

Plan de thèse

1875

Contents

1 Simulation des algorithmes sur des réseaux théoriques	5
1.1 Objectifs	5
1.2 Définitions	5
1.3 Données: Génération aléatoires de graphes	6
1.3.1 Génération de réseaux de flots	6
1.3.2 Génération de line graphes non orientés sous jacent au réseau de flots	6
1.3.3 Processus de simulations	7
1.4 Résultats	8

Chapter 1

Simulation des algorithmes sur des réseaux théoriques

1.1 Objectifs

Les travaux réalisées dans cette partie ont pour but de montrer que les algorithmes proposés (couverture et correction) fournissent un graphe non orienté de distance de Hamming minimale lorsque la matrice d'adjacence de ce graphe contient peu d'erreurs de correlations c'est-à-dire le nombre d'erreurs inférieure à 6. Pour se faire, nous montrons que les sommets, n'appartenant à aucune couverture (labellisés à -1) doivent être traités avec la méthode de correction de degré minimum avec permutation afin d'avoir de meilleurs résultats.

1.2 Définitions

Définition 1 Une erreur de corrélation est l'existence d'une corrélation entre deux arêtes (ajout de 1 dans la matrice) lorsqu'il n'en existe pas. De même, l'absence de corrélation entre 2 arêtes (ajout de 0 dans la matrice) lorsqu'il en existe est aussi une erreur de corrélation.

Définition 2 métrique: distance de Hamming

La métrique utilisée pour différencier deux graphes est la distance de Hamming. La distance de Hamming est le rapport du nombre d'arêtes différentes dans le graphe produit (en comparaison avec le line graphe généré) sur le nombre total d'arêtes dans le line graphe généré.

On considère deux types d'arêtes différentes:

- Les arêtes inexistantes dans le line graphe généré. Elles correspondent à 1 dans la matrice d'adjacence du line graphe produit et à 0 dans celle du line graphe généré.
- Les arêtes inexistantes dans le line graphe produit. Elles correspondent à 1 dans la matrice d'adjacence du line graphe généré et à 0 dans celle du line graphe produit.

Ainsi, une distance de Hamming égale à 0 signifie que le line graphe produit est le line graphe généré tandis que une distance de Hamming supérieure ou égale à 1 signifie que les line graphes généré et produit ne sont pas isomorphes en arêtes.

Définition 3 Un ordre de traitement de sommets est une permutation de l'ensemble des sommets

Exemple 1 Soit $E = (A, B, C, D, E)$ un ensemble de sommets.

Un ordre de traitement est B, A, D, C, E

1.3 Données: Génération aléatoires de graphes

1.3.1 Génération de réseaux de flots

La structure de données utilisée, pour le graphe du réseau de flots, est une *matrice d'adjacence*. Cette matrice d'adjacence est une matrice creuse carrée de n sommets. On définit manuellement le nombre de sommets n et le degré moyen α du graphe. La probabilité d'existence d'une arête est de $prob = \frac{\alpha}{N}$. Toutefois, si le graphe obtenu par la matrice d'adjacence n'est pas connexe, on choisit aléatoirement un sommet de chaque composante connexe et on ajoute une arête entre ces sommets.

Pour orienter les arêtes, on choisit certains sommets comme étant des sources et on les marque puis on leur attribue un numéro d'ordre qui est l'ordre de parcours des sommets (soit un parcours en largeur *Breadth First Search BFS*). À partir des sommets sources, on considère les sommets adjacents non marqués des sources qu'on ajoute dans une *file*. On ajoute des arcs entre les sources et ces sommets adjacents, ensuite on marque les sommets adjacents et enfin on leur ajoute un numéro d'ordre. On reprend l'étape précédente à partir des sommets en tête de file et on s'arrête lorsque la file est vide. S'il existe des sommets adjacents marqués n'ayant aucun arc entre eux, alors on ajoute un arc du sommet de numéro d'ordre minimal vers le sommet de numéro d'ordre maximal. Le graphe obtenu est alors orienté (un *Directed Acyclic Graph DAG*).

L'ajout des flots sur chaque arc se fait par un parcours en largeur (BFS). On définit les valeurs minimales et maximales des grandeurs physiques, sélectionnées selon le réseau énergétique à modéliser. À titre d'exemple, les valeurs choisies pour des grandeurs électriques sont: les intensités $I = [150, 200]$, les tensions $U = [220, 250]$, les puissances $P = [33000, 62500]$.

On débute par les sommets sources dont on génère une valeur aléatoire comprise dans l'un des intervalles de ces grandeurs. Chaque arc sortant du sommet source reçoit un flot égal à la somme des valeurs aléatoires sur les arcs entrants du sommet source multipliée par le facteur ϵ (désignant les pertes joules) et divisée par le degré sortant de ce sommet si nous avons comme grandeurs les intensités et les puissances. Dans le cas de grandeurs comme les tensions, le flot de chaque arc sortant est la valeur aléatoire multipliée par le facteur ϵ . On propage les valeurs des grandeurs physiques jusqu'à ce qu'on arrive aux sommets puits. L'application de ces règles de flots permettent de vérifier la *loi de conservation des noeuds*.

1.3.2 Génération de line graphes non orientés sous jacent au réseau de flots

Nous considérons le réseau de flots généré comme un graphe non orienté parce que l'orientation des arêtes n'a aucune influence sur les noeuds en commun des arcs.

Nous nommons les arêtes du graphe et désignons par *arêtes corrélées*, des arêtes ayant un sommet commun. Si des arêtes sont correlées alors nous leur attribuons 1 comme valeur de corrélation. Dans le cas contraire, la valeur de corrélation vaut 0.

La matrice binaire de corrélation est la *matrice d'adjacence du line graphe* non orienté de notre réseau de flots généré. Cette matrice est notée *matE*.

Rappelons qu'un line graphe admet une couverture en cliques c'est-à-dire que chaque sommet est contenu dans deux cliques au maximum et chaque arête est couverte par une seule clique.

1.3.3 Processus de simulations

On génère 500 graphes de flots de $n = 30$ sommets ayant un degré moyen $\bar{d} = 5$. On en déduit également 500 line graphes non orientés de 150 sommets. Soit la probabilité p suivant la loi normale $N(0, 1)$ correspondant à la modification de la corrélation entre arêtes dans un line graphe.

- si $p = 0 \rightarrow$ on supprime uniquement des arêtes dans le line graphe.
- si $p = 0.5 \rightarrow$ on ajoute et supprime équiprobablement des arêtes.
- si $p = 1.0 \rightarrow$ on ajoute uniquement des arêtes dans le line graphe.

Nous décidons de modifier $k = \{1, 2, 3, \dots, 9\}$ corrélations dans la matrice $matE$ et nous appliquons les algorithmes de couverture et de correction. À la fin de l'algorithme de couverture, s'il existe des sommets du line graphe non couverts par *une ou deux cliques*, on les étiquette avec la valeur -1 et nous appliquons l'algorithme de correction sur l'ensemble de sommets à -1 , noté *sommets_1*, selon les méthodes suivantes:

- méthode 1 : degré minimum avec remise.
Elle consiste à sélectionner le sommet de degré minimum dans l'ensemble *sommets_1*, à appliquer l'algorithme de correction afin de modifier *matE* et enfin à re-exécuter les deux algorithmes sur la matrice *matE* modifiée.
- méthode 2 : coût minimum avec remise.
Elle consiste à sélectionner le sommet de coût minimum dans l'ensemble *sommets_1*, à appliquer l'algorithme de correction afin de modifier *matE* et enfin à re-exécuter les deux algorithmes sur la matrice *matE* modifiée.
- méthode 3 : coût minimum avec permutation des sommets de *sommets_1*.
Elle consiste à choisir un nombre fini d'ordre de traitements de sommets à -1 puis de choisir l'ordre qui a le coût de modification de la matrice *matE* minimale.
- méthode 4 : degré minimum avec permutation des sommets de *sommets_1*.
Elle consiste à choisir un nombre fini d'ordre de traitements de sommets à -1 puis de traiter chaque sommet en fonction de son degré et enfin de choisir l'ordre qui a le coût de modification de la matrice *matE* minimale.

Sur chaque line graphe, on modifie $k = \{1, 2, 3, \dots, 9\}$ corrélations d'arêtes $\alpha = \{1, \dots, 5\}$ fois et on note la matrice obtenue $matE_{k,\alpha}$. On applique les algorithmes de couverture et de correction sur la matrice *matE_{k,α}* et on compare

1. le line graphe généré dont la matrice est *matE* et le graphe produit (qui est aussi un line graph) dont la matrice *matE_{k,α}* a été modifiée par l'algorithme de correction si cette matrice n'était pas un line graphe. On obtient la distance de Hamming (notée $DH_{k,\alpha}$) entre *matE* et *matE_{k,α}* qui est le nombre d'arêtes différentes entre ces deux line graphes.
2. le line graphe généré modifié de k corrélations dont la matrice est *matE_k* et le graphe produit dont la matrice *matE_{k,α}* a été modifiée par l'algorithme de correction si cette matrice n'était pas un line graphe. On obtient la distance line (notée $DL_{k,\alpha}$) entre *matE_k* et *matE_{k,α}* qui est le nombre d'arêtes différentes entre ces deux line graphes.

On définit par les variables moy_DH et moy_DL , les moyennes respectives des distances de Hamming (notée $DH_{k,\alpha}$) et des distances line (notée $DL_{k,\alpha}$) pour une valeur donnée de k et pour tout $\alpha = \{1, \dots, 5\}$.

$$moy_DH_k = \sum_{\alpha=1}^5 DH_{k,\alpha} \quad moy_DL_k = \sum_{\alpha=1}^5 DL_{k,\alpha} \quad (1.1)$$

1.4 Résultats

Dans cette partie, nous présentons les distributions des distances line et de Hamming moyennées selon les méthodes de degré minimum avec remise et de coût minimum avec permutation des sommets de *sommets_1* (sommets labellisés à -1) afin de comprendre les expérimentations effectuées. Nous expliquons le choix de la méthode de degré minimum avec permutation et montrons que les algorithmes (couverture et correction) proposent de meilleurs résultats lorsque la matrice de corrélation possède plus de corrélations *faux positives* que de corrélations *faux négatives* c'est-à-dire peu d'erreurs de corrélations.

Rappelons que les tests sur les figures 1.3 et 1.1 considèrent qu'on ajoute et supprime uniformement des arêtes c'est-à-dire que la probabilité de modification des corrélations est égale à $p = 0.5$. Nous ne décrirons que 2 méthodes (degré minimum avec remise et coût minimum avec permutation) afin que le lecteur puisse comprendre les tests et courbes réalisées.

Commençons par la méthode de degré minimum avec remise. La figure 1.1 représente les distributions des distances line moy_DL (à gauche) et de Hamming moy_DH (à droite) pour $k = \{1, 2, 3, \dots, 9\}$ corrélations modifiées. Chaque batonnet correspond au pourcentage de line graphs associé à une distance line ou une distance de Hamming. À titre d'exemple, le pourcentage de line graphs proposés dont $moy_DL = 2$ (noté $X_{moy_DL==2}$) est égal à 43% pour $k = 2$ corrélations modifiées c'est-à-dire $X_{moy_DL==2} = 43\%$ des line graphs proposés ont une arête différente. Alors que ce même pourcentage pour $k = 4$ est de 13%. Les distributions de distances line et de Hamming sont respectivement asymétrique pour $k = 1, 2, 3$ corrélations modifiées et symétrique (gaussienne) pour $k > 3$ corrélations modifiées. En effet, le pic de l'histogramme pour une distance line moy_DL est de 60% pour $k = 1$ corrélation modifiée alors qu'il est de 12% dans la courbe de $k = 5$ corrélations modifiées (voir figure 1.1). Pour comprendre la différence de propriété des histogrammes, nous allons étudier les corrélations entre les distances line moy_DL et de Hamming moy_DH en définissant la fonction F comme ci dessous:

$$F_{k,\alpha} = \frac{|moy_DL_{k,\alpha} - moy_DH_{k,\alpha}|}{\max(moy_DL_{k,\alpha}, moy_DH_{k,\alpha})} \quad (1.2)$$

La fonction de répartition cumulée de F est dans la figure 1.2. Si $F = 1$ alors les corrélations modifiées correspondent aux arêtes différentes entre les deux line graphs ($moy_DH = 0$ et $moy_DL = k$). Par contre $F = 0$ signifie que l'algorithme de correction ajoute ou supprime des arêtes, différentes des corrélations modifiées. Par exemple dans la figure 1.2, les corrélations entre moy_DL et moy_DH sont proches de 0 ($F \approx 0$) pour $k < 3$. Cela signifie que la courbe est en escalier avec une grosse marche. Cependant, pour $k > 2$, la courbe tend vers une fonction racine carrée traduisant que les corrélations sont dans le même ordre.

Quant à la méthode du coût minimum avec permutation des sommets de *sommets_1*, nous affichons les histogrammes de deux catégories de distances. À gauche, nous avons la distribution

des distances lines et à droite celle des distances de Hamming (fig 1.3). Chaque batonnet correspond au pourcentage de line graphes associé à une distance line ou une distance de Hamming. À titre d'exemple, le pic de l'histogramme pour une distance line moy_DL est de 75% pour $k = 1$ corrélations modifiées tandis que celui pour $k = 5$ corrélations modifiées est de 12.5% (voir figure 1.3). Les distributions de distances line et de Hamming sont respectivement asymétrique pour $k = 1, 2, 3$ corrélations modifiées et symétrique (gaussienne) pour $k > 3$ corrélations modifiées. En effet, dans la distribution asymétrique, le pic est assimilé au pourcentage de line graphes proposés dont la distance line est égale au nombre de corrélations modifiées $moy_DL = k$. Cela se vérifie dans la courbe des distances de Hamming (figure à droite) en ce sens que le pic pour la représentation de distance de Hamming correspond à $moy_DH = 0$ et le pourcentage $X_{moy_DH=0}$ de line graphes proposés dont $moy_DH = 0$ est identique au pourcentage $X_{moy_DL=k}$ (pourcentage de line graphes dont $moy_DL = k$). Cela s'explique par le fait que l'algorithme de correction retrouve les corrélations modifiées dans la matrice $matE_k$. Dans la distribution gaussienne, le pic est au alentours de la moyenne des distances line. Le pic est environ trois fois le nombre $k > 3$ de corrélations modifiées pour les distances moy_DH et moy_DL . Nous cherchons à savoir l'évolution de la distance line moy_DL par rapport à celle de Hamming moy_DH . Pour cela, on calcule leur corrélation par la fonction F définie en 1.2 et dont la fonction de répartition cumulée est dans la figure 1.4. Si moy_DL et moy_DH évoluent de manière opposé (l'un est supérieur à 0, l'autre est égal à 0), la fonction F est égale à 1. Par contre, s'ils sont très corrélés (i.e évoluent identiquement) alors $F = 0$. Ainsi lorsqu'elles évoluent dans le même ordre, on obtient la courbe racine carrée. Tel est le cas dans la figure 1.4 pour $k > 3$. Les courbes tendent vers la fonction exponentielle quand k devient grand. Par ailleurs $F \approx 0$ quand k est très petit (fig 1.4, $k < 3$).

Nous recherchons la meilleure méthode de correction parmi quatre méthodes qui sont : degré minimum avec remise, coût minimum avec remise, degré minimum avec permutation et coût minimum avec permutation. Notre objectif étant de trouver le line graphe le plus proche du line graphe initial, nous allons utiliser la moyenne des distributions de Hamming pour comparer les méthodes parce que le calcul de la distance line se base sur la matrice $matE_k$ et cette matrice ne forme pas un line graphe après la suppression de k arêtes. La figure 1.5 affiche les courbes des distances de Hamming moyennées pour chaque méthode lorsqu'on modifie k corrélations. Nous remarquons que la pire des méthodes est celle de coût minimum avec remise (couleur en rouge avec un rond) car elle est au dessus des autres et la meilleure est celle de *degré minimum avec permutation* (couleur en rouge avec un carré) car elle propose des line graphes ayant le nombre minimum d'arêtes différentes pour $k > 5$.

Nous retenons, pour la suite, la méthode de degré minimum avec permutation comme méthode de correction des sommets n'appartenant à aucune couverture (sommets de *sommets_1*).

Rappelons que la variable p est la probabilité de modification des corrélations dans la matrice $matE$. Faisons varier cette variable p de 0 à 1 par pas de 0.1 dans le but de visualiser l'impact de corrélations *fausses positives* et *fausses négatives* dans l'exécution des algorithmes. La figure 1.6 résume l'évolution des distances de Hamming moy_DH en fonction de $k \in \{1, \dots, 9\}$ pour différentes valeurs de $p \in \{0, 0.1, \dots, 1\}$. Nous constatons que les algorithmes donnent de meilleurs résultats pour $p = 1$ et de mauvais résultats pour $p = 0$. En d'autres termes, lorsqu'on ajoute que des arêtes (au nombre de 9) i.e $p = 1$ dans le line graphe, Ces arêtes sont supprimées pour retrouver les line graphes générés. Cependant, le nombre de arêtes différentes est multiplié par 2 lorsqu'on ne supprime que des arêtes. Tel est le cas pour $k = 9$ arêtes supprimées ($p = 0$) ou l'algorithme de correction ajoute environ 15 arêtes de plus dans les line graphes. Nous pensons

que le meilleur compromis est la probabilité $p = 0.8$ parce que pour peu de corrélations modifiées ($k < 5$) les graphes produits et générés sont différents de $k < 5$ arêtes correspondant aux k corrélations effectuées et au-delà $k \geq 5$, le nombre d'arêtes différentes est fonction du nombre de corrélations faites multiplié par 1.5. Cela signifie qu'il faut, dans la matrice de corrélation, 20% de corrélations *fausses negatives* et 80% de corrélation *fausses positives*.

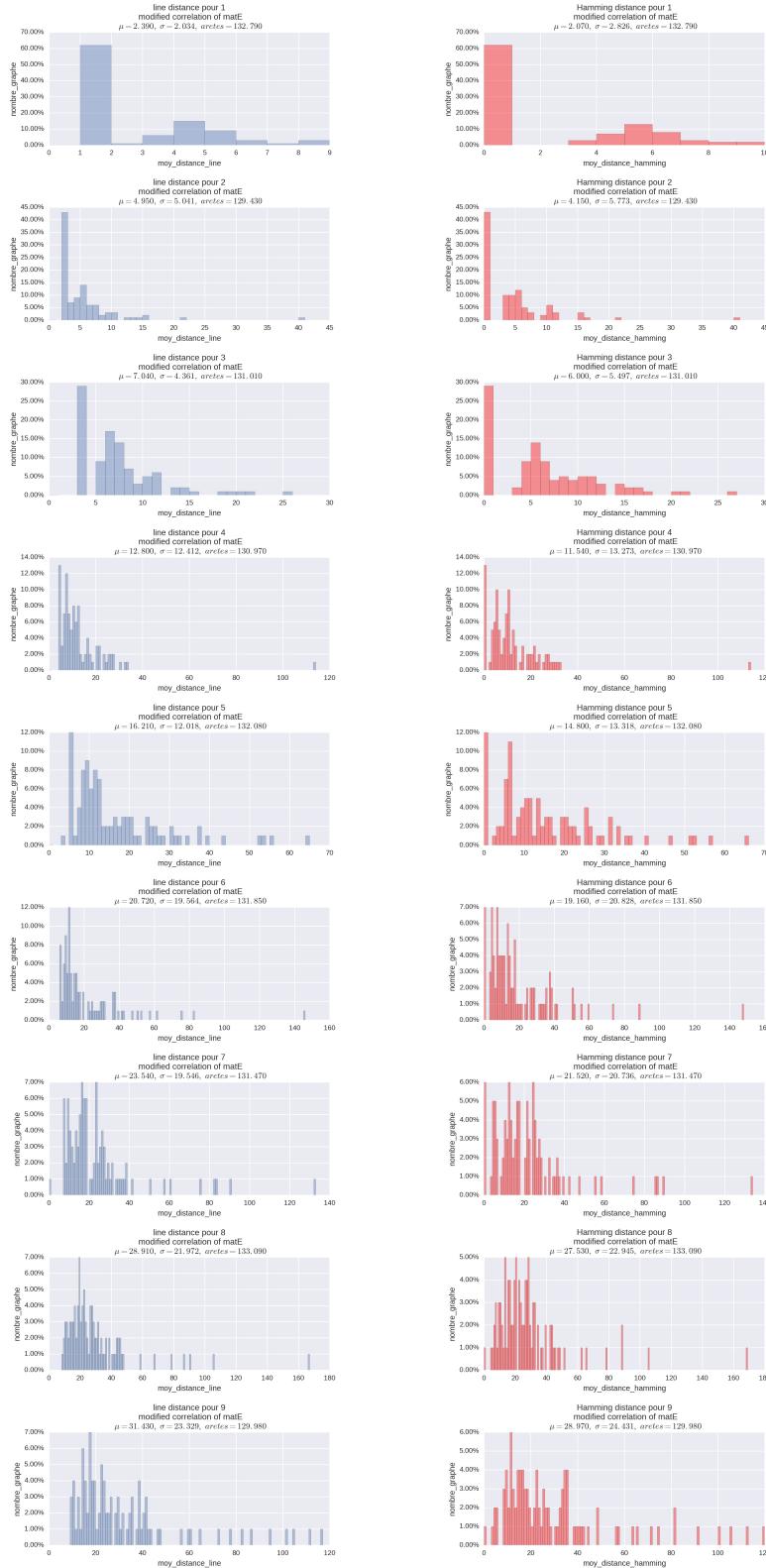


Figure 1.1: Méthode coût minimum avec remise : distribution des distances line moy_DL et de Hamming moy_DH pour $k \in [1, \dots, 9]$ de corrélations alterées

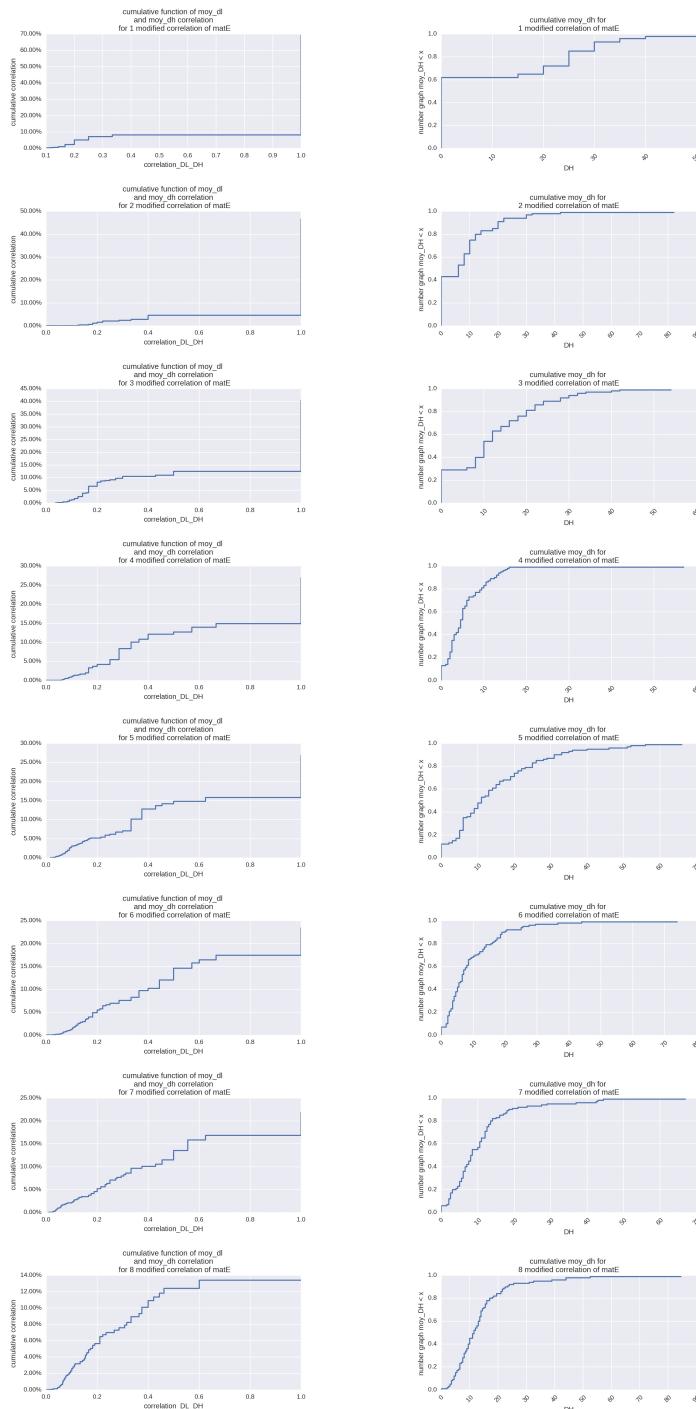


Figure 1.2: Méthode coût minimum avec remise : fonction de répartition cumulée des distances line moy_DL et de Hamming moy_DH pour $k \in [1, \dots, 9]$ de corrélations alterées

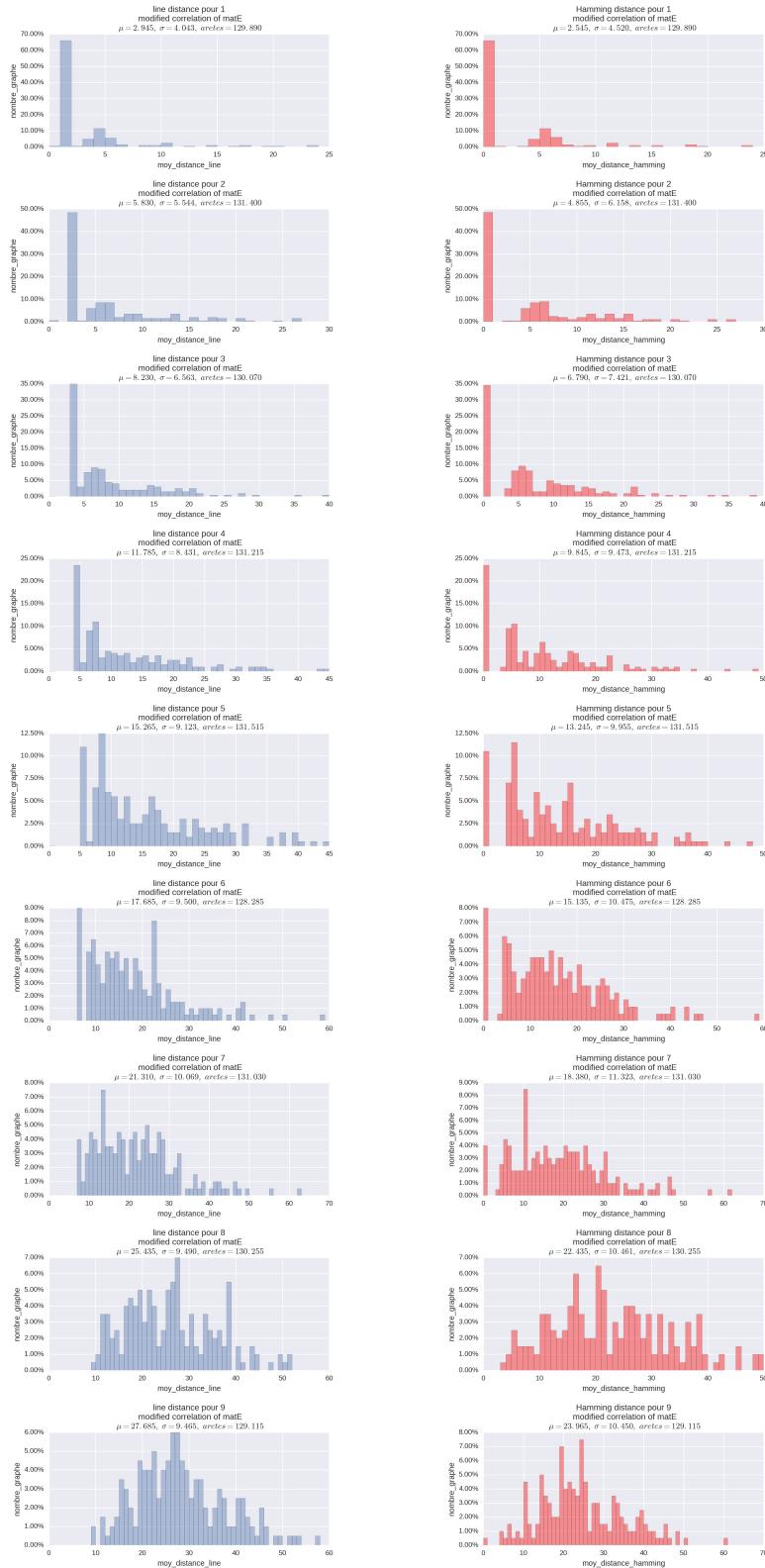


Figure 1.3: Méthode degré minimum avec permutation : distribution des distances line *moy_DL* et de Hamming *moy_DH* selon le nombre $k \in [1, \dots, 9]$ de corrélations alterées

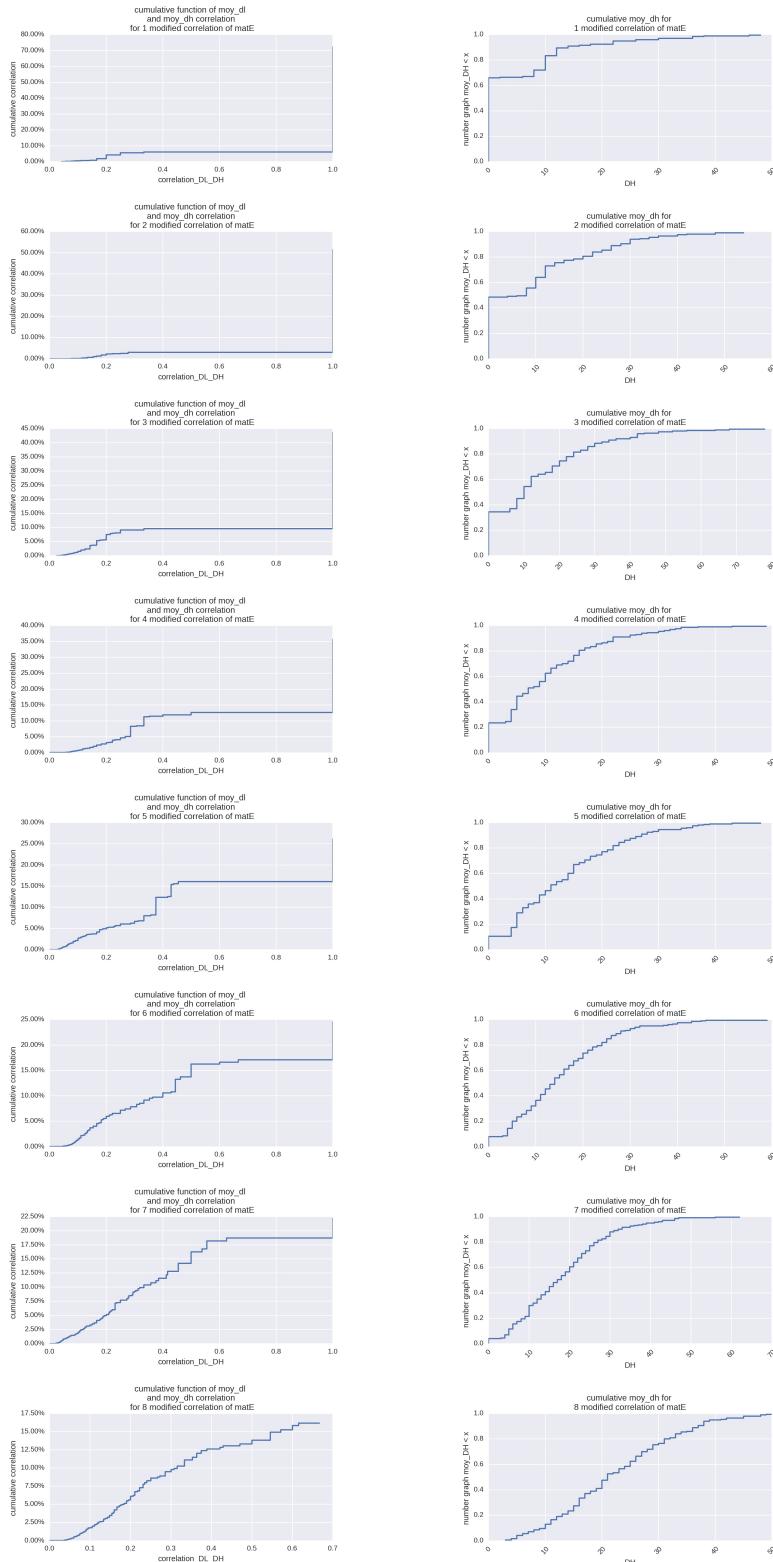


Figure 1.4: Méthode degré minimum avec permutation : fonction de répartition cumulée des distances line moy_DL et de Hamming moy_DH selon le nombre $k \in [1, \dots, 9]$ de corrélations alterées

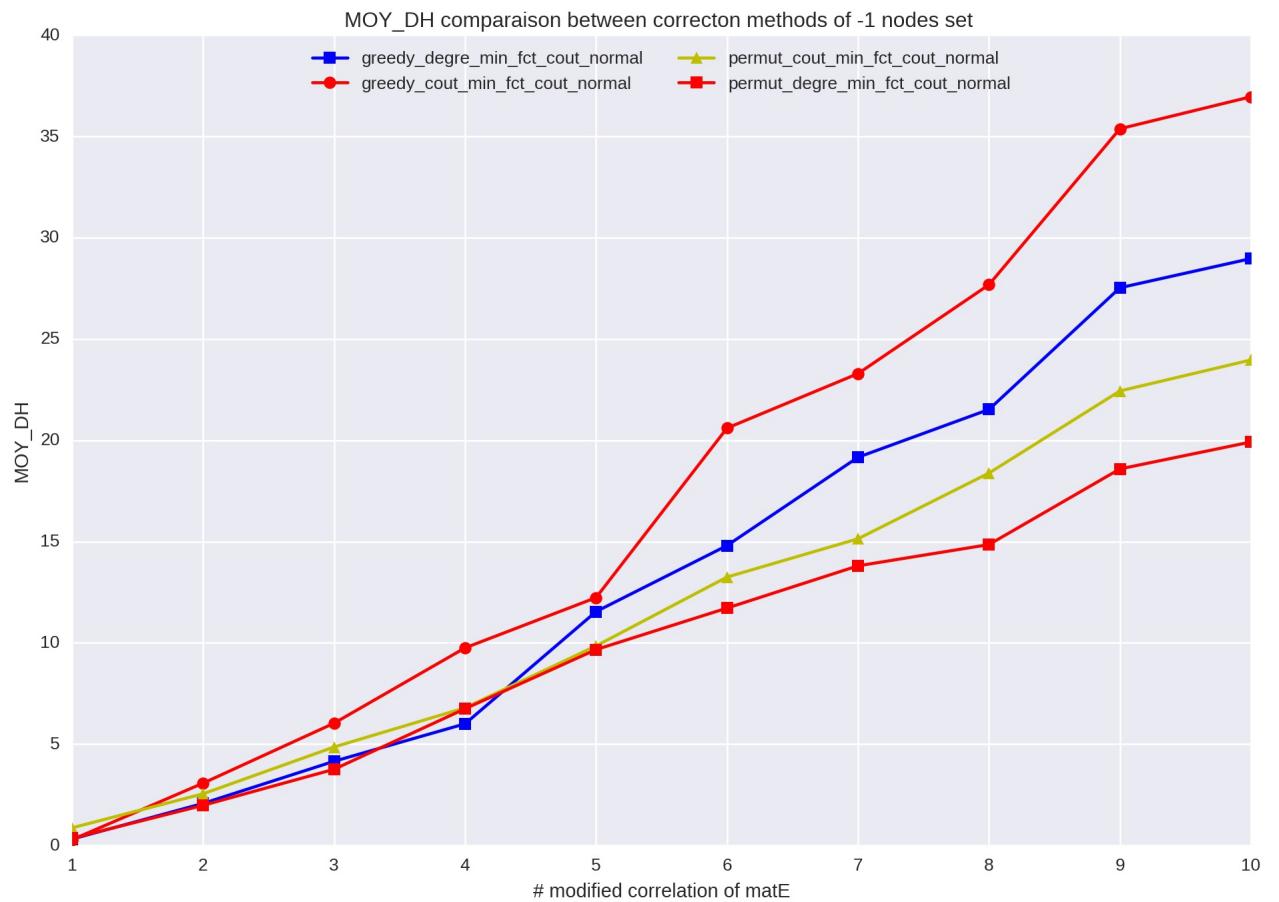
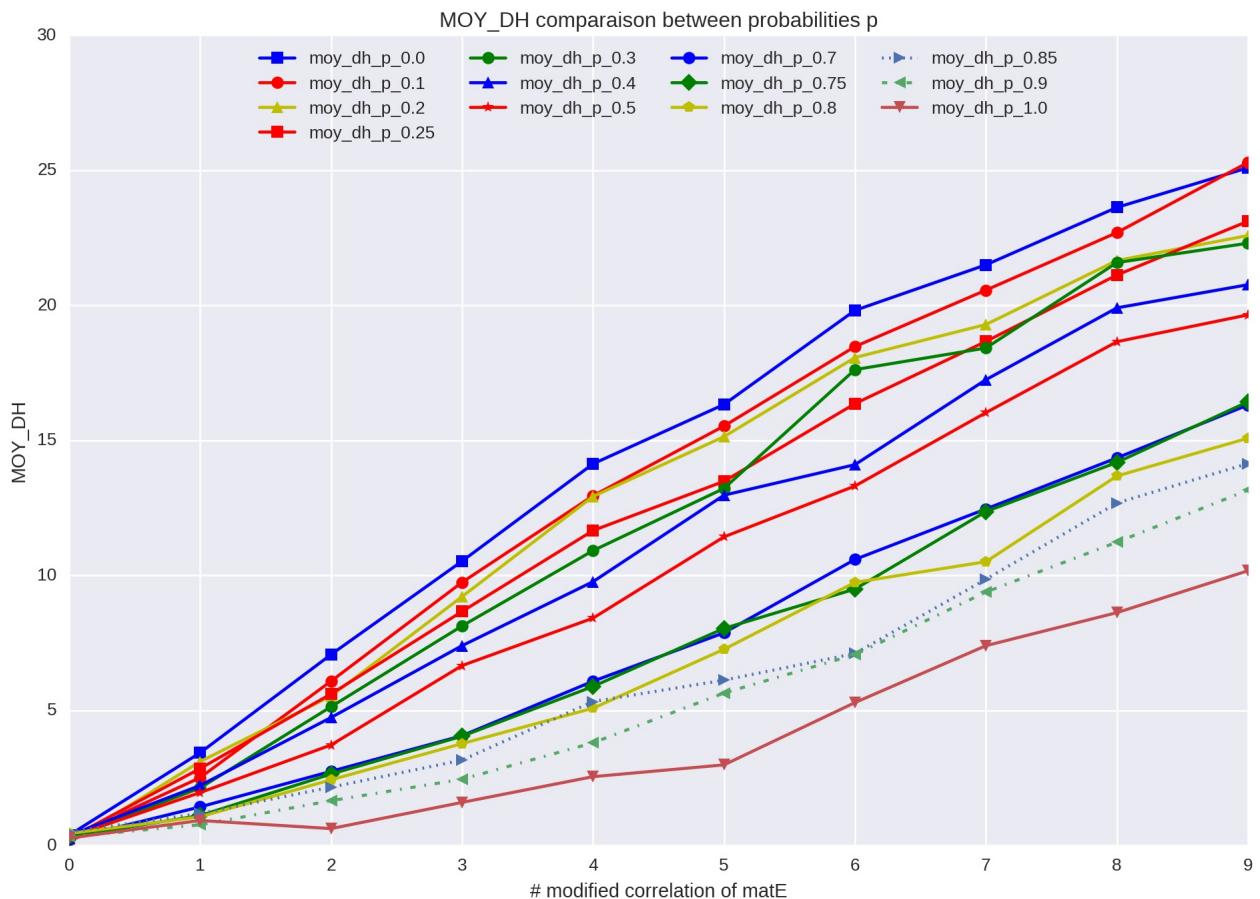


Figure 1.5: Comparaison des différentes méthodes de correction de sommets pour $k \in [1, \dots, 9]$

Figure 1.6: Impact des différentes probabilités p sur les distances de Hamming pour $k \in [1, \dots, 9]$