Aix*Marseille université
Socialement engagée

# Bordures de motifs fréquents et transversaux minimaux d'hypergraphes

Travail d'Etudes et de Recherche
Master 2 Informatique (SID)
2023-2024

*Auteurs:*
Idriss FELLOUSSI
Aghilas SMAIL

*Encadrant:*
Nicolas DURAND
Mohamed QUAFAFOU

February 26, 2024

# Table de matières

# 1 Remerciements

Nous tenons à remercier chaleureusement M. Nicolas DURAND et M. Mohamed QUAFAFOU, d'avoir encadré ce travail, de leurs conseils et de leurs encouragements. Nous les remercions encore plus de nous avoir offert la chance de bénéficier de cette opportunité du sujet de TER en fouille de données qui certes nous a ouvert et nous ouvrira des portes vers de nouveaux horizons.

Nos remerciements et notre reconnaissance vont également à tous les professeurs que nous avons rencontré tout le long de nos parcours universitaires pour leur aide, leur soutien, leur patience et leur disponibilité.

À nos camarades et nos amis, les étudiants cette promotion du Master Sciences et Ingénierie de Données, qui ont traversé ce chemin avec nous merci d'avoir fait cette expérience inoubliable.

À nos parents, nos sœurs et nos frères, quoique nous disons, ça ne sera jamais suffisant pour exprimer notre gratitude et notre reconnaissance envers vous.

Enfin, à tous ceux qui ont contribué de près ou de loin à l'achèvement de ce travail, merci !

# 2 Abstract

This work explores a novel probabilistic approach for efficiently mining frequent patterns in hypergraphs, a key data mining challenge often hampered by scalability and complexity issues. We propose an improved hypergraph reduction algorithm with integrated sampling for efficiency, leading to smaller yet representative hypergraphs. Additionally, a novel probabilistic algorithm efficiently identifies minimal transversals in the reduced graph, crucial for pattern characterization. Through experiments on diverse datasets, our approach demonstrates a promising efficiency and scalability compared to traditional methods, opening doors for effective pattern mining in various domains, particularly those dealing with complex data.

# 3 Introduction

L'exploration des motifs fréquents constitue une problématique fondamentale dans le domaine de la fouille de données, visant à identifier les ensembles d'items cooccurrents au sein de transactions. Cette quête a conduit à l'étude approfondie des motifs fréquents et de leurs dérivés, tels que les motifs maximaux, émergents ou séquentiels, offrant des applications diverses, de la génération de règles d'association, à la classification et à la recommandation

Les motifs fréquents maximaux, en particulier, jouent un rôle crucial, représentant une sélection restreinte de motifs. Cette sélection, associée aux motifs non fréquents minimaux, délimite respectivement la bordure positive et négative de l'ensemble des motifs fréquents. La dualisation, illustrée par le calcul de transversaux minimaux d'un hypergraphe, permet de passer d'une bordure à l'autre. Un transversal se définit comme un ensemble minimal de sommets intersectant chaque hyperarête d'un hypergraphe

Dans cette optique, une méthodologie basée sur la dualisation et le calcul de transversaux minimaux approximatifs a été développée, comme présenté dans la publication de N. Durand et M. Quafafou intitulée "Frequent Itemset Border Approximation by Dualization" [**1**].

L'essence de ce travail consiste à explorer et comprendre la méthode susmentionnée ainsi que son implémentation, avec une focalisation particulière sur l'identification de pistes d'amélioration, notamment en ce qui concerne les temps d'exécution. L'objectif principal demeure d'appréhender la pertinence de cette approche méthodologique dans le contexte de l'exploitation efficace des motifs fréquents, tout en mettant en lumière les aspects amélioratifs envisageables pour optimiser ses performances.

En ce qui concerne l'organisation du reste du rapport, la section 4 présentera la méthode existante, jetant les bases nécessaires pour la compréhension approfondie du sujet. Dans la section 5, nous procéderons à un comparatif expérimental, incluant une version multithreadée pour évaluer les performances sous diverses conditions. La section 6 sera consacrée à la proposition d'une méthode probabiliste, offrant une alternative novatrice. Les expérimentations liées à cette approche seront détaillées dans la section 7. Enfin, nous synthétiserons nos résultats et perspectives dans la section 8, concluant ainsi notre exploration des méthodes de réduction des hypergraphes pour l'efficacité de l'extraction des motifs fréquents.

# 4 Méthode existante

La méthode existante présente une approche novatrice dans le calcul des traverseaux minimales approximées, spécifiquement adaptée à la fouille de données. Elle s'inscrit dans le contexte d'un hypergraphe $H = (V, E)$, où $V$ représente l'ensemble des items et $E$ les hyperarêtes connectant les transactions à leurs items respectifs.

L'introduction des bordures positive $Bd^+(S)$ et négative $Bd^-(S)$ apporte une dimension cruciale à cette méthodologie. Ces concepts, définis en termes de motifs fréquents maximaux et motifs non fréquents minimaux, sont liés par des propriétés fondamentales. Les propriétés $Bd^-(S) = MinTr(\overline{Bd^+(S)})$ et $Bd^+(S) = \overline{MinTr(Bd^-(S))}$ formalisent le passage entre les bordures positive et négative, établissant ainsi la base de la dualisation.

La méthode d'approximation des bordures repose sur le concept de dualisation, impliquant deux fonctions, $f$ et $g$, qui facilitent le passage entre la bordure positive et la bordure négative, ainsi que vice versa. Plus précisément, la fonction $f$ effectue le passage de la bordure positive ($Bd^+(S)$) à la bordure négative ($Bd^-(S)$), tandis que la fonction $g$ opère dans le sens inverse, passant de la bordure négative ($Bd^-(S)$) à la bordure positive ($Bd^+(S)$).

L'idée clé de cette approche réside dans la substitution de $f$ par $\tilde{f}$, où $\tilde{f}$ représente une version approximative de $f$. Cette substitution s'effectue à partir de la bordure positive, permettant le calcul d'une bordure négative approximative ($\tilde{Bd}^-(S)$) en se basant sur des transversaux minimaux approximatifs.

La réduction de l'hypergraphe $H$ constitue une étape clé, exploitant les intersections d'hyperarêtes. La construction d'un graphe valué permet de représenter graphiquement ces intersections, où chaque arête reflète la relation entre deux hyperarêtes. L'algorithme glouton intervient ensuite pour sélectionner les arêtes les plus significatives, conduisant ainsi à la formation de l'hypergraphe réduit $HR = (VR, ER)$.

Cette méthode offre une approche intégrée et systématique pour le calcul des traverseaux minimales approximées, tout en maintenant un équilibre judicieux entre la précision et la complexité computationnelle.

# 5 Comparatif expérimental

Nous avons étudié une version multithreadée (HRmulti) de la méthode de réduction d'hypergraphes monothreadée (HR) déjà existante. Cette version, bien qu'en phase bêta et non testée, a été soumise à une série de tests. Nous avons notamment vérifié que les résultats produits par cette version étaient parfaitement identiques à ceux de la version monothreadée. De plus, nous avons comparé les temps d'exécution des deux versions afin d'évaluer les éventuels gains de performance offerts par la version multithreadée.

Pour mener à bien cette évaluation, nous avons exploité six ensembles de données distincts, chacun offrant des perspectives uniques et contribuant à une compréhension globale des méthodes de réduction. Ces données, soigneusement sélectionnées pour leur diversité et leur pertinence, proviennent du réputé répertoire FIMI [**2**] dédié à la fouille de motifs fréquents et le dépot "Hypergraph Dualization Repositry " [**3**]. Les ensembles de données incluent :

1. **ac (accidents)** : comprend l'ensemble complémentaire des ensembles d'items fréquents maximaux avec un seuil de support $t$ du jeu de données "accident". "ac_90k" signifie que le seuil de support est de 90 000, mettant en lumière des motifs infrequents minimaux dans les données d'accidents.
2. **bms (BMS-WebView2)** : contient l'ensemble complémentaire des ensembles d'items fréquents maximaux du jeu de données "BMS-WebView2", avec un seuil de support $t$. "bms2_100" indique un seuil de support de 100, explorant des motifs infrequents dans les données de navigation web.
3. **Mushroom** : fournit des données sur 23 espèces de champignons à branchies, permettant l'analyse de motifs fréquents dans les caractéristiques biologiques des champignons.
4. **Chess** : contient des stratégies pour les jeux d'échecs, représentant un ensemble de données stratégique où l'analyse de motifs peut révéler des schémas de jeu fréquents et gagnants.
5. **Connect** : inclut des stratégies pour le jeu de connect-4, un autre ensemble de données stratégique qui permet l'étude de configurations gagnantes à travers l'analyse de motifs fréquents.
6. **Kosarak** : comporte des données anonymes de parcours d'un portail d'information en ligne hongrois, offrant un aperçu des motifs de navigation fréquents parmi les utilisateurs.

Le protocole expérimental a été conçu pour évaluer en détail les performances de la méthode de réduction d'hypergraphes. La flexibilité des datasets permet de les adapter en fonction des capacités de la machine, offrant la possibilité d'explorer des datasets plus vastes tels que ac_90k.dat, bms2_100.dat, etc. Chaque dataset représente un hypergraphe dérivé du complément des motifs fréquents maximaux $Bd^+$, avec la valeur minimale de support spécifiée dans le nom du fichier.

Nous commencons par le calcul de l'hypergraphe réduit à l'aide de la methode monothreadée et multithreadée tout en spécifiant le nombre de threads. Ensuite, nous calculons la distance entre les deux hypergraphes réduits.
Il est impératif de fixer le nombre de threads pour HRmulti, en veillant à ce qu'il reste inférieur ou égal au nombre de cœurs du processeur. Cette procédure permet d'assurer une évaluation rigoureuse tout en évitant les problèmes potentiels liés à la mémoire pendant l'exécution.

Les figures 1, 2, 3, 4, 5 et 6 présentent les résultats pour les jeux de données chess, kosarak, bms, accidents, connect et mushroom.

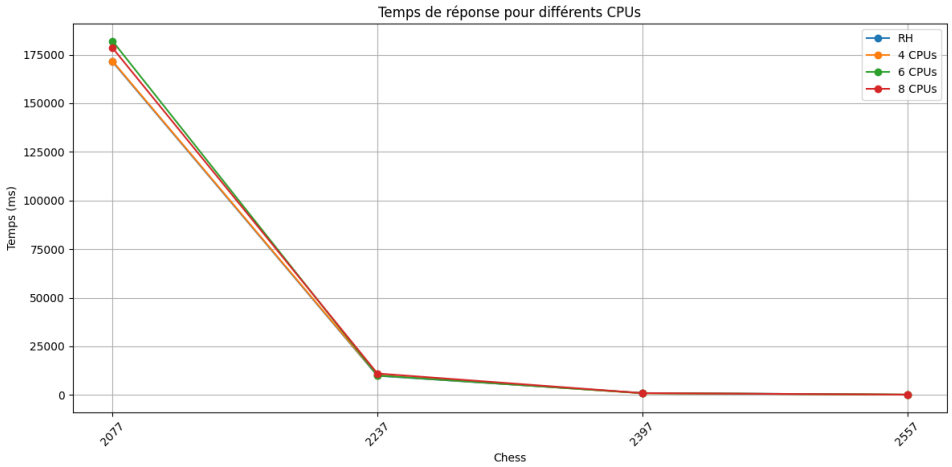


Figure 1: Comportement de HR et HRmultithread avec variation de nombre de CPUs sur le dataset Chess.

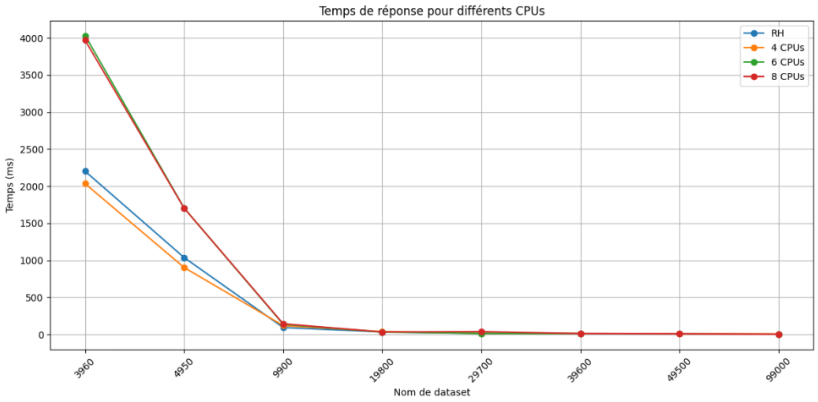


Figure 2: Comportement de HR et HRmultithread avec variation de nombre de CPUs sur le dataset Kosarak.

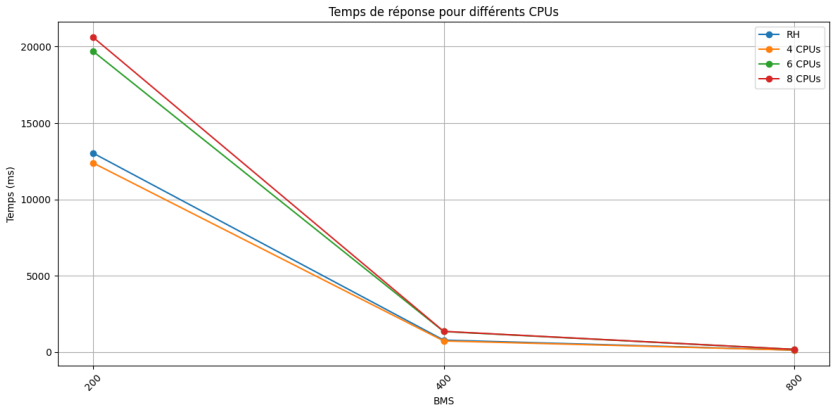


Figure 3: Comportement de HR et HRmultithread avec variation de nombre de CPUs sur le dataset bms.

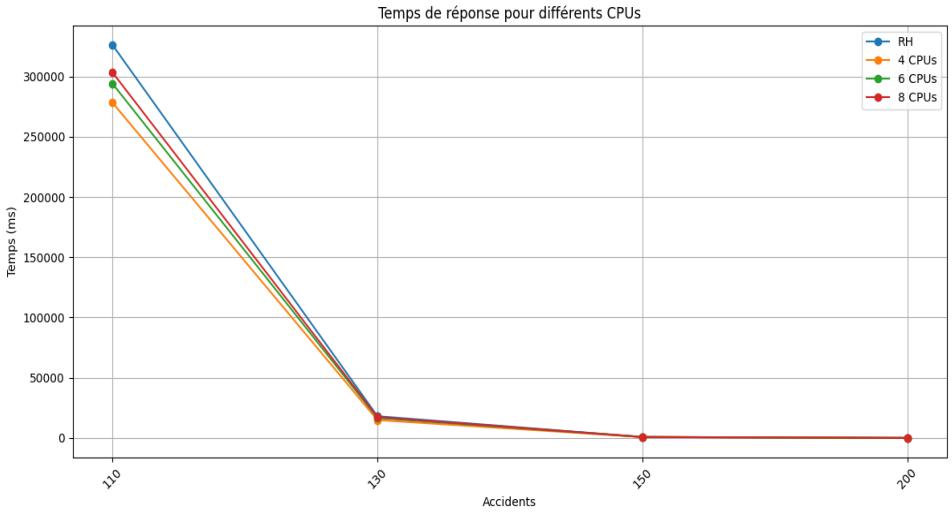


Figure 4: Comportement de HR et HRmultithread avec variation de nombre de CPUs sur le dataset accidents.

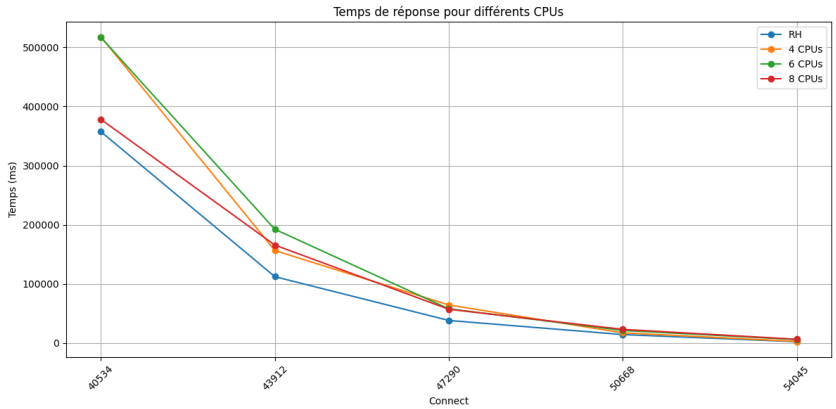


Figure 5: Comportement de HR et HRmultithread avec variation de nombre de CPUs sur le dataset connect.

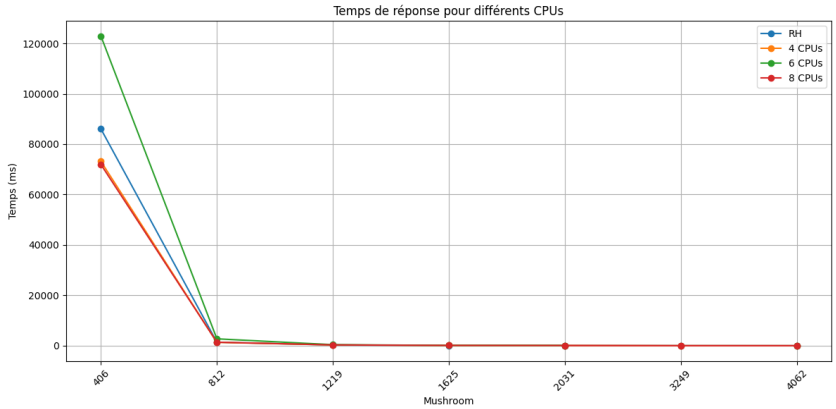


Figure 6: Comportement de HR et HRmultithread avec variation de nombre de CPUs sur le dataset Mushroom.

En comparant les temps d'exécution des deux versions, nous avons constaté que, bien que HRmulti offre un avantage en termes de performance grâce à l'utilisation de plusieurs threads, le gain de temps par rapport à la version HR n'était pas aussi significatif que prévu.

Cela peut-être dû aux caractéristiques du matériel vu que la performance de la parallélisation dépend également des caractéristiques du matériel, telles que le nombre de cœurs du processeur, la mémoire disponible, etc. Si le matériel ne prend pas en charge efficacement le multithreading, les gains peuvent être limités. Dans notre cas il s'agit d'un ordinateur portable à 8 coeurs physiques avec une RAM de 1638Mo avec le système d'exploitation Windows avec Java comme langage de programmation.

# 6 Méthode probabiliste proposée

Dans le cadre de l'évolution continue de la recherche sur les motifs fréquents, une nouvelle approche prometteuse se distingue : l'approche probabiliste. Cette méthode innovante se propose de révolutionner le calcul des motifs fréquents en s'affranchissant des limites des techniques traditionnelles, souvent confrontées à des difficultés d'échelle et de complexité avec l'augmentation du volume de données. L'approche probabiliste, en se basant sur des modèles statistiques et des probabilités, vise à estimer la fréquence des motifs de manière plus flexible et moins coûteuse en ressources. En introduisant un cadre de calcul qui accepte une certaine marge d'erreur contrôlée, cette méthode permet de traiter des ensembles de données massifs de manière plus efficace, offrant ainsi une solution viable pour les applications nécessitant une analyse rapide et à grande échelle. Cette évolution conceptuelle pourrait non seulement accélérer le processus de découverte de motifs fréquents, mais également ouvrir la voie à de nouvelles perspectives dans l'exploration de données

## 6.1 Une approche probabiliste

L'approche probabiliste pour le calcul des motifs fréquents, en particulier dans le contexte des hypergraphes, s'enrichit d'une stratégie de réduction innovante visant à optimiser le traitement des données. Cette méthode consiste à appliquer une réduction ciblée de l'hypergraphe initial par des pourcentages prédéfinis, tels que **5%**, **10%**, et ainsi de suite. L'objectif est de simplifier la structure de l'hypergraphe en diminuant le nombre d'hyperarêtes et de sommets, tout en préservant une représentation fidèle des relations complexes qu'il encapsule.

En réduisant systématiquement l'hypergraphe de manière contrôlée, cette approche permet de réduire significativement la complexité du calcul des transversaux, c'est-à-dire des ensembles de sommets qui intersectent chaque hyperarête de l'hypergraphe. Le calcul des transversaux, étant au cœur de l'identification des motifs fréquents dans de nombreux domaines d'application, se trouve ainsi grandement facilité. La réduction de l'hypergraphe permet non seulement d'accélérer le processus de calcul en réduisant la taille de l'espace de recherche, mais offre également la possibilité d'atteindre une approximation des transversaux minimales avec une efficacité accrue.

Cette méthode utilise des processus aléatoires ou probabilistes pour sélectionner les sommets ou les hyperarêtes à fusionner. L'aspect stochastique de l'approche permet de gérer la complexité et la diversité des structures des hypergraphes en introduisant une flexibilité qui peut être avantageuse pour explorer différentes configurations de réduction tout en préservant certaines propriétés structurelles clés de l'hypergraphe original.

L'utilisation d'une approche stochastique pour la réduction d'hypergraphes est particulièrement utile dans plusieurs contextes :

- **Grande variabilité des propriétés :** quand il y a une grande variabilité dans les propriétés des sommets et des hyperarêtes, une approche déterministe pourrait être biaisée ou insuffisante pour capturer la complexité de l'hypergraphe.

- **Préservation des propriétés statistiques :** l'objectif de préserver certaines propriétés statistiques de l'hypergraphe original, comme la distribution des degrés des sommets, les motifs de connexion, etc., peut être mieux atteint avec une approche probabiliste.

- **Obtention de différentes instances :** pour obtenir différentes instances de l'hypergraphe réduit pour des analyses de sensibilité ou pour évaluer l'impact de la réduction sur des tâches spécifiques (par exemple, classification et clustering).

## 6.2 Algorithme et déroulement du processus

Comme évoqué auparavant, nous commencons par charger notre hypergraphe de base afin de pouvoir effectuer un échantillonage aléatoire dessus, tout en fixant un seuil approprié pour avoir un hypergraphe réduit. À partir de ce dernier, on commence le calcul des transversaux minimales à l'aide d'une approche stochastique pour trouver une couverture minimale de sommets dans un hypergraphe. Elle itère m fois, échantillonnant de manière aléatoire un sous-ensemble (Hr_i) de taille N pour chaque hyperarête de l'hypergraphe. À chaque itération, elle calcule l'ensemble des arêtes couvertes par l'ensemble de candidats (TM_i), l'union de ces ensembles est accumulée dans (TM_old). Après les itérations, la fonction identifie le plus petit ensemble d'arêtes (min_TM) dans (TM_old). Le résultat final est une liste de sommets représentant la couverture minimale de sommets dans l'hypergraphe. L'aspect stochastique de l'algorithme permet d'explorer différentes configurations de couverture de sommets.

**Algorithm 1:** Réduction probabiliste d'un hypergraphe

**Entrée:** Un pourcentage de réduction (seuil)
**Sortie :** Un hypergraphe réduit (graphe_reduit)

```
Function reduire_hypergraphe(seuil: nombre réel):
    graphe_reduit ← liste vide;
    for chaque hyper-arête dans hypergraphe do
        nombre_elements_a_tirer ← (taille de hyper-arête) × (seuil / 100);
        hyperarête_réduite ← selectionner_aleatoirement(hyperarête,
          nombre_elements_a_tirer);
        Ajouter hyperarête_réduite à graphe_reduit;
    end
    return graphe_reduit;
```

**Algorithm 2:** Transversal minimal d'un hypergraphe

**Entrée:** hypergraphe
**Entrée:** seuil: seuil de réduction
**Sortie :** Un transversal minimal de l'hypergraphe d'entrée

```
1  Function calculer_transversal_minimal(hypergraphe,seuil):
2      TM_old ← ∅;
3      for i ← 1 to m do
4          Hr_i ← reduire_hypergraphe(hypergraphe,seuil);
5          for chaque transversal candidat ensemble_candidat dans Hr_i do
6              arêtes_couvertes ← ∅;
7              for chaque hyperarête dans l'hypergraphe do
8                  if est_recouvrement_sommets(ensemble_candidat, hyperarête) then
9                      arêtes_couvertes ← arêtes_couvertes∪ toutes les paires de
                         sommets de l'hyperarête;
10                 end
11             end
12             TM_old ← TM_old ∪ arêtes_couvertes;
13         end
14     end
15     if TM_old ≠ ∅ then
16         min_TM ← min(TM_old);
17     end
18     else
19         min_TM ← ∅;
20     end
21     return transversal minimal converti en liste;
```

La complexité de cette algorithme visant à trouver les transversaux minimaux est presenté dessous :

**Complexité temporelle:** $O(m \times |E| \times N)$ dans le pire des cas, où:

- $m$ est le nombre d'itérations.
- $|E|$ est le nombre d'hyperarêtes dans l'hypergraphe.
- $N$ est le nombre de sommets dans l'hypergraphe.

**Complexité spatiale:** $O(|V| + |E|)$, où:

- $|V|$ est le nombre de sommets dans l'hypergraphe.
- $|E|$ est le nombre d'hyperarêtes dans l'hypergraphe.

# 7 Experimentations avec l'approche probabiliste

Les expérimentations avec cette approche consiste à calculer l'ensemble des traverseaux minimales k fois en prenant à titre d'instance un hypergraphe de base et le réduire à un premier hypergraph réduit puis un deuxième hypergraphe réduit avec des differents seuils en utilisant l'algorithme de réduction proposé précédement .Ensuite, nous calculons les transversaux minimaux pour le premier hypergraphe l'ensemble trouvé sera renforcé par les nouveaux transversaux minimaux trouvés à partir du deuxième hypergraphe réduit, nous calculons par la suite le gain probable de tranversaux minimaux calculés.

L'approche expérimentale poursuit l'objectif de calculer les transversaux minimaux de manière itérative jusqu'à atteindre un point où la découverte de nouveaux transversaux minimaux devient négligeable. À chaque itération, la réduction des hypergraphes avec des seuils différents vise à explorer la dynamique des transversaux minimaux dans des contextes variés. L'idée sous-jacente est de déterminer comment la variation des seuils de réduction influe sur la composition de l'ensemble des transversaux minimaux. En analysant le gain probable de transversaux minimaux entre deux seuils distincts, nous cherchons à comprendre le comportement de l'algorithme lorsqu'il converge vers une liste plus exhaustive de transversaux minimaux. Ce processus itératif permet d'ajuster la sensibilité de l'algorithme et d'explorer la stabilité des transversaux minimaux jusqu'à ce que le point de saturation soit atteint, indiquant ainsi une convergence vers la liste complète des transversaux minimaux.

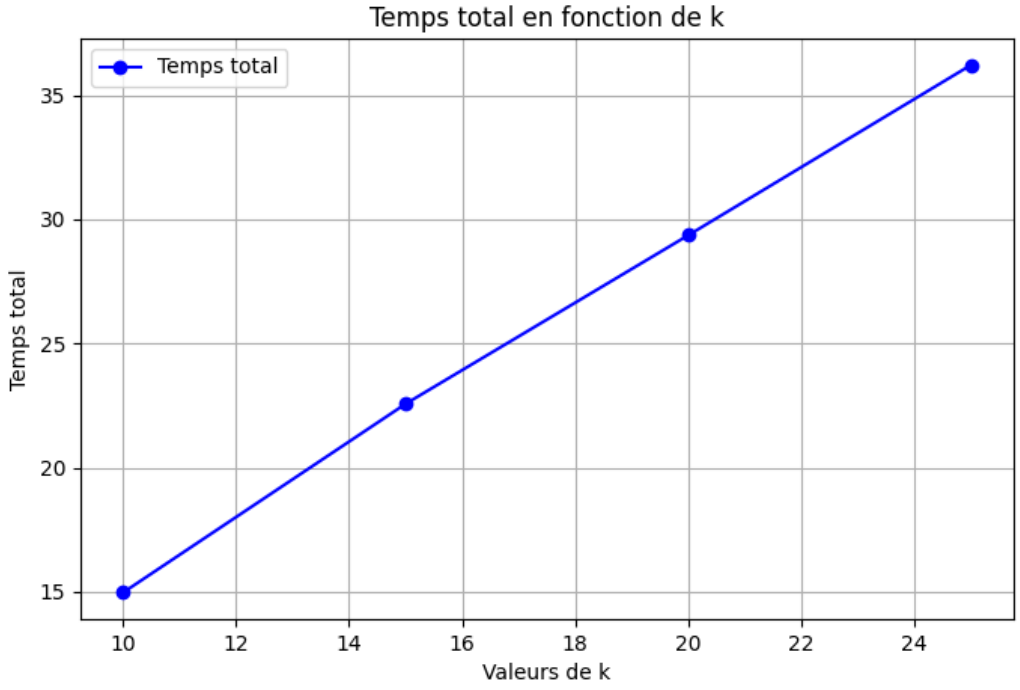


Figure 7: Temps d'exécution en fonction de k en secondes

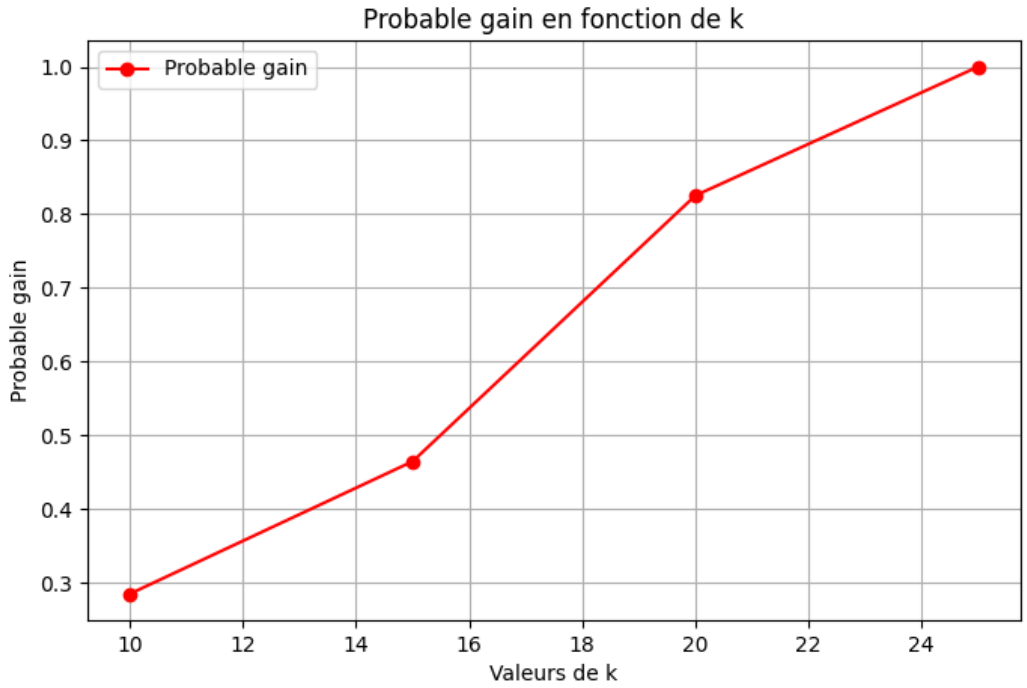


Figure 8: Gain probable en fonction du nombre d'itérations k

L'objectif de cette approche est d'évaluer la stabilité des transversaux minimaux dans des scénarios de réduction variés. L'analyse se poursuit en comparant la taille de l'ensemble des transversaux minimaux obtenus avec un seuil de 20% avec ceux obtenus avec un seuil de 10%. Pour k = 25, le gain probable est de 1.0. Cela signifie que toutes les transversales minimales trouvées avec la réduction de 20% sont également présentes dans la réduction de 10%. Cette comparaison permet de quantifier le gain probable de transversaux minimaux résultant de la réduction avec un seuil plus élevé. Ainsi, cette méthodologie expérimentale offre une perspective approfondie sur la sensibilité de l'algorithme aux variations des seuils de réduction, tout en explorant la robustesse de l'algorithme dans la découverte des transversaux minimaux dans des contextes de réduction d'hypergraphes.

# 8 Conclusion

Ce travail s'est penché sur l'efficacité d'une approche probabiliste visant à calculer les transversaux minimaux dans des hypergraphes soumis à des seuils de réduction variables.

L'algorithme de réduction a démontré sa capacité à ajuster la sensibilité de la détection des transversaux minimaux en fonction des seuils appliqués. L'analyse du gain probable entre deux hypergraphes réduits a fourni des indications précieuses sur la stabilité et la convergence de l'algorithme vers la liste complète des transversaux minimaux. Cette approche offre une flexibilité dans l'exploration des structures d'hypergraphes, permettant une adaptation aux différentes caractéristiques des ensembles de données. Les résultats obtenus contribuent à une compréhension approfondie des mécanismes de réduction et de calcul des transversaux minimaux, ouvrant la voie à des applications potentielles dans divers domaines tels que l'extraction de motifs dans les données complexes.

# Références

[**1**] Nicolas Durand, Mohamed Quafafou. Frequent Itemset Border Approximation by Dualization. Transactions on Large-Scale Data- and Knowledge-Centered Systems, 2016, 26, pp.32-60. Springer

[**2**] Frequent Itemset Mining Dataset Repository, `http://fimi.uantwerpen.be/data/`

[**3**] Hypergraph Dualization Repository, `https://research.nii.ac.jp/~uno/dualization.html`