

# Plan de thèse

1875



# Contents

<b>1 Simulation des algorithmes sur des réseaux théoriques</b>	<b>5</b>
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 sous jacent au réseau de flots non orienté . . . . .	7
1.3.3 Prise en compte de l'erreur de corrélation dans la matrice <i>matE</i> . . . . .	7
1.4 Résultats . . . . .	8
1.4.1 Distribution de la méthode de permutation aléatoire . . . . .	9
1.4.2 Comparaison des méthodes de correction . . . . .	9
1.4.3 Influence des erreurs de corrélations sur les distributions . . . . .	10
1.4.4 Impact de la fonction de coût sur les distributions . . . . .	10



# 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 plus de corrélations *fausses positives* que de corrélations *fausses négatives* et peu d'erreurs de corrélations. Le nombre d'erreurs doit être inférieur à 6 pour une probabilité de corrélations fausses positives supérieure à 0.8 . Pour se faire, nous montrons que les sommets, n'appartenant à aucune couverture (sommets  $\in$  *sommets\_1*) doivent être corrigés avec la méthode de **permutation aléatoire minimum** afin d'avoir de meilleurs résultats.

### 1.2 Définitions

**Définition 1** Une erreur de corrélation est l'*existence ou l'absence de corrélation entre deux arêtes (ou arcs)* lorsque, respectivement, il n'existe pas de corrélations ou il en existe une.

L'*absence de corrélation (corrélation fausse négative)* est désignée par 0 dans la matrice d'adjacence tandis que l'*existence de corrélation (corrélation fausse positive)* a une valeur 1 dans cette matrice.

**Définition 2** Une corrélation entre arêtes (ou arcs) est l'*existence d'un sommet commun aux arêtes (ou arcs)*. Ce sommet commun peut être source, destination ou intermédiaire comme présenté dans la figure 1.1

**Définition 3** métrique: *distance de Hamming*

La métrique utilisée pour différencier deux graphes est la distance de Hamming. La distance de Hamming est le nombre d'arêtes (ou arcs) différentes entre deux graphes ayant le même ensemble de sommets.

Ainsi, une distance de Hamming égale à 0 signifie que les deux graphes sont identiques. Tandis que une distance de Hamming égale à  $k$  signifie qu'il a  $k$  arêtes différentes entre ces deux graphes.

**Définition 4** fonction de coût d'un sommet

La fonction de coût d'un sommet est le coût de chaque arête ajoutée ou supprimée lorsqu'on applique l'algorithme de correction sur ce sommet.

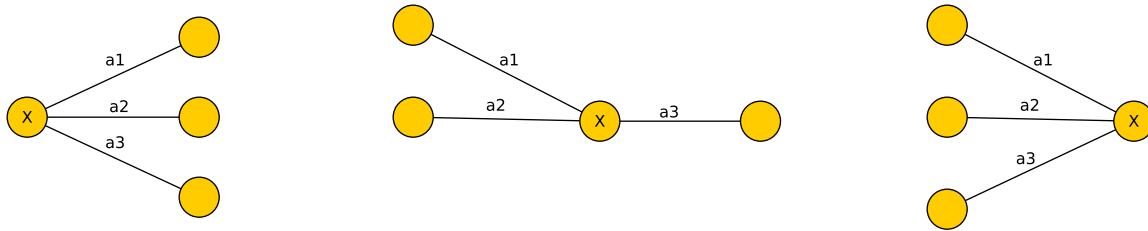


Figure 1.1: De la gauche à la droite: sommet  $X$  source, sommet  $X$  intermédiaire, sommet  $X$  destination

Le coût d'une arête peut être

- unitaire : l'ajout et la suppression valent 1.
- normal: la suppression est la probabilité de l'arête et l'ajout vaut 1 moins cette probabilité.
- carré : la suppression est la probabilité au carré de l'arête et l'ajout vaut 1 moins cette probabilité au carré.
- quadratique : la suppression est la probabilité de l'arête à la puissance 4 et l'ajout vaut 1 moins cette probabilité à la puissance 4.

## 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 carrée de  $n$  sommets. On définit le nombre de sommets  $n$  et le degré moyen  $\alpha$  du graphe. La probabilité d'existence d'une arête est de  $proba = \frac{\alpha}{N}$ . Toutefois, si le graphe obtenu à partir de 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 réalise un tri topologique avec un parcours en largeur *Breadth First Search (BFS)* du graphe non orienté généré, à partir de certains sommets choisis comme des sources. Chaque sommet  $x$  a un ordre topologique  $D_x$  et l'arête  $a_{xy}$  devient soit l'arc  $a_{xy}$  si  $D_x < D_y$  soit l'arc  $a_{yx}$  si  $D_x > D_y$ . 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. Ces valeurs sont 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 flots sur les arcs entrants du sommet source multipliée par le facteur  $\epsilon$  (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 le flot multiplié par le facteur  $\epsilon$ . 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 sous jacent au réseau de flots non orienté

Nous nommons les arcs du graphe et nous construisons la matrice de corrélation selon la notion de corrélation entre arcs (0 aucun sommets en commun sinon 1). La matrice binaire de corrélation est la *matrice d'adjacence du line graphe* sous jacent au graphe non orienté de notre réseau de flots généré. Cette matrice est symétrique et est notée  $matE$ .

Si la matrice  $matE$  ne contient *aucune erreur de corrélation* alors elle 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. Dans le cas contraire, nous utilisons l'algorithme de correction afin de proposer une couverture pour chaque sommet n'appartenant à aucune clique.

Notre but est de proposer une couverture en cliques des matrices  $matE$  erronées afin que la distance de Hamming entre le line graphe de matrice d'adjacence  $matE$  et le line graphe générée soit minimum.

### 1.3.3 Prise en compte de l'erreur de corrélation dans la matrice $matE$

On génère 500 graphes de flots de  $n = 30$  sommets ayant un degré moyen  $\bar{d} = 5$ . On en déduit également 500 line graphes de 150 sommets. On introduit trois paramètres  $k, p, prob$ :

1.  $k$  désigne le nombre de corrélations erronées à ajouter dans la matrice  $matE$ . Dans nos simulations,  $k \in [1, 9]$ .
2.  $p$  désigne le type d'erreurs à réaliser, soit corrélation *fausses positives* (ajout d'arêtes) soit corrélation *fausses négatives* (suppression d'arêtes) soit les deux. Cette variable  $p \in [0, 1]$  varie par pas de 0.1. Par exemple
  - 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.
3.  $prob$  désigne la probabilité associée à une corrélation selon le type d'erreurs effectué dans  $matE$ . Ce paramètre est important car les valeurs de corrélations calculées ne sont pas binaires mais probabilistes. Nous en reparlerons dans la section 1.4.4.

Pour ajouter des erreurs à notre matrice  $matE$  de corrélation correcte, on tire aléatoirement  $k$  cases non encore modifiées. Nous mettons chaque case et sa case symétrique à 1 si la probabilité de la case (généré selon la loi uniforme) est inférieure ou égale à  $p$ . Si cette case est déjà à 1, on choisit une autre case.

Nous décidons de modifier  $k \in [1, 9]$  corrélations, selon  $p = 0.5$ , 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 ajoute à l'ensemble *sommets\_1* et nous appliquons l'algorithme de correction sur chaque sommet de *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 une permutation dont les sommets sont classés par ordre croissant de leur coût de modification de la matrice *matE* et à appliquer l'algorithme de correction sur cette permutation.

- méthode 4 : degré minimum avec permutation des sommets de *sommets\_1*.

Elle consiste à choisir une permutation dont les sommets sont classés par ordre croissant de leur degré et à appliquer l'algorithme de correction sur cette permutation.

- méthode 5 : permutation aléatoire des sommets de *sommets\_1*.

Elle consiste à choisir aléatoirement  $N$  permutations puis à appliquer l'algorithme de correction et à sélectionner la permutation ayant un coût et une distance de Hamming minimum.

Considérons  $k \in [1, 9]$  le nombre d'erreurs de corrélations dans *matE*,  $\alpha \in [1, 5]$  le nombre de fois qu'on applique  $k$  erreurs sur la matrice *matE<sub>k</sub>*,  $LG_{k,\alpha}$  le line graphe dont on a modifié  $k$  corrélations  $\alpha$  fois et  $LG'_{k,\alpha}$  le line graphe fourni par les algorithmes de couverture et de correction à partir du line graphe  $LG_{k,\alpha}$ .

On note les matrices d'adjacence de  $LG_{k,\alpha}$  et  $LG'_{k,\alpha}$  respectivement *matE<sub>k,\alpha</sub>* et *matE'<sub>k,\alpha</sub>*. En comparant

1.  $LG$  et  $LG'_{k,\alpha}$ , on obtient la distance de Hamming notée  $DH_{k,\alpha}$ .

2.  $LG_{k,\alpha}$  et  $LG'_{k,\alpha}$ , on a la distance line notée  $DL_{k,\alpha}$ .

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  $\alpha \in [1, 5]$ .

$$moy\_DH_k = \sum_{\alpha=1}^5 DH_{k,\alpha} \quad moy\_DL_k = \sum_{\alpha=1}^5 DL_{k,\alpha} \quad (1.1)$$

## 1.4 Résultats

Nous décrirons les distributions des distances line et de Hamming moyennées (*moy\_DL* ou *moy\_DH*) pour une méthode de correction (aléatoire) puis nous comparons les cinq méthodes de correction en se basant sur les distances de Hamming moyennées. Nous expliquons le choix de la méthode de *permutation aléatoire* 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* et aussi peu d'erreurs de corrélations ( $k < 6$ ).

Nous présentons également l'impact de la fonction de coût dans les distributions de distances de Hamming.

### 1.4.1 Distribution de la méthode de permutation aléatoire

La figure 1.2 représente les histogrammes de la méthode de permutation aléatoire pour une probabilité  $p = 0.5$  et une fonction de coût normal  $F_1$ . À gauche de cette figure, nous avons les distributions des distances lines et à droite celle des distances de Hamming. Ces histogrammes partent d'une corrélation modifiée (en haut de la figure 1.2) à 9 corrélations modifiées (en bas de la figure 1.2).

Dans la représentation d'une distribution, chaque batonnet correspond au pourcentage de graphes associé à une distance moyennée (line ou Hamming). À titre d'exemple, pour  $k = 2$  corrélations modifiées, le pic de l'histogramme est à  $moy\_DL = 2$  pour les distances lines et de  $moy\_DH = 0$  pour les distances de Hamming. Les pourcentages de graphes pour  $moy\_DL = 2$  et  $moy\_DH = 0$  sont identiques c'est-à-dire 50%.

Nous avons, dans les histogrammes, une droite verticale  $y = k$  (en rouge) correspondant à  $k$  erreurs de corrélations. On constate que les pics sont à gauche sur la droite  $y = k$  pour  $k \leq 6$  corrélations modifiées. Cela signifie que le pourcentage de graphe dont  $moy\_DL = k$  et  $moy\_DH = 0$  est le plus élevé (on retrouve en général le line graphe initial) et entraîne une asymétrie de la distribution des distances pour  $k \leq 6$ . En revanche, au-delà de  $k > 6$ , l'algorithme de correction retrouve peu de corrélations modifiées. Ce qui place le pic à droite de la droite  $y = k$  et fournit une distribution gaussienne des distributions.

Le fait que les distributions de distance line et de Hamming soient, toutes les deux, asymétriques (pour  $k \leq 6$ ) soient symétriques (pour  $k > 6$ ) nous interrogent sur l'évolution des distributions de distance line par rapport à celle de Hamming. 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 1.3.

$$F_{k,\alpha} = \frac{|moy\_DL_{k,\alpha} - moy\_DH_{k,\alpha}|}{\max(moy\_DL_{k,\alpha}, moy\_DH_{k,\alpha})} \quad (1.2)$$

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 1.3 pour  $k > 6$ . Les courbes tendent vers la fonction racine carrée quand  $k$  devient grand. Par ailleurs  $F \approx 0$  quand  $k$  est très petit (fig 1.3,  $k \leq 6$ ).

### 1.4.2 Comparaison des méthodes de correction

Nous recherchons la meilleure méthode de correction parmi les cinq méthodes énumérées plus haut. Pour se faire, on dispose des distributions de distances line et de Hamming, des histogrammes, des fonctions de répartitions de ces distributions et aussi des moyennes de distances line/Hamming associées à  $k$  corrélations érronées pour chacune des méthodes. Nous utilisons la moyenne des distances de Hamming pour la comparaison de méthodes parce que la distance de Hamming permet d'évaluer la différence entre le graphe de base et celui prédit par nos algorithmes.

Rappelons qu'on a  $p = 0.5$  et la fonction de coût est unitaire ( $F_0$ ). La figure 1.4 affiche les courbes des différentes méthodes pour des distances de Hamming moyennées en fonction des  $k$  erreurs de corrélations.

Nous constatons que la pire des méthodes est celle de degré minimum avec remise (en rouge avec un rond) car elle est au dessus des autres et la meilleure est celle de *de permutation aléatoire* (en jaune avec un triangle) car elle propose des line graphs ayant le nombre minimum d'arêtes

différentes pour  $\forall k$ .

Nous retenons, pour la suite, la méthode de permutation aléatoire comme méthode de correction des sommets n'appartenant à aucune couverture (sommets  $\in sommets\_1$ ).

### 1.4.3 Influence des erreurs de corrélations sur les distributions

Faisons varier la variable  $p \in [0, 1]$  par pas de 0.1 dans le but de visualiser l'impact de corrélations *fausses positives* et *fausses négatives* dans l'exécution des algorithmes. Rappelons que l'ajout et la suppression d'arêtes ont le même coût de traitement c'est-à-dire 1. La figure 1.5 résume l'évolution des types d'erreurs de corrélations ( $p$ ) pour des distances de Hamming *moy-DH* en fonction de  $k \in [1, 9]$  erreurs de corrélations.

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 corrélations *fausses positives* i.e  $p = 1$  dans la matrice *matE*, les algorithmes proposent, dans la majorité des cas, un line graphe dont ces arêtes sont supprimées. Cela s'illustre dans la figure 1.6 où l'ajout de 5 corrélations *fausses positives* influencent très peu les line graphes proposés puisqu'ils sont identiques aux line graphes initiales dans 45% des cas. En revanche, ce pourcentage baisse avec beaucoup de corrélations *fausses positives*. Tel est le cas pour  $k = 9$  corrélations où le taux est de 24%.

Par ailleurs, le mauvais résultat obtenu pour des probabilités  $p < 0.5$  s'explique par la correction de très peu de corrélations (inférieure à 1/3 des corrélations modifiées) entraînant l'ajout d'arêtes au voisinage des sommets à traiter (sommets  $\in sommets\_1$ ).

Nous pensons que le meilleur compromis est la probabilité  $p = 0.8$  parce que, pour peu de corrélations modifiées ( $k < 5$ ), les lines graphes produits et générés diffèrent 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*.

Que se passe-t-il si on priorise l'ajout de corrélations *fausses negatives* à chaque traitement c'est-à-dire la suppression d'arêtes. Dans ce cas, l'ajout d'arêtes *fausses positives* a un coût moins important que celui des *fausses negatives*.

### 1.4.4 Impact de la fonction de coût sur les distributions

Nous définissons quatre fonctions de coût: unitaire ( $n=0$ ), normal ( $n=1$ ), carré ( $n=2$ ), quadratique ( $n = 4$ ) selon l'expression suivante

$$F_n = \begin{cases} prob[(a_i, a_j)]^n \\ (1 - prob[(a_i, a_j)])^n \end{cases} \quad (1.3)$$

ou  $prob[(a_i, a_j)]^n$ , la corrélation entre les arêtes  $a_i$  et  $a_j$ , correspond au coût d'ajout de corrélations fausses négatives,  $(1 - prob[(a_i, a_j)])^n$  au cout d'ajout de corrélations fausses positives.

Étant donnée que nous avons utilisé des matrices binaires dans la génération de line graphes, nous assignons des probabilités pour chaque type de corrélation tel que:

- $prob[(a_i, a_j)] = [0, 0.5[$  : corrélation vrai négative i.e  $0 \rightarrow 0$
- $prob[(a_i, a_j)] = [0.5, 0.79[$ : corrélation fausse négative i.e  $1 \rightarrow 0$

- $\text{prob}[(a_i, a_j)] = 0.8$  : corrélation fausse positive i.e  $0 \rightarrow 1$
- $\text{prob}[(a_i, a_j)] = [0.8, 1]$  : corrélation vrai positive i.e  $1 \rightarrow 1$

La figure 1.7 affiche les courbes des différentes fonctions de coût en fonction de  $k$  corrélations modifiées pour  $p = 0.5$ .

Les 4 courbes sont superposées pour  $k \leq 5$  et au-delà de  $k > 5$ , les courbes carrée ( $n=2$ ) et normal ( $n=1$ ) ont les plus petites distances moyennées de Hamming. Pour  $k > 9$ , la courbe unitaire ( $n=0$ ) évolue de manière exponentielle tandis que les courbes carrée ( $n=2$ ) et normal ( $n=1$ ) ont un rapport de 2 entre les distances de Hamming et le nombre  $k$  de corrélations modifiées. On en déduit que la pire fonction de coût est la fonction unitaire ( $n=0$  en jaune).

Toutefois, nous ne pouvons pas choisir la meilleure fonction de coût car les fonctions des courbes *normal* et *carré* évoluent identiquement. On peut conclure qu'utiliser les probabilités de corrélations pour le calcul du coût améliore significativement les distances moyennées de Hamming.

On choisit pour la suite la fonction de cout normal  $F_1$  pour le coût de chaque arête.

## 12 CHAPTER 1. SIMULATION DES ALGORITHMES SUR DES RÉSEAUX THÉORIQUES

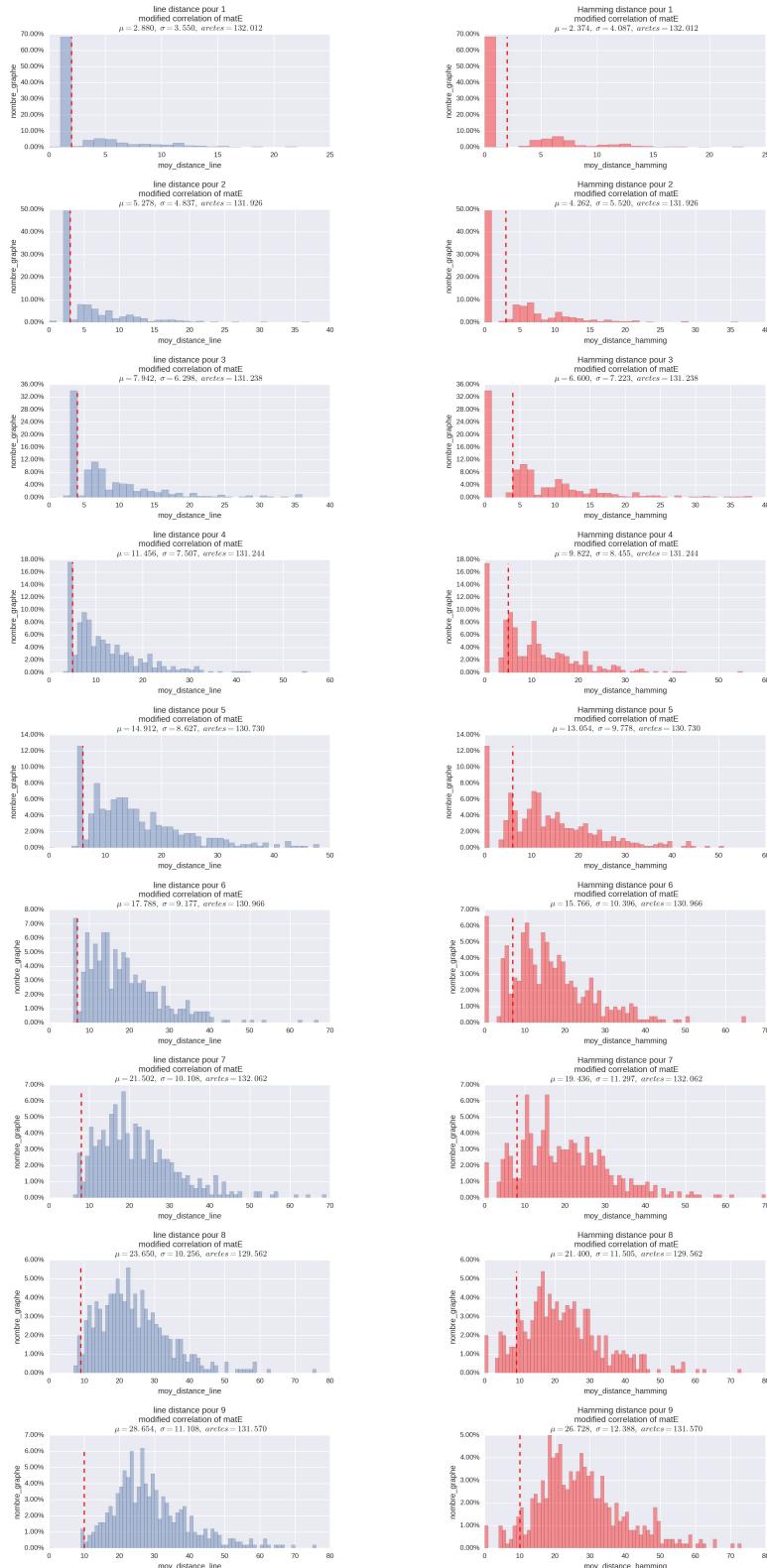


Figure 1.2: Méthode permutation aléatoire avec coût unitaire : distribution des distances line  $moy\_DL$  et de Hamming  $moy\_DH$  pour  $k \in [1, 9]$  corrélations alterées

Figure 1.3: 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

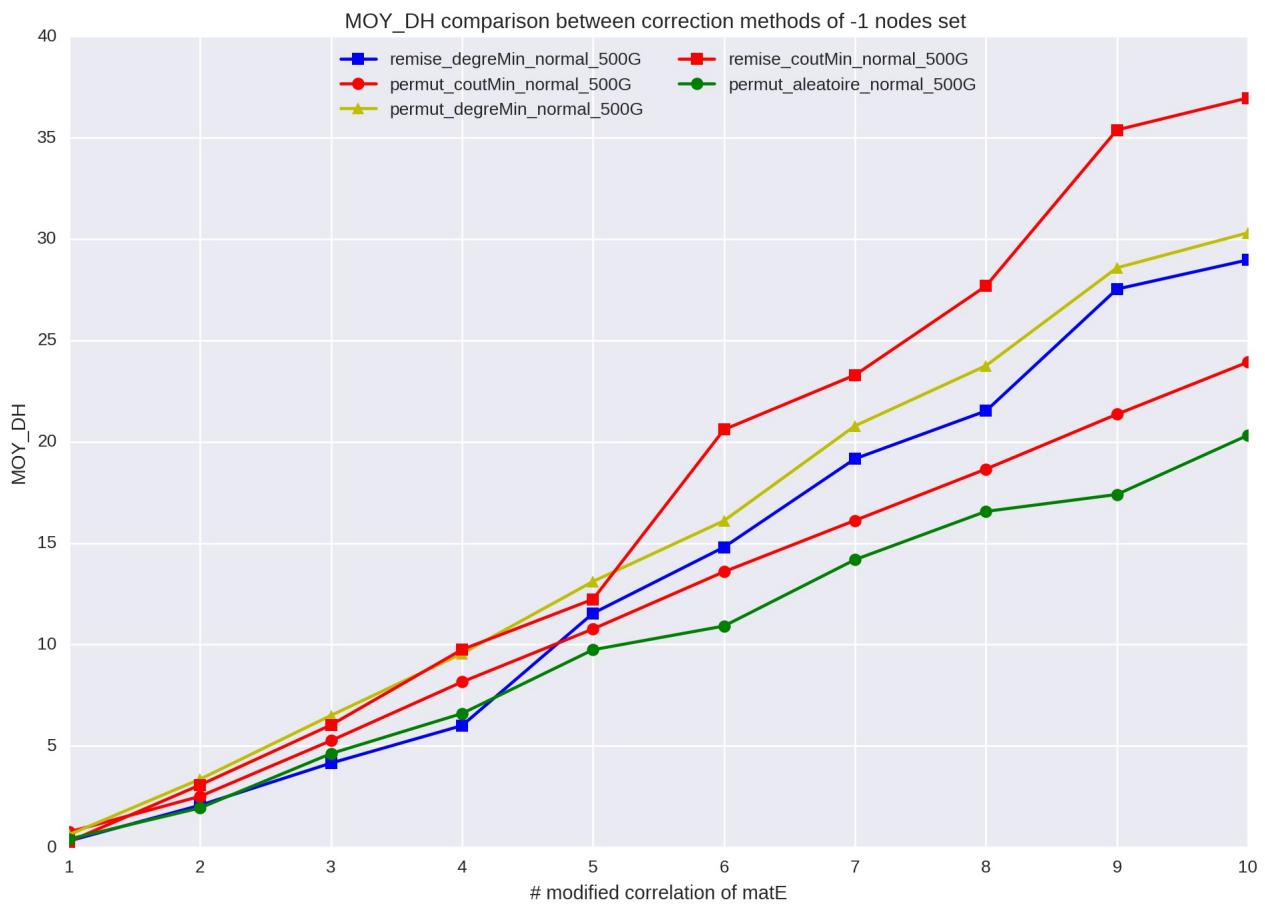


Figure 1.4: Comparaison des différentes méthodes de correction de sommets pour  $k \in [1, 9]$  corrélations modifiées

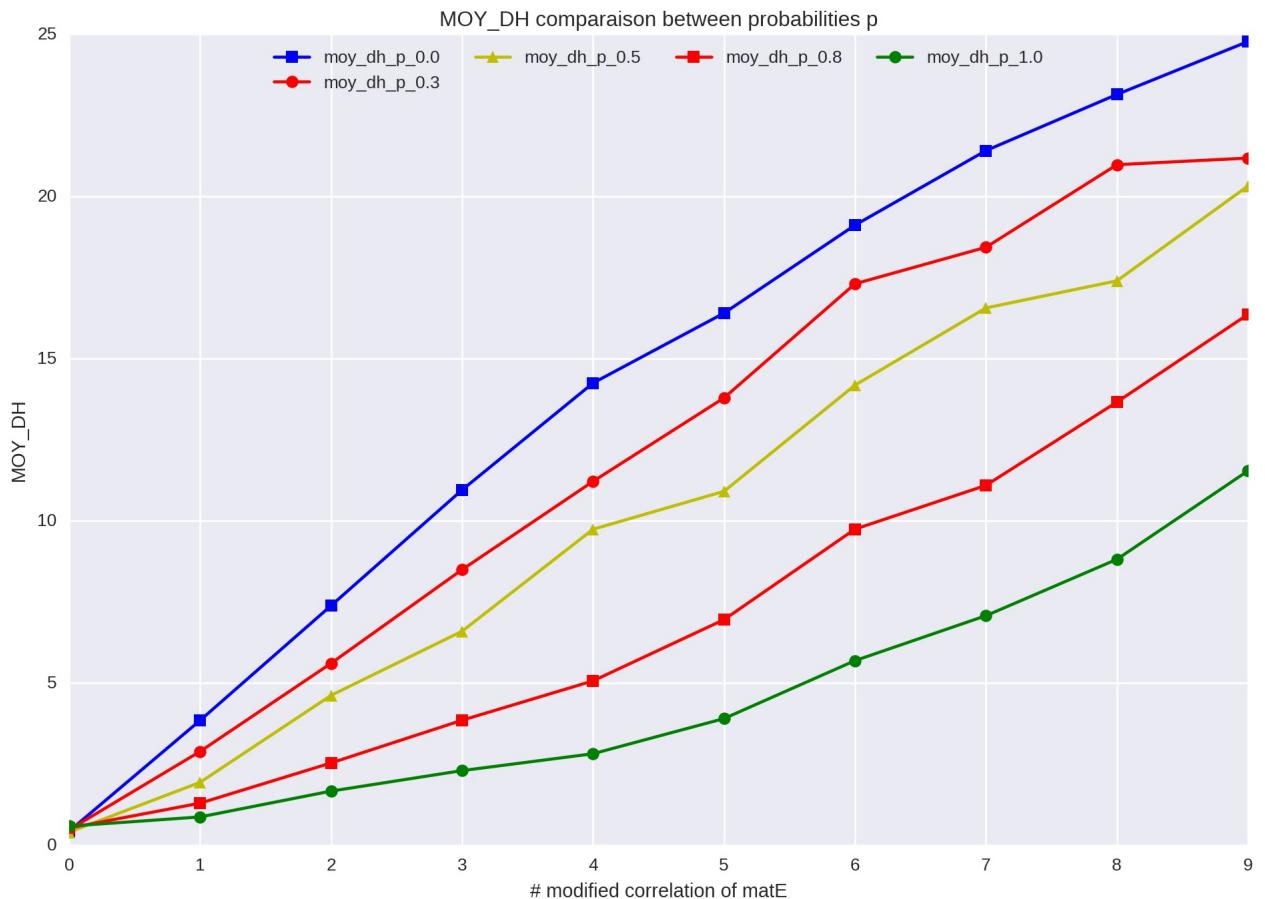


Figure 1.5: Comparaison des différentes probabilités d'ajout  $k \in [1, 9]$  de corrélations fausses positives et fausses négatives sur la méthode de permutation aléatoire

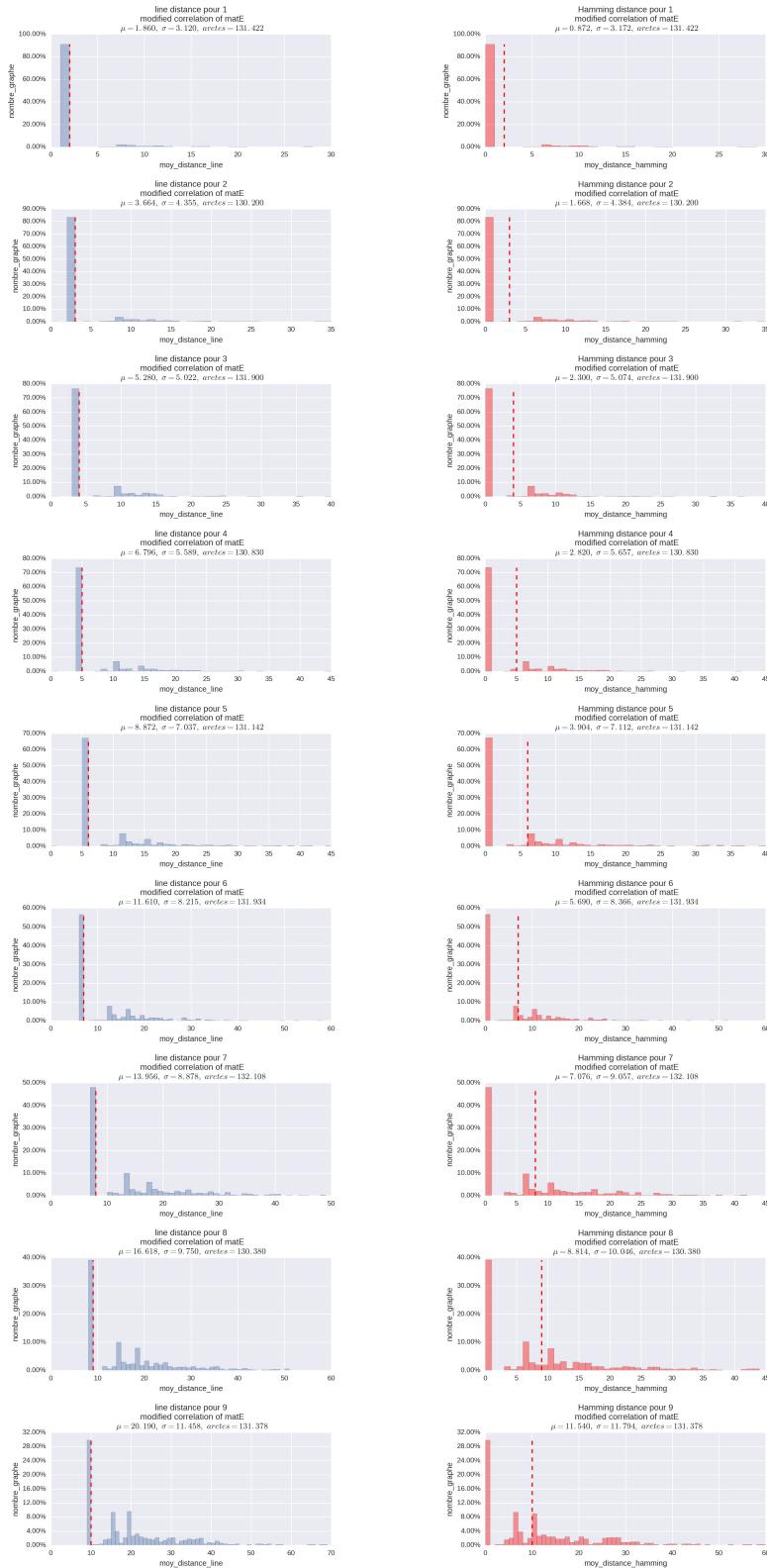


Figure 1.6: distribution des distances line et de Hamming pour  $p = 1.0$  et  $k \in [1, 9]$  correlations modifiées

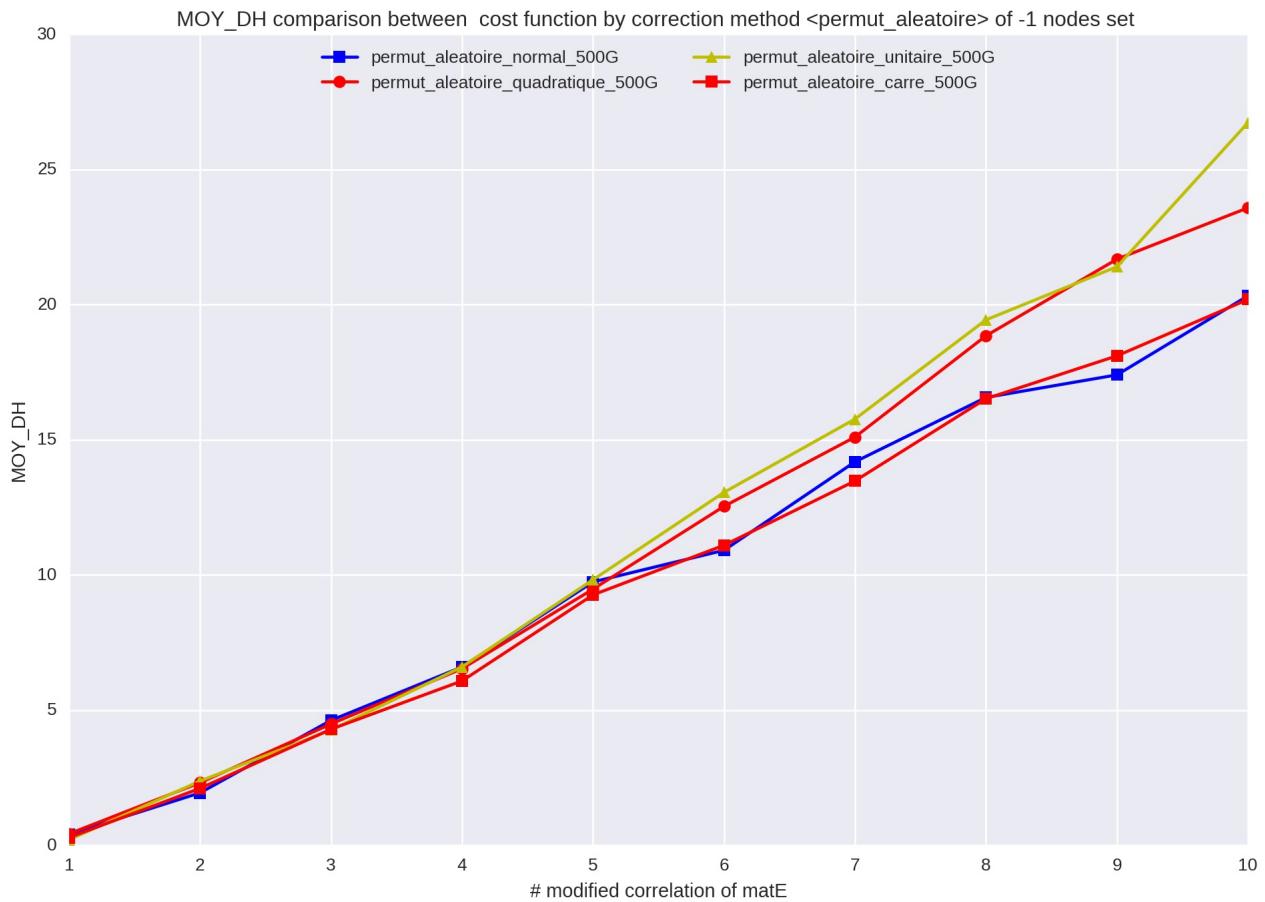


Figure 1.7: Comparaison des différentes fonction de coût sur l'ajout de  $k \in [1, 9]$  de corrélations fausses positives et fausses négatives pour une probabilité  $p = 0.5$  avec la méthode de permutation aléatoire