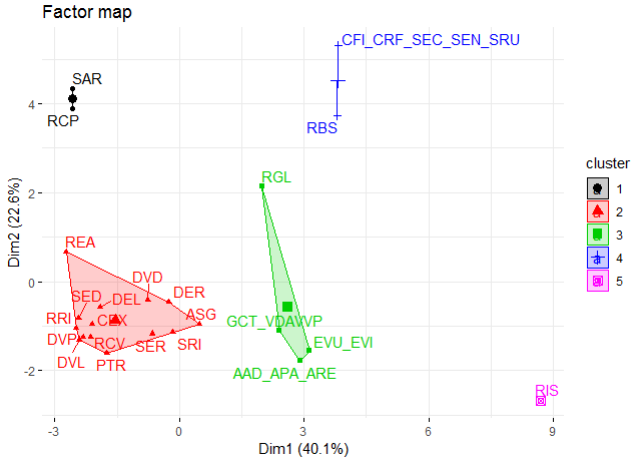



FIGURE 7.33 – Représentation des réseaux des systèmes du modèle EPR 2 sur le premier plan factoriel

Le premier plan factoriel de l'ACP représente environ 63% de l'information. Nous remarquons que les réseaux "RIS" est isolé et éloigné des autres. Les réseaux représentés par chaque couleurs sont proches entre eux et loins des autres. les groupes en rouge et en vert sont légèrement proches. La fusion de groupes peut être envisagée ou non. La figure 7.34 représente le résultat du clustering sur les composantes principales identifiées par l'ACP.

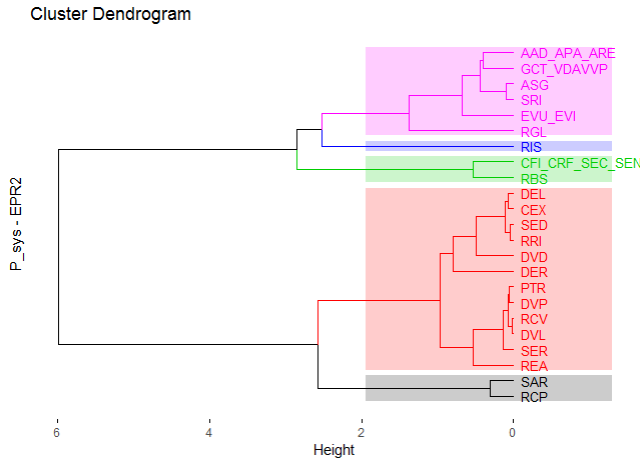



FIGURE 7.34 – Clustering des réseaux des systèmes du modèle EPR 2 par la CAH sur ACP

les clusters obtenus pour un partitionnement en $K = 5$ groupes aussi suggérer par le saut d'inertie sont résumés dans le tableau 7.21.

| Nom du cluster | Eléments | Taille du cluster |
|---|---|---|
| Clust1 | "SAR","RCP" | 2 |
| Clust2 | "PTR", "RRI","ASG", "CEX","DER", "DVD","DEL","DER", "DVL","SER","REA", "SRI","RCV","SED" | 14 |
| Clust3 | "AAD_APA_ARE", "RGL","EVU_EVi", "GCT_VDAVVP" | 4 |
| Clust4 | "CFI_CRF_SEC_SEN_ SRU", "RBS" | 2 |
| Clust5 | "RIS" | 1 |

TABLE 7.21 – Clustering des réseaux de systèmes EPR2 par la CAH après l'ACP

Nous remarquons que les clusters sont de tailles différentes avec un grand cluster (Clust2) qui contient 14 réseaux. En réalisant une CAH sur les données brutes nous obtenons également 5 clusters.
La confrontation des résultats obtenus par la CAH sur les composantes principales de l'ACP et celle sans ACP montre que les clusters obtenus sont identiques. L'ACP aura servi à visualiser les données.
Ces résultats doivent être discutés avec le constructeur nucléaire pour le design des systèmes et les experts en fonctionnement.

### 7.4.5 Classification du jeu de données $P_{sys-EPR2}$

Dans ce paragraphe nous souhaitons réaliser une classification du jeu données $P_{sys-EPR2}$. Pour se faire, nous étiquettons les différents réseaux par la classe trouvée dans le pa-

ragraphe de clustering ci-dessus. Le jeu de données $P_{sys-EPR2}$ est donc composé de 23 réseaux appartenant à 5 classes tels que décrit dans le tableau 7.21. Vu que le nombre de réseaux du jeu de données est très petit, nous utilisons comme méthode de validation de l'apprentissage supervisé le Leave-One-Out Cross-Validation présenté dans le paragraphe 2 de l'annexe Validation croisée (Cross-Validation).
Le tableau 7.22 présente les différentes erreurs de prédictions réalisées par chacune des méthodes à base d'arbre de décision.

| Méthode | Erreur de Prédiction |
|---|---|
| CART | 0.35 |
| Bagging CART | 0.21 |
| Random Forest | 0.18 |

TABLE 7.22 – Erreurs de prédiction des méthodes à base d'arbres de décision sur $P_{sys-EPR2}$

Nous remarquons que bien que la taille du jeu de données est très faible et les classes à prédire sont loin d'être équi-représentées comme c'est le cas dans les autres jeux de données, la méthode d'arbre de décision (CART) donne des résultats acceptables. Ses performances ont été améliorées grâce aux méthodes ensemblistes Bagging et Random Forest.

## 7.5 Conclusion

Dans les paragraphes précédents, nous avons défini une centralité synthétique qui regroupe l'information relative aux sommets d'un réseau. Cette mesure est basée sur les interdépendences entre leurs vecteurs de centralités. L'interdépendance est capturée par la corrélation des rangs de Kendall. Ensuite, nous avons établi une mesure de similarité entre réseaux, indépendamment de leurs tailles, densité, et autres types de propriétés. Nous avons appliqué ces deux notions pour effectuer un clustering sur différents types de graphes théoriques. Les résultats produits sont satisfaisants. Cette méthode est donc efficace pour discriminer les différents types de réseaux dans chaque jeu de données.
Enfin, nous avons appliqué cette nouvelle mesure de similarité sur des réseaux réels qui sont les réseaux systèmes de sûreté $P_{sys-EPR2}$. La validation croisée classique ne convient pas pour l'ensemble des données $P_{sys-EPR2}$ puisqu'il est de taille réduite. Pour y remédier, une validation croisée de type Leave-one-out a permis de prouver que les résultats obtenus par les méthodes à base d'arbres de décision sont satisfaisants.

# Chapitre 8

# Conclusion et perspectives

Ce travail de thèse s'intéresse dans un premier temps à la modélisation en graphes des études de sûreté nucléaire et plus précisément les EPS. Nous proposons dans le chapitre 4 une méthode générique de construction de réseaux à partir d'un système de sûreté puis d'une séquence accidentelle. Nous appliquons cette méthode à une séquence accidentelle relative à l'initiateur "Baisse incontrôlée du niveau primaire" dans les états d'arrêt du réacteur, ensuite nous présentons son utilisation sur systèmes de sûreté réels du modèle EPR 2. Les réseaux construits sont dirigés et attribués (types de sommets et types de liens) et ont des propriétés topologiques semblables à celles des réseaux complexes.

Dans le chapitre 5, nous nous intéressons à la prédiction du Risk Increase Factor, un facteur d'importance qui caractérise un composant dans la mesure où il permet d'évaluer l'impact de sa défaillance sur la valeur du risque et par conséquent sur la sûreté de l'installation. Comme variable de prédiction, nous utilisons les différentes mesures de centralités décrites dans chapitre 3. La prédiction de cette mesure est un problème à deux classes (faible impact $RIF = 0$ et impact élevé $RIF = 1$ sur la sûreté). De plus, il s'agit d'une classification avec déséquilibre de classes sur échantillon de petite taille car le calcul du RIF est coûteux. Nous procédons tout d'abord à une réduction de variables sur l'échantillon d'apprentissage pour ne retenir que celles qui sont fortement corrélées au RIF (la variable cible) et moins corrélées entre elles. Cette étape permet de garder les variables du degré sortant (degOut), la proximité entrante (cIn), le PageRank (pRank) et le degré de hubité (hub). Ensuite nous utilisons les méthodes à base d'arbres de classification. Cela produit un classifieur faiblement performant. De plus, il est sensible à l'échantillonnage car les résultats dépendent du choix aléatoire de l'échantillon d'apprentissage. Pour y remédier, nous utilisons les méthodes ensemblistes, Random Forest et Bagging, qui améliorent pratiquement tous les paramètres, en particulier la précision passe de 0.5 à 1. La dernière méthode ensembliste utilisée est le Gradient Boosted Machine. Celle-ci donne une précision parfaite semblable aux deux méthodes précédentes, un rappel (sensitivity) double et qui devient égale à 0.667

ce qui est intéressant, vu le fort déséquilible de classes.
Pour pallier au déséquilibre de classes ($RIF = 1$ ne représente que 5% des données), nous utilisons dans un premier temps l'arbre de classification sur des échantillons aléatoires stratifiés. Les résultats obtenus sont comparables à ceux obtenus par la méthode Gradient Boosted Machine mais restent néanmoins moins bons. Dans un deuxième temps, les principales méthodes de rééquilibre de classes sont explorées pour améliorer résultats obtenus par l'arbre de classification. La méthode de sur-échantillonnage (over-sampling) rend le rappel parfait mais dégrade la précision et donne une F-mesure meilleure que celle obtenue sur les données brutes, l'AUC est également améliorée. Cette méthode donne de meilleurs résultats que les autres et améliore presque tous les indicateurs de performances pour l'arbre de classification. Nous réalisons, ensuite, la prédiction par la régression logistique. Ainsi, nous construisons plusieurs modèles utilisant une variable à la fois, deux, trois et finalement les quatre variables. Nous étudions le modèle qui minimise le Critère d'Information d'Akaike (AIC) connu pour l'évaluation de la qualité de modèles. Le modèle sélectionné est `RIF` $\sim$ `degOut+cIn+hub`. Ce modèle donne un rappel parfait et une précision faible contrairement à l'arbre de classification. Nous construisons des modèles logistiques sur des échantillons stratifiés. Le modèle `RIF` $\sim$ `cIn+pRank+hub` est le meilleur selon (AIC), la précision s'améliore légèrement par rapport au cas précédent mais le rappel s'est considérablement dégradé. Ainsi, contrairement à ce qui était attendu, la stratification n'améliore pas les résultats de la régression logistique.
Comme évoqué précédemment, le choix du modèle de prédiction dépendra certainement de l'application attendue. Une précision parfaite est attendue quand des moyens financiers sont déployés par exemple pour modifier l'installation. Dans ce cas l'arbre de classification avec GBM est la méthode la plus pertinente. Cependant, si l'objectif est de fournir une liste exhaustive de composants à revérifier par les analystes EPS ou les concepteurs de nouveaux systèmes, le rappel doit être maximal et donc nous pensons que la méthode de régression logistique est plus adaptée. Finalement, cette prédiction sert aussi comme outil d'aide à la décision quand les systèmes sont en cours de conception, les séquences pas encore définitives et les EPS ne sont pas encore réalisées car les données d'entrée ne sont pas disponibles.

La seconde partie de ce manuscrit regroupe deux chapitres et s'intéresse à la comparaison entre graphes. L'homogénéité d'une centralité est étudiée sur plusieurs réseaux. L'homogénéité est une similarité mono-variable. Cette notion a permis de comparer des graphes théoriques par rapport à une mesure de centralité. Ensuite, l'homogénéité d'une centralité est analysée sur les réseaux réels de systèmes de sûreté précédemment décrits. Nous avons identifié, par exemple, que les deux réseaux "AAD_APA_ARE" et "ASG" de $P_{sysEPR2}$ sont homogènes par rapport au degré entrant (degIn), mais

non-homogènes par rapport à la proximité sortante (cOut).
Par la suite, nous étudions la similarité entre deux centralités dans un même réseaux. Cette similarité correspond à la relation de dépendance évaluée à l'aide des corrélations de Spearman et de Kendall. Comme dans l'étude d'homogénéité, nous validons l'approche sur les graphes théoriques et nous l'utilisons sur les réseaux réels. Nous montrons, par exemple, que presque toutes les centralités dans le réseau "AAD_APA_ARE" sont liées.
Le chapitre 7 introduit une nouvelle mesure de similarité qui permet de comparer de manière globale les structures des réseaux : Une nouvelle représentation d'un réseau y est proposée. Elle est basée sur les corrélations de Kendall entre les vecteurs de centralités que nous nommons "coefficient synthétique de centralité". Par la suite, une mesure de similarité utilisant ce coefficient synthétique de centralité est introduite afin de comparer des réseaux qui peuvent être de tailles différentes, de propriétés topologiques différentes (densité, diamètre, coefficient de clustering globale,...). Dans un premier temps nous montrons l'efficacité de cette mesure de similarité sur le clustering de différents jeux de données de graphes appartenant aux familles classiques décrites dans le chapitre 3, à savoir Erdös-Rényi, Small-World et Barabasi-Albert. Ensuite, nous vérifions que cette mesure de similarité permet aussi la classification à base d'arbre de décision, des modèles de prédiction pour chaque jeu de données. Cette nouvelle mesure de similarité est également appliquée pour le clustering d'un jeu de données composé des réseaux réels qui sont les réseaux systèmes de sûreté $P_{sys-EPR2}$ présenté dans chapitre 4. La classification par les méthodes à base d'arbres de décision donne des résultats satisfaisants sur ce jeu de données. Bien entendu ces résultats doivent être discutés avec le constructeur nucléaire pour le design des systèmes et les experts en fonctionnement.

## Perspectives et applications potentielles

**1- Intégration au logiciel Andromeda**
L'industrialisation de la construction de réseaux dirigés attribués pour toutes les séquences accidentelles et son intégration au logiciel Andromeda d'EDF. Aujourd'hui Andromeda est l'outil de référence pour réaliser des AQS [1] à EDF, avec l'integration de KB3 dans la chaine de production des missions systèmes : Ces missions se retrouvent dans les répertoires andromeda des modèles pour être gérés en version et pour plus de transparence pour les différents utilisateurs des modèles EPS. Pour donner à l'analyste un autre point de vue grâce à l'approche de réseaux complexes, EDF a décidé d'industrialiser le processus de génération des réseaux complexes dans

1. AQS : Analyse Qualitative de Séquences (accidentelles)

un premier temps des systèmes de sureté impliqués dans les séquences accidentelles et ensuite des réseaux plus globaux correspondant à ces dernières ou à une conséquence particulière. Au niveau système, ces réseaux vont permettre de bien mettre en évidence les différentes dépendances entre systèmes et en particulier, les liaisons systèmes supports/systèmes frontaux. En outre, de part leurs centralités les composants avec des centralités importantes doivent susciter un intérêt particulier au regard de leurs données de fiabilité pour faire en sorte que leur contribution ne soit pas source d'instabilité du modèle.

**2- Comparaison des analyses**

La comparaison des analyses des réseaux complexes des différentes séquences accidentelles que ce soit par notre nouvelle mesure de similarité structurelle ou en comparant par exemple les types de systèmes ou composants qui apparaissent importants dans chacune des séquences modélisées. Cela permet d'approfondir les connaissances des analystes EPS sur leurs modélisations mais aussi des ingénieurs chargés de la conception de nouveaux systèmes en particulier pour les nouveaux réacteurs.

**3- Réseaux attribués pour les agressions**

La prise en compte des agressions (séisme, incendie, tornade, innondation, ...) pose un certain nombre de challenges aussi bien pour leur modélisation que pour la quantification. En effet, les couches agression viennent engraisser des modèles déjà bien lourds à calculer. En terme d'analyse quantitative, un nombre important d'évènements de base (feuilles) viennent se greffer aux arbres de mission pour modéliser les pertes de composants/systèmes liées aux agressions. Le cadre des réseaux complexes fournit un moyen pour tenir compte des agressions de façon plus élégante avec l'introduction d'"attributs agression". Ces attributs (local, altitude, fragilité, exposition, ...) peuvent révèler de façon très rapide la structure restante en cas d'agression et fournir des éléments qualitatifs très importants. D'autre part, la propagation d'effets liés aux agressions peuvent se faire de façon assez naturelle dans la structure de réseaux attribués. Ce qui peut se faire via des méthodes de simulation ; par exemple les méthodes multi-agents (cf. netlogo [81]).

**4- Prise en compte de phénomènes dynamiques**

La prise en compte des aspects dynamiques pose aussi un certain nombre de questions difficiles à traiter dans les modèles EPS qui sont dans la majorité des cas (si l'on excepte les modèles de très petite taille) des modèles statiques. En effet, plusieurs aspects sont à considérer à ce niveau :

— il y a un besoin de créditer les récupérations de certains matériels durant le temps de scrutation (cf. [29]). Dans ce cas, il serait intéressant de voir dans quelle mesure

le cadre des réseaux dynamiques peut-il aider à faire un éclairage sur ce genre de situation à la lumière des effets cascades potentiels.

— le besoin de réalisme pour une gestion des ordres d'enclenchement des systèmes de sauvegarde entre les systèmes en fonctionnement et les systèmes en standby. Cet ordre et la cinétique des évènements peuvent avoir leur importance pour la gestion de l'accident et donc un éclairage de l'analyse des réseaux serait le bienvenu (ex. La propagation d'aggressions en utilisant les réseaux dynamiques comme ceux utilisés pour les études épidémiologiques (cf. travaux de Barthélémy, M. tels que [8]).

— la prise en compte des phénomènes physiques (en particulier pour les EPS de niveau 2).

Bien entendu, on ne s'attend pas à ce que les réseaux complexes puissent fournir des éléments quantitatifs à toutes ces questions, néanmoins la simulation et l'analyse pourrait révéler des structures cachées dans quant à la fragilité ainsi qu'à la résilience des parades mises en oeuvre.

# Bibliographie


[1] Daily Estimated Size of the World Wide Web internet. `http://www.worldwidewebsize.com`. Accessed : 2018-09-18.

[2] *Random Graphs*. Wiley-Blackwell, 2011.

[3] D Ait-Ferhat, T Friedlhuber, and M Hibti. An andromeda extension for network based safety assessment. In *American Nuclear Society Winter meeting on Risk Management for Complex Socio-technical Systems*, 2013.

[4] H. Akaike. Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, 60(2) :255–265, 1973.

[5] R. Albert, H. Jeong, and A. L. Barabási. Internet : Diameter of the world-wide web. *nature*, 401(6749) :130, 1999.

[6] G. H. Ball and D. J. Hall. Isodata, a novel method of data analysis and pattern classification. Technical report, Stanford research inst Menlo Park CA, 1965.

[7] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439) :509–512, 1999.

[8] M. Barthélemy, A. Barrat, R. Pastor-Satorras, and A. Vespignani. Dynamical patterns of epidemic outbreaks in complex heterogeneous networks. *Journal of theoretical biology*, 235(2) :275–288, 2005.

[9] B. Bollig and I. Wegener. Improving the variable ordering of obdds is np-complete. *IEEE Transactions on computers*, 45(9) :993–1002, 1996.

[10] B. Bollobás, O. Riordan, J Spencer, and G. Tusnády. The degree sequence of a scale-free random graph process. *Random Structures & Algorithms*, 18(3) :279–290, 2001.

[11] P. Bonacich. Factoring and weighting approaches to status scores and clique identification. *The Journal of Mathematical Sociology*, 2(1) :113–120, 1972.

[12] P. Bonacich. Some unique properties of eigenvector centrality. *Social Networks*, 29(4) :555 – 564, 2007.

[13] S. P. Borgatti and M. G. Everett. A graph-theoretic perspective on centrality. *Social Networks*, 28(4) :466–484, 10 2006.

[14] M. Bouissou, N. Villatte, H. Bouhadana, and M. Bannelier. Knowledge modelling and reliability processing : presentation of the figaro language and associated tools. Technical report, Electricite de France (EDF), 1991.

[15] U. Brandes. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2) :136–145, 2008.

[16] L. Breiman. Bagging predictors. *Machine learning*, 24(2) :123–140, 1996.

[17] L. Breiman. Random forests. *Machine learning*, 45(1) :5–32, 2001.

[18] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984. new edition [**?**] ?

[19] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and regression trees. *Wadsworth International Group*, 1984.

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

[21] T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1) :1–27, 1974.

[22] E. Castrillo, E. León, and J. Gómez. Fast heuristic algorithm for multi-scale hierarchical community detection. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 982–989. ACM, 2017.

[23] L. Chang, W. Li, L. Qin, W. Zhang, and S. Yang. pscan : Fast and exact structural graph clustering. *IEEE Transactions on Knowledge and Data Engineering*, 29(2) :387–401, 2017.

[24] J. Chen and Y. Saad. Dense subgraph extraction with application to community detection. *IEEE Transactions on Knowledge and Data Engineering*, 24(7) :1216–1230, 2012.

[25] W. J. Conover. *Practical nonparametric statistics*. John Wiley  Sons, New York, 1971.

[26] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2) :224–227, 1979.

[27] P. Diaconis and R. L. Graham. Spearman's Footrule as a Measure of Disarray. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(2), 1977.

[28] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3) :297–302, 1945.

[29] EPRI. Treatment of time interdependencies in fault tree generated cutset results. Technical report, Palo Alto, 2003.

[30] Kayhan Erciyes. *Complex Networks : An Algorithmic Perspective*. 09 2014.

[31] P. Erdös and A. Rényi. On random graphs i. *Publicationes Mathematicae Debrecen*, 6 :290, 1959.

[32] P. Erdos and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1) :17–60, 1960.

[33] C. A. Ericson. Fault tree analysis. In *System Safety Conference, Orlando, Florida*, volume 1, pages 1–9, 1999.

[34] L. C. Freeman, S. P. Borgatti, and D. R. White. Centrality in valued graphs : A measure of betweenness based on network flow. 1991.

[35] L.C. Freeman. Centrality in social networks : Conceptual clarification. *Social Networks*, 1(3) :215–239, 1979.

[36] T. Friedlhuber and A. Hibti, M .and Rauzy. Reinforcement of qualitative risk assessment-proposals from computer science. In *Proceedings of PSAM Topical Conference in Tokyo*, 2013.

[37] H. P. Friedman and J. Rubin. On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62(320) :1159–1178, 1967.

[38] J. H. Friedman. Greedy function approximation : a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[39] S. J. Friedman and K. J. Supowit. Finding the optimal variable ordering for binary decision diagrams. In *Proceedings of the 24th ACM/IEEE Design Automation Conference*, pages 348–356. ACM, 1987.

[40] J. B Fussell. How to hand-calculate system reliability and safety characteristics. *IEEE Transactions on Reliability*, 24(3) :169–174, 1975.

[41] J. D. Gibbons. Nonparametric statistical inference, 1971.

[42] M. Halkidi, M. Vazirgiannis, and Y. Batistakis. Quality scheme assessment in the clustering process. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 265–276. Springer, 2000.

[43] J. A. Hartigan. Clustering algorithms, new york : John willey and sons. *Inc. Pages113129*, 1975.

[44] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning : data mining, inference and prediction*. Springer, 2 edition, 2009.

[45] M Hibti. What if we revisit evaluation of psa models with network algorithm. In *PSA, International Topical Meeting on Probabilistic Safety Assessment and Analysis*, 2013.

[46] M Hibti, A Marechal, and A Oudjit. Exploring relations between graph metrics and importance measures in pra sequences. In *International Conference on Probabilistic Safety Assessment and Management (PSAM 13)*, 2016.

[47] M. O. Hill and A. J. E. Smith. Principal component analysis of taxonomic data with multi-state discrete characters. *Taxon*, pages 249–255, 1976.

[48] C. H. Hubbell. An input-output approach to clique identification. *Sociometry*, 28(4) :377–399, 1965.

[49] L. J. Hubert and J. R. Levin. A general statistical framework for assessing categorical clustering in free recall. *Psychological bulletin*, 83(6) :1072, 1976.

[50] L. J. Hubert and J. R. Levin. A general statistical framework for assessing categorical clustering in free recall. *Psychological bulletin*, 83(6) :1072, 1976.

[51] L. Ifrah. *L'information et le renseignement par Internet :«Que sais-je ?» n 3881.* Presses universitaires de France, 2010.

[52] P. Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37 :547–579, 1901.

[53] G. Jeh and J. Widom. Simrank : a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543. ACM, 2002.

[54] P. Jensen, M. Morini, M. Karsai, T. Venturini, A. Vespignani, M. Jacomy, J. P. Cointet, P. Mercklé, and E. Fleury. Detecting global bridges in networks. *Journal of Complex Networks*, 4(3) :319–329, 2015.

[55] H. Jiang, J. Gao, Z. Gao, and G. Li. Safety analysis of process industry system based on complex networks theory. In *Mechatronics and Automation, 2007. ICMA 2007. International Conference on*, pages 480–484. IEEE, 2007.

[56] I. T. Jolliffe. Graphical representation of data using principal components. *Principal component analysis*, pages 78–110, 2002.

[57] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1) :39–43, Mar 1953.

[58] L. Kaufman and P. Rousseeuw. Finding groups in data : An introduction to cluster analysis. 1990.

[59] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '98, pages 668–677, Philadelphia, PA, USA, 1998. Society for Industrial and Applied Mathematics.

[60] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5) :604–632, September 1999.

[61] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins. The web as a graph : measurements, models, and methods. In *International Computing and Combinatorics Conference*, pages 1–17. Springer, 1999.

[62] W. J. Krzanowski and Y. T. Lai. A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, pages 23–34, 1988.

[63] L. Lebart, A. Morineau, and M. Piron. *Statistique exploratoire multidimensionnelle*, volume 3. Dunod Paris, 1995.

[64] H. W. Lilliefors. On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318) :399–402, 1967.

[65] N. Lunardon, G. Menardi, and N. Torelli. Rose : A package for binary imbalanced learning. *R Journal*, 6(1), 2014.

[66] F. H. C Marriott. Practical problems in a method of cluster analysis. *Biometrics*, pages 501–514, 1971.

[67] G. W. Milligan. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, 45(3) :325–342, 1980.

[68] G. W. Milligan. A monte carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, 46(2) :187–199, 1981.

[69] G. W. Milligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2) :159–179, 1985.

[70] M. Newman, A. L. Barabasi, and D. J. Watts. *The structure and dynamics of networks*, volume 19. Princeton University Press, 2011.

[71] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64 :026118, Jul 2001.

[72] N Rasmussen. Wash-1400 (reactor safety study). *An assessment of accident risks in US Commercial Nuclear Power Plants*, 8, 1975.

[73] D. A. Ratkowsky and G. N. Lance. Criterion for determining the number of groups in a classification. 1978.

[74] I. Rogers. The google pagerank algorithm and how it works. 01 2002.

[75] C Ruiz-Martin, A. L. Paredes, and G. A. Wainer. Applying complex network theory to the assessment of organizational resilience. *IFAC-PapersOnLine*, 48(3) :1224–1229, 2015.

[76] A. J. Scott and M. J. Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, pages 387–397, 1971.

[77] J. Scott. *Social Network Analysis : A Handbook*. Sage Publications, London, 2nd edition, 2000.

[78] T. Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biol. Skr.*, 5 :1–34, 1948.

[79] R. L. Thorndike. Who belongs in the family ? *Psychometrika*, 18(4) :267–276, 1953.

[80] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 63(2) :411–423, 2001.

[81] S. Tisue and U. Wilensky. Netlogo : A simple environment for modeling complexity. In *International conference on complex systems*, volume 21, pages 16–21. Boston, MA, 2004.

[82] J. L. Torres Sanchez. *Vulnerability, interdependencies and risk analysis of coupled infrastructures : power distribution network and ICT*. PhD thesis, Université de Grenoble, 2013.

[83] J. Travers and S. Milgram. The small world problem. *Phychology Today*, 1(1) :61–67, 1967.

[84] M. Van der Borst and H. Schoonakker. An overview of psa importance measures. *Reliability Engineering & System Safety*, 72 :241–245, 2001.

[85] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301) :236–244, 1963.

[86] S. Wasserman and K. Faust. *Social network analysis : Methods and applications*, volume 8. Cambridge university press, 1994.

[87] D. J. Watts. Networks, dynamics, and the small-world phenomenon. *American Journal of sociology*, 105(2) :493–527, 1999.

[88] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684) :440, 1998.

# Annexe

# A Clustering

Le clustering (ou partitionnement) de données ou de réseaux est une appoche fondamentale pour détecter les patterns sous-jacents. Dans le cas particulier des réseaux, l'objectif de cette approche est d'explorer les structures cachées communément appelées structures communautaires. Le clustering attire de plus en plus l'attention par ses nombreuses applications parmi lesquelles on peut citer : dans le domaine médicale notamment pour détecter les cellules cancéreuses en imagerie médicale, dans le domaine du marketing grâce aux les systèmes de recommandations (Netflix, Amazon, et autres), dans le domaine de sondage des flux sociaux et opinion publique par le biais de l'analyse des réseaux sociaux (Facebook, Tweeter et autres).
Dans cette partie nous rappelons les différentes notions et méthodes de clustering utilisées dans ce manuscrit.

## A1 Méthode de Classification Ascendente Hiérarchique (CAH)

L'objectif d'une méthode de Classification Hiérarchique est de trouver une hiérarchie valuée. Cela s'appuie sur la notion de distance (mesure de similarité/dissimilarité) entre éléments (réseaux dans cette étude) qui induit une mesure de l'hétérogénéité d'un groupe fondée sur la distance entre un élément qui y appartient et un autre n'y appartenant pas. On distingue deux types d'approches :

— Agglomerative, qui est une approche de bas en haut ("bottom up") ou ascendante : la CAH. Elle commence avec $n$ clusters, chacun composé par un seul élément. A chaque itération, deux clusters "proches" sont fusionnés jusqu'à satisfaire le critère d'arrêt choisi.

— Divisive, qui est plûtot une approche allant du haut vers le bas ("top down") ou descendante. Elle commence par un seul cluster contenants les $n$ éléments. L'objectif est de diviser un clusters en deux partitions plus petites jusqu'à atteindre le critère d'arrêt fixé.

La hiérarchie produite est représentée par un arbre appelé dendrogramme.

### A1.1 Etapes de réalisation d'une CAH

Une Classification Ascendante Hiérarchique se déroule de la manière suivante :

(a) En premier lieu, la dissimilarité entre les $n$ éléments (réseaux) est calculée, cela suppose bien évidemment d'avoir choisi la mesure de distance à utiliser.

(b) Ensuite, il est question de regrouper les deux éléments dont le regroupement permet de minimiser l'un des critères d'agrégation considéré (cf. 2), créant ainsi une classe comprenant ces deux éléments (réseaux).

(c) Par la suite, l'algorithme calcule la dissimilarité entre cette classe et les $n-2$ autres éléments (réseaux) en utilisant le critère d'agrégation considérés. Puis, les deux réseaux ou groupes de réseaux sont regroupés afin de minimiser le critère d'agrégation.

Le processus continue jusqu'à ce que tous les réseaux soient regroupés dans un seul cluster. Ces regroupements successifs permettent d'obtenir un arbre binaire de classification (dendrogramme) comme illustré dans la figure A1, dont la racine est la classe regroupant l'ensemble des réseaux. Il est possible de choisir une partition en procédant à une troncature (horizontale) de l'arbre à un niveau donné. Ce niveau correspond à un choix soit fixé par des contraintes utilisateurs (le nombre de classes à obtenir est connu) ou par des critères plus objectifs tels que les sauts d'inertie, où l'utilisateur doit repérer des sauts importants dans les valeurs, en analysant l'histogramme des indices de niveau.

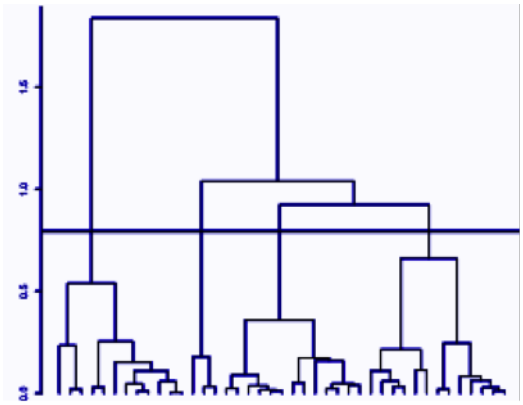


FIGURE A1 – Exemple de dendrogramme

### A1.2 Méthodes d'agrégation de clusters

Le choix de la méthode d'agrégation permet d'indiquer à l'algorithme la logique à suivre pour fusionner deux clusters. Les méthodes les plus utilisées dans la littérature sont les suivantes :

— **Liaison maximale ou complète :** La distance entre deux clusters ($C_1$ et $C_2$) est définie comme étant la valeur maximale de toutes les distances entre les éléments du cluster $C_1$ et les éléments du cluster $C_2$ pris par paire. En d'autres termes, l'idée est de regrouper les deux éléments présentant la plus grande distance entre

éléments des deux clusters. Elle a donc pour objectif de produire des clusters plus compacts.

— **Liaison minimale ou unique :** La distance entre deux clusters est définie comme la valeur minimale de toutes les distances entre les éléments du cluster $C_1$ et les éléments du cluster $C_2$ pris par paire. Il est donc question de regrouper les deux éléments présentant la plus petite distance entre éléments des deux clusters. Ainsi, elle tend à produire de grands clusters (plus allongés).

— **Liaison moyenne ou moyenne :** La distance entre deux clusters est définie comme la distance moyenne entre les éléments du cluster $C_1$ et les éléments du cluster $C_2$.

— **Méthode de Ward [85] :** Il s'agit de la méthode la plus connue et utilisée. Elle minimise la variance totale intra-clusters, c'est-à-dire que chaque étape fusionne la paire de clusters qui minimise la distance considérée. Ainsi, l'objectif est réaliser un gain minimum d'inertie intra-classe à chaque agrégation.

La figure A2 illustre quelques méthodes d'agrégation présentées ci-dessus. En rouge est représentée la méthode de liaison complète, en vert la liaison unique et finalement en noir la centroïde.

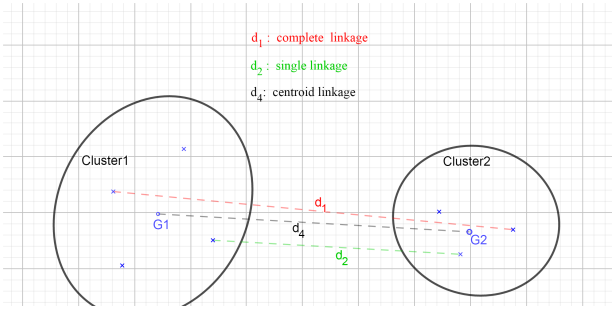



FIGURE A2 – Méthode d'agrégation de clusters

La méthode de liaison moyenne représentée dans la figure A3 donne un résultat moyen entre celui des deux méthodes complète et unique.

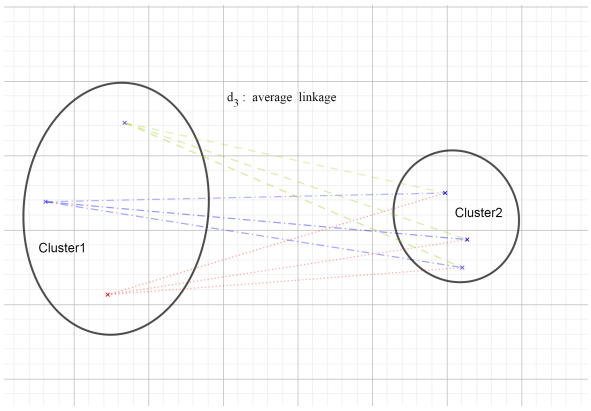



FIGURE A3 – Méthode d'agrégation de clusters (liaison moyenne)

## A2 Analyse en Composantes Principales (ACP)

L'analyse en composantes principales (ACP)[47] est une procédure statistique qui utilise une transformation orthogonale pour convertir un ensemble d'observations de variables éventuellement corrélées en un ensemble de valeurs de variables linéairement non corrélées appelées composantes principales.
S'il y a $n$ observations avec $p$ variables, le nombre de composantes principales distinctes est alors $\min(n-1, p)$.
Cette transformation est définie de manière à ce que la première composante principale présente la variance la plus grande possible (c'est-à-dire qu'elle représente le plus possible la variabilité des données), et que chaque composante suivante présente à son tour la variance la plus élevée possible sous la contrainte qu'elle est orthogonale aux composantes précédentes. Les vecteurs résultants (chacun étant une combinaison linéaire des variables et contenant n observations) constituent un ensemble de bases orthogonales non corrélées.

Combinant les méthodes de l'ACP, le clustering ascendant hiérarchique, permet de mieux visualiser les données. Les méthodes de composantes principales peuvent être utilisées comme une étape de prétraitement pour débruiter les données, transformer les variables catégorielles en variables continues, en équilibrer l'influence de plusieurs groupes de variables. Il peut également être utile de représenter graphiquement le clustering hiérarchique [56].

## A3 Méthodes de détermination du nombre de clusters

Il existe de nombreuses méthodes de détermination du nombre optimal de clusters pour un algorithme de clustering. Nous en présentons quelques unes.

— La méthode "Elbow" [79] : Les méthodes de partitionnements telles que le clustering par K-means s'appuient sur l'idée de définir des clusters (ou partitions) de telle sorte que la variation totale intra-cluster soit minimale. La variation totale intra-cluster se mesure par la somme totale des carrés intra-clusters nommés WWS (en anglais : "within-cluster sum of square"). Etant donné que la WSS totale mesure la compacité des clusters, il est nécessaire qu'elle ait une valeur la plus faible possible.

La méthode Elbow s'intéresse au WSS total en tant que fonction du nombre de clusters. Le nombre idéal de clusters est choisi de telle sorte que l'ajout d'un cluster ne produit pas d'amélioration significative de la valeur totale de WWS. Cette méthode semble parfois ambigue, une alternative serait d'utiliser la méthode de silhouette moyenne qui, elle aussi, peut être appliquée pour toute approche de clustering.

— La méthode "Silhouette" [58] : Il s'agit d'une méthode statistique utilisée notamment pour la validation du résultat d'un clustering. Elle a pour objectif de mesurer à quel point un individu se situe dans son cluster. La méthode de silhouette moyenne calcule donc la valeur de la moyenne des silhouettes des observations pour différentes valeurs de $K$ (nombre de clusters), et considère le nombre optimal de clusters $K$ celui qui maximise cette silhouette moyenne. Ce qui assure ainsi une bonne qualité de clustering [58].

— La méthode "gap-statistic" : Cette méthode a été publiée par Tibshirani, Walther, et Hastie dans [80]. Cette approche, comme les deux précédentes, peut être appliquée indépendemment du choix de la méthode de clustering. Elle compare sur les différentes valeurs du nombre potentiel de clusters $K$, la variation totale intra-clusters observées pour différentes valeurs de $K$ à leurs valeurs attendues dans la distribution de référence nulle des données. L'estimation du nombre optimal de clusters sera une valeur qui maximise le "gap-statistic" ou statistique d'écart (c'est-à-dire qui produit la statistique d'écart la plus grande). Cela signifie que la structure de clustering est loin de la distribution aléatoire uniforme des points.

# B Méthodes d'apprentissage supervisé

Nous présentons les principales approches de classification qui sont utilisées pour la prédiction de classe d'appartenance dans les divers chapitres de ce travail. Plus de détails sur les méthodes d'apprentissage peuvent être trouvés dans le livre de Hastie, Tibshirani et Friedman[44].

## B1 Arbre de Classification et de Régression (CART)

Il s'agit d'un partitionnement recursif de l'ensemble des attributs (variables de prédictions) selon les valeurs qui minimise une fonction de coût, telle que la somme des erreurs quadratiques. Cela a pour objectif de mieux séparer les individus et ainsi, les classer. Il existe plusieurs méthodes dont CART (Classification And Regression Tree) [19] et CHAID (CHi-square Adjusted Interaction Decision), le choix se fait par rapport aux critères de partitionnement voulus. Elles s'appliquent peu importe la nature de la variable à prédire : qualitative (discrète) ou quantitative (continue). Ainsi, il existe deux types d'arbres de décision :

— Arbres de classification : comme leur nom l'indique, il s'agit d'expliquer une variable de type nominal (facteur). L'idée est qu' à chaque partitionnement, on cherche parmi toutes les coupures possibles celles qui séparent au mieux les classes [18] ; en donnant deux nœuds fils les plus homogènes possibles et en minimisant une certaine fonction d'impureté considérée des deux nœuds fils par rapport au nœud père ;

— Arbres de régression : la variable expliquée est de type numérique et donc continue et l'objectif est de prédire une valeur la plus proche possible de la valeur observée. La construction d'un tel arbre implique la définition d'une suite de nœuds où chacun permettrait de faire une partition des observations en 2 groupes en se basant sur des variables prédictives.

L'approche inclut donc la définition en premier lieu d'un critère pour sélectionner le meilleur nœud possible à une étape donnée, ensuite établit une condition d'arrêt de la phase découpage et donc un nœud terminal (feuille), l'affectation au nœud terminal de la classe ou la valeur la plus probable, l'élagage de l'arbre dans le cas où les nœuds deviennent trop nombreux et cela en procédant à une sélection d'un sous arbre optimal à partir de l'arbre entier, ensuite il est question de valider l'arbre à l'aide d'une des techniques de validation présentées dans le paragraphe 2.

## B2 Régression logistique

La régression logistique ou régression binomiale estime la probabilité qu'une caractéristique $Y$ (variable à prédire) soit présente (par exemple, la probabilité d'estimation du "succès") étant données les valeurs des variables prédictives ($X$).
Dans cette partie nous présentons le cas d'une seule variable prédictive catégorielle $X$. $\pi = P(Y = 1|X = x)$ représente la chance d'obtenir une valeur $Y = 1$ de la variable à prédire sachant qu'on observe la modalité $x$ de la variable prédictive $X$.

Bien entendu, en régression logistique les observations utilisées sont supposées indépendantes. Bien que n'assumant pas de relation linéaire entre la variable à prédire et les variables prédictives, elle suppose, néanmoins, une relation linéaire entre le logit (cf. plus bas) des variables prédictives et la réponse (la variable à prédire).

Les variables indépendantes utilisées peuvent être des transformations non linéaires des variables indépendantes originales.

Concernant la variable à prédire, il n'est pas impératif qu'elle obéisse à une loi normale, mais elle doit plutôt obéir à une distribution exponentielle (binomiale, Poisson, multinomiale, normale).

La régression logistique binaire suppose la distribution binomiale de la réponse (variable à prédire), de plus l'homogénéité de la variance n'a pas besoin d'être satisfaite.

Elle utilise l'estimation du maximum de vraisemblance (MLE) , comme les modèles linéaires généralisés, plutôt que la méthode des moindres carrés pour estimer les paramètres, et s'appuie donc sur des approximations à grande échelle.

Les modèles logit sont un cas particulier des modèles log-linéaires. Dans le cas où une variable à prédire binaire dans le modèle log-linéaire, il est possible de construire les logits pour aider à l'interprétation du modèle log-linéaire. Certains modèles logit avec seulement des variables catégorielles ont des modèles log-linéaires équivalents.

Le modèle Logit modélise comment la variable de réponse binaire dépend d'un ensemble de variables explicatives il partage le même composant aléatoire : $Y$ qui est binomial et la même composante systématique (fonction linéaire des variables explicatives) avec le modèle Probit, cependant, ils diffèrent dans la fonction de lien.

Le modèle de régression logistique en considérant variable prédictive $X$ comme l'un des facteurs de risque pouvant contribuer au phénomène étudié. La probabilité de succès dépendra des niveaux du facteur de risque. le Logit s'écrit donc sur la forme suivante :

$$\text{logit}(\pi_i) = \log\frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 X_i$$

donc

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

Les valeurs ajustées entre le modèle logit et probit sont souvent très similaires. L'un est rarement mieux adapté (ou pire) que l'autre, bien que l'on puisse observer davantage de différences avec des données faiblement denses.

## B3 Méthodes ensemblistes d'apprentissage

Ensemble learning helps improve machine learning results by combining several models. This approach allows the production of better predictive performance compared to a single model

### B3.1 Random Forest

Dans les modèles de forêts d'arbres aléatoires [17], chaque arbre est construit à partir d'un échantillon prélevé avec remise de l'ensemble d'apprentissage. De plus, au lieu d'utiliser toutes les observations, on en sélectionne aléatoirement un sous-ensemble afin de poursuivre la randomisation de l'arbre. En conséquence, le biais de la forêt augmente légèrement, mais en raison de la moyenne d'arbres moins corrélés, sa variance diminue, ce qui se traduit par un meilleur modèle global.
Dans un algorithme d'arbres complétement aléatoires, le caractère aléatoire est plus accentué car les seuils de fractionnement sont randomisés : au lieu de rechercher le seuil le plus discriminant, des seuils sont tirés aléatoirement pour chaque caractéristique candidate et le meilleur de ces seuils générés aléatoirement est sélectionné en tant que règle de fractionnement. Cela permet généralement de réduire un peu plus la variance du modèle, au détriment d'une augmentation légèrement plus importante du biais.

### B3.2 Bagging

La méthode "Bootstrapped Aggregation" (Bagging) [16] est une des méthodes d'ensembles qui crée plusieurs modèles de même type (utilisant la même méthode de classifiaction) sur différents sous-échantillons du même jeu de données. Les prédictions faites sur chaque modèle sont combinées pour proposer un résultat global. Cette approche est particulièrement intéressante pour les méthodes de classification et de régression sensibles à l'échantillonage telles que les arbres de décison.

### B3.3 Gradient Boosted Machine

Le boosting est une technique itérative qui ajuste le poids d'une observation en fonction de la dernière classification. Si une observation a été mal classée, elle tente d'en

augmenter le poids et inversement. Le boosting réduit en général l'erreur de biais et construit de puissants modèles prédictifs. Cependant, ils peuvent parfois être trop ajustés sur les données de l'échantillon d'apprentissage.

Le "Boosting" a été généralisé et adapté sous la forme de "Gradient Boosted Machine" (GBM) [38] pour être utilisée avec les arbres de décision CART.

# C Méthodes de validation d'un modèle de classification

Il existe plusieurs méthodes de validation de modèles qui régissent le mode expérimental à adopter entre l'apprentissage et le test, parmi lesquelles :

## C1 Hold-out

Cette méthode réserve en général 2/3 des données pour l'apprentissage et le 1/3 restant pour le test. En effet, cette méthode propose de construire le modèle de prédiction sur les 2/3 des données prélevées aléatoirement, et réalise les prédictions sur le 1/3 restant.

## C2 Validation croisée (Cross-Validation)

La validation croisée est une procédure de ré-échantillonnage utilisée pour évaluer des modèles d'apprentissage automatique sur un échantillon de données limité. Cet échantillon doit être subdivisé en $k$ groupes distincts et exhaustives, c'est à dire une partition en $k$-sous ensembles non-vides.
La validation croisée est principalement utilisée dans l'apprentissage automatique appliqué pour estimer les performances de prédiction d'un modèle d'apprentissage automatique sur des données méconnues. C'est-à-dire d'utiliser un échantillon limité afin d'estimer comment le modèle devrait fonctionner en général lorsqu'il est utilisé pour faire des prédictions sur des données non utilisées pendant la formation du modèle.
Cette méthode très connue parce qu'elle est simple à comprendre et qu'elle aboutit généralement à une estimation moins biaisée ou moins optimiste. La procédure générale est la suivante : Commencer par diviser le jeu de données aléatoirement en $k$ groupes ; ensuite pour chaque groupe unique :
Prendre le groupe comme un ensemble de données à retenir ou à tester, et prendre les groupes restants comme un ensemble de données d'apprentissage. Il faut ajuster un modèle sur le groupe d'apprentissage et l'évaluer sur le groupe de test. Conserver le score d'évaluation (prédiction) sans le modèle. Enfin il faut résumer les performances du modèle en utilisant l'échantillon de prédiction du modèle. Il est important de noter que chaque observation de l'échantillon de données est affectée à un unique groupe et y reste pendant la durée de la procédure. Cela signifie que chaque échantillon a la possibilité d'être utilisé 1 fois et de former le modèle $k - 1$ fois. Il existe une variante de cette méthode qui s'appelle le Leave-one-out, il s'agit d'un cas particulier de la Cross validation avec $k = n$ (le nombre d'observations).

# D Évaluation des performances d'un classifieur

Quand nous parlons d'évaluation des performances d'un modèle de classification, on s'intéresse plutôt à ses capacités de prédiction plutôt qu'au temps de calcul, ou temps de convergence pour la classification ou construction de modèles, ou passage à l'échelle. Plusieurs indicateurs sont considérés selon l'objectif de l'étude. Nous en présentons quelques uns qui sont utilisés dans ces travaux.

— **Matrice de confusion** elle sert à confronter la valeur observée confrontée à la prédiction. Elle a la forme suivante :

| Observée | Prédite | | TOTAL |
|---|---|---|---|
| | 0 | 1 | |
| 0 | VN | FP | VN+FP |
| 1 | FN | VP | FN+VP |
| Total | VN+ FN | FP+VP | n |

Table D1 – Matrice de Confusion

— Vrais positifs ($VP$) : représentent le nombre d'observations prédites comme "1" (positives) et qui sont réellement des "1" (positives).

— Faux positifs ($FP$) : représentent le nombre d'observations prédites comme "1" (positives) mais qui sont en réalité des "0" (négatives).

— Taux d'erreur : représente la proportion d'observations mal classées par rapport au nombre total d'observations.

$$error = \frac{(FP + FN)}{n}$$

.

— **Rappel** : Aussi appelé sensibilité (sensitivity), elle correspond au taux de vrais positifs ($TVP$). Cette mesure s'intéresse à pénaliser les mauvaises classifications des observations qui sont en réalité "1" mais qui ont été considérées comme "0". En d'autre terme, elle évalue la sensibilité du modèle à détecter les vrais positifs (VP) quitte à avoir des faux positifs (FP). Elle se calcule de la manière suivante :

$$sensitivity = \frac{VP}{VP + FN}$$

— **Précision** : L'intérêt de cette mesure est de pénaliser le modèle qui n'a pas été précis qui a considéré des "0" en "1" (i.e : On veut avoir le plus de VP et le moins de FP).

Elle se calcule par la formule suivante :

$$precision = \frac{VP}{VP + FP}$$

— **Spécificité** : Elle se calcule par la formule suivante :

$$specificity = \frac{VN}{VN + FP}$$

— **ROC** (Receiver Operating Characteristic) : Il s'agit de la courbe qui représente le taux vrais positifs ($TVP$) par rapport au taux faux positifs ($TFP$). Son utilisation a plusieurs avantages dont le faite de ne pas dépendre de la distribution des classes. Ainsi, elle est robuste car s'affranchit de la connaissance des coûts de classification et de la distribution des classes. L'idée de la courbe ROC est de faire varier le « seuil » de 1 à 0 et, pour chaque cas, calculer le $TVP$ et le $TFP$ que l'on reporte dans un graphique : en abscisse le $TFP$, en ordonnée le $TVP$.

— **AUC** (Aire Under Curve) : C'est l'aire sous la courbe de ROC. Elle indique la probabilité pour que le modèle place un positif devant un négatif (dans le meilleur des cas $AUC = 1$).
Celle-ci informe sur la probabilité que le résultat du modèle, face à deux observations (une positive "1" et une négative "0"), permette de un classement correct. Quand le modèle est parfaitement discriminant, la surface sous la courbe (AUC) vaut 1. Cela signifie donc que, face à deux observations (une positive "1" et l'autre négative "0"), le modèle permet de distinguer dans 100% des cas l'observation positive de celle qui ne l'est pas.
A l'inverse, lorsque le modèle n'est pas discriminant, la probabilité de distinguer l'observation positive de l'observation négative est de 50% (hasard). Dans ce cas, la surface sous la courbe ROC est égale à $0,5$.