

Mémoire

Pour Obtention du diplôme de Master En Informatique

Option : Système Informatique (SIQ)

Étude et classification des méthodes de compression de graphes par extraction de motifs et k2-trees

Réalisé par :

Mlle. Hafsa Bousbiat

eh_bousbiat@esi.dz

ESI

Mlle. Sana Ihadadene

es_ihadadene@esi.dz

ESI

Encadré par :

Dr. Karima Amrouche

k_amrouche@esi.dz

ESI

Dr. Hamida Seba

hamida.seba@univ-lyon1.fr

Université de Lyon

Remerciement

La construction de ce mémoire n'aurait été possible sans l'intervention de plusieurs personnes. Qu'elles trouvent ici l'expression de nos plus sincères remerciements pour leurs précieux conseils.

Nous tenons d'abord à exprimer notre gratitude à l'égard de nos deux encadrantes, Madame Seba Hamida, professeur à l'Université Claude Bernard de Lyon et Madame Amrouche Karima, enseignante à l'École Nationale Supérieure d'Informatique, pour leur précieuse aide, leurs judicieux conseils et pour le temps qu'elles nous ont consacré tout au long de ce travail.

Nous remercions également tout le corps professoral et administratif de l'équipe pédagogique de l'École Nationale Supérieure d'Informatique, et à leur tête nos enseignants, pour la richesse et la qualité de la formation qu'ils nous ont offerte tout au long de notre cursus.

Nous tenons aussi à remercier les membres du jury qui nous ont fait l'honneur d'accepter de juger cet humble travail.

Enfin, nous adressons nos plus sincères remerciements à tous nos proches et amis, qui nous ont toujours encouragés au cours de la réalisation de ce mémoire.

Merci à tous et à toutes ...

Résumé

Nous vivons dans un monde où la quantité d'informations ne cesse d'augmenter et dont la bonne gestion implique l'utilisation des graphes qui se sont répandus dans différents domaines allant des réseaux sociaux et de communication jusqu'aux domaines de la chimie et de la biologie. Cette abondance de données générées fait appel à une technique aussi vieille que la discipline de traitement de données mais qui connaît de nouveaux défis aujourd'hui : la compression. La compression de graphes est un domaine dans lequel le graphe initial subit des transformations pour en obtenir une version plus réduite et compacte permettant, dans la majorité des cas, d'effectuer les traitements dans un temps nettement meilleur.

Dans ce travail, nous étudierons les différentes méthodes de compression existantes dans la littérature dans le but de compléter et d'affiner davantage une classification entamée dans un travail de Master précédent. Nous mettrons l'accent sur les méthodes de compression basées sur l'extraction de motifs ainsi que les méthodes basées sur les k^2 -trees. Par conséquent, nous proposons d'établir une synthèse bibliographique sur ces deux classes et nous établirons une étude comparative entre les méthodes de ces deux classes selon divers critères.

Mots Clés : *Compression de graphes, Big Data, Extraction de motifs, K2-trees, Graphe du Web.*

Abstract

We live in a world where the amount of data is constantly increasing and whose good management involves the use of graphs that have spread in different fields from social and communication networks to the fields of chemistry and biology. This abundance of generated data calls for a technique that is as old as the discipline of data processing but which is facing new challenges today : compression. Graph compression is a field in which the initial graph undergoes transformations in order to obtain a smaller and more compact version allowing, in the majority of cases, to perform the processings in a much better time.

In this work, we will study the different methods of compression existing in the literature in order to complete and further refine a classification started in a previous Master's work. We will focus on compression methods based on pattern extraction as well as k^2 -trees methods. Therefore, we propose to establish a bibliographic synthesis on these two classes and we will establish a comparative study between the methods of these two classes according to various criteria.

Key words : *Graph compression, Big Data, Pattern extraction, K2-trees, Web graph.*

ملخص

إننا نعيش في وقتنا الحالي واقعا يعرف تزايد في الكم المعرفي و المعلوماتي مما اقتضى و أوجب استعمال المنحنيات التي اصبحت واسعة الانتشار و اكتسحت عدة ميادين مختلفة انطلاقا من مواقع التواصل الاجتماعي و الاتصالات وصولا لميادين الكيمياء و البيولوجيا . إن هذه الكمية الهائلة من المعلومات تلزم الرجوع إلى تقنية قديمة قدم مجال معالجة البيانات و التي تواجه تحديات جديدة : الضغط ... ضغط المنحنيات هو مجال يخضع فيه الرسم البياني الأولي لتحويلات من اجل الحصول على نسخة أصغر تسمح في غالبية الحالات بتحسين الوقت اللازم لمعالجة و تحليل البيانات.

من خلال هذا العمل سنقدم دراسة مفصلة عن الطرق المختلفة للضغط الموجودة بهدف استكمال تصنيفات و تحسينها في اطار التحضير لرسالة الماجستير. سنركز على طرق الضغط المعتمدة على استخراج الأنماط ، و كذلك الطرق المعتمدة على الأشجار ، و من ثم نقترح انشاء منشورات ببيوغرافية على هاتين الفئتين ، كما سننشيء دراسات مقارنة بين أساليب هاتين الفئتين وفقا لمعايير مختلفة.

كلمات مفتاحية: ضغط الرسم البياني ، البيانات الكبيرة ، استخراج الأنماط ، أشجار ، الرسم البياني على الويب.

Table des matières

Remerciement	I
Résumé	II
Liste des figures	VIII
Liste des tableaux	IX
Introduction générale	1
1 Théorie des graphes	2
1.1 Graphe non orienté	2
1.1.1 Définitions et généralités	2
1.1.2 Représentation graphique	3
1.1.3 Propriétés d'un graphe	3
1.2 Graphe orienté	4
1.2.1 Définitions et généralités	4
1.2.2 Représentation graphique	4
1.2.3 Quelques Propriétés :	5
1.3 Notion de Connexité	5
1.4 Quelques types de graphe	5
1.5 Graphe partiel et sous-graphe :	6
1.5.1 Définitions :	6
1.5.2 Quelques Types de sous graphes :	6
1.6 Représentation Structurale d'un graphe	6
1.6.1 Matrice d'adjacence	7
1.6.2 Matrice d'incidence	8
1.6.3 Liste d'adjacence	8
1.7 Les domaines d'application	9
1.7.1 Graphes des réseaux sociaux :	9
1.7.2 Graphes en Bioinformatique :	9
1.7.3 Le Graphe du web :	10
1.8 Conclusion	10

2	Compression de graphes	11
2.1	Compression de données :	11
2.2	Compression appliquée aux graphes :	12
2.2.1	Les types de compression :	12
2.2.2	Les métriques d'évaluation des algorithmes de compression :	13
2.2.3	Classification des méthodes de compression :	14
2.3	Compression par les arbres K^2 -Trees	19
2.4	Compression par extraction de motifs	36
2.4.1	Compression basée vocabulaire	36
2.4.2	Compression basée sur l'agrégation des motifs	45
2.4.3	Synthèse des méthodes de compression par extraction de motifs	50
2.5	Bilan général	53
	Conclusion générale	57
A	Description des graphes de test	58

Table des figures

1.1	Exemple de représentation graphique d'un graphe non orienté	3
1.2	Exemple de représentation graphique d'un digraphe.	4
1.3	Graphe orienté G	7
1.4	Matrice d'adjacence du graphe G	7
1.5	Graphe orienté G	8
1.6	Matrice d'incidence du graphe G	8
1.7	Graphe orienté G	9
1.8	Liste d'adjacence du graphe G	9
2.1	Compression sans perte.	13
2.2	Compression avec perte.	13
2.3	Classification des méthodes de compression (W. GUERMAH, 2018)	17
2.4	Classification proposées des méthodes de compression	18
2.5	Exemple de représentation k^2 -tree	20
2.6	Exemple d'une représentation dk^2 -tree	22
2.7	Exemple d'une représentation k^3 -tree	23
2.8	Exemple d'une représentation k^2 -tree1 avec les quatre encodages	24
2.9	Exemple d'une représentation Delta- k^2 -tree	25
2.10	Exemple d'une représentation k^2 -treap d'un graphe pondéré	26
2.11	Structures de données pour une représentation k^2 -treap	27
2.12	Exemple d'une représentation Ik^2 -tree	27
2.13	Exemple d'une représentation diff Ik^2 -tree	28
2.14	Exemple d'un graphe étiqueté, attribué, orienté et multiple	29
2.15	Exemple d'une représentation Att_k^2 -tree (1/2)	30
2.16	Exemple d'une la représentation Att_k^2 -tree (2/2)	31
2.17	Comparaison des performances des méthodes basées sur les k^2 -trees appliquées sur le graphe eu2005 selon le nombre de bits par liens	32
2.18	Comparaison entre les méthodes basées sur les k^2 -trees en fonction du taux de compression	33
2.19	Exemple illustrant le principe de fonctionnement (Asano et al., 2008)	44
2.20	Exemple d'exécution de gRepair sur G	47
2.21	Exemple d'exécution de VNM	49
2.22	Exemple d'exécution de DSM (Hernández and Navarro, 2014).	50

2.23	Comparaison des performances des méthodes de compression basées sur l'agrégation des liens des motifs	51
2.24	Comparaison des performances des méthodes de compression basées vocabulaire	51
2.25	Comparaison des performances des méthodes par extraction de motifs et les méthodes basées sur les arbres k^2 -trees	56

Liste des tableaux

2.1	Synthèse des méthodes de compression par k^2 -trees.	34
2.2	Synthèse des méthodes de compression par k^2 -trees.	35
2.3	Tableau comparatif entre les méthodes de clustering avec n = nombre de nœuds, m = nombre d'arêtes, k = nombre de clusters, t = nombre d'itérations, d = degré moyen de nœuds, $h(m_h)$ = nombre de nœuds (arêtes) dans la structure hyperbolique.	38
2.4	Synthèse des méthodes de compression par extraction de motifs.	52
2.5	Comparaison entre les méthodes basées sur k^2 -trees et basées sur l'extraction de motifs.	54
A.1	Graphes de test	58

Acronymes

BFS Breadth First Search. 35

CONDENSE CONditional Diversified Network Summarization. 38, 47

DAC Directly Addressable Codes. 19

DFS Deapth First Search. 19

DSM Dense SubGraph Mining. 44

ECWG Efficient Compression of web graph. 39

GCUPMT Graph Compression Using Pattern Matching. 39

MDL Minimum Description Length. 15, 33, 34, 37, 38

OLAP Online Analytical Processing. 23

RDF Resource Description Framework. 14, 15, 17

VNM Virtual Node Miner. 42–44

VOG Vocabulary Graph. 34–36

Introduction générale

Avec l'énorme quantité de données produites par les activités humaines de nos jours, le problème de données massives (Big data) est devenu un enjeu essentiel. Un des outils les plus efficaces pour traiter ces données est l'utilisation des graphes. Les graphes sont des outils de modélisation utilisés dans beaucoup de domaines pour la représentation des données : réseaux sociaux et de communication (entités reliées entre elles par des liens physiques ou communautaires), chimie (relations entre les atomes), biologie (interactions entre protéines par exemple), ..., etc.

Il y a plusieurs solutions pour contrer le volume de données mais nous nous intéressons à une solution aussi vieille que la discipline de traitement de données elle-même mais qui connaît de nouveaux défis de nos jours : la compression.

La compression de graphes est un domaine dans lequel le graphe initial subit des transformations pour en obtenir une version plus réduite. Différentes techniques permettent cette compression, avec ou sans perte d'information, et génèrent de nouveaux graphes sur lesquels il est intéressant d'effectuer les différents traitements.

Deux types de méthodes de compression de graphes se sont distingués parmi tout les autres types de méthodes : les méthodes de compression en utilisant les k2-trees et les méthodes de compression par extraction de motifs. En effet, elles permettent de trouver dans la majorité des cas un bon compromis entre l'espace mémoire et les temps de traitement. Nous présenterons dans ce travail une synthèse bibliographique sur les différentes méthodes de compression dérivant de ces deux grandes classes.

Nous avons hiérarchisé notre mémoire en deux grands chapitres. Le premier est une introduction au domaine de la théorie des graphes incluant ainsi les notions fondamentales de ce domaine, les types de graphes, leurs représentations structurelles et leurs domaines d'application. Dans le second chapitre, nous introduisons les définitions de base du domaine de compression de données appliquées aux graphes à savoir : les motivations, les types de compression et les métriques d'évaluation. Enfin, nous présenterons les classifications des méthodes existantes de compression de graphe ainsi que la classification que nous proposons. Nous mettrons l'accent après sur les deux classes de méthodes qui forment la problématique de notre sujet de Master : la compression par extraction de motifs et la compression par les k2-trees. Nous concluons par une synthèse où nous comparerons entre les méthodes des deux classes.

Chapitre 1

Théorie des graphes

Pour faciliter la compréhension d'un problème, nous avons tendance à le dessiner ce qui nous amène parfois même à le résoudre. La théorie des graphes est fondée à l'origine sur ce principe. De nombreuses propriétés et méthodes ont été pensées ou trouvées à partir d'une représentation schématique pour être ensuite formalisées et prouvées.

La théorie des graphes est historiquement un domaine mathématique qui s'est développé au sein d'autres disciplines comme la chimie, la biologie, la sociologie et l'industrie. Elle constitue aujourd'hui un corpus de connaissances très important et un instrument efficace pour résoudre une multitude de problèmes.

Dans ce chapitre, nous présenterons les notions et les concepts clés relatifs aux graphes, à savoir : la définition d'un graphe, ses types et sa représentation structurelle. Nous clôturons le chapitre avec quelques domaines d'application des graphes.

1.1 Graphe non orienté

1.1.1 Définitions et généralités

Un graphe non orienté G est la donnée d'un couple (V, E) où $V = \{v_1, v_2, \dots, v_n\}$ est un ensemble fini dont les éléments sont appelés sommets ou nœuds (Vertices en anglais) et $E = \{e_1, e_2, \dots, e_m\}$ est un ensemble fini d'arêtes (Edges en anglais). Toute arête e de E correspond à un couple non ordonné de sommets $(v_i, v_j) \in E \subset V \times V$ représentant ses extrémités (Müller, 2012) (Fages, 2014).

Soient $e = (v_i, v_j)$ et $e' = (v_k, v_l)$ deux arêtes de E , On dit que :

- v_i et v_j sont les extrémités de e et e est incidente à v_i et v_j (Hennecart et al., 2012).
- v_i et v_j sont voisins ou adjacents, s'il y a au moins une arête entre eux dans E (IUT, 2012).
- L'ensemble des sommets adjacents aux deux extrémités de e est appelé le voisinage de e (Müller, 2012).
- e et e' sont voisins s'ils ont une extrémité commune (Lopez, 2003).
- L'arête e est une boucle si ses extrémités coïncident, i.e, $v_i = v_j$ (IUT, 2012).
- L'arête e est multiple si elle a plus d'une seule occurrence dans l'ensemble E .

1.1.2 Représentation graphique

Un graphe non orienté G peut être représenté par un dessin sur un plan comme suit (Müller, 2012) :

- Les nœuds $v_i \in V$ de G sont représentés par des points distincts.
- Les arêtes $e = (v_i, v_j) \in E$ de G sont représentés par des lignes, pas forcément rectilignes, qui relient les extrémités de chaque arête e .

Exemple : Soit $g=(V1, E1)$ un graphe non orienté tel que : $V1=\{ 1,2,3,4,5 \}$ et $E1=\{(1,2), (1,4), (2,2), (2,3), (2,5), (3,4)\}$. La représentation graphique de g est alors donnée par le schéma de la figure 1.1.

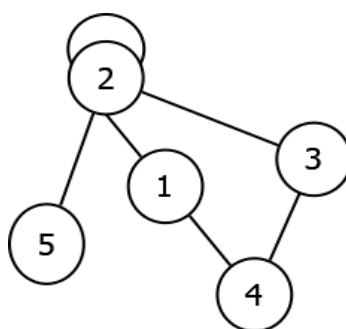


FIGURE 1.1 – Exemple de représentation graphique d'un graphe non orienté

1.1.3 Propriétés d'un graphe

- **Ordre d'un graphe :** On appelle ordre d'un graphe le nombre de ses sommets, i.e, $\text{Card}(V)$ (Roux, 2014).
- **Taille d'un graphe :** On appelle taille d'un graphe le nombre de ses arêtes, i.e, $\text{Card}(E)$ (Roux, 2014).
- **Degré dans un graphe :**
 - **Degré d'un sommet :** Le degré d'un sommet noté $d(v_i)$ est le nombre d'arêtes incidentes à ce sommet, sachant qu'une boucle compte pour deux (Müller, 2012). Dans l'exemple de la figure 1.1, le degré du sommet (1) est : $d(1)=2$.
 - **Degré d'un graphe :** Le degré d'un graphe est le degré maximum de ses sommets, i.e, $\max(d(v_i))$ (Müller, 2012). Dans l'exemple de la figure 1.1, le degré du graphe g est $d(2)=5$.
- **Rayon et diamètre dans un graphe :**
 - **Distance :** La distance entre deux sommets v et u est le plus petit nombre d'arêtes qu'on doit parcourir pour aller de v à u ou de u à v (Müller, 2012).
 - **Diamètre d'un graphe :** C'est la plus grande distance entre deux sommets de ce graphe (Müller, 2012).

- **Rayon d'un graphe** : C'est la plus petite distance entre deux sommets de ce graphe (Parlebas, 1972).

1.2 Graphe orienté

1.2.1 Définitions et généralités

Un graphe orienté G est la donnée d'un couple (V, E) où V est un ensemble fini dont les éléments sont appelés les sommets de G et $E \subset V \times V$ est un ensemble de couples ordonnés de sommets dits arcs (Müller, 2012). G est appelé dans ce cas digraphe (directed graph).

Pour tout arc $e = (v_i, v_j) \in E$:

- v_i est dit extrémité initiale ou origine de e et v_j est l'extrémité finale de e (Müller, 2012).
- v_i est le prédécesseur de v_j et v_j est le successeur de v_i (IUT, 2012).
- les sommets v_i, v_j sont des sommets adjacents (Jean-Charles Régin, 2016).
- e est dit sortant en v_i et incident en v_j (Jean-Charles Régin, 2016).
- e est appelé boucle si $v_i = v_j$, i.e, l'extrémité initiale et finale sont identiques (IUT, 2012).

1.2.2 Représentation graphique

Un graphe $G = (V, E)$ peut être projeté sur le plan en représentant :

- Dans un premier temps les nœuds $v_i \in V$ par des points disjoints du plan.
- Et dans un second temps les arcs $e = (v_i, v_j) \in E$ par des lignes orientées reliant par des flèches les deux extrémités de e .

Exemple :

Soit $g = (V_1, E_1)$ un digraphe tel que : $V_1 = \{1, 2, 3, 4\}$ et $E_1 = \{(1, 2), (1, 3), (3, 2), (3, 4), (4, 3)\}$. La représentation graphique de g est alors donnée par le schéma de la figure 1.2.

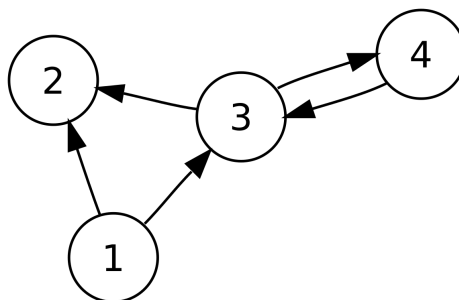


FIGURE 1.2 – Exemple de représentation graphique d'un digraphe.

1.2.3 Quelques Propriétés :

- **Ordre d'un digraphe** : est le nombre de sommets $n = \text{Card}(V)$ (Roux, 2014).
- **taille d'un digraphe** : est le nombre d'arcs $m = \text{Card}(A)$ (Roux, 2014).
- **Degré dans un digraphe** : Le degré d'un sommet $v_i \in V$ dans un digraphe $G=(V,E)$ est donné par la formule :

$$d(v_i) = d^+(v_i) + d^-(v_i)$$

où $d^+(v_i)$ est le nombre d'arcs sortants du sommet v_i et est appelé degré extérieur et $d^-(v_i)$ représente le nombre d'arcs incidents et est appelé degré intérieur (Müller, 2012).

- **Voisinage dans un digraphe** : Le voisinage d'un sommet $v_i \in V$, noté $V(v_i)$, dans un digraphe $G = (V, E)$ est :

$$V(v_i) = \text{succ}(v_i) \cup \text{pred}(v_i),$$

où $\text{succ}(v_i)$ est l'ensemble des successeurs de v_i et $\text{pred}(v_i)$ est l'ensemble de ses pré-décesseurs (Rigo, 2010), i.e, le voisinage de v_i est l'ensemble des sommets qui lui sont adjacents.

1.3 Notion de Connexité

Les graphes sont généralement exploitables à travers leur interrogation qui permet de fournir des réponses aux problèmes modélisés. L'une des informations les plus importantes dans un graphe est la notion de relations (directes ou indirectes) entre deux nœuds ou plus formellement *la connexité* dans un graphe. Dans cette partie, nous allons définir les concepts relatifs à cette notion dans le cas d'un graphe non orienté (resp. orienté).

- **Chaine (resp. Chemin)** : est une liste de sommets $S = (v_0, v_1, v_2, \dots, v_k)$ tel qu'il existe une arête (resp. un arc) entre chaque couple de sommets successifs (Müller, 2012).
- **Cycle (resp. Circuit)** : est une chaine (resp. chemin) dont le premier et le dernier sommet sont identiques (Roux, 2014).
- **Graphe connexe** : Un graphe non orienté (resp. orienté) est dit connexe (resp. fortement connexe) si pour toute paire de sommets (v_i, v_j) , il existe une chaine (resp. chemin) S les reliant (Müller, 2012).

1.4 Quelques types de graphe

Avec les avancées technologiques au fil du temps, plusieurs types de graphes ont vu le jour. En effet, la complexité et la variété des problèmes scientifiques existants modélisés par ces derniers ont poussé les chercheurs à adapter leur structure selon le problème auquel ils font face. Durant cette section, nous allons définir les principaux types existants.

- **Graphe Complet** : Un graphe $G = (V, E)$ est un graphe complet si tous les sommets $v_i \in V$ sont adjacents (Jean-Charles Régin, 2016). Il est souvent noté K_n où $n = \text{card}(V)$ (Roux, 2014).
- **Graphe étiqueté et graphe pondéré** : Un graphe étiqueté $G = (V, E, W)$ est un graphe non orienté (resp. orienté) dont chacune des arêtes (resp. arcs) $e_i \in E$ est doté d'une étiquette w_i . Si de plus, w_i est un nombre alors G est dit graphe pondéré (valué) (Roux, 2014).
- **Graphe simple et graphe multiple** : Un graphe $G = (V, E)$ non orienté (resp. orienté) est dit simple s'il ne contient pas de boucles et toute paire de sommets sont reliés par au plus une arête (resp. un arcs). Dans le cas contraire, G est dit multiple (IUT, 2012).

1.5 Graphe partiel et sous-graphe :

La quantité de données disponible aujourd'hui et sa croissance de manière exponentielle ont favorisé la décomposition des graphes en des entités plus petites afin de garantir une facilité de compréhension et d'analyse dans le but d'extraire l'information la plus pertinente. Dans cette partie, nous allons définir de manière plus formelle ce que ces entités sont, ainsi que leurs types.

1.5.1 Définitions :

Soient $G = (V, E)$, $G' = (V', E')$ et $G'' = (V'', E'')$ trois graphes.

- Le graphe G' est appelé **graphe partiel** de G si : $V' = V$ et $E' \subset E$ (Roux, 2014). En d'autres termes, un graphe partiel est obtenu en supprimant une ou plusieurs arêtes de G .
- Le graphe G'' est dit **sous-graphe** de G si : $V'' \subset V$ et $E'' \subset E \cap (V'' \times V'')$ (Rigo, 2010), i.e, un sous-graphe est obtenu en enlevant un ou plusieurs nœuds du graphe initial ainsi que les arêtes dont ils représentent l'une des deux extrémités.

1.5.2 Quelques Types de sous graphes :

- **Une Clique** : est un sous-graphe complet de G (Rigo, 2010).
- **Biparti** : G' est un sous-graphe biparti si il existe une partition de V' en deux sous ensembles notés V_1 et V_2 , i.e $V' = V_1 \cup V_2$ et $V_1 \cap V_2 = \emptyset$, tel que $E' = V_1 \times V_2$ (Rigo, 2010).
- **Étoile** : est un cas particulier de sous-graphe biparti où V_1 est un ensemble contenant le sommet central (dit *hub*) uniquement et V_2 contient le reste des nœuds (dits *spokes*) (Koutra et al., 2015) .

1.6 Représentation Structurale d'un graphe

Bien que la représentation graphique soit un moyen pratique pour définir un graphe, elle n'est clairement pas adaptée ni au stockage du graphe dans une mémoire, ni à son traitement.

Pour cela, plusieurs structures de données ont été utilisées pour représenter un graphe, ces structures varient selon l'usage du graphe et la nature des traitements appliqués. Nous allons présenter dans cette partie les structures les plus utilisées.

Soit un graphe $G(V, E)$ d'ordre n et de taille m dont les sommets v_1, v_2, \dots, v_n et les arêtes (ou arcs) e_1, e_2, \dots, e_m sont ordonnés de 1 à n et de 1 à m respectivement.

1.6.1 Matrice d'adjacence

La matrice d'adjacence de G est une matrice booléenne carrée d'ordre n : $(m_{ij})_{(i,j) \in [0;n]^2}$, dont les lignes (i) et les colonnes (j) représentent les identifiants des sommets de G . Les entrées (ij) prennent une valeur de "1" s'il existe un arc (une arête dans le cas d'un graphe non orienté) allant du sommet i au sommet j et un "0" sinon, i.e, (Lehman et al., 2010) (SABLIK, 2018) (IUT, 2012) :

$$m_{ij} := \begin{cases} 1 & \text{si } (v_i, v_j) \in E \\ 0 & \text{sinon} \end{cases}$$

Dans le cas d'un graphe non orienté, la matrice est symétrique par rapport à la première diagonale, i.e, $m_{ij} = m_{ji}$. Dans ce cas le, graphe peut être représenté avec la composante triangulaire supérieure de la matrice d'adjacence (Müller, 2012).

Note :

- Cette représentation est valide pour le cas d'un graphe non orienté et orienté.
- Dans le cas d'un graphe pondéré, les "1" sont remplacés par les poids des arêtes (ou arcs) (Lopez, 2003).
- Ce mode de représentation engendre des matrices très creuses (comprenant beaucoup de zero) (Hennecart et al., 2012).

Exemple : La figure 1.4 représente un exemple de matrice d'adjacence pour le graphe G ci-contre (figure 1.3) :

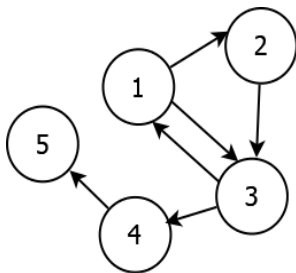


FIGURE 1.3 – Graphe orienté G

$$M = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

FIGURE 1.4 – Matrice d'adjacence du graphe G

Place occupée en mémoire : n^2 pour un graphe d'ordre $|n|$ (Lopez, 2003).

1.6.2 Matrice d'incidence

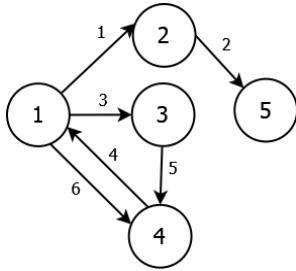
La matrice d'incidence d'un graphe orienté G est une matrice de taille $n \times m$ dont les lignes représentent les identifiants des sommets ($i \in V$) et les colonnes représentent les identifiants des arcs ($j \in E$) et dont les coefficients (m_{ij}) sont dans $\{-1, 0, 1\}$, tel que (Hennecart et al., 2012) (SABLIK, 2018) :

$$m_{ij} := \begin{cases} 1 & \text{si le sommet } i \text{ est l'extrémité final de l'arc } j \\ -1 & \text{si le sommets } i \text{ est l'extrémité initial de l'arc } j \\ 0 & \text{sinon} \end{cases}$$

Pour un graphe non orienté, les coefficients (m_{ij}) de la matrice sont dans $\{0, 1\}$, tel que (Hennecart et al., 2012) :

$$m_{ij} := \begin{cases} 1 & \text{si le sommet } i \text{ est une extrémité de l'arête } j \\ 0 & \text{sinon} \end{cases}$$

Exemple : La figure 1.6 représente un exemple d'une matrice d'incidence pour le graphe G ci-contre (figure 1.5) :

FIGURE 1.5 – Graphe orienté G

$$M = \begin{pmatrix} -1 & 0 & -1 & 1 & 0 & -1 \\ 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & -1 & 0 \\ -1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

FIGURE 1.6 – Matrice d'incidence du graphe G

Place occupée en mémoire : $n \times m$

1.6.3 Liste d'adjacence

La liste d'adjacence d'un graphe G est un tableau de $|n|$ listes, où chaque entrée (i) du tableau correspond à un sommet et comporte la liste $T[i]$ des successeurs (ou prédécesseurs) de ce sommet, c'est à dire tous les sommets j tel que $(i,j) \in E$ (SABLIK, 2018).

Dans le cas d'un graphe non orienté, on aura : $j \in \text{la liste } T[i] \iff i \in \text{la liste } T[j]$ (IUT, 2012).

Exemple : La figure 1.8 représente un exemple d'une liste d'adjacence pour le graphe G ci-contre (figure 1.7) :

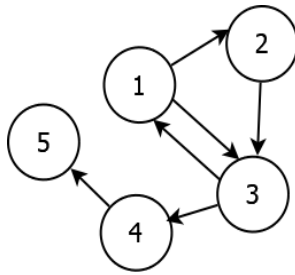


FIGURE 1.7 – Graphe orienté G

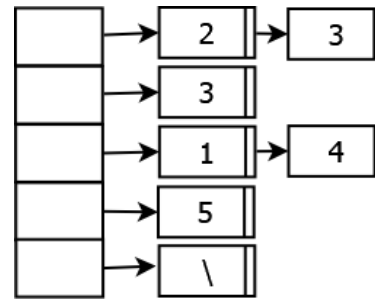


FIGURE 1.8 – Liste d'adjacence du graphe G

Place occupée en mémoire : Dans le cas d'un graphe orienté, l'espace occupé par le graphe est de $n + m$. Dans le cas d'un graphe non orienté, l'espace est de $n + 2m$ car une arête est représentée deux fois.

1.7 Les domaines d'application

La diversité des domaines faisant appel à la modélisation par des graphes ne cesse d'augmenter, allant des réseaux sociaux aux réseaux électriques et réseaux biologiques et arrivant jusqu'aux World Wide Web. Dans cette partie, nous allons décrire trois domaines d'application les plus répandus des graphes.

1.7.1 Graphes des réseaux sociaux :

Les réseaux sociaux représentent un lieu d'échange et de rencontre entre individus (entités) et dont l'utilisation est devenue de nos jours une nécessité. Pour représenter les interactions entre ces individus, nous avons généralement besoin d'avoir recours aux graphes où les sommets sont des individus ou des entités et les interactions entre eux sont représentées par des liens. Vue la diversité des interactions sociales, la modélisation de ces réseaux nécessite différents types de graphes : graphes non orientés pour les réseaux sociaux avec des relations symétriques, graphes orientés pour représenter des relations non symétriques comme c'est le cas dans les réseaux de confiance, graphes pondérés pour les réseaux sociaux qui contiennent différents niveaux d'intensité dans les relations, ..., etc. (Lemmouchi, 2012)

1.7.2 Graphes en Bioinformatique :

La bio-informatique est un domaine qui se trouve à l'intersection de deux grands domaines celui de l'informatique et celui de la biologie. Elle a pour but d'exploiter la puissance de calcul des équipements informatiques pour effectuer des traitements sur des données moléculaires massives (Pellegrini et al., 2004).

Elle est largement utilisée dans l'analyse des séquences d'ADN et des protéines à travers leur modélisation sous forme de graphe. A titre d'exemple, les graphes non orientés multiples sont un outil de modélisation des réseaux d'interaction protéine-protéine (Pellegrini et al., 2004), le but dans ce cas est l'étude du comportement d'une protéine par rapport à une autre.

1.7.3 Le Graphe du web :

Le graphe du Web est un graphe orienté dont les sommets sont les pages du web et les arcs modélisent l'existence d'un lien hypertexte dans une page vers une autre (Brisaboa et al., 2009). Il représente l'un des graphes les plus volumineux : en juillet 2000 déjà, on estimait qu'il contenait environ 2,1 milliards de sommets et 15 milliards d'arêtes avec 7,3 millions de pages ajoutées chaque jour (Guillaume and Latapy, 2002). De ce fait, ce graphe a toujours attiré l'attention des chercheurs. En effet, l'étude de ses caractéristiques a donné naissance à plusieurs algorithmes intéressants, notamment l'algorithme PageRank de classement des pages web qui se trouve derrière le moteur de recherche le plus connu de nos jours : Google.

1.8 Conclusion

Dans ce chapitre nous avons présenté les notions et les concepts généraux qui touchent à la théorie des graphes : définitions des graphes, leurs principales propriétés, leurs représentations ainsi que leurs domaines d'application.

Le point important qu'on a pu tirer de cette partie est que les graphes sont devenus un moyen crucial et indispensable dans la modélisation des problèmes dans plusieurs domaines. Cependant ils deviennent de plus en plus complexes et volumineux avec la grande quantité de données disponible de nos jours. De ce fait, leur stockage, visualisation et traitement sont devenus difficiles. La compression de graphe est née comme solution à ce problème. Dans le chapitre suivant nous allons présenter la compression de graphes, son rôle et ses différentes méthodes.

Chapitre 2

Compression de graphes

La puissance des processeurs, de nos jours, augmente plus vite que les capacités de stockage ce qui engendre un déséquilibre entre le volume des données qu'il est possible de traiter et de stocker. Dès lors, la réduction de la taille des données, plus formellement la compression de données, a été un domaine de recherche très actif.

Dans ce chapitre, nous allons ouvrir un champ de réflexion en introduisant tout d'abord le domaine de compression des données et son application dans la théorie des graphes. Puis dans un second temps, nous allons présenter une étude bibliographique sur les méthodes de compression existantes aujourd'hui et qui s'inscrivent dans l'une des deux classes de méthodes de compression : les méthodes de compression par les k2-trees et les méthodes de compression par extraction de motifs, pour finir avec une étude comparative entre elles.

2.1 Compression de données :

La compression de données est principalement une branche de la théorie de l'information qui traite des techniques et méthodes liées à la minimisation de la quantité de données à transmettre et à stocker. Sa caractéristique de base est de convertir une chaîne de caractères vers un autre jeu de caractères occupant un espace mémoire le plus réduit possible tout en conservant le sens et la pertinence de l'information (Lelewer and Hirschberg, 1987).

Les techniques de compression de données sont principalement motivées par la nécessité d'améliorer l'efficacité du traitement de l'information. En effet, la compression des données en tant que moyen peut rendre l'utilisation des ressources existantes beaucoup plus efficace.

De ce fait, une large gamme d'applications usant de ce domaine tel que le domaine des télécommunications et le domaine du multimédia est apparue offrant une panoplie d'algorithmes de compression (Sethi et al., 2014). Sans les techniques de compression, Internet, la télévision numérique, les communications mobiles et les communications vidéo, qui ne cessaient de croître, n'auraient été que des développements théoriques.

2.2 Compression appliquée aux graphes :

La compression des graphes a été proposée comme solution pour le traitement et le stockage des graphes volumineux. Elle permet de transformer un grand graphe en un autre plus petit tout en préservant ses propriétés générales et ses composants les plus importants. Les principaux avantages de la compression sont (Liu et al., 2018a) :

- **La réduction de la taille des données et de l'espace de stockage :** De nos jours, les graphes représentant les bases de données, les réseaux sociaux et tous types de données numériques sont caractérisés par une croissance exponentielle de leurs volumes, ce qui rend leur stockage difficile et coûteux en terme d'espace mémoire. Les techniques de compression produisent des graphes plus petits qui nécessitent moins d'espace. Cela permet aussi de réduire le nombre d'opérations d'E/S ainsi que les communications entre nœuds dans un environnement distribué et de charger le graphe en mémoire centrale.
- **L'exécution rapide des algorithmes de traitement et des requêtes sur les graphes :** L'exécution des différents algorithmes de traitement sur des graphes volumineux peut s'avérer coûteuse en terme de temps et peut ne pas donner les résultats attendus. La compression permet d'obtenir de petits graphes qui peuvent être traités, analysés et interrogés plus efficacement et dans un temps raisonnable.
- **La Facilité d'analyse et de visualisation des graphes :** Les techniques de compression permettent de représenter les données et les structures des graphes massives d'une manière plus significative permettant ainsi leur analyse et leur visualisation contrairement au graphes d'origines qui ne peuvent même pas être chargés en mémoire.
- **L'élimination du bruit :** l'un des exemples les mieux illustrant cet avantage sont les grands graphes du web qui sont considérablement bruités. Ce bruit peut perturber l'analyse en faussant les résultats et en augmentant la charge de travail liée au traitement des données. La compression permet donc de filtrer le bruit et de ne mettre en évidence que les données importantes.

2.2.1 Les types de compression :

La compression de graphes est définie comme l'ensemble des méthodes et techniques permettant de réduire l'espace mémoire occupé par ces derniers tout en gardant la même signification que le graphe d'origine. Dès lors, deux approches se présentent : la compression avec ou sans perte, que nous allons détailler dans ce qui suit.

a) Compression Sans Perte :

Certains domaines d'application de la compression nécessitent un niveau élevé d'exactitude et une restitution exacte, donc une compression sans perte. Dans cette catégorie, le graphe G subit des transformations pour avoir une représentation compacte G' qui lors de la décompression donne exactement G . La figure ci-dessous illustre cette définition.



FIGURE 2.1 – Compression sans perte.

b) Compression Avec Perte :

Contrairement à la compression sans perte, la compression avec perte permet la suppression permanente de certaines informations jugées inutiles (redondantes) pour améliorer la qualité de la compression. En d'autres termes, le graphe G subit des transformations pour avoir une représentation compacte G' qui lors de la décompression donne un graphe G'' probablement différent de G mais l'approximant le plus possible. La figure ci-dessous illustre cette définition.



FIGURE 2.2 – Compression avec perte.

2.2.2 Les métriques d'évaluation des algorithmes de compression :

Devant la panoplie d'algorithmes et de techniques de compression de graphe disponibles dans la littérature, des critères de comparaison et d'évaluation entre ces méthodes doivent être bien définis. Dans cette partie, nous présenterons les principales mesures de performances.

a) Le temps de compression :

C'est une métrique qui donne le temps d'exécution de l'algorithme de compression. Elle est généralement mesurée en secondes (ou ms).

b) Le ratio de compression :

Le ratio de compression (CR) est la mesure la plus courante pour calculer l'efficacité d'un algorithme de compression. Il est défini comme le rapport entre le nombre total de bits requis pour stocker les données non compressées et le nombre total de bits nécessaires pour stocker les données compressées.

$$CR = \frac{\text{Nbre. de bits du graphe originale}}{\text{Nbre. de bits du graphe finale}}$$

Le CR est parfois appelé bit par bit (bpb) et il est défini alors comme étant le nombre moyen de bits requis pour stocker les données compressées (Uthayakumar et al., 2018). Dans le cas des algorithmes de compression de graphe on a :

- **Le nombre de bits par nœud** : représente l'espace mémoire nécessaire pour stocker un nœud (bpn pour « *bits per node* » en Anglais).
- **Le nombre de bits par lien** : représente l'espace mémoire nécessaire pour stocker un arc dans le cas d'un graphe orienté ou une arête dans le cas d'un graphe non orienté (bpe pour « *bits per edge* » en Anglais).

c) Le taux de compression :

Exprimée en pourcentage, cette métrique permet de mesurer la performance de la méthode de compression. Elle peut être exprimée de deux manières différentes :

- **Le taux de compression** : Le rapport entre volume du graphe après compression et le volume initial du graphe.

$$t = \frac{\text{La taille du graphe finale}}{\text{La taille du graphe originale}}$$

- **Le gain d'espace** : Le gain d'espace représente la réduction de la taille du graphe compressé par rapport à la taille du graphe original.

$$G = 1 - \frac{\text{La taille du graphe finale}}{\text{La taille du graphe originale}}$$

2.2.3 Classification des méthodes de compression :

Le domaine de compression des graphes est un domaine qui a connu une grande évolution vu son importance. Une multitude de méthodes ont été proposées au cours des dernières années. Elles diffèrent l'une de l'autre dans plusieurs points : le type de graphe en entrée, le type de structure en sortie, le type de compression et la technique utilisée pour la compression. En se basant sur ces différences, plusieurs classifications ont été suggérées. Nous allons dans ce qui suit présenter les plus importantes parmi ces classifications.

Dans (Maneth and Peternek, 2015), les auteurs proposent une classification basée tout d'abord sur le type de compression. Ils regroupent les méthodes en deux catégories principales : les méthodes de compression sans perte et les méthodes de compression avec perte. La première catégorie est subdivisée à son tour selon le type de représentation du graphe en sortie : représentation succincte, représentation structurelle ou une représentation sous forme de fichier RDF¹. Les méthodes donnant une représentation succincte représentent le graphe sous forme d'une chaîne de bits succincte irréversible. La sortie de ces méthodes est ainsi une structure compacte du graphe original. Parmi les méthodes de cette classe, nous trouvons : Web Framework de Boldi et Vigna (Boldi and Vigna, 2004). La deuxième classe est la représentation structurelle.

1. Resource Description Framework (RDF) : est un modèle de graphe destiné à décrire de façon formelle les ressources Web et leurs métadonnées, de façon à permettre le traitement automatique de telles descriptions.

Contrairement à l'approche précédente, les méthodes de cette classe modifient la structure du graphe initial, sachant que les modifications apportées sont réversibles. La sortie sera donc une structure réduite et non pas compacte de la version initiale. Parmi ces méthodes, nous citons : RePair de Claude et Navarro (Claude and Navarro, 2010b). La dernière classe est la compression des fichiers RDF et qui comporte des méthodes assez récentes. Nous trouvons parmi ces techniques : Dcomp de (Martínez-Prieto et al., 2012). Les méthodes de compression avec perte quant à elles apportent des modifications irréversibles sur le graphe en supprimant les informations redondantes et le bruit. Comme exemple, nous citons : ASSG de Zhang et al. (Zhang et al., 2014a).

Une autre classification a été exposée par Lui et al. dans (Liu et al., 2018a) qui classe les méthodes sur trois niveaux. Au premier et deuxième niveaux, les techniques de compression sont regroupées en fonction du type de graphe en entrée selon deux critères : graphe statique ou dynamique et graphe simple ou étiqueté. Pour le troisième niveau, les auteurs catégorisent les méthodes selon la technique de traitement utilisée. Quatre catégories sont définies : les méthodes de regroupement ou d'agrégation, ces méthodes permettent d'agréger de manière récursive un ensemble de nœuds, liens ou carrément un cluster en un super nœud (appelé parfois nœud virtuel), comme exemple de ces techniques, nous trouvons Grass (LeFevre and Terzi, 2010). Le deuxième type de méthodes englobe les méthodes de compression de bits. Ces méthodes minimisent le nombre de bits nécessaires au stockage du graphe en se basant sur le principe de description minimal (Minimum Description Length (MDL) en Anglais). Elles peuvent être avec ou sans perte. Parmi elles, nous citons LSH-based (Khan et al., 2014). La troisième classe comporte les méthodes de simplification qui suppriment les arêtes les moins importantes selon un certain critère. Parmi ces méthodes, nous trouvons celle proposée dans pqr Shen et al. (Shen et al., 2006). La dernière catégorie est la classe des méthodes basées sur l'influence, les méthodes de cette catégorie décrivent le graphe par les flux d'influence les plus importants ce qui permet de l'analyser plus facilement. Ces méthodes permettent de formuler le problème de compression comme un processus d'optimisation dans lequel la quantité de données liée à l'influence est maintenue en sortie. Parmi ces techniques, nous mentionnons (Shi et al., 2015).

La dernière classification que nous allons présenter est la classification proposée dans un master de l'année dernière par le binôme : Mlle. Belhocine et Mr. Guermah (W. GUERMAH, 2018). Cette classification se base sur le principe utilisé dans le processus de compression. Elle regroupe six classes de méthodes : 1) compression basée sur l'ordre des nœuds en exploitant le principe de similarité et de localité du graphe, les méthodes de cette catégorie cherchent à trouver un ordre des nœuds, qui doit répondre à deux propriétés essentielles : la similarité² et la localité³, cet ordre est ensuite utilisé dans la construction d'une structure de données qui compresse le graphe en entrée, comme exemple de ces méthodes, nous trouvons Layered Label Propagation (Boldi et al., 2011) qui utilise l'ordre LLP⁴ et Recursive Graph Bisection de

2. Deux nœuds proches ont tendance à avoir des voisins similaires.

3. Les liens sortants d'un nœud ont tendance à se diriger vers un ensemble de nœuds qui sont proches.

4. LLP est un algorithme itératif qui produit une séquence d'ordres de nœuds en se basant sur les étiquettes

(Dhulipala et al., 2016) qui utilise un ordre BP⁵. 2) compression basée sur l'ordre des nœuds en exploitant la linéarisation du graphe, les méthodes de cette classe se basent sur une nouvelle structure de données intitulée structure de données eulérienne, elles sont conçues principalement pour les grands graphes, parmi ces méthodes nous trouvons Neighbor Query Friendly Compression (Maserrat and Pei, 2012). 3) compression basée sur l'étiquetage des nœuds par des intervalles, ces méthodes visent à construire une structure d'index à partir du graphe initial qui permet de répondre aux requêtes de voisinage, parmi ces méthodes nous citons DAGGER (Yıldırım et al., 2012). 4) compression basée sur la structure d'arbre K^2 , les méthodes de cette catégorie s'appuient sur la représentation K^2 -trees pour compresser le graphe, ces méthodes font l'objet de notre travail et seront détaillées par la suite. 5) compression basée sur l'agrégation des nœuds, ces méthodes sont parmi les méthodes les plus populaires dans le domaine de compression, elles cherchent à compresser le graphe initial en agrégeant un certain nombre de nœuds en un seul nœud appelé super-nœud, cette agrégation se fait selon différentes manières, elle peut être orientée par une fonction objectif représentant l'espace de stockage optimal généralement établie par le principe de la longueur de description minimale MDL ou par la similarité des caractéristiques des nœuds tels que les labels et les attributs ou par l'extraction de motifs en utilisant des techniques de pattern mining, cette dernière représente l'une des matières de notre recherche. 6) compression basée sur l'agrégation des liens, contrairement aux techniques de la classe précédente, les méthodes de cette classe visent à compresser le graphe initial en fusionnant certains ensembles de ses liens en un seul liens appelé super-lien, elle est établie selon deux façons, elle est orientée soit par le biais des règles de grammaire, ou bien par l'extraction de motifs que nous allons étudier par la suite. la figure 2.3 représente le schéma de cette classification.

affectées aux clusters après avoir partitionner le graphe initial.

5. BP est un ordre basé sur le problème de bissection de graphes, qui cherche à trouver une meilleure partition des nœuds du graphe tout en minimisant une fonction objectif.

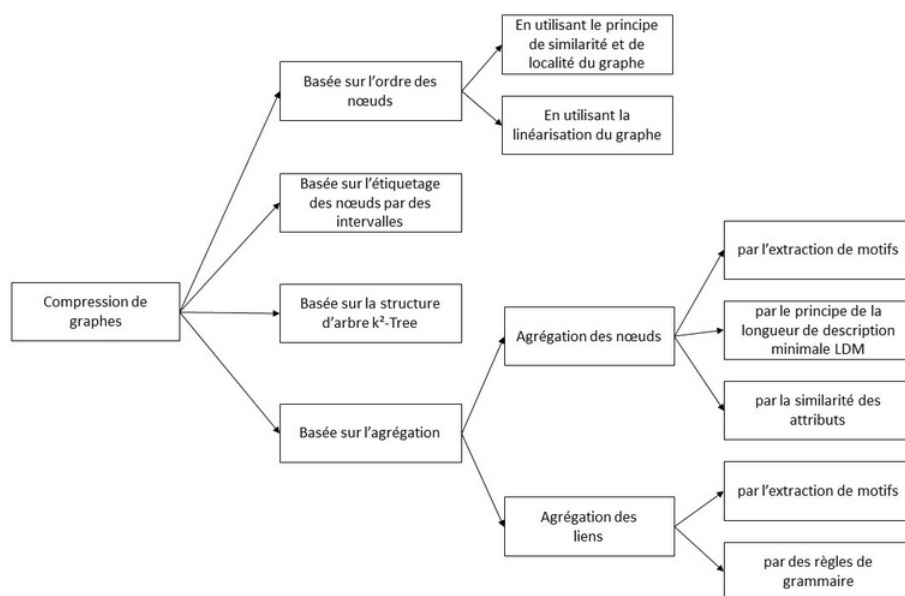


FIGURE 2.3 – Classification des méthodes de compression (W. GUERMAH, 2018)

Après l'étude des méthodes basées sur l'agrégation par extraction de motifs, nous avons constaté que certaines classes ne sont pas bien définies. Nous avons remarqué que la classification a proposé d'inclure les méthodes basées sur l'extraction de motifs dans deux classes de méthodes qui sont : les méthodes basées sur l'agrégation des liens et les méthodes basées sur l'agrégation des nœuds, or ce sont les méthodes d'extraction de motifs qui englobent certaines méthodes d'agrégation et non pas le contraire. En effet, notre étude bibliographique a clairement montré que les méthodes de compression de graphe par extraction de motifs ne sont pas toutes des méthodes agrégatives. Nous trouvons, par exemple, certaines de ces méthodes qui donnent en sortie une liste des motifs les plus pertinents du graphe avec une matrice d'erreurs sans avoir recours à l'agrégation. Donc nous avons proposé de dissocier la classe des méthodes basées sur l'extraction de motifs de la classe des méthodes basées sur l'agrégation. En outre, nous avons considéré, contrairement à la classification précédente, que les méthodes de compression basées sur les règles de grammaire font parties des méthodes de compression par extraction de motifs. Nous justifions cela par le fait que ces méthodes utilisent l'un des deux principes suivants :

1. La transformation de la liste d'adjacence en une chaîne de caractères et le remplacement, de manière itérative, de la sous-chaîne de longueur deux la plus fréquente par une règle de production et qui représente donc dans ce cas un motif.
2. La recherche d'un motif, de type *digram*⁶, satisfaisant une certaine condition et le remplacement de toutes ses occurrences par une production dans la grammaire.

Le fait que les motifs dans le cas de cette classe n'ont pas une structure de sous-graphes connus (clique, étoile, ...) n'empêche pas que le processus appliqué s'intègre toujours dans la discipline d'extraction de motifs. Nous avons essayé donc de rectifier ces imperfections tout en raffinant davantage les deux classes qui font l'objet de notre Master.

6. Un digramme : est sous-graphe composé de deux arêtes ayant un sommet en commun.

La classe des méthodes de compression par les arbres k2-trees ne peut être encore raffinée car elles partagent toutes le même principe et diffèrent dans le type de graphe en entrée ou dans le codage en sortie.

Pour la classe des méthodes de compression par extraction de motifs, nous distinguons cinq classes : 1) les méthodes de compression basées vocabulaire faisant appel aux méthodes de clustering, 2) les méthodes de compression basées vocabulaire exploitant les propriétés de la matrice d'adjacence, 3) les méthodes de compression basées agrégation des nœuds des motifs, 4) les méthodes de compression basées agrégation des liens des motifs utilisant des règles de grammaire, 5) les méthodes de compression basées agrégation des liens des motifs faisant appel à des heuristiques de partitionnement de graphes. La première et la deuxième classe se caractérisent par l'ensemble des motifs qui est prédéfini au départ. Cependant, elles se distinguent l'une de l'autre dans le principe de fonctionnement : l'une utilise des méthodes de clustering et de détection de communautés tant dis que l'autre utilise directement la matrice d'adjacence en exploitant ses propriétés. Les trois (03) dernières classes forment l'ensemble des méthodes d'extraction de motifs basées sur l'agrégation de ces derniers.

La figure 2.4 représente la classification de l'an dernier après raffinement où nous avons mis en évidence nos apports et nos modifications en pointillé.

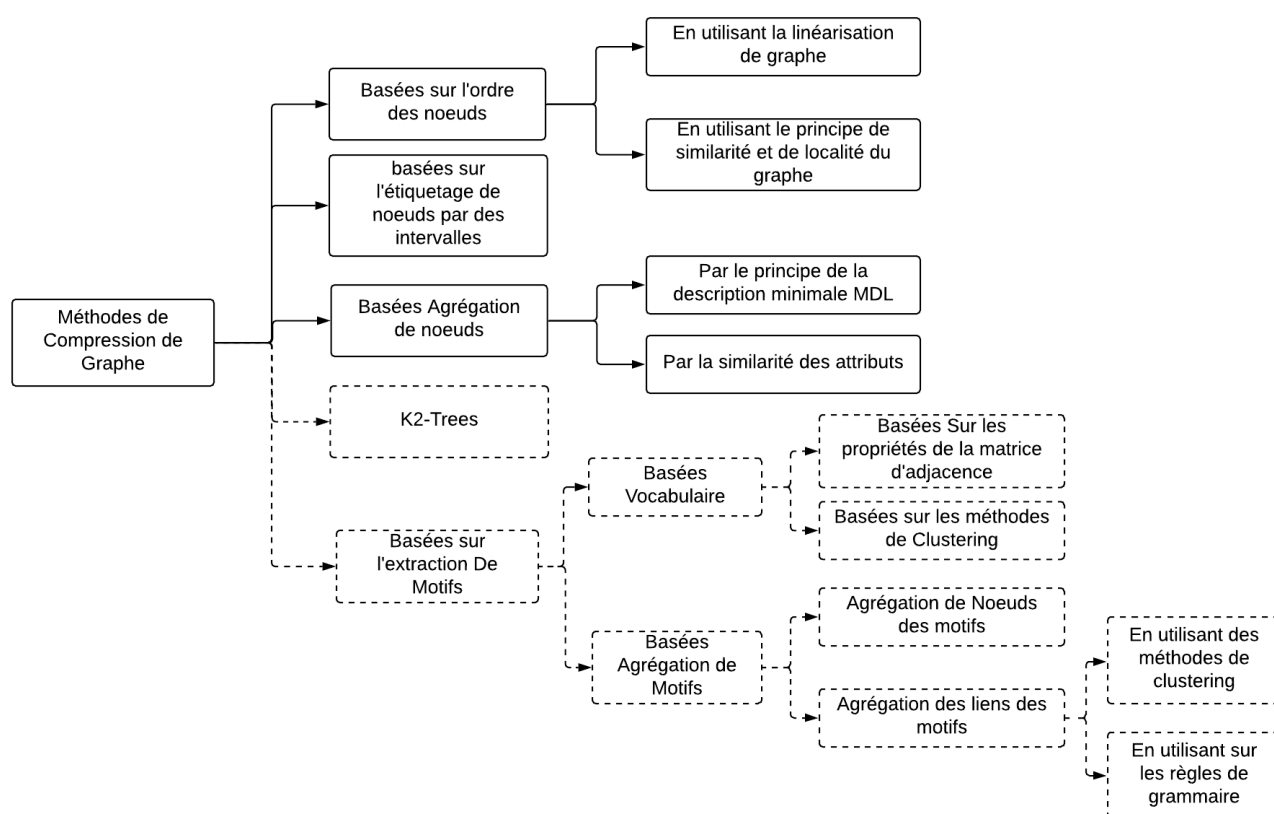


FIGURE 2.4 – Classification proposées des méthodes de compression

2.3 Compression par les arbres K^2 -Trees

k^2 -tree est une structure de données conçue à l'origine pour la compression des graphes du web. L'algorithme de base a été proposé par Brisaboa et al. dans leur article *k^2 -trees for Compact Web Graph* (Brisaboa et al., 2009). Elle a été appliquée ensuite dans d'autres types de données tel que les réseaux sociaux (Shi et al., 2012), les données rasters (De Bernardo et al., 2013) et les bases de données RDF (Alvarez-Garcia et al., 2017).

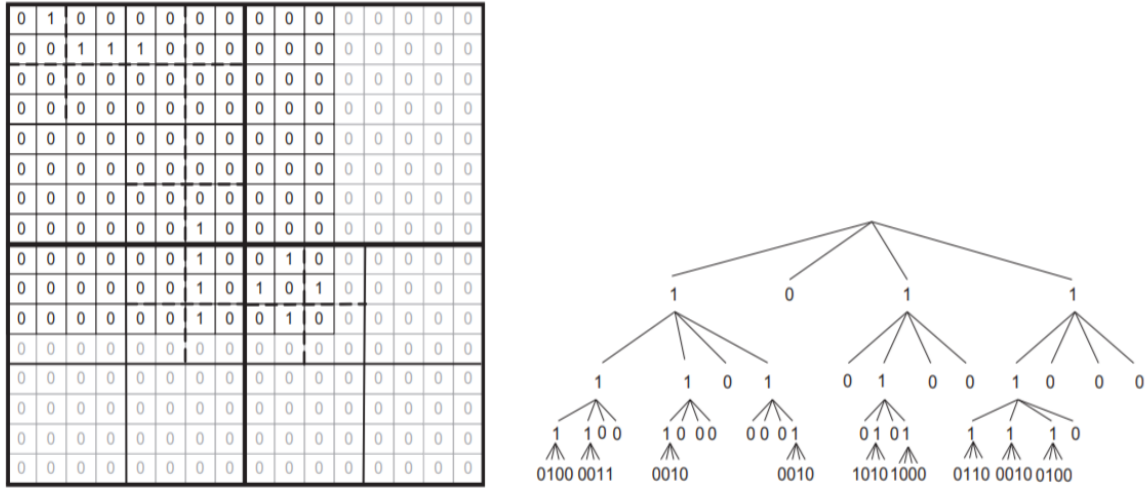
En général, l'algorithme peut être appliqué à n'importe quelle matrice binaire. Dans le cadre de notre étude nous nous intéressons seulement à la matrice d'adjacence d'un graphe. La compression par les k^2 -trees exploite les propriétés de la matrice d'adjacence et tire parti des zones vides pour réduire l'espace de stockage et permettre au graphe de tenir en mémoire centrale. Il offre aussi la possibilité de naviguer dans le graphe sans le décompresser, et de répondre aux requêtes de voisinage direct et inverse.

Étant donné une matrice d'adjacence A d'ordre $|n|$, k^2 -tree la représente sous forme d'un arbre de recherche k^2 -aires⁷ de hauteur $h = \lceil \log_k n \rceil$. Chaque nœud de l'arbre contient un seul bit avec deux valeurs possibles : 1 pour les nœuds internes et 0 pour les feuilles, sauf le dernier niveau où les feuilles représentent les cases de la matrice A et peuvent prendre une valeur 0 ou 1. Chaque nœud interne de l'arbre a exactement k^2 fils. Avant la construction de l'arbre, il faut s'assurer que n est une puissance de k . Dans le cas contraire, l'algorithme étend la matrice en rajoutant des zéros à droite et en bas de la matrice. L'ordre de la matrice devient donc $n' = k^{\lceil \log_k n \rceil}$.

Pour construire l'arbre, k^2 -tree commence par diviser la matrice en k^2 sous matrices d'ordre $|n/k|$. La racine correspond à la matrice complète. Chaque sous matrice représente un nœud dans le premier niveau de l'arbre, elle est ajoutée comme un fils à la racine suivant un ordre de gauche à droite et de haut en bas. Le nœud est à 1 si la sous matrice qu'il représente contient au moins un 1, et à 0 si elle ne contient que des 0. Le processus est répété de manière récursive sur les sous matrices représentées par des 1. k^2 sous matrices sont considérées à chaque subdivision. L'opération est répétée jusqu'à ce que la subdivision atteigne les cases de la matrice qui représenteront les feuilles de l'arbre au dernier niveau.

La figure 2.5 illustre la représentation k^2 -tree d'une matrice de taille 10×10 , étendue à une taille 16×16 pour un $k=2$ (Brisaboa et al., 2015).

7. Les arbres n -aires sont une généralisation des arbres binaires : chaque nœud a au plus n fils.

FIGURE 2.5 – Exemple de représentation k^2 -tree

Pour le stockage de l'arbre, l'algorithme utilise deux tableaux binaires : un tableau T (Tree) contenant tous les nœuds de l'arbre à l'exception du dernier niveau et un tableau L (Leaves) contenant les feuilles du dernier niveau. Les nœuds et les feuilles sont ordonnés selon un parcours en largeur de l'arbre. Ci-dessous les deux tableaux T et L de l'exemple précédent (figure 2.5) :

$$T = 1011 \ 1101 \ 0100 \ 1000 \ 1100 \ 1000 \ 0001 \ 0101 \ 1110$$

$$L = 0100 \ 0011 \ 0010 \ 0010 \ 1010 \ 1000 \ 0110 \ 0010 \ 0100$$

Dans le pire des cas, l'espace total pour la description de la structure est $k^2 * m(\log_{k^2} \frac{n^2}{m} + O(1))$, où n est le nombre de nœuds du graphe et m le nombre de liens. Cependant, pour les graphes réels, l'espace nécessaire pour le stockage est bien meilleur.

Dans le même article (Brisaboa et al., 2009), et dans le but d'obtenir un compromis entre la taille de l'arbre et le temps de parcours, les auteurs proposent une hybridation qui consiste à changer la valeur du paramètre k en fonction du niveau de l'arbre en donnant à k une grande valeur au début pour réduire le nombre de niveaux et améliorer ainsi le temps de recherche, et une petite valeur à la fin pour avoir des petites sous matrices et réduire l'espace de stockage. Pour le stockage de l'arbre, un tableau T_i est utilisé pour chaque valeur k_i , le tableau L reste le même.

Plusieurs variantes de l'algorithme de base ont été proposées dans la littérature dont le but était soit d'obtenir un meilleur résultat de compression, soit d'appliquer la méthode sur d'autres types de graphes. Nous allons dans ce qui suit présenter les principaux travaux qui traitent ce sujet.

Dans (Shi et al., 2012), les auteurs proposent deux techniques d'optimisation de l'algorithme : la première consiste à trouver un certain ordre des nœuds qui permet de regrouper les 1 de la matrice d'adjacence dans une seule sous matrice au lieu qu'ils soient dispersés de manière aléatoire. La recherche d'un ordre optimal des nœuds n'est pas envisageable. Avec $k=2$, le problème peut être réduit à un autre problème (min bisection⁸) qui est NP-difficile. Les auteurs utilisent alors un parcours Depth First Search (DFS)⁹ avec des heuristiques pour trouver une approximation de l'ordre optimal. Cette optimisation permet de réduire le nombre de nœuds internes de l'arbre et produit ainsi un arbre optimal. La deuxième optimisation est de trouver la valeur de k la plus adéquate pour chaque nœud interne, calculer cette valeur pour chaque nœud peut engendrer un temps de calcul très important. Pour éviter cela, les auteurs affectent la même valeur k pour les nœuds ayant le même parent.

Dans un travail ultérieur (Brisaboa et al., 2014b), les auteurs de l'article de base apportent deux améliorations principales de leur méthode dans le but d'optimiser l'espace et le temps de parcours de l'arbre produit. La première est de construire k^2 arbres distincts pour les k_0^2 sous matrices du premier niveau. Elle présente plusieurs avantages : (1) l'espace est réduit étant donné que la taille de chaque arbre est en fonction de $\frac{n^2}{k^2}$, (2) le temps de parcours s'améliore puisque T et L sont plus petits. La deuxième amélioration est la compression de L qui consiste à construire un vocabulaire V de toutes les sous matrices du dernier niveau sous forme de séquences de bits, les classer par fréquence d'apparition et remplacer leurs occurrences dans L par des pointeurs. Cela permet d'éviter la redondance et de réduire la taille de la structure. Les pointeurs sont représentés par des codes de longueur variable ordonnés, le plus petit correspond à la sous matrice la plus fréquente. Néanmoins, cette représentation ne permet pas un accès direct dans L étant donné qu'une décompression séquentielle est nécessaire pour récupérer une position. Pour remédier à ce problème, les auteurs utilisent le principe de Directly Addressable Codes (DAC)¹⁰ (Brisaboa et al., 2013) pour garantir un accès rapide au pointeur et conserver ainsi une navigation efficace.

Exemple : Pour la figure 2.5, le vocabulaire et L sont représentés comme suit :

$$V = [0010\ 0100\ 0011\ 1010\ 1000\ 0110]$$

$$L = c_1c_2c_0c_0c_3c_4c_5c_0c_1$$

Dans (Brisaboa et al., 2012), les mêmes auteurs développent la représentation k^2 -trees pour les graphes dynamiques. Ils proposent une nouvelle structure nommée dk^2 -trees pour *Dynamique k^2 -trees* qui offre les mêmes capacités de compression et fonctionnalités de navigation que le cas statique et qui permet également d'avoir des mises à jour sur le graphe. Pour atteindre ces objectifs, dk^2 -tree remplace la structure statique de k^2 -tree par une implémentation dynamique. Dans cette nouvelle implémentation, les deux tableaux T et L sont remplacés par deux

8. Le problème de bisection minimale (MBP) est un problème NP-hard bien connu, qui est destiné à diviser les sommets d'un graphe en deux moitiés égales afin de minimiser le nombre de ces arêtes avec exactement une extrémité dans chaque moitié.

9. DFS : Parcours en profondeurs du graphe

10. DAC est une technique qui encode une séquence d'entiers ou mots en utilisant une structure à longueur variable, son avantage principale est l'accès direct au mot sans passer par le décodage.

arbres, nommés T_{tree} et L_{tree} respectivement. Les feuilles de T_{tree} et L_{tree} stockent des parties des bitmaps T et L. La taille de ces feuilles est une valeur paramétrable. Les nœuds internes des deux arbres permettent d'accéder aux feuilles et de les modifier. Chaque nœud interne de T_{tree} contient trois(03) éléments : deux compteurs b et o, qui contiennent respectivement le nombre de bits et le nombre de "1" stockés dans les feuilles descendantes de ce nœud, et un pointeur P vers le nœud fils. Les nœuds internes de L_{trees} sont similaires sauf qu'ils ne contiennent que b et P. Avec cette structuration, T_{tree} et L_{tree} permettent la mise à jour directe de l'arbre $k_2 - tree$ dans le cas l'ajout et la suppression des liens dans le graphe.

La figure 2.6 présente une représentation dk^2 -tree (Brisaboa et al., 2012) :

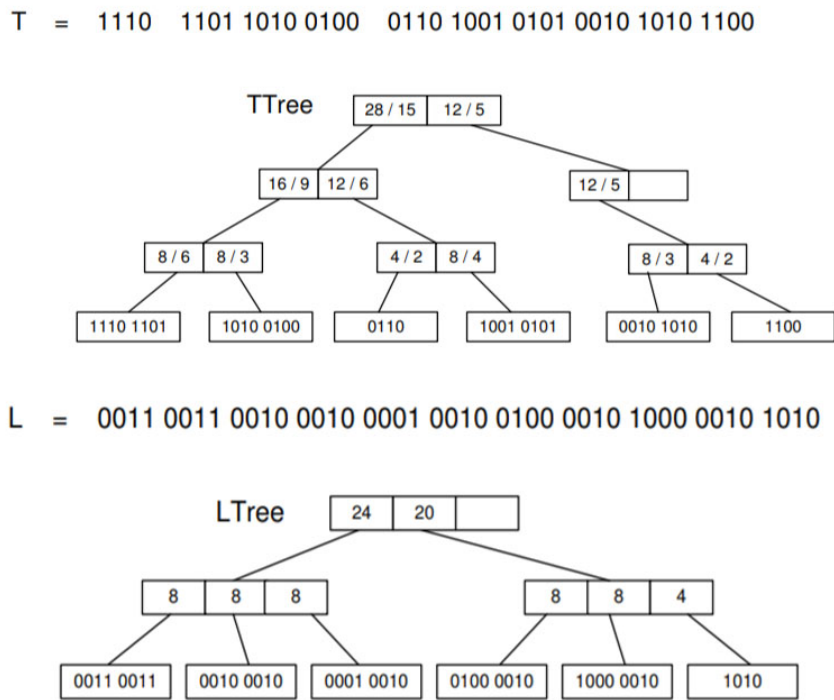


FIGURE 2.6 – Exemple d'une représentation dk^2 -tree

Sandra et al. (De Bernardo et al., 2013) présentent k^n -tree, une généralisation des k^2 -tree pour les problèmes multidimensionnels. Cette méthode possède plusieurs applications. Elle est utilisée pour représenter les bases de données multidimensionnelles, les données rasters et les graphes dynamiques. k^n -tree repose sur k^2 -tree pour représenter une matrice à n-dimensions (dites *tensor* en Anglais). La matrice est décomposée en k^n sous-matrices de même taille, comme suit : sur chaque dimension, $K-1$ hyperplans divisent la matrice dans les positions $i * \frac{n}{K}$, pour $i \in [1, K - 1]$. Une fois les dimensions partitionnées, k^n sous-matrices sont induites, elles sont représentées par des nœuds dans l'arbre comme dans l'algorithme de base. Les structures utilisées pour le stockage sont aussi les mêmes (T et L). En posant $n=3$, la méthode peut être appliquée sur les graphes dynamiques ou temporels. Ce type de graphe est représenté par une grille à 3 dimensions $X \times Y \times T$, où les deux premières dimensions représentent les nœuds de départ et de destination, et la troisième dimension représente le temps. La figure 2.7 présente

une représentation k^3 -tree d'un graphe dynamique (de Bernardo Roca, 2014).

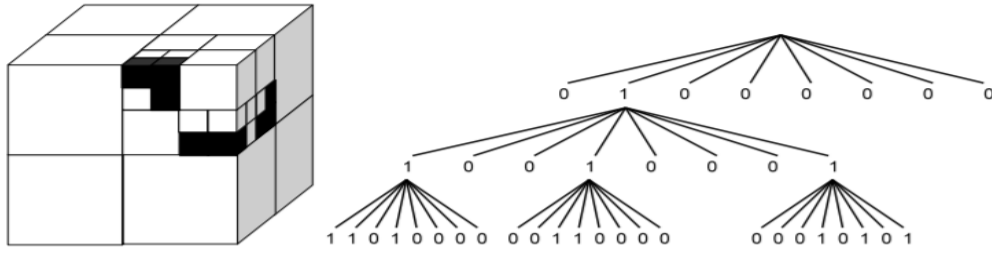


FIGURE 2.7 – Exemple d'une représentation k^3 -tree

La représentation de base des arbres k^2 -trees regroupe seulement les zones de zéros, puisqu'elle a été conçue au début pour les graphes du web qui possèdent une matrice d'adjacence extrêmement creuse. Dans (de Bernardo Roca, 2014), Les auteurs proposent d'étendre cette représentation en regroupant les zones de "1" également. L'idée générale est d'arrêter la décomposition de la matrice d'adjacence quand une zone unie est trouvée, à savoir des zéros ou des uns. Pour distinguer entre les différents nœuds, une représentation quadtree est utilisée (De Berg et al., 1997). Une couleur est attribuée à chaque nœud, blanc pour une zone de zéro, noir pour une zone de uns et gris pour les nœuds internes, i.e, les zones contenant des uns et des zéros. Pour le stockage des nœuds, les auteurs proposent quatre encodages présentés dans ce qui suit :

- $k^2\text{-trees}^{1^{2\text{-bits-naive}}}$: Dans cet encodage, deux bits sont utilisés pour représenter chaque type de nœud. L'attribution des bits n'est pas arbitraire, le premier bit du poids fort indique si le nœud est un nœud interne (0) ou une feuille (1), le deuxième détermine si les feuilles sont blanches (0) ou noires (1). Nous aurons donc : "10" pour les nœuds gris, "01" pour les nœuds noirs et "00" pour les nœuds blancs. Notant que les feuilles du dernier niveau sont représentées par un seul bit. Après le codage, le premier bit de chaque nœud sauf ceux du dernier niveau est stocké dans T , un autre tableau T' est créé pour sauvegarder le deuxième bit. Les nœuds du dernier niveau sont stockés dans L .
- $k^2\text{-tree}^{1^{2\text{-bits}}}$: Le même principe que l'encodage précédant sauf que les nœuds gris sont représentés par un seul bit, toujours à "1". Le tableau T' va contenir dans ce cas la couleur des feuilles ce qui va réduire la taille de la structure.
- $k^2\text{-tree}^{1^{DF}}$: Cet encodage est similaire à $k^2\text{-trees}^{1^{2\text{-bits}}}$, mais il utilise un seul bit pour les nœuds blancs et deux bits pour les nœuds noirs et gris, compte tenue de la fréquence des nœuds blancs dans les graphes du monde réel par rapport aux autres. Nous aurons donc : "0" pour les nœuds blancs, "10" pour les nœuds gris et "11" pour les nœuds noirs.
- $k^2\text{-tree}^{1^{5\text{-bits}}}$: Le dernier encodage repose sur la représentation de base. Un nœud blanc est représenté par "0", un nœud noir ou gris par "1", exactement comme le k^2 -tree d'origine. Pour identifier un nœud noir (zone de uns), il sera représenté par une combinaison impossible : k^2 fils de "0" sont ajoutés aux nœuds noirs pour les distinguer.

La figure 2.8 illustre une représentation k^2 -tree1 avec les quatre encodages (de Bernardo Roca, 2014) :

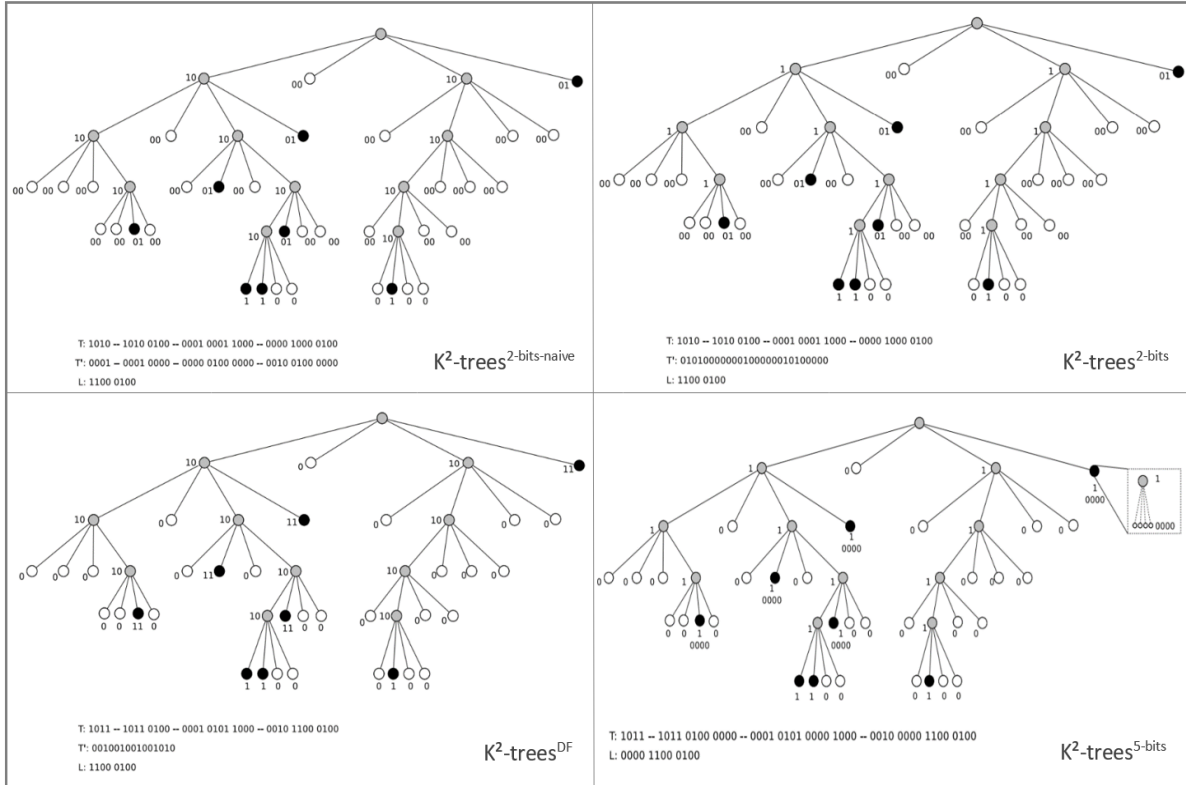


FIGURE 2.8 – Exemple d'une représentation k^2 -tree1 avec les quatre encodages

Dans (Zhang et al., 2014b), les auteurs proposent Delta- k^2 -tree, une variante qui exploite la propriété de similarité entre les nœuds voisins du graphe pour réduire le nombre de uns dans la matrice d'adjacence. Notons par *Matrix* la matrice d'adjacence, Delta- k^2 -trees construit une nouvelle matrice appelée *Delta-matrix*, une ligne i de *Delta-matrix* va contenir la différence entre les deux lignes *Matrix*[i] et *Matrix*[$i-1$] si cela décroît le nombre de uns sinon elle sera égale à *Matrix*[i], comme suit :

$$\left\{ \begin{array}{l} \text{Si } \text{count1s}(\text{matrix}[i]) < \text{countDif}(\text{matrice}[i], \text{matrix}[i-1]) \\ \quad \text{Delta} - \text{matrix}[i] := \text{matrix}[i] \\ \text{Sinon} \\ \quad \text{Delta} - \text{matrix}[i] := \text{matrix}[i] \oplus \text{matrix}[i-1] \end{array} \right.$$

Où *count1s* compte le nombre de 1 dans une ligne, *countDif* compte le nombre de bits différents entre deux lignes et \oplus représente le "ou exclusif".

Pour déterminer si une ligne est identique à celle de la matrice d'adjacence ou non, un tableau D est utilisé : Si $D[i]=1$ la ligne est identique, sinon c'est une ligne de différence. La matrice *Delta-matrix* contient moins de uns que la matrice d'adjacence, d'où elle est plus creuse, ce qui permet de réduire la taille de la structure et avoir un meilleur taux de compression. Cependant le temps de parcours est plus grand car pour accéder à certaines lignes (lignes de

différence), le graphe doit être décompressé et la matrice d'adjacence reconstituée.

La figure 2.9 présente un exemple de la représentation Delta- k^2 -tree (Zhang et al., 2014b).

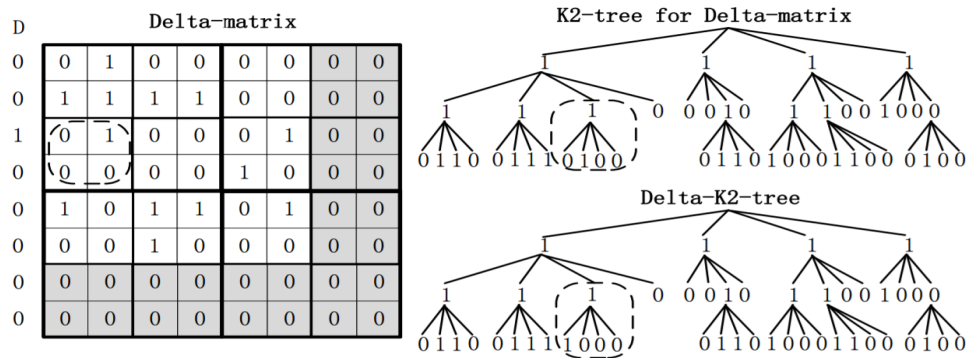


FIGURE 2.9 – Exemple d'une représentation Delta- k^2 -tree

K^2 -treap est une autre variante de k^2 -tree, elle a été proposée dans (Brisaboa et al., 2014a). Cette variante combine les k^2 -trees avec une autre structure de données appelée treap¹¹ (Aragon and Seidel, 1989). Les auteurs appliquent cette méthode sur des grilles multidimensionnelles comme Online Analytical Processing (OLAP) pour pouvoir les stocker et répondre efficacement aux requêtes top-K (Badr, 2013). La méthode peut être également appliquée sur les graphes pondérés, où chaque case de la matrice d'adjacence du graphe comporte le poids de l'arête qu'elle représente au lieu d'un 1. Comme dans l'algorithme de base, une décomposition récursive en k^2 sous-matrices est appliquée sur la matrice d'adjacence et un arbre k^2 -air est construit comme suit : la racine de l'arbre va contenir les coordonnées ainsi que la valeur de la cellule ayant le plus grand poids de la matrice. La cellule ajoutée à l'arbre est ensuite supprimée de la matrice. Si plusieurs cellules ont la même valeur maximale, l'une d'elles est choisie au hasard. Ce processus est répété récursivement sur chaque sous matrice. La procédure continue sur chaque branche de l'arbre jusqu'à ce qu'on tombe sur les cellules de la matrice d'origine ou sur une sous matrice complètement vide (contient que des zéros).

La figure 2.10 suivante illustre la représentation k^2 -treap d'un graphe pondéré (Badr, 2013) :

11. Les Treaps sont des arbres de recherche binaire avec des nœuds ayant deux attributs : clé et priorité. La recherche dans ces arbres s'effectue selon ces attributs.

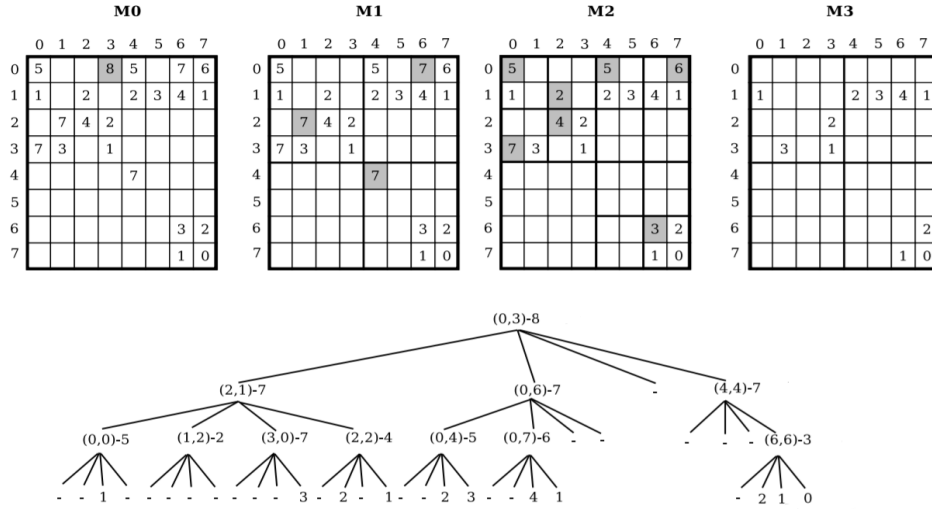


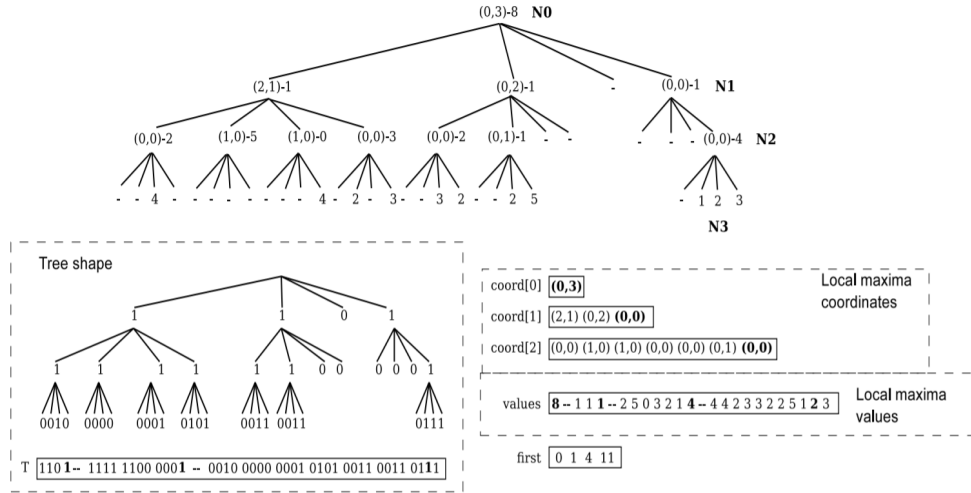
FIGURE 2.10 – Exemple d’une représentation k^2 -treap d’un graphe pondéré

Pour avoir une bonne compression, k^2 -treap effectue des transformations sur les données stockées. La première transformation consiste à changer les coordonnées représentées dans l’arbre en des coordonnées relatives par rapport à la sous matrice actuelle. La deuxième est de remplacer chaque poids dans l’arbre par la différence entre sa valeur et celle de son parent.

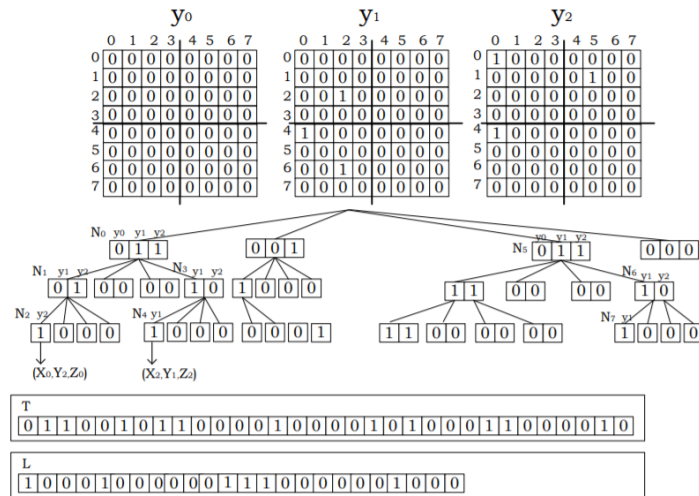
Trois structures de données sont utilisées pour sauvegarder les coordonnées et les valeurs des cellules ainsi que la topologie de l’arbre. Chaque structure est détaillée dans ce qui suit :

- *Listes de coordonnées locales* : La séquence de coordonnées de chaque niveau l de l’arbre est stockée dans une liste $coord[l]$.
- *Liste des valeurs* : Le parcours de l’arbre se fait en largeur, la séquence des valeurs récupérées est stockée dans une liste nommée $values$. Un tableau nommé $first$ est utilisé pour sauvegarder la position début de chaque niveau dans $values$.
- *L’arbre* : La structure de l’arbre k^2 -treap est sauvegardée avec un arbre k^2 -tree, les nœuds contenant des valeurs dans k^2 -treap sont représentés par des uns et les nœuds vides par des zéros. Pour le stockage de l’arbre, un seul tableau T est utilisé.

La figure 2.11 représente les structures de données utilisées pour le stockage de l’arbre de la figure 2.10 précédente (Badr, 2013) :

FIGURE 2.11 – Structures de données pour une représentation k^2 -treap

Garcia et al. (Garcia et al., 2014) proposent Ik^2 -tree pour Interleaved k^2 -tree. Elle est appliquée sur les bases de données RDF ainsi que sur les graphes dynamiques. Dans les graphes dynamiques, les deux premières dimensions correspondent aux nœuds source et destination et la troisième dimension reflète le temps. Le graphe est donc défini par $|T|$ matrices d'adjacence prises à des instants t_k différents. Ik^2 -tree représente les matrices simultanément. Chaque matrice est représentée par un arbre k^2 -tree et ils sont par la suite regroupés dans un seul arbre. Chaque nœud de l'arbre obtenu représente une sous-matrice comme dans l'algorithme de base, sauf qu'au lieu d'utiliser un seul bit, Ik^2 -tree utilise 1 à $|T|$ bits pour représenter le nœud. Le nœud racine contient $|T|$ bits. Le nombre de bits de chaque fils dépend du nombre de uns de son parent. L'arbre finale est toujours stocké avec deux tableaux : T et L. La figure 2.12 est un exemple de Ik^2 -tree appliqué sur un graphe dynamique représenté dans trois instants (Garcia et al., 2014) :

FIGURE 2.12 – Exemple d'une représentation Ik^2 -tree

Une variante de Ik^2 -tree appelée Differential Ik^2 -tree a été étudiée dans (Alvarez-Garcia et al., 2017). Elle vise à améliorer le taux de compression en représentant uniquement les changements survenus sur le graphe à un instant t_i au lieu d'une instance complète : à l'instant t_0 , une capture complète du graphe (matrice d'adjacence) est stockée. À l'instant t_k , pour $k > 0$, seules les arrêtes qui changent de valeurs entre t_{k-1} et t_k sont stockées. Les matrices sont représentées à la fin de la même manière que Ik^2 -tree. La limite de cette représentation est que la structure doit être décompressée lors d'une requête. La figure 2.13 montre un exemple d'une représentation $\text{diff}Ik^2$ -trees (Alvarez-Garcia et al., 2017).

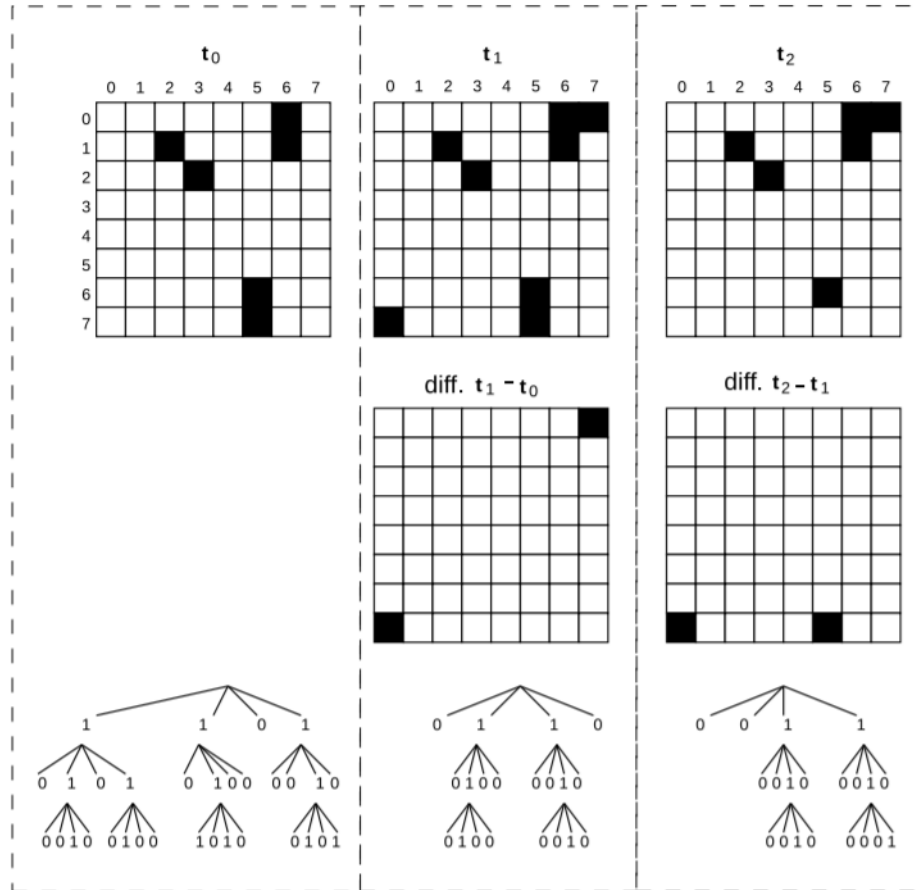


FIGURE 2.13 – Exemple d'une représentation $\text{diff}Ik^2$ -tree

Dans (Álvarez-García et al., 2018), les auteurs étendent la représentation k^2 -tree pour les bases de données orientées graphes. Ces graphes sont étiquetés, attribués, orientés et ont des arêtes multiples. Ils présentent le graphe sous forme d'une nouvelle structure intitulée $\text{Att}k^2$ -tree pour *Attributed k^2 -trees*. La figure 2.14 montre un exemple de graphe pris en compte par $\text{Att}K^2$ -tree (Álvarez-García et al., 2018).

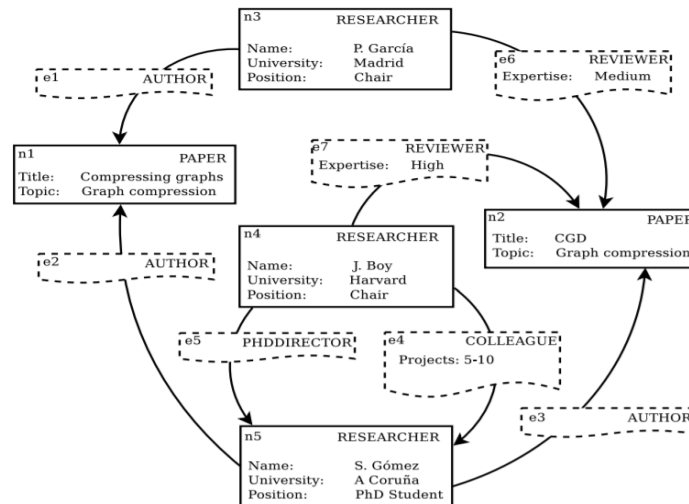


FIGURE 2.14 – Exemple d'un graphe étiqueté, attribué, orienté et multiple

Structures de données : La représentation obtenue par la compression est composée d'un ensemble d'arbres k^2 -tree et d'autres structures supplémentaires. Le graphe est représenté par trois composantes : un schéma de données, les données incluses dans les nœuds et les liens et finalement la relation entre les éléments du graphe. Chaque composant est présenté dans ce qui suit :

- *Schéma* : Ce composant gère les étiquettes et les attributs de chaque type d'éléments, il joue le rôle d'un index dans la représentation. Il est composé de :

Un schéma de nœuds : représenté par un tableau qui contient les étiquettes des nœuds ordonnées lexicographiquement. Un identifiant est attribué à chaque nœud du graphe selon l'ordre du tableau, les m_1 nœuds possédant la première étiquette du tableau vont avoir des identifiants de 1 à m_1 , les m_2 nœuds avec la deuxième étiquette du tableau vont avoir des identifiants de $m_1 + 1$ à $m_1 + m_2$ et ainsi de suite. Chaque entrée du tableau va ainsi stocker le plus grand identifiant portant son étiquette, cela permet de trouver l'étiquette d'un nœud à travers son identifiant.

Un schéma d'arêtes : Comme dans le cas des nœuds, un tableau est utilisé pour stocker les étiquettes des arêtes avec le même principe.

Le schéma est le point de départ de la représentation, il permet d'obtenir l'étiquette d'un nœud ou d'une arête, et d'accéder à ses attributs.

- *Données* : Ce composant contient toutes les valeurs que peut prendre un attribut dans le graphe. Un attribut peut être représenté de deux façons différentes selon sa fréquence d'apparition, on distingue donc deux types d'attributs :

Attributs rares : Ce sont les attributs qui prennent généralement des valeurs différentes à chaque apparition, ils sont stockés dans des listes et indexés avec l'identifiant de l'élément.

Attributs fréquents : Ce type d'attributs est sauvegardé dans deux matrices, une pour les attributs des nœuds et l'autre pour les attributs des liens. Les matrices sont stockées sous forme d'arbres k^2 -trees.

La figure 2.15 illustre les deux composants schéma et données de la représentation Att_k^2 -trees de la figure 2.14 (Álvarez-García et al., 2018) :

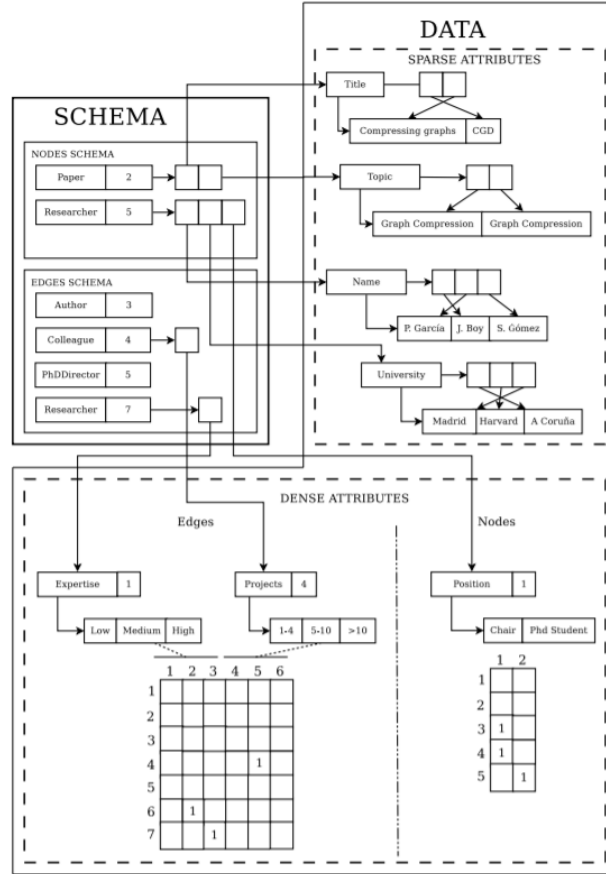


FIGURE 2.15 – Exemple d'une représentation Att_k^2 -tree (1/2)

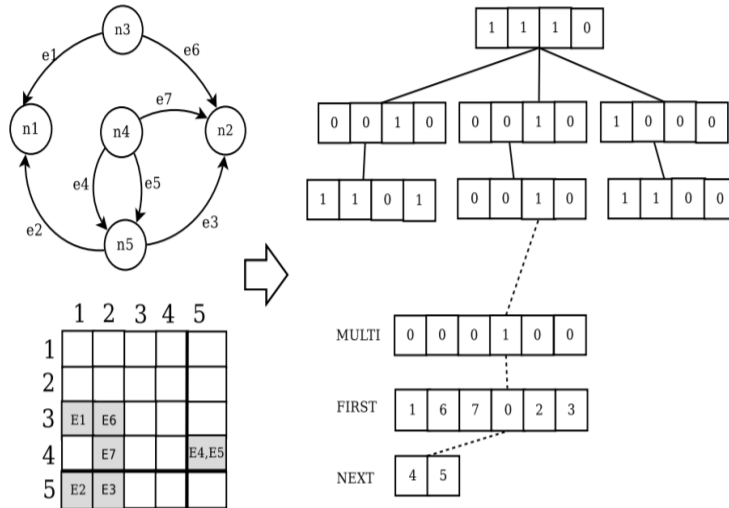
— **Relations :** C'est le dernier composant de Att_k^2 -tree, il stocke les relations entre les nœuds et les arêtes du graphe en utilisant un arbre k^2 -tree et d'autres structures pour sauvegarder les identifiants des arêtes ainsi que les arêtes multiples. Les structures supplémentaires sont les suivantes :

Multi : Un tableau qui indique si l'arête est multiple ou non.

Firt : Un tableau qui donne l'identifiant de l'arête, ou de celui de la première dans le cas d'une arête multiple.

Next : Un tableau qui contient les identifiants des arêtes multiples restantes.

La figure 2.16 donne la représentation Att_k^2 -tree des relations du graphe de la figure 2.14 (Álvarez-García et al., 2018) :

FIGURE 2.16 – Exemple d’une la représentation Attk^2 -tree (2/2)

Dans le même article (Álvarez-García et al., 2018), les auteurs étendent Attk^2 -tree pour les graphes dynamiques. Ils proposent une nouvelle variante appelée dynAttk^2 -tree qui supporte le changement dans les attributs et les liens du graphe. Comme Attk^2 -tree, dynAttk^2 -tree représente le graphe avec trois composantes : Schémas, données et relations. Les composantes sont semblables à ceux de Attk^2 -tree mais avec certaines amélioration vue la nature dynamique du graphe.

Structure de données :

- *Schéma* : En ce qui concerne les nœuds, leurs étiquettes sont stockées dans une liste dynamique ordonnée lexicographiquement. En outre, une séquence dynamique est utilisée pour sauvegarder le type de chaque nœud. Elle est stockée ensuite sous forme d’un arbre d’ondelettes¹² (Grossi et al., 2003). Le même principe est appliqué sur les arêtes.
- *Données* : Les attributs rares sont stockés dans des listes dynamiques, quant aux attributs fréquents, ils sont sauvegardés avec des arbres dk^2 -trees (un arbre pour chaque attribut).
- *Relations* : Le stockage des relations se fait à l’aide d’un dk^2 -tree et des tableaux dynamiques pour stocker les identifiants des arêtes et les arêtes multiples.

Synthèse des méthodes de compression basée sur les K^2 -trees

Nous avons présenté dans les parties précédentes les différentes méthodes de compression existantes basées sur la représentation k^2 -trees, et expliqué le principe de fonctionnement de chaque une d’elles. Nous allons par la suite comparer entre ces méthodes et résumer nos recherches.

Une présentation synthétique de notre étude est fournie dans le tableau 2.2. Chaque ligne du tableau représente une méthode, tandis que chaque colonne représente un aspect susceptible

12. Ou wavelet en anglais, est un arbre binaire équilibré qui contient des données compressées dans une représentation optimale.

d'être utilisé dans la méthode (type de graphe, type de compression, structure en sortie). Nous constatons que toutes les méthodes de cette classe sont des méthodes de compression sans perte qui supportent les graphes orientés et non orientés, cependant nous remarquons une variation dans l'aspect temporelle, certains méthodes sont destinées aux graphes statiques comme k^2 -trees de base et k^2 -trees1, tandis que d'autres sont appliquées aux graphes dynamiques comme k^n -trees et dk^2 -trees. Certains méthodes se distinguent aussi par le type de graphe comme k^2 -treaps qui s'applique aux graphes pondérés et Att_k^2 -trees qui accepte les graphes étiquetés, attribués et multiples. Nous observons aussi que toutes les méthodes donnent en sortie une représentation succincte. Le tableau résume aussi les résultats de l'application des méthodes sur des graphes de test. Les figures 2.17 et 2.18 sont des représentations visuelles des ces résultats.

La figure 2.17 représente les résultats obtenues par l'application des méthodes de compression basées sur les k^2 -trees sur le graphe eu-2005 en tenant compte de l'ordre chronologique de leurs apparitions. Les résultats sont sous forme de nombre de bits par liens. Nous remarquons que la première méthode de base donne un résultat de compression de 5.2bpe, ce nombre a diminué avec l'apparition des améliorations et des nouvelles méthodes. Nous constatons alors une amélioration contenue de l'efficacité de ces méthodes de manière générale surtout avec les améliorations proposées par (Brisaboa et al., 2014b) et Δk^2 -trees où le nombre de bits atteint approximativement 3.22 bits par liens.

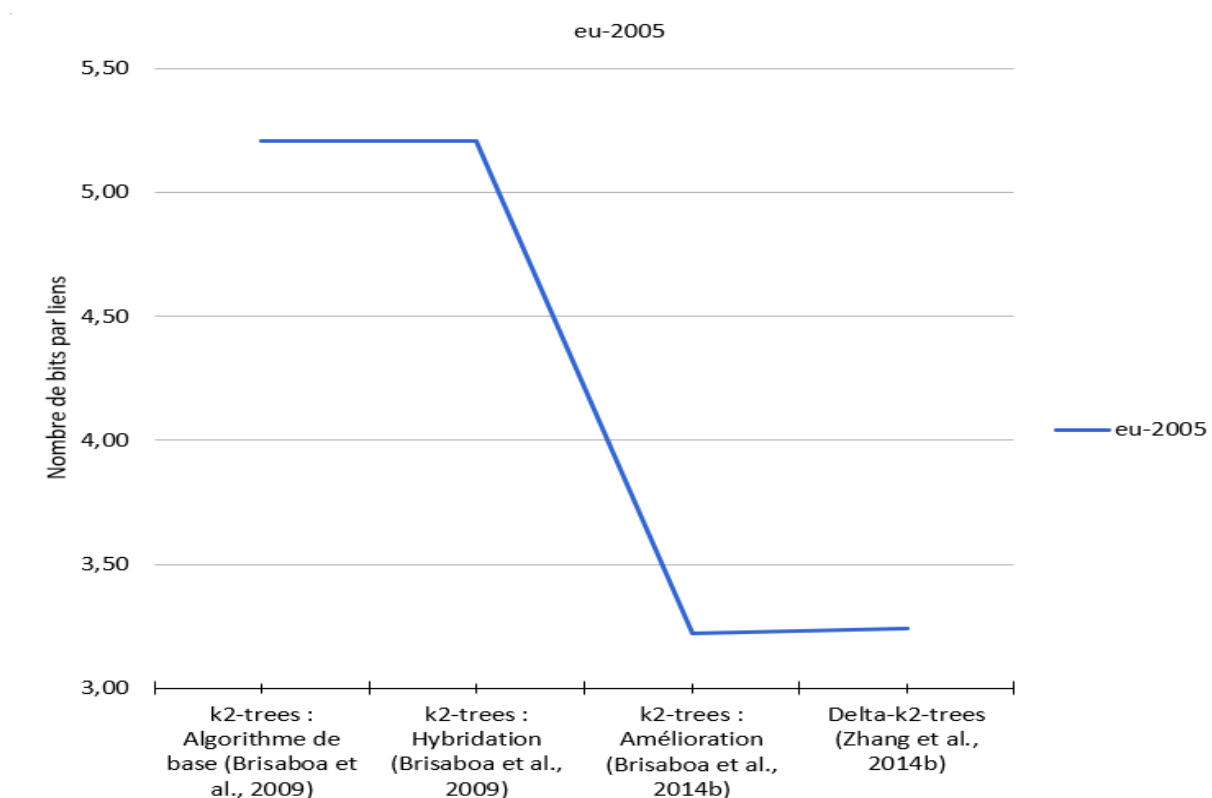


FIGURE 2.17 – Comparaison des performances des méthodes basées sur les k^2 -trees appliquées sur le graphe eu2005 selon le nombre de bits par liens

A partir de la figure 2.18 qui présente les résultats de l'application de différentes méthodes

sur quatre graphes de test (CommNet, cond-mat, eu-2005, movielen10M) en fonction du taux de compression. Nous remarquons que entre k^2 -trees et k^2 -trees1, la méthode k^2 -trees est nettement meilleure, cela s'explique par le fait que les graphes du web sont représentés par des matrices creuses et k^2 -trees1 est mieux adaptée au matrices avec des zones de uns comme les matrices d'images par exemple. Pour les graphes dynamiques, nous observons que k^n -trees donne de bons résultats par rapport à Ik^2 -trees mais avec une légère différence. Pour les graphes attribués, $Att k^2$ -trees donne un taux de compression de 93%.

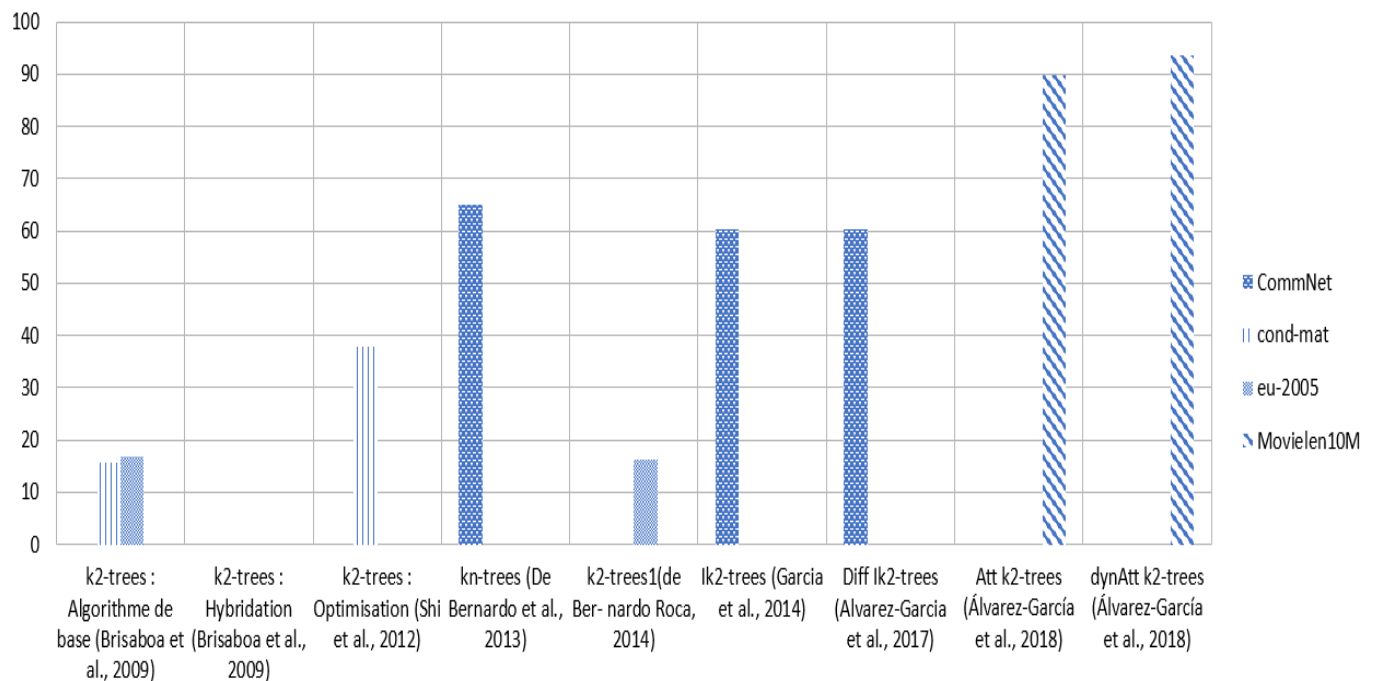


FIGURE 2.18 – Comparaison entre les méthodes basées sur les k^2 -trees en fonction du taux de compression

Article	Graphe en entrée					Compression		Structure en sortie		Graphe de test	Résultat
	Orienté	Non orienté	Statique	Dynamique	Autre Propriétés	Avec perte	Sans perte	Succincte	Structurale		
k^2 -trees : Algorithme de base (Brisaboa et al., 2009)	✓	✓	✓	✗		✗	✓	✓	✗	eu-2005 cond-mat	5.21 bits/lien taux de compression : 16.88% taux de compression : 15.58%
k^2 -trees : Hybridation (Brisaboa et al., 2009)	✓	✓	✓	✗		✗	✓	✓	✗	eu-2005	5.21 bits/lien
k^2 -trees : Optimisation (Shi et al., 2012)	✓	✓	✓	✗		✗	✓	✓	✗	cond-mat	taux de compression : 37.96%
k^2 -trees : Amélioration (Brisaboa et al., 2014b)	✓	✓	✓	✗		✗	✓	✓	✗	eu-2005	3.22 bits/lien
dk^2 -trees (Brisaboa et al., 2012)	✓	✓	✗	✓		✗	✓	✓	✗	eu-2005	6.2 bits/lien
k^n -trees (De Bernardo et al., 2013)	✓	✓	✗	✓		✗	✓	✓	✗	CommNet	taux de compression : 65.16%

TABLE 2.1 – Synthèse des méthodes de compression par k^2 -trees.

Article	Graphe en entrée					Compression		Structure en sortie		Graphe de test	Résultat
	Orienté	Non orienté	Statique	Dynamique	Autre Propriétés	Avec perte	Sans perte	Succincte	Structurale		
k^2 -trees I (de Ber-nardo Roca, 2014)	✓	✓	✓	✗		✗	✓	✓	✗	eu-2005	taux de compression : 16.23%
Delta- k^2 -trees (Zhang et al., 2014b)	✓	✓	✓	✗		✗	✓	✓	✗	eu-2005	3.24 bits/liens
k^2 -treaps (Brisaboa et al., 2014a)	✓	✓	✓	✗	Pondéré	✗	✓	✓	✗	SalesDay	2.48 bits/liens
Ik^2 -trees (Garcia et al., 2014)	✓	✓	✗	✓		✗	✓	✓	✗	CommNet	taux de compression : 60.43%
Diff Ik^2 -trees (Alvarez-Garcia et al., 2017)	✓	✓	✗	✓		✗	✓	✓	✗	CommNet	taux de compression : 60.43%
Att k^2 -trees (Álvarez-García et al., 2018)	✓	✓	✗	✓	Étiqueté Attribué Multiple	✗	✓	✓	✗	Movielen-10M	taux de compression : 89.97%
dynAtt k^2 -trees (Álvarez-García et al., 2018)	✓	✓	✗	✓	Étiqueté Attribué Multiple	✗	✓	✓	✗	Movielen-10M	taux de compression : 93.75%

TABLE 2.2 – Synthèse des méthodes de compression par k^2 -trees.

2.4 Compression par extraction de motifs

Les motifs fréquents sont des connaissances extraites des données. Leur but est de fournir à l'utilisateur des informations non triviales, implicites, présumées non connues. Ils lui offrent ainsi une meilleure appréhension des données. Dès lors, l'extraction de motifs fréquents est devenue une tâche importante dans la fouille de données et un thème très étudié par la communauté. Elle a aussi été vastement utilisée dans le domaine de compression des graphes vu qu'elle permet de ne garder que l'information utile et d'éliminer les redondances de manière efficace. En effet, nous trouvons plusieurs méthodes basées sur ce principe où nous pourrions clairement distinguer deux grandes classes : (i) les méthodes de compression basées vocabulaire (ii) les méthodes de compression basées Agrégation.

Dans cette section, nous allons expliquer le principe de base de chaque classe où nous allons subdiviser chacune en plusieurs sous-classes en se basant sur ce dernier.

2.4.1 Compression basée vocabulaire

Les méthodes de compression par extraction de motifs basées vocabulaire sont des méthodes qui ont attirées l'attention des chercheurs ces dernières années car elles permettent une meilleure compréhension du graphe. Elles partent toujours d'un ensemble de structures prédéfinies qui ont été prouvées fréquentes dans les graphes réels. Deux sous classes de cette dernières peuvent être identifiées :

a. Les Méthodes basées sur des techniques de clustering

Les méthodes de cette classe s'appuient sur le fait qu'on ne peut pas comprendre facilement les graphes denses, alors que quelques structures simples sont beaucoup plus faciles à comprendre et souvent très utiles pour analyser le graphe. Elles se basent sur des algorithmes de détection de communautés. La question suivante peut alors se poser : pourquoi ne pas appliquer l'un des nombreux algorithmes de détection de communautés ou de partitionnement de graphes pour compresser le graphe en termes de communautés ? La réponse est que ces algorithmes ne servent pas tout à fait le même objectif que la compression. Généralement, ils détectent de nombreuses communautés sans ordre explicite, de sorte qu'une procédure de sélection des sous-graphes les plus «importants» est toujours nécessaire. En plus de cela, ces méthodes renvoient simplement les communautés découvertes, sans les caractériser (par exemple, clique, étoile) et ne permettent donc pas à l'utilisateur de mieux comprendre les propriétés du graphe.

Parmi les méthodes de détection de communautés les plus importantes dans la littérature nous trouvons :

- (a) **METIS(Karypis and Kumar, 2000)** : est un schéma de partitionnement de graphe multi-niveau basé sur la bisection récursive (dichotomie). Il se compose de trois phases essentielles : 1) une phase de Grossissement qui consiste en le groupement des nœuds (ou d'arêtes) selon un critère pour former un hypergraphe, 2) une phase de partitionnement qui consiste en le partitionnement du graphe en respectant les contraintes du problème et en optimisant une certaine fonction objective, une façon

d'obtenir une k-way partition est de répéter la phase (01) jusqu'à obtention de k sommets, 3) finalement, le partitionnement de l'hypergraphe est projeté successivement vers le niveau suivant et un algorithme de raffinement est appliqué pour optimiser la fonction objective.

- (b) **SPECTRAL(Hespanha, 2004)** : partitionne un graphe en effectuant une classification en k-means sur les k premiers vecteurs propres du graphe d'entrée. L'idée derrière cette classification est que les nœuds avec une connectivité similaire ont des valeurs propres similaires dans les k premiers vecteurs.
- (c) **LOUVAIN(Blondel et al., 2008)** : est une méthode de partitionnement basée sur la modularité pour détecter la structure de communauté hiérarchique. LOUVAIN est une méthode itérative : (i) Chaque nœud est placé dans sa propre communauté. Ensuite, les voisins j de chaque nœud i sont pris en compte et i est déplacé vers la communauté j si le déplacement produit le gain de modularité maximum. Le processus est appliqué à plusieurs reprises jusqu'à ce qu'aucun gain supplémentaire ne soit possible. (ii) Un nouveau graphe est créé dont les super nœuds représentent les communautés et les super arêtes sont pondérées par la somme des poids des liens entre les deux communautés. L'algorithme converge généralement en quelques itérations.
- (d) **SLASHBURN(Kang and Faloutsos, 2011)** : est un algorithme de ré-ordonnancement de nœud initialement développé pour la compression de graphes. Il effectue deux étapes de manière itérative : (i) il supprime les nœuds de haute centralité du graphe (ii) Il réorganise les nœuds de manière à ce que les identificateurs les plus petits soient attribués aux nœuds de degré élevé et les nœuds des composants déconnectés obtiennent les identificateurs les plus grands. Le processus est répété sur le composant connecté le plus large.
- (e) **BIGCLAM(Yang and Leskovec, 2013)** : est une méthode de détection de communautés à chevauchement évolutive. Elle est construite sur le constat que les chevauchements entre les communautés sont étroitement liés. En modélisant explicitement la force d'affiliation de chaque couple nœud-communauté, un facteur latent non négatif est attribué à cette dernière, qui représente le degré d'appartenance à la communauté. Ensuite, la probabilité d'un bord est modélisée en fonction des affiliations des communautés partagées.
- (f) **HYCOM(Araujo et al., 2014)** : est un algorithme qui détecte les communautés à structure hyperbolique. Il se rapproche de la solution optimale en détectant de manière itérative les communautés importantes. L'idée clé est de trouver dans chaque étape une communauté unique qui minimise une fonction objective basée sur le principe MDL en fonction des communautés précédemment détectées.

Nous présentons dans le tableau 2.3 une synthèse entre ces méthodes dans le but de mieux comprendre leur impact sur les méthodes de compression. Nous constatons que le choix de l'algorithme de clustering peut orienter le processus de recherche vers un ou plusieurs types de structures ce qui influence le résultat et les performances de la compression .

	Chevauchement	Clique	Étoile	sous-graphe Bipartie	Chaîne	Structure Hyperbolique	Complexité
Metis	✗	Beaucoup	certaines	certaines	peu	peu	$O(m \cdot k)$
Spectral	✗	Beaucoup	certaines	Beaucoup	peu	peu	$O(n^3)$
Louvain	✗	Beaucoup	certaines	peu	peu	peu	$O(n \log n)$
SlashBurn	✓	Beaucoup	Beaucoup	certaines	peu	peu	$O(t(m + n \log n))$
Bigclam	✓	Beaucoup	certaines	peu	peu	peu	$O(d \cdot n \cdot t)$
Hycom	✓	certaines	Beaucoup	certaines	peu	Beaucoup	$O(k(m + h \log h^2 + hm_h))$
KCBC	✓	Beaucoup	certaines	peu	peu	peu	$O(t(m + n))$

TABLE 2.3 – Tableau comparatif entre les méthodes de clustering avec n = nombre de nœuds, m = nombre d’arêtes, k = nombre de clusters, t = nombre d’itérations, d = degré moyen de nœuds, $h(m_h)$ = nombre de nœuds (arêtes) dans la structure hyperbolique.

Une première technique de compression usant de ces méthodes s’intitule Vocabulary Graph (VOG) (Koutra et al., 2015). C’est une méthode de base sur laquelle s’appuient toutes les autres méthodes de cette classe. Elle permet de compresser un graphe statique non orienté G à l’aide d’un vocabulaire de sous-structures qui apparaissent fréquemment dans les graphes réels et ayant une signification sémantique, tout en minimisant le coût du codage en utilisant le principe de longueur de description minimale (MDL)¹³. Le vocabulaire Ω utilisé est composé de six structures qui sont : clique (fc) et quasi-clique (nc), noyau bipartie (cb) et quasi-noyau bipartie (nb), étoile (st) et chaîne (ch). On peut avoir un chevauchement au niveau des nœuds, les liens quand à eux sont pris selon un ordre FIFO et ne peuvent pas se chevaucher, i.e, la première structure $s \in M$ qui décrit l’arête dans A détermine sa valeur.

On note par C_x l’ensemble de tous les sous-graphes possibles de type $x \in \Omega$, et C l’union de tous ces ensembles, $C = \cup_x C_x$. La famille de modèles noté \mathcal{M} représente toutes les permutations possibles des éléments de C . Par MDL, on cherche $M \in \mathcal{M}$ qui minimise le mieux le coût de stockage du modèle et de la matrice d’adjacence. En d’autre terme, VOG formule le problème de compression comme un problème d’optimisation dont la fonction objective est :

$$\min(D, M) = L(M) + L(E) \text{ avec } E = A \oplus M \text{ représentant l’erreur.}$$

Pour l’encodage du modèle, on a pour chaque $M \in \mathcal{M}$:

$$L(M) = L_{\mathbb{N}}(|M|+1) + \log \binom{|M| + |\Omega| - 1}{|\Omega| - 1} + \sum_{s \in M} (-\log \Pr(x(s)|M) + L(s))$$

Le premier terme représente le nombre de structures dans le modèle, le second terme encode le nombre de structures par type $x \in \Omega$ tant dis que le troisième terme permet pour chaque structure $s \in M$, d’encoder son type $x(s)$ avec un code de préfixe optimal et d’encoder sa structure. Le codage des structures se fait selon leurs types :

13. MDL est un concept de la théorie de l’information permettant de trouver le modèle ayant une longueur minimale : $\min(D, M) = L(M) + L(D | M)$ où $L(M)$ est la longueur du modèle et $L(D | M)$ est la longueur en bits de la description des données en utilisant le modèle M .

Clique : Pour l'encodage d'une clique, on calcule le nombre de nœuds de celle-ci, et on encode leurs ids : $L(fc) = L_{\mathbb{N}}(|fc|) + \log \binom{n}{|fc|}$

Quasi-Clique : Les quasi-cliques sont encodées comme des cliques complètes, tout en identifiant les arêtes ajoutées dont le nombre est $\|nc\|$ et manquantes dont le nombre est noté $\|nc\|'$ en utilisant des codes de préfixe optimaux : $L(nc) = L_{\mathbb{N}}(|nc|) + \log \binom{n}{|nc|} + \log(|nc|) + \|nc\|l_1 + \|nc\|'l_0$ où $l_1 = -\log(\|nc\|/(\|nc\|+\|nc\|'))$ et analogue à l_0 sont les longueurs des codes de préfixe optimaux des arêtes ajoutées et manquantes.

Noyau biparti : notant par A et B les deux ensembles du noyau bipartie. On encode leurs tailles, ainsi que les identifiants de leurs sommets : $L(fb) = L_{\mathbb{N}}(|A|) + L_{\mathbb{N}}(|B|) + \log \binom{n}{|A|} + \log \binom{n-|A|}{|B|}$.

Quasi-Noyau biparti : Comme les quasi-cliques, les quasi-noyaux bipartie sont codés comme suit : $L(nb) = L_{\mathbb{N}}(|A|) + L_{\mathbb{N}}(|B|) + \log \binom{n}{|A|} + \log \binom{n-|A|}{|B|} + \log(|area(nb)|) + \|nb\|l_1 + \|nb\|'l_0$.

Étoile : L'étoile est un cas particulier d'un noyau bipartie. D'abord on calcule le nombre de nœuds des extrémités de l'étoile, ensuite on identifie le nœud central parmi les n sommets et les nœuds des extrémités parmi les n-1 restants. $L(st) = L_{\mathbb{N}}(|st|-1) + \log n + \log \binom{n-1}{|st|-1}$.

Chaîne : On calcule d'abord le nombre d'éléments de la chaîne, ensuite on encode les identifiants des nœuds selon leurs ordre dans la chaîne : $L(ch) = L_{\mathbb{N}}(|ch| - 1) + \sum_{i=0}^{|ch|} (n - i)$

Matrice d'erreur : la matrice d'erreur E est encodée en deux parties E^+ et E^- . E^+ correspond à la partie de A que M modélise en rajoutant des liens non existants contrairement à E^- qui représente les parties de A que M ne modélise pas. Notons que les quasi-cliques et les quasi-noyaux bipartie ne sont pas inclut dans la matrice d'erreur puisqu'ils sont encodés avec exactitude. Le codage de E^+ et E^- est similaire à celui des quasi-cliques, on a :

$$\begin{aligned} L(E^+) &= \log(|E^+|) + \|E^+\|l_1 + \|E^+\|'l_0 \\ L(E^-) &= \log(|E^-|) + \|E^-\|l_1 + \|E^-\|'l_0 \end{aligned}$$

Pour la recherche du meilleur modèle $M \in \mathcal{M}$, VOG procède sur trois étapes :

- (a) **Génération des sous-structures :** Dans cette phase, Les méthodes de détection de communautés et de clustering sont utilisées pour décomposer le graphe en sous-graphes pas forcément disjoints. La méthode de décomposition utilisée dans VOG est SlashBurn.
- (b) **Étiquetage des sous-graphes :** L'algorithme cherche pour chaque sous-graphe généré dans l'étape précédente la structure $x \in \Omega$ qui le décrit le mieux, en tolérant un certain seuil d'erreur.

- a **Étiquetage des structures parfaites :** Tout d'abord, le sous-graphe est testé pour une similarité sans erreur par rapport aux structures complètes du vocabulaire :

- si tous les sommets d'un sous graphe d'ordre n ont un degré égale à $n-1$, il s'agit alors d'une clique.
- si tous les sommets ont un degré de 2 sauf deux sommets ayant le degré 1, le sous-graphe est une chaîne.
- si les amplitudes de ses valeurs propres maximales et minimales sont égales, le sous-graphe est un noyau bipartie où les sommet de chaque ensemble du noyau sont identifiés à travers un parcours Breadth First Search (BFS)¹⁴ avec coloration des sommets.
- Quant à l'étoile, elle est considérée comme un cas particulier d'un noyau bipartie, il suffit donc que l'un des ensembles soit composé d'un seul sommet.

b Étiquetage des structures approximatives : Si le sous graphe ne correspond pas à une structure complète, on cherche la structure qui l'approxime le mieux en terme du principe MDL.

Après avoir représenté le sous graphe sous forme d'une structure, on l'ajoute à l'ensemble des structures candidates C , en l'associant à son coût.

(c) **Assemblage du modèle :** Dans cette dernière étape, une sélection d'un ensemble de structures parmi ceux de C est réalisée. Des heuristiques de sélections sont utilisées car le nombre de permutations est très grand ce qui implique des calculs exhaustifs. Les heuristiques permettent d'avoir des résultats approximatifs et rapides, parmi les heuristiques utilisées dans VOG on trouve :

- PLAIN : Cette heuristique retourne toutes les structures candidates, e.i, $M = C$.
- TOP-K : Cette heuristique sélectionne les k meilleurs candidats en fonction de leur gain en bits.
- GREEDY'N FORGET(GNF) : Parcourt structure par structure dans l'ensemble C ordonné selon la qualité (gain en bits), ajoute la structure au modèle tant qu'elle n'augmente pas le coût total de la représentation, sinon l'ignore.

Comme nous l'avons déjà précisé, VOG formule le problème de compression de graphe en tant que problème d'optimisation basé sur la théorie de l'information, l'objectif étant de rechercher les sous structures qui minimisent la longueur de description globale du graphe. Un élément clé de VOG est la méthode de décomposition utilisée qui peut donner en sortie des sous-graphes ayant des nœuds et/ou des arêtes en commun et dont VOG (Koutra et al., 2015) ne suppose que le premiers cas. En partant de ce constat, les auteurs de (Liu et al., 2015) proposent VoG-overlapp, une extension de VOG prenant en compte les chevauchements des structures sous forme d'une étude expérimentale de l'effet de diverses méthodes de décomposition sur la qualité de la compression.

L'idée de base de VoG-overlapp est d'inclure une pénalité pour les chevauchements importants dans la fonction objective ce qui oriente le processus de sélection des structures vers la sortie souhaitée. Elle devient alors :

14. BFS : est un algorithme de parcours en largeur.

$$\min L(G, M) = \min \{L(M) + L(E) + \mathbf{L}(\mathbf{O})\}$$

Le principe de calcul de $L(M)$ et $L(E)$ demeure le même, avec \mathbf{O} une matrice cumulant le nombre de fois que chacune des arêtes a été couverte par le modèle. Le cout du codage de la matrice des chevauchements est donné par la formule (2.1).

$$L(\mathbf{O}) = \log(|\mathbf{O}|) + \|\mathbf{O}\| l_1 + \|\mathbf{O}\|' l_0 + \sum_{o \in \varepsilon(\mathbf{O})} L_N(|o|) \quad (2.1)$$

Où :

- $|\mathbf{O}|$ est le nombre d'arêtes (distinctes) qui se répètent dans le modèles \mathbf{M} .
- $\|\mathbf{O}\|$ et $\|\mathbf{O}\|'$ représentent respectivement le nombre d'arêtes présentes et manquantes dans \mathbf{O} .
- $l_1 = -\log(\frac{\|\mathbf{O}\|}{\|\mathbf{O}\| + \|\mathbf{O}\|'})$, de manière analogue l_0 , sont les longueurs des codes de préfixe optimaux pour les arêtes actuelles et manquantes, respectivement.
- $\varepsilon(\mathbf{O})$ est l'ensemble des entrées non nulles dans la matrice \mathbf{O} .

Durant la même année, Shah et al.(Shah et al., 2015) ont proposé une autre variation de VoG, TimeCrunch, pour le cas des graphes simples (sans boucles) non orientés dynamiques représentés par un ensemble de graphes associés chacun à un timestamp. En d'autres termes, ils considèrent les graphes $G = \bigcup_{t_i} G_{t_i}(V, E_{t_i})$ $1 \leq i \leq t$ où $G_{t_i} = G$ a l'instant t_i . Un nouveau vocabulaire est proposé pour décrire proprement l'évolution des sous-structures dans le temps. En effet, ils partent du même vocabulaire de structures statiques $\Omega = \{st(etoile), fc(clique), nc(quasi - clique), bc(bipartie), nb(quasi - bipartie), ch(chaine)\}$ dont ils affectent une signature temporelle $\delta \in \Delta$ où $\Delta = \{o(oneshot), r(ranged), p(periodique), f(flickering), c(constante)\}$.

Comme les éléments du modèle sont modifiés, son cout est alors aussi modifié pour inclure pour chaque structure s non seulement sa connectivité $c(s)$ correspondant aux arêtes des zones induites par s mais aussi sa présence temporelle $u(s)$ correspondant aux timestamps dans lesquels s apparait dans le graphe G .

$$L(M) = L_N(|M| + 1) + \log\left(\frac{|M| + |\Phi| - 1}{|\Phi| - 1}\right) + \sum_{s \in M} (\log P(v(s)|M) + L(c(s)) + \mathbf{L}(u(s)))$$

Le cout de l'encodage de la présence temporelle diffère selon ses caractéristiques. Nous présenterons dans ce qui suit la formule correspondant à chaque signature.

- **Oneshot** : cette signature décrit les sous-structures qui apparaissent dans un seul timestamp, i.e, $|u(s)| = 1$. Donc le cout de l'encodage se réduit aux nombre de bits nécessaires pour sauvegarder le timestamp : $L(u(s)) = \log(t)$.
- **Ranged** : dans ce cas la sous-structure apparait dans tous les graphes se trouvant entre deux timestamps t_{debut} et t_{fin} . Le cout englobe le nombre de timestamps dans lesquels elle apparait ainsi que les identifiants des deux timestamp t_{debut} et t_{fin} : $L(u(s)) = L_N(|u(s)|) + \log\left(\binom{t}{2}\right)$.

- **Periodic** : cette catégorie est une extension de la précédente d'où : $L(p) = L(r)$. En effet, la périodicité peut être déduite à partir des marqueurs début et de fin ainsi que du nombre de pas de temps $|u(s)|$ entre chaque deux timestamps, permettant ainsi de reconstruire $u(s)$
- **Flickering** : ce type décrit les structures qui apparaissent dans n timestamps de manière aléatoire. Le coût doit englober donc le nombre de timestamps ainsi que leurs identifiants d'où : $L(u(s)) = L_N(|u(s)|) + \log \binom{t}{|u(s)|}$.
- **Constant** : dans ce cas la sous-structure apparait dans tout les timestamps et donc elle ne dépend pas du temps d'où $L(c)=0$.

Nous notons que décrire $u(s)$ est encore un autre problème de sélection de modèle pour lequel les auteurs tirent parti du principe MDL . En effet, juste comme pour le codage de la connectivité, il peut ne pas être précis avec une signature temporelle donnée. Toutefois, toute approximation entraînera des coûts supplémentaires pour l'encodage de l'erreur qui englobent dans ce cas l'erreur de l'encodage de la connectivité ainsi que l'erreur de l'encodage de la signature temporelle.

Algorithm 1 TIMECRUNCH

- 1: **Génération des sous-structures candidates** : Génération de sous-graphes pour chaque G_{t_i} en utilisant un des algorithmes de décomposition de graphes statiques.
 - 2: **Étiquetage des sous-structures candidates** : Associer chaque sous-structure à une étiquette $x \in \Omega$ minimisant son MDL .
 - 3: **Assemblage des sous-structures candidates temporelles** : Assembler les sous-structures des graphes G_{t_i} pour former des structures temporelles avec un comportement de connectivité cohérent et les étiqueter conformément en minimisant le coût de codage de la présence temporelle. Enregistrer le jeu de candidats $C_x \in C$.
 - 4: **Composition du graphe compressé** : Composition du modèle M d'importantes structures temporelles non redondantes qui résument G à l'aide des méthodes heuristiques VANILLA, TOP-10, TOP-100 et STEPWISE. Choisir M associé à l'heuristique qui génère le coût de codage total le plus faible.
-

Une dernière variante, s'intitulant CONditional Diversified Network Summarization (CONDENSE), a été présentée par Liu et al. (Liu et al., 2018b) où ils abordent efficacement trois contraintes principales des méthodes précédentes : (i) leurs dépendance à la méthode d'extraction de motifs (ii) l'incapacité de certaines à gérer les motifs qui se chevauchent (iii) leur dépendance vis-à-vis de l'ordre dans lequel les structures candidates sont considérées lors de la phase d'assemblage. En effet, pour résoudre le premier problème, ils combinent plusieurs méthodes d'extraction de motifs ce qui améliore la qualité des structures candidates en dépit du temps d'exécution. Tant dis que pour répondre à la deuxième contrainte ils utilisent la fonction objective proposée dans (Liu et al., 2015). Arrivant à la dernière phase de l'algorithme, ils proposent quatre nouvelles heuristiques : (1) STEP : choisie les K meilleures structures, (2) STEP-P : partitionne le graphe et affecte chaque motif à la partition ayant un chevauchement maximal de nœuds avec lui. Ces partitions sont parcourues parallèlement pour ne prendre que la meilleure de toutes les structures dans chacune

des partitions, (3) STEP-PA : amélioration de STEP-P en désignant chaque partition du graphe comme étant active, puis si une partition échoue x fois pour trouver une structure qui réduit le coût MDL, cette partition est déclarée inactive et n'est pas visitée dans les prochaines itérations, (4) K-STEP : combinaison des trois premières heuristiques. Il transforme par la suite chaque motif trouvé en un super-nœud.

b. Les méthodes basées sur les propriétés de la matrice d'adjacence

Les graphes peuvent avoir différentes représentations. Chacune des structures de données présente des avantages et des inconvénients en ce qui concerne la quantité de mémoire nécessaire pour stocker les données et la facilité d'accès aux données. Selon les besoins, il est parfois utile de stocker les données dans des structures de données plus grandes, qui nécessitent plus d'espace mais offrent un accès efficace aux données. En se basant sur ce constat plusieurs méthodes ont été proposées dans la littérature pour compresser la matrice d'adjacence en exploitant les propriétés des graphes réels pour trouver les motifs les plus fréquents dans cette dernière.

(Asano et al., 2008) ont exploité les propriétés du graphe du web pour présenter une nouvelle méthode de compression, appelée Efficient Compression of web graph (ECWG), sans perte permettant d'extraire les motifs à partir de la matrice d'adjacence. Ils proposent un vocabulaire composé de six types de blocs (Motifs) : un bloc horizontal de 1, un bloc vertical de 1, un bloc diagonal de 1, un rectangle de 1, un bloc de 1 sous forme de L et le singleton 1. Avant de procéder à l'extraction des motifs, la liste d'adjacence du graphe est partitionnée selon les domaines (ex : esi.dz, usthb.dz, ...). Une nouvelle matrice d'adjacence est donc construite pour chaque hôte(domaine) contenant les liens existants entre ses pages, auxquels les liens inter-hôtes sont concaténés. Les blocs B sont détectés par la suite et chacun est représenté par un quadruplets (i, j, type(B), dim(B)) où i,j représentent les coordonnées du premier élément du bloc dans la matrice d'adjacence de l'hôte, type(B) représente le type du bloc et dim(B) représente les dimensions du bloc (ce champ est omi dans le cas du singleton).

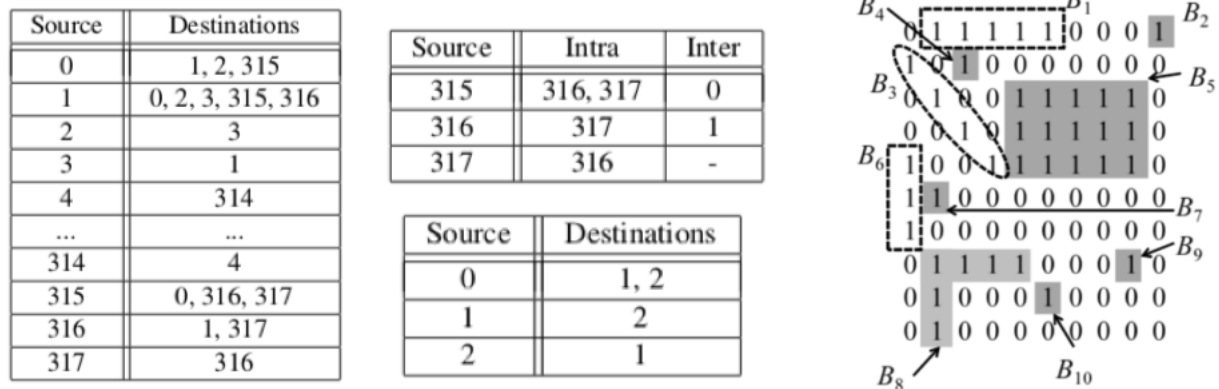


FIGURE 2.19 – Exemple illustrant le principe de fonctionnement (Asano et al., 2008)

Dans une méthode toute récente s'intitulant Graph Compression Using Pattern Matching (GCUPMT), Shah et Rushabh (Shah, 2018) partitionnent les lignes de la matrice d'adjacence en plusieurs blocs ayant la même taille des motifs qui sont dans ce cas sous forme de vecteurs prédéfinies. Les blocs sont comparés avec l'ensemble des motifs ce qui entraîne, en cas de correspondance, le remplacement du bloc par un indicateur du motif précédé par un 1 indiquant que les bits suivants appartiennent à un indicateur de motif. Dans le cas contraire, les données brutes sont stockées directement précédées par un 0.

2.4.2 Compression basée sur l'agrégation des motifs

Les méthodes de compression par extraction de motifs basées sur l'agrégation sont des méthodes qui agrègent plusieurs nœuds ou liens d'un motif en un seul nœud, appelé super-nœud. Le graphe en sortie, dit super-graphe, devient dès lors plus simple et moins complexe offrant ainsi une aisance et une facilité de traitement, d'exploration et de visualisation. Nous présenterons dans ce qui suit les deux sous-classes de cette classe qui se distinguent selon que l'agrégation concerne les nœuds ou les liens.

a. Les méthodes de compression basées l'agrégation de nœuds

Les techniques de compression basées sur l'agrégation des nœuds des motifs sont des méthodes qui ont existé depuis plusieurs décennies offrant plusieurs avantages. Elles visent à résumer le graphe initial en agrégeant les nœuds des motifs découvert dans le but de diminuer le nombre de nœuds existants et d'offrir une meilleure visibilité et analyse du graphe.

Une première méthode de cette classe s'intitule Subdue (Ketkar et al., 2005). Elle effectue une recherche *Branch&Bound* qui commence à partir des sous-structures composées de tous les sommets avec des étiquettes uniques. Les sous-structures sont prolongées de toutes les manières possibles par un sommet et une arête ou par une arête afin de générer des sous-structures candidates. Subdue conserve les instances des sous-structures et utilise l'isomorphisme de graphe pour déterminer les instances de la sous-structure candidate. Les sous-structures sont ensuite évaluées en fonction de leur compression de la longueur de description (DL) du jeu de données. Cette procédure se répète jusqu'à ce que toutes les sous-structures soient prises en compte ou que les contraintes imposées par l'utilisateur ne soient plus vérifiées. A la fin de la procédure, Subdue indique les meilleures sous-structures de compression. Le système Subdue fournit également la possibilité d'utiliser la meilleure sous-structure trouvée lors d'une étape de découverte pour compresser le graphe d'entrée en remplaçant les instances de la sous-structure par un seul sommet et en effectuant le processus de découverte sur le compressé. Cette fonctionnalité génère une description hiérarchique du jeu de données du graphe à différents niveaux d'abstraction en termes de sous-structures découvertes.

Dans (Rossi and Zhou, 2018), les auteurs partent de l'observation que les graphes réels sont formés souvent de nombreuses cliques de grande taille. En utilisant ceci comme base, GraphZip décompose le graphe en un ensemble de grandes cliques, qui est ensuite utilisé pour compresser et représenter le graphe de manière succincte.

b. Les méthodes de compression basées sur l'agrégation de liens

Les méthodes de compression par extraction de motifs basées sur l'agrégation de liens sont parmi les méthodes les plus populaires. Leur objectif est de produire un graphe compressé à partir du graphe initial en remplaçant les liens denses du graphe par un nouveau

super-nœud. Elles se divisent selon le principe en deux grandes classes : celles utilisant les règles de grammaire et celles utilisant des heuristiques de clustering. Nous détaillerons dans ce qui suit ces deux classes.

b.1. Les méthodes de compression basées sur les règles de grammaire La classe des méthodes de compression basées sur les règles de grammaire est une généralisation d'une méthode de compression des dictionnaires s'intitulant *Re-pair*. Son principe de base consiste en la recherche, à chaque itération, de la paire de symboles la plus fréquente dans une séquence de caractères et de la remplacer par un nouveau symbole, jusqu'à ce qu'il ne soit plus commode de les remplacer. Nous notons que dans ce cas le motif est sous forme de deux arêtes ayant un sommet en commun, nommé *digraph*.

Une première méthode suivant ce principe a été proposée dans (Claude and Navarro, 2010b) baptisée *Approximate Re-pair*. Dans cette méthode un graphe $G=(V,E)$ est représenté sous forme d'une séquence de caractères T :

$$T=T(G)= \overline{v_1} v_{1,1} \dots v_{1,a} \overline{v_2} v_{2,1} \dots v_{2,a_2} \dots \overline{v_n} v_{n,1} \dots v_{n,a_n}$$

où $\overline{v_i}$ représente l'identificateur du sommet v_i . Elle procède en trois étapes essentielles expliquer dans l'algorithme 2 . Lorsqu'il n'y a plus de paires à remplacer,

Algorithm 2 Approximate Re-pair

- 1: **Calcule des fréquences** : T est parcourue séquentiellement et chaque pair $t_i t_{i+1}$ est ajoutée à un tableau de hachage H avec leur nombre d'occurrences.
 - 2: **Recherche des k meilleurs paires** : H est parcourue et les k paires les plus fréquentes sont retenues, en utilisant k pointeurs vers les cellules de H .
 - 3: **Le remplacement simultané** : les k paires identifiées dans l'étape précédente sont simultanément remplacées par un nouveau identifiant et une règle de production est ajoutée.
-

Approximate Re-pair s'arrête donnant comme résultat une représentation compacte C de la chaîne T . Pour finaliser le processus, tous les indicateurs de nœuds $\overline{v_i}$ seront supprimés de C . De plus, l'algorithme crée une table qui contiendra des pointeurs vers le début de la liste d'adjacence de chaque nœud dans C . Grâce à cette table l'algorithme pourra répondre aux requêtes de recherche de successeurs en un temps optimal.

Dans un travail ultérieur (Claude and Navarro, 2010a) , les même auteurs s'intéressent aux requêtes de recherche des nœuds prédécesseurs et successeurs à partir du graphe compressé de *Approximate Re-pair* directement. Il proposent alors de combiner leur méthode avec une représentation basée sur les relations binaires de (Barbay et al., 2006). En effet, ce dernier consiste à représenter les listes d'adjacences à l'aide d'une représentation séquentielle permettant de rechercher les occurrences d'un symbole puis de rechercher les voisins inverses à l'aide de cette primitive.

Claude et Ladra (Claude and Ladra, 2011) partageaient les mêmes préoccupations des auteurs de la méthode précédente et ont proposé comme solution de combiner la méthode *Re-pair* avec la représentation *k2-tree*. Ils obtiennent alors une compression

de 2,27 (pbe) sur le graphe UK2002, tout en conservant la possibilité d'interroger les voisins entrants et sortants (Maneth and Peternek, 2015).

Une dernière méthode de cette classe s'instituant gRepair a été proposée dans (Maneth and Peternek, 2018). Ce nouveau algorithme de compression détecte de manière récursive des sous-structures répétées et les représente via des règles de grammaire. Des requêtes spécifiques telles que l'accessibilité entre deux nœuds ou des requêtes de chemin normal peuvent ainsi être évaluées en temps linéaire (ou en temps quadratique, respectivement), sur la grammaire, permettant ainsi des accélérations proportionnelles au taux de compression. La figure 2.20 présente le résultat de cette méthode sur un exemple.

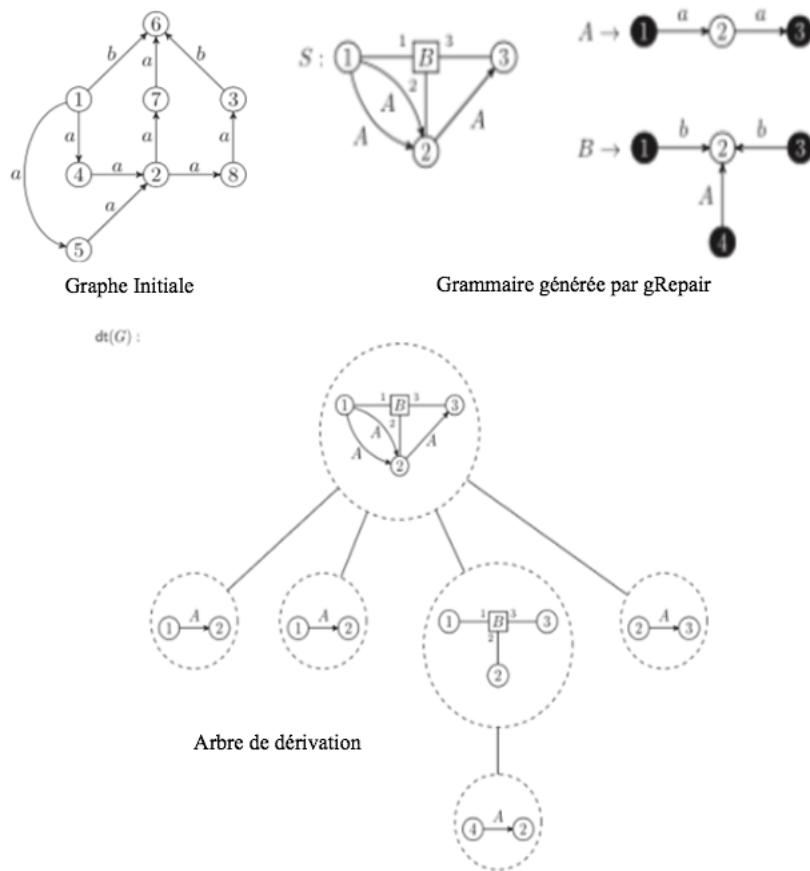


FIGURE 2.20 – Exemple d'exécution de gRepair sur G (Maneth and Peternek, 2018).

b.2. Les méthodes de compression basées sur des heuristiques de clustering

Les méthodes de compression appartenant à cette classe sont des méthodes basées sur la recherche des sous-graphes denses (ayant des nœuds fortement connectés). Elles sont destinées principalement aux graphes du Web et les graphes des réseaux sociaux dans le but de faciliter leur exploration et analyse.

(Buehrer and Chellapilla, 2008) ont exploité l'existence de plusieurs ensembles de pages web qui ont les mêmes liens sortants. S'intitulant Virtual Node Miner (VNM), leur approche est basée sur la réduction du nombre de liens en créant des nouveaux

sommets virtuels qui sont ajoutés au graphe. Soit $G = (V, E)$ un graphe orienté, l'algorithme proposé se compose de deux phases essentielles :

i. **Phase de Clustering :**

Le but de cette première étape est de contourner la tâche presque impossible d'extraction simultanée de centaines de millions de points de données en groupant d'abord les sommets similaires dans le graphes dans des clusters. Pour cela k fonctions de hachage indépendantes sont utilisées pour obtenir une matrice de taille $V * k$. Par la suite, les lignes de la matrice sont triées lexicographiquement et elle est parcourue colonne par colonne en regroupant les lignes ayant la même valeur. Lorsque le nombre total de lignes chute au-dessous d'un seuil ou que le bord de la matrice de hachage est atteint, les identifiants des sommets associés aux lignes sont renvoyés au processus d'extraction (Phase 02).

- ii. **Phase d'Extraction de Motifs :** Le but de cette étape est de localiser des sous-ensembles communs de liens sortants dans les sommets donnés. Ainsi les ensembles plus grands et fréquents présentent un intérêt, car ils peuvent représenter des motifs plus pertinents et une meilleure compression. En effet, les performances de compression d'un motif sont calculés en fonction de sa fréquence dans la liste d'adjacence, et de sa taille qui est le nombre de liens qu'il contient (Formule 2.2).

$$Compression(P) = (P.frquence - 1)(P.taille - 1) - 1 \quad (2.2)$$

Afin d'extraire ces motifs, VNM utilise une heuristique gloutonne. Cette heuristique procède comme suit :

- A. Extraire un histogramme des identifiants de liaison sortante à partir de la liste d'adjacence des sommets données.
- B. Les listes sont réorganisées dans l'ordre décroissant des fréquences des liens sortants en éliminant ceux qui apparaissent une seul fois uniquement.
- C. Chaque lien sortant est ajouté à un arbre de préfixes avec l'ensemble trié de ces extrémités initiales selon leurs identifiants.
- D. L'arbre est par la suite parcouru afin d'identifier les motifs qui maximisent la formule de performance de la compression (Formule 2.2). Ces motifs sont ensuite convertis en nœuds virtuels et les identificateurs de sommets de leurs listes sont remplacés par les ids des nœuds virtuels dans la liste d'adjacence.

L'algorithme est appliqué jusqu'à ce que la réduction n'apporte pas un gain significatif. La figure 2.21 illustre le principe de fonctionnement de cette méthode sur un exemple.

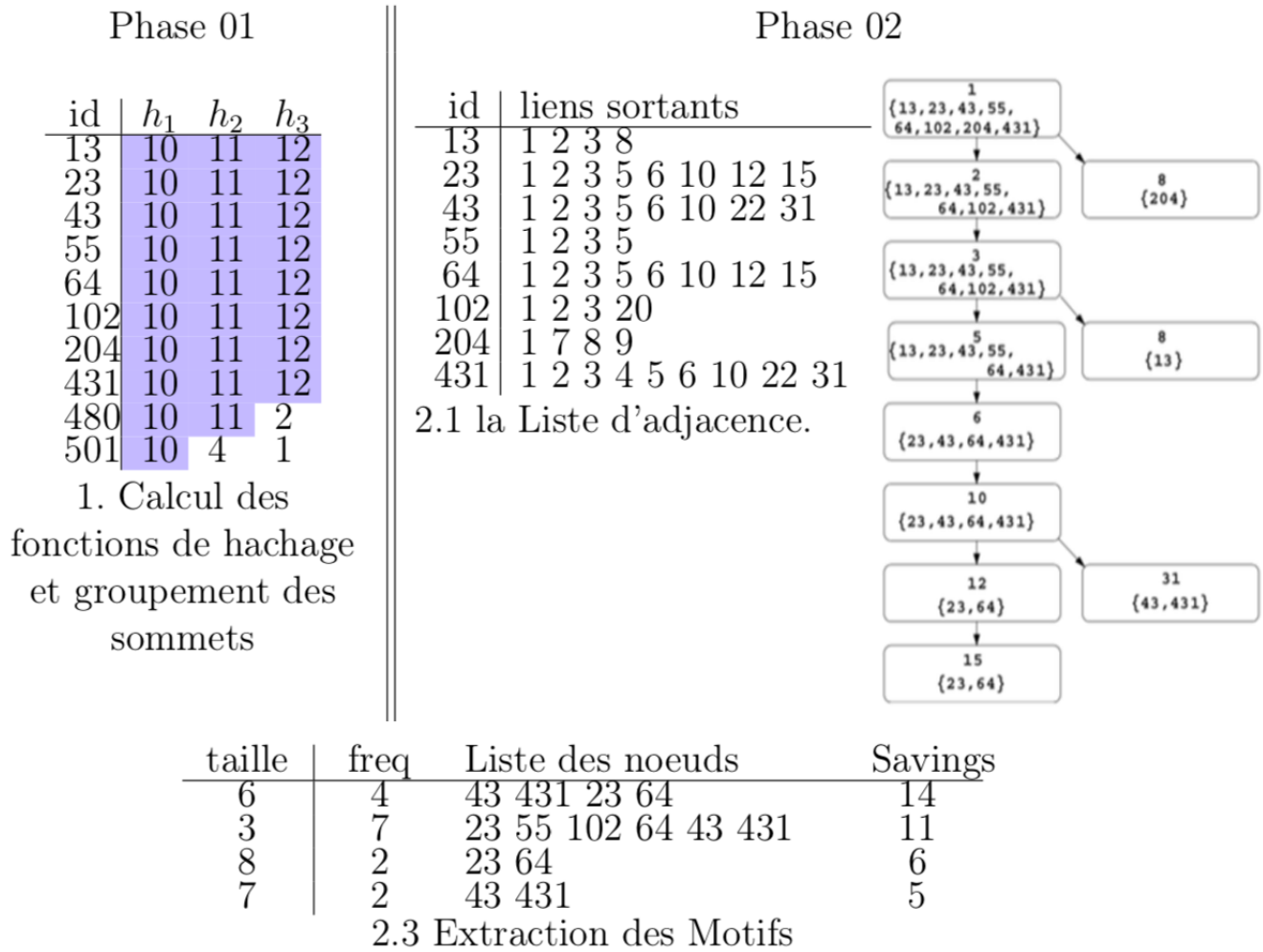


FIGURE 2.21 – Exemple d'exécution de VNM (Buehrer and Chellapilla, 2008).

Une variante de VNM, Dense SubGraph Mining (DSM), a été proposée par Hernandez et Navarro (Hernández and Navarro, 2014). Comme première contribution, ils augmentent les types de structures découvertes dans la phase de clustering pour englober aussi : les cliques, les bi-cliques. L'extraction de motifs cette fois-ci n'est pas basée sur un parcours des feuilles vers la racine mais l'inverse où l'ensemble des sommets finaux des liens du motifs est constitué des étiquettes des nœuds de l'arbre inclus dans le chemin de la racine vers la feuille et les sommets initiaux sont la liste des sommets inclus dans le nœud feuille. Leur deuxième contribution consiste en une hybridation dans le but de représenter le graphe en sortie à l'aide de structures compactes. Une première approche proposée est d'utiliser les k2-trees (Brisaboa et al., 2009) qui donnent la représentation la plus compacte. La deuxième hybridation consiste en une nouvelle structure proposée par les auteurs.

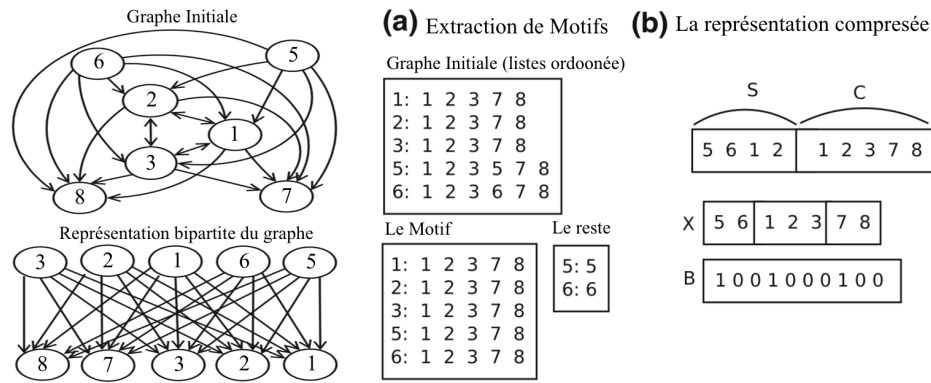


FIGURE 2.22 – Exemple d'exécution de DSM (Hernández and Navarro, 2014).

2.4.3 Synthèse des méthodes de compression par extraction de motifs

Durant les sections précédentes nous avons expliqué de manière générale les fondements de base de chaque sous-classe de la classe des méthodes de compression par extraction de motifs ainsi que le principe de fonctionnement de leurs méthodes.

Nous regroupant dans le tableau 2.4 les différentes caractéristiques des méthodes de ces sous-classes. Nous observons qu'elles sont toutes sans perte et destinées aux graphes statiques, à l'exception de la classe des méthodes de compression Basées vocabulaire en utilisant les méthodes de clustering qui contiennent des méthodes de compression supportant les graphes dynamiques. On remarque aussi que la majorité des méthodes de compression par extraction de motifs fournissent en sortie une représentation succincte. Le tableau 2.4 contient aussi un résumé sur les résultats de l'application de ces méthodes que nous représentons dans les figures 2.23 et 2.24.

La figure 2.23 illustre les résultats des différentes méthodes de compression basées sur l'agrégation des liens des motifs sous forme de nombre de bits par lien sur le même graphe uk-2002. Elle montre que la sous-classe des méthodes utilisant les techniques de clustering donnent des résultats nettement meilleurs que la sous-classe des méthodes utilisant les règles de grammaire.

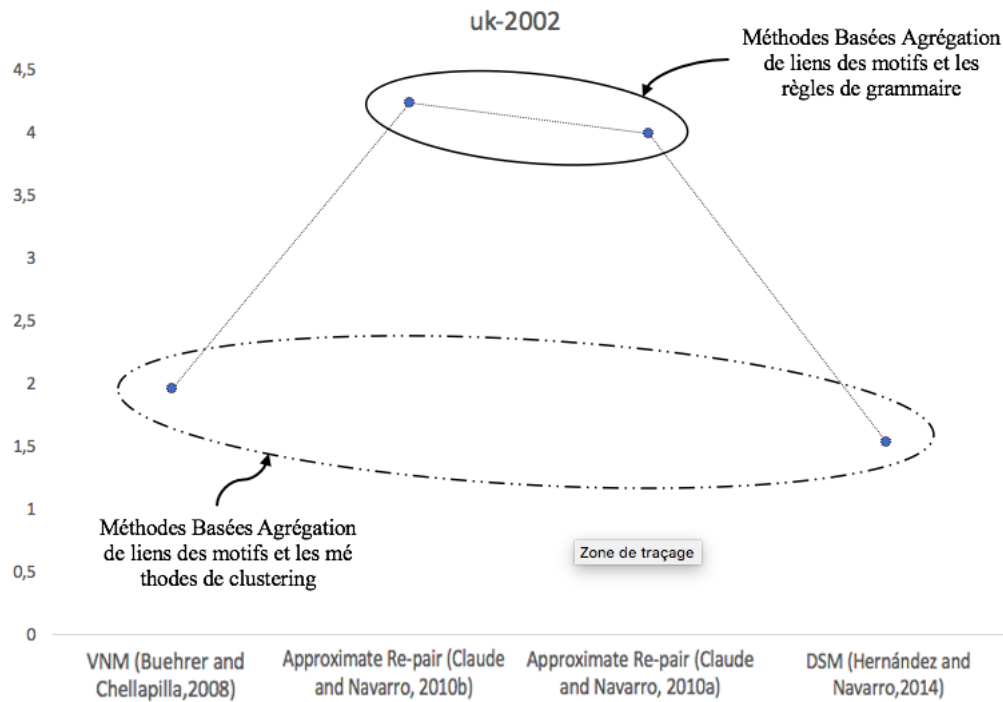


FIGURE 2.23 – Comparaison des performances des méthodes de compression basées sur l'agrégation des liens des motifs

Tant dis que la figure 2.24 montre les résultats des taux de compression de la sous-classe des méthodes de compression basées vocabulaire sur deux graphes Enron et uk-2002. Nous remarquons que les résultats sont presque équivalents pour toutes les méthodes et sont autour de 75%.

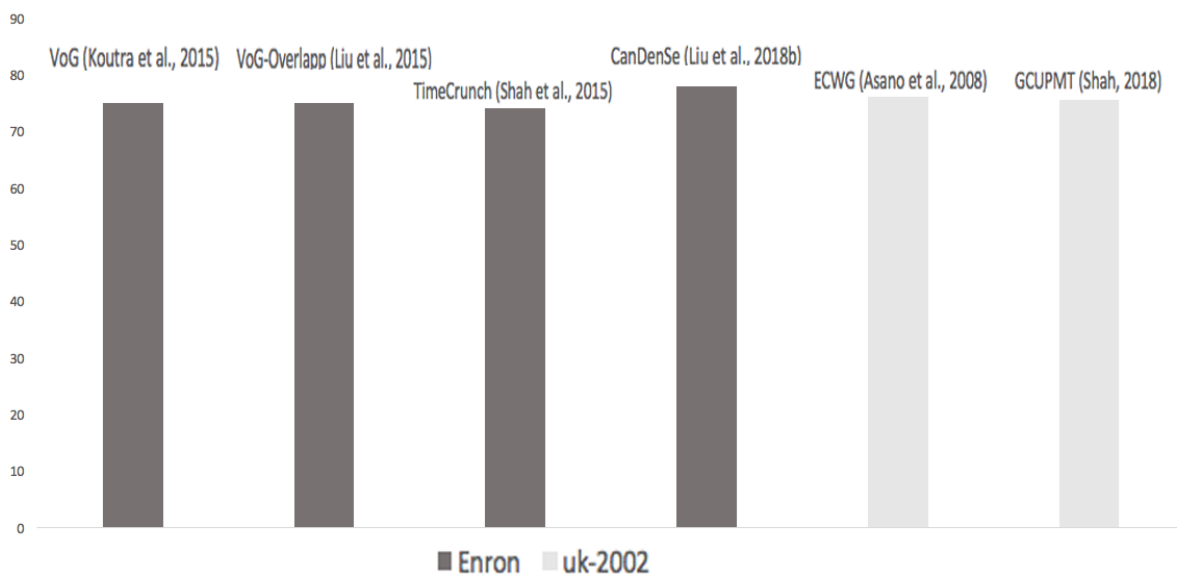


FIGURE 2.24 – Comparaison des performances des méthodes de compression basées vocabulaire

Classe	Méthode	Graphe en entrée				Compression		Structure en sortie		Graphe de test	Résultat
		Orienté	Non orienté	Statique	Dynamique	Avec perte	Sans perte	Succincte	Structurale		
Basée vocabulaire en utilisant les méthodes de clustering	VoG (Koutra et al., 2015)	×	✓	✓	×	×	✓	✓	×	Enron	75%
	VoG-Overlapp (Liu et al., 2015)	×	✓	✓	×	×	✓	✓	×	Enron	75%
	TimeCrunch (Shah et al., 2015)	×	×	✓	✓	×	✓	✓	×	Enron	74%
	CanDenSe (Liu et al., 2018b)	×	✓	×	✓	×	✓	✓	✓	Enron	78%
Basée vocabulaire en utilisant les propriétés de la matrice d'adjacence	ECWG (Asano et al., 2008)	✓	✓	✓	×	×	✓	✓	×	uk-2002	76.1%
	GCUPMT (Shah, 2018)	✓	✓	✓	×	×	✓	✓	×	graphes 8192 nœuds	70%
Basée Agrégation de nœuds des motifs	Subdue (Ketkar et al., 2005)	×	✓	✓	×	×	✓	×	✓	Graphe des composantes chimiques	16%
	GraphZip (Rossi and Zhou, 2018)	×	✓	✓	×	×	✓	×	✓	Web-Google	19%
Basée Agrégation de liens des motifs en utilisant les règles de grammair	Approximate Re-pair (Claude and Navarro, 2010b)	✓	×	✓	×	×	✓	×	✓	uk-2002	4.23
	Approximate Re-pair (Claude and Navarro, 2010a)	✓	×	✓	×	×	✓	×	✓	uk-2002	3.98 bpe
	gRe-pair (Maneth and Peternek, 2018)	✓	✓	✓	×	×	✓	×	✓	NotreDame	4.84 bpe
	VNM (Buehrer and Chellapilla, 2008)	✓	×	✓	×	×	✓	×	✓	uk-2002	1.95 bpe
Basée Agrégation de liens des motifs et les méthodes de clustering	DSM (Hernández and Navarro, 2014)	✓	×	✓	×	×	✓	✓	✓	uk-2002	1.53 bpe

TABLE 2.4 – Synthèse des méthodes de compression par extraction de motifs.

2.5 Bilan général

De nos jours, les graphes sont omniprésents. Cependant, leur taille présente un obstacle presque insurmontable à la compréhension du caractère essentiel des données. D'où la nécessité de la compression qui permet de réduire la taille des graphes tout en gardant le caractère utile de l'information incluse dans ces derniers.

Nous avons étudié dans ce chapitre différentes méthodes de compression de graphes dans le but d'établir une classification des méthodes basée sur deux approches : l'extraction de motifs et les arbres k2-trees. Nous nous sommes appuyés pour cela sur le principe de fonctionnement de chacune d'elles. Nous proposons six classes de méthodes :

- i Les méthodes de compression par les k2-trees
- ii Les méthodes de compression par extraction de motifs basées vocabulaires en utilisant des méthodes de clustering.
- iii Les méthodes de compression par extraction de motifs basées vocabulaires exploitant les propriétés de la matrice d'adjacence.
- iv Les méthodes de compression par extraction de motifs basées agrégation de nœuds.
- v Les méthodes de compression par extraction de motifs basées agrégation de liens en utilisant des règles de grammaire.
- vi Les méthodes de compression par extraction de motifs basées agrégation de liens en utilisant des heuristiques de clustering.

Nous avons présenté le principe de fonctionnement des méthodes relatif à chaque classe tout en comparant leurs caractéristiques. Ces classes diffèrent les unes des autres dans leurs fondements théoriques, le type de compression considéré, la complexité des algorithmes, les objectifs et les domaines d'application. Dans le tableau A.1 , nous allons essayer de synthétiser les principales différences et similitudes entre les deux approches, compression par extraction de motifs et compression par les arbres k2-trees, que nous avons pu constater à travers notre recherche bibliographique. Ils ont un fort impact sur le choix de la méthode de compression du moment qu'ils ont une influence directe sur les performances.

	k^2 -trees	Extraction de motifs
Type de compression	toujours sans perte	toujours sans perte
Structure en sortie	Toujours succincte	Peut être succincte où structurelle où les deux en même temps
technique utilisée	exploitation de la matrice d'adjacence du graphe	exploitation des motifs fréquents dans le graphe
Dépendance	Dépend du paramètre k	Dépend selon la méthode de l'algorithme de clustering ou du vocabulaire de motifs utilisé
Objectif	Compression, Réduire l'espace de stockage et le temps de parcours	- Compression, - Réduire l'espace de stockage et le temps de parcours, - Extraire les informations pertinentes, - Visualisation
Domaine d'application	tous les domaines	tous les domaines
Type de graphes supporté	- Statique orienté, - Attribué, - Dynamique,	- Statique (orienté et non orienté), - étiqueté, - Dynamique,

TABLE 2.5 – Comparaison entre les méthodes basées sur k^2 -trees et basées sur l'extraction de motifs.

Après avoir comparer les caractéristiques des deux classes de méthodes de compression, nous sommes intéressées à comparer les résultats expérimentales obtenues dans leur différents travaux et utilisant les même graphes de test. À ce stade, notre but est d'établir une étude comparative plus approfondie sur les performances des deux classes. Les deux courbes de la figure 2.25 présentent le nombre de bits par liens obtenues sur deux graphes de test uk-2002 et eu-2005 (Nous fournissons plus de détails sur ses deux graphes en annexe A).

D'après la figure, on constate que les deux courbes atteignent leur minimum dans le cas de la méthode DSM (Hernández and Navarro, 2014) avec 2,11 (bpe) pour eu-2005 et 1,53 (bpe) pour uk-2002. Cette méthode donne ainsi le meilleur résultat dans les deux classes. Nous remarquons aussi que les méthodes basées vocabulaire en utilisant les propriétés de la matrice d'adjacence et la méthode de compression VNM (Buehrer and Chellapilla, 2008) s'inscrivant dans les méthodes de compression basées agrégation des liens des motifs usant des méthodes de clustering donnent des résultats presque équivalent aux deux méthodes de compression par les arbres k^2 -trees : k^2 -trees : Amélioration (Brisaboa et al., 2014b) et Delta- k^2 -trees (Zhang

et al., 2014b). D'autre part, la première méthode (Claude and Navarro, 2010b) des méthodes de compression par extraction de motifs basées agrégation de liens des motifs en utilisant des règles de grammaire donne le même nombre de bits que l'algorithme de base k2-trees (Brisaboa et al., 2009). Tant dis que la deuxième méthode (Claude and Navarro, 2010a) de cette même classe améliore de manière significative le résultat donnant ainsi de meilleurs performances.

Nous avons aussi constaté que malgré le nombre considérable de travaux dans le domaine de la compression des graphes, il y a encore de nombreux problèmes ouverts importants sur le terrain. Nous mentionnons ci-dessous quelques un des plus remarquables.

Le premier est l'aspect récursive des méthodes de construction des arbres k2-trees qui est un obstacle envers les performances (complexité temporelle et spatiale) de cette classe de méthodes. Certes l'utilisation de l'approche récursive permet souvent d'avoir un programme naturel, plus simple et plus facile à comprendre mais le nombre important d'appels imbriqués générés dans le cas des graphes du monde réel augmente de manière considérable le temps d'exécution et la mémoire nécessaire pour la construction du graphe compressé.

Le deuxième aspect est la forte influence des méthodes de clustering sur les performances de la compression. Certes, CONDENSE (Liu et al., 2018b) améliore le taux de compression mais elle détériore le temps d'exécution .

Un autre aspect à considérer est l'aspect temporel des graphes. En effet, tant dis que plusieurs méthodes de compression par les arbres k2-trees existent pour ce type de graphes, nous n'avons trouver que quelques méthodes dans le cas des méthodes par extraction de motifs basées vocabulaire et aucune méthodes dans la sous-classe des méthodes agrégatives malgré qu'elles offrent un taux de compression compétitif dans le cas statique. De plus, il existe peu de travaux sur la compression par les arbres k2-trees des graphes temporels attribués et aucun travail, à notre connaissance, sur la compression par extraction de motifs pour ce type de graphe. Cependant, de nombreux réseaux du monde réel, tels que les réseaux sociaux où le suivie des communautés (ou motifs) est très utile, peuvent facilement être modélisés sous la forme de graphes attribués temporels. Nous avons même remarqué que dans le cas de certains types de graphes statiques, tel que les graphes multi-vue, aucun schéma de compression n'a été proposé. Une solution dans le cas de l'exemple précédent serait peut être de généraliser le schéma de compression de VoG. Cette piste de recherche est motivé par l'apparition de nouvelles méthodes de clustering, comme celle proposée (Wang et al., 2019), adaptées pour les graphes multi-vue .

Finalement, nous pensons qu'une nouvelle direction prometteuse est l'utilisation des méthodes d'apprentissage dans le processus d'extraction de motifs des méthodes de compression. Le caractère prometteur de cette direction est due aux succès apporté par le domaine d'apprentissage ces dernières années, notamment le *Deep Learning*.

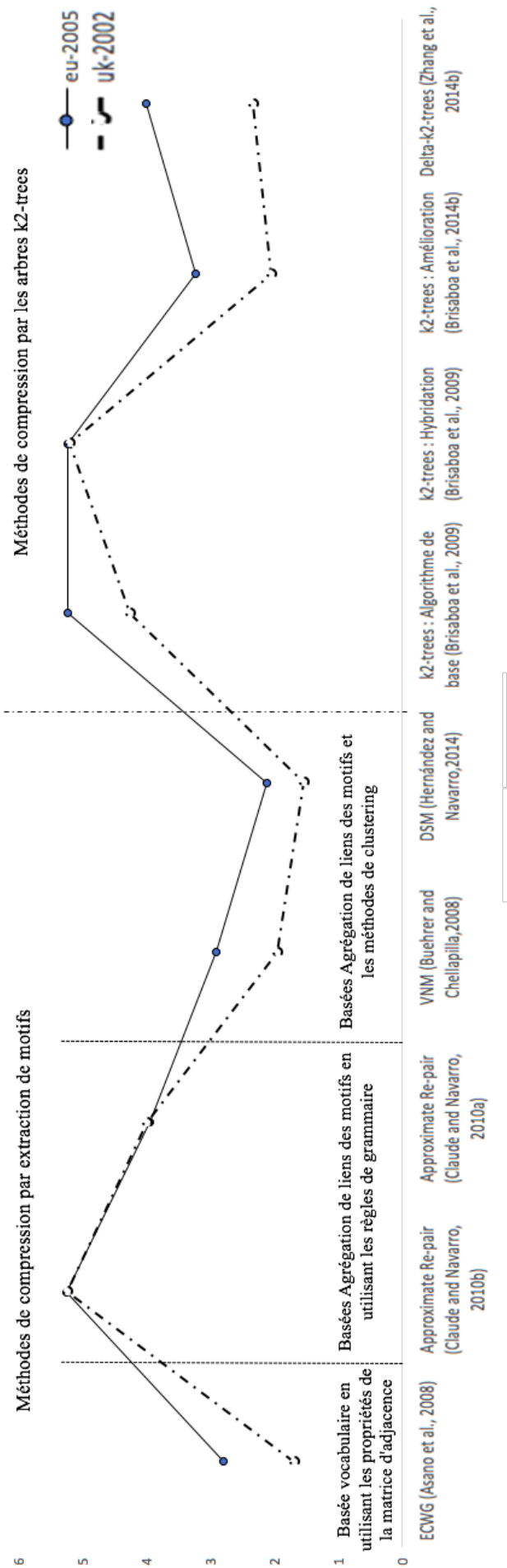


FIGURE 2.25 – Comparaison des performances des méthodes par extraction de motifs et les méthodes basées sur les arbres k^2 -trees

Conclusion générale

À l'ère du Big Data, les informations sont générées rapidement et leur analyse est devenue un défi. Diverses méthodes ont été proposées pour accélérer l'analyse de ces données et en extraire les informations. En effet, la situation d'abondance d'information aujourd'hui a favorisé l'incapacité d'assimilation et de réemploi de l'information et a montré l'inadaptation des outils de gestion classiques. Une des solutions les plus connus pour pallier à ces problèmes est la compression. Comme les graphes représentent la plupart des données du monde réel, les méthodes de compression des graphes ont été à la pointe de ces efforts.

Nous nous sommes intéressées dans ce travail au domaine de compression des graphes, et plus particulièrement aux méthodes de compression par les k2-trees et les méthodes de compression par extraction de motifs qui sont des méthodes adaptées en générale pour les graphes du web et les graphes des réseaux sociaux. Nous avons présenté les différentes méthodes de compression existantes dans chacune des deux classes où nous avons essayé de mettre en relief les points forts de chacune. Nous avons ainsi établi une étude comparative entre elles en se basant sur leurs caractéristiques qui peuvent avoir une forte influence sur leur performance et oriente le choix de la méthode. Nous avons également raffiné la classification des méthodes de compression des graphes proposée dans un travail de Master précédent. Nous nous sommes basées pour cela sur le principe de fonctionnement et les fondements théoriques de chacune en respect de la classification précédente.

Cette étude nous a amenée à constater que malgré le progrès existant dans ce domaine plusieurs défis, tels que : la compression temps réel et les temps de traitements sont encore présents. Ces défis constituent de nouvelles problématiques de recherche qui sont toujours ouvertes.

Comme perspective à ce travail, il serait intéressant d'implémenter ces méthodes et leurs variantes. Cette implémentation nous permettra de mieux comparer entre ces méthodes en employant le même jeux de données. De ce fait, nous pourrons établir une comparaison plus objective entre elles.

Annexe A

Description des graphes de test

Graphe de test	Caractéristiques	
	Nombre de nœuds	Nombre de liens
eu-2005	862 milles de nœuds	19M de liens
Cond-mat	36 milles de nœuds	171 milles de liens
CommNet	taille : 225.9MB	
SalesDay	1 milles de nœuds	
Movielens10M	10 milles de nœuds	10M de liens
ASOregon	13 milles nœuds	37 milles liens
Enron	80 milles nœuds	288 milles liens
uk-2002	18 milles nœuds	298 milles liens
graphe test de GCUPMT	8 192 milles nœuds	
Composante chimique	21 étiquettes	422 transactions
NotreDame	325 milles de nœuds	1M de liens

TABLE A.1 – Graphes de test

Bibliographie

- (2012). *Quelques rappels sur la théorie des graphes*. IUT Lyon Informatique.
- Alvarez-Garcia, S., de Bernardo, G., Brisaboa, N. R., and Navarro, G. (2017). A succinct data structure for self-indexing ternary relations. *Journal of Discrete Algorithms*, 43 :38–53.
- Álvarez-García, S., Freire, B., Ladra, S., and Pedreira, Ó. (2018). Compact and efficient representation of general graph databases. *Knowledge and Information Systems*, pages 1–32.
- Aragon, C. R. and Seidel, R. G. (1989). Randomized search trees. In *Foundations of Computer Science, 1989., 30th Annual Symposium on*, pages 540–545. IEEE.
- Araujo, M., Günnemann, S., Mateos, G., and Faloutsos, C. (2014). Beyond blocks : Hyperbolic community detection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 50–65. Springer.
- Asano, Y., Miyawaki, Y., and Nishizeki, T. (2008). Efficient compression of web graphs. In *International Computing and Combinatorics Conference*, pages 1–11. Springer.
- Badr, M. (2013). *Traitement de requêtes top-k multicritères et application à la recherche par le contenu dans les bases de données multimédia*. PhD thesis, Cergy-Pontoise.
- Barbay, J., Golynski, A., Munro, J. I., and Rao, S. S. (2006). Adaptive searching in succinctly encoded binary relations and tree-structured documents. In *Annual Symposium on Combinatorial Pattern Matching*, pages 24–35. Springer.
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics : theory and experiment*, 2008(10) :P10008.
- Boldi, P., Rosa, M., Santini, M., and Vigna, S. (2011). Layered label propagation : A multiresolution coordinate-free ordering for compressing social networks. In *Proceedings of the 20th international conference on World wide web*, pages 587–596. ACM.
- Boldi, P. and Vigna, S. (2004). The webgraph framework i : compression techniques. In *Proceedings of the 13th international conference on World Wide Web*, pages 595–602. ACM.
- Brisaboa, N. R., de Bernardo, G., Gutiérrez, G., Ladra, S., Penabad, M. R., and Troncoso, B. A. (2015). Efficient set operations over k2-trees. In *Data Compression Conference (DCC), 2015*, pages 373–382. IEEE.

- Brisaboa, N. R., De Bernardo, G., Konow, R., and Navarro, G. (2014a). K 2-treaps : Range top-k queries in compact space. In *International Symposium on String Processing and Information Retrieval*, pages 215–226. Springer.
- Brisaboa, N. R., De Bernardo, G., and Navarro, G. (2012). Compressed dynamic binary relations. In *Data Compression Conference (DCC), 2012*, pages 52–61. IEEE.
- Brisaboa, N. R., Ladra, S., and Navarro, G. (2009). k 2-trees for compact web graph representation. In *International Symposium on String Processing and Information Retrieval*, pages 18–30. Springer.
- Brisaboa, N. R., Ladra, S., and Navarro, G. (2013). Dacs : Bringing direct access to variable-length codes. *Information Processing & Management*, 49(1) :392–404.
- Brisaboa, N. R., Ladra, S., and Navarro, G. (2014b). Compact representation of web graphs with extended functionality. *Information Systems*, 39 :152–174.
- Buehrer, G. and Chellapilla, K. (2008). A scalable pattern mining approach to web graph compression with communities. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 95–106. ACM.
- Claude, F. and Ladra, S. (2011). Practical representations for web and social graphs. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1185–1190. ACM.
- Claude, F. and Navarro, G. (2010a). Extended compact web graph representations. In *Algorithms and Applications*, pages 77–91. Springer.
- Claude, F. and Navarro, G. (2010b). Fast and compact web graph representations. *ACM Transactions on the Web (TWEB)*, 4(4) :16.
- De Berg, M., Van Kreveld, M., Overmars, M., and Schwarzkopf, O. (1997). Computational geometry. In *Computational geometry*, pages 1–17. Springer.
- De Bernardo, G., Álvarez-García, S., Brisaboa, N. R., Navarro, G., and Pedreira, O. (2013). Compact queriable representations of raster data. In *International Symposium on String Processing and Information Retrieval*, pages 96–108. Springer.
- de Bernardo Roca, G. (2014). *New data structures and algorithms for the efficient management of large spatial datasets*. PhD thesis, Citeseer.
- Dhulipala, L., Kabiljo, I., Karrer, B., Ottaviano, G., Pupyrev, S., and Shalita, A. (2016). Compressing graphs and indexes with recursive graph bisection. *arXiv preprint arXiv :1602.08820*.
- Fages, J.-G. (2014). *Exploitation de structures de graphe en programmation par contraintes*. PhD thesis, Ecole des Mines de Nantes.

- Garcia, S. A., Brisaboa, N. R., de Bernardo, G., and Navarro, G. (2014). Interleaved k2-tree : Indexing and navigating ternary relations. In *Data Compression Conference (DCC), 2014*, pages 342–351. IEEE.
- Grossi, R., Gupta, A., and Vitter, J. S. (2003). High-order entropy-compressed text indexes. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 841–850. Society for Industrial and Applied Mathematics.
- Guillaume, J.-L. and Latapy, M. (2002). The web graph : an overview. In *Actes d’ALGOTEL’02 (Quatrièmes Rencontres Francophones sur les aspects Algorithmiques des Télécommunications)*.
- Hennecart, F., Bretto, A., and Faisant, A. (2012). *Eléments de théorie des graphes*.
- Hernández, C. and Navarro, G. (2014). Compressed representations for web and social graphs. *Knowledge and information systems*, 40(2) :279–313.
- Hespanha, J. P. (2004). An efficient matlab algorithm for graph partitioning. *Santa Barbara, CA, USA : University of California*.
- Jean-Charles Régin, A. M. (2016). *Théorie des graphes*. Technical report.
- Kang, U. and Faloutsos, C. (2011). Beyond ‘caveman communities’ : Hubs and spokes for graph compression and mining. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 300–309. IEEE.
- Karypis, G. and Kumar, V. (2000). Multilevel k-way hypergraph partitioning. *VLSI design*, 11(3) :285–300.
- Ketkar, N. S., Holder, L. B., and Cook, D. J. (2005). Subdue : Compression-based frequent pattern discovery in graph data. In *Proceedings of the 1st international workshop on open source data mining : frequent pattern mining implementations*, pages 71–76. ACM.
- Khan, K. U., Nawaz, W., and Lee, Y.-K. (2014). Set-based unified approach for attributed graph summarization. In *Big Data and Cloud Computing (BdCloud), 2014 IEEE Fourth International Conference on*, pages 378–385. IEEE.
- Koutra, D., Kang, U., Vreeken, J., and Faloutsos, C. (2015). Summarizing and understanding large graphs. *Statistical Analysis and Data Mining : The ASA Data Science Journal*, 8(3) :183–202.
- LeFevre, K. and Terzi, E. (2010). Grass : Graph structure summarization. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pages 454–465. SIAM.
- Lehman, E., Leighton, F. T., and Meyer, A. R. (2010). *Mathematics for computer science*. Technical report, Technical report, 2006. Lecture notes.

- Lelewer, D. A. and Hirschberg, D. S. (1987). Data compression. *ACM Computing Surveys (CSUR)*, 19(3) :261–296.
- Lemmouchi, S. (2012). *Etude de la robustesse des graphes sociaux émergents*. PhD thesis, Université Claude Bernard-Lyon I.
- Liu, Y., Safavi, T., Dighe, A., and Koutra, D. (2018a). Graph summarization methods and applications : A survey. *ACM Computing Surveys (CSUR)*, 51(3) :62.
- Liu, Y., Safavi, T., Shah, N., and Koutra, D. (2018b). Reducing large graphs to small super-graphs : a unified approach. *Social Network Analysis and Mining*, 8(1) :17.
- Liu, Y., Shah, N., and Koutra, D. (2015). An empirical comparison of the summarization power of graph clustering methods. *arXiv preprint arXiv :1511.06820*.
- Lopez, P. (2003). Cours de graphes.
- Maneth, S. and Peternek, F. (2015). A survey on methods and systems for graph compression. *arXiv preprint arXiv :1504.00616*.
- Maneth, S. and Peternek, F. (2018). Grammar-based graph compression. *Information Systems*, 76 :19–45.
- Martínez-Prieto, M. A., Fernández, J. D., and Cánovas, R. (2012). Compression of rdf dictionaries. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pages 340–347. ACM.
- Maserrat, H. and Pei, J. (2012). Community preserving lossy compression of social networks. In *2012 IEEE 12th International Conference on Data Mining*, pages 509–518. IEEE.
- Müller, D. (2012). *Introduction à la théorie des graphes*. Commission romande de mathématique (CRM).
- Parlebas, P. (1972). Centralité et compacité d’un graphe. *Mathématiques et sciences humaines*, 39 :5–26.
- Pellegrini, M., Haynor, D., and Johnson, J. M. (2004). Protein interaction networks. *Expert review of proteomics*, 1(2) :239–249.
- Rigo, M. (2010). *Théorie des graphes*. Université de liège, Faculté des sciences Département de mathématiques.
- Rossi, R. A. and Zhou, R. (2018). Graphzip : a clique-based sparse graph compression method. *Journal of Big Data*, 5(1) :10.
- Roux, P. (2014). *Théorie des graphes*.
- SABLIK, M. (2018). Graphe et langage.

- Sethi, G., Shaw, S., Vinutha, K., and Chakravorty, C. (2014). Data compression techniques. *International Journal of Computer Science and Information Technologies*, 5(4) :5584–6.
- Shah, N., Koutra, D., Zou, T., Gallagher, B., and Faloutsos, C. (2015). Timecrunch : Interpretable dynamic graph summarization. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1055–1064. ACM.
- Shah, R. J. (2018). Graph compression using pattern matching techniques. *arXiv preprint arXiv :1806.01504*.
- Shen, Z., Ma, K.-L., and Eliassi-Rad, T. (2006). Visual analysis of large heterogeneous social networks by semantic and structural abstraction. *IEEE transactions on visualization and computer graphics*, 12(6) :1427–1439.
- Shi, L., Tong, H., Tang, J., and Lin, C. (2015). Vegas : Visual influence graph summarization on citation networks. *IEEE Transactions on Knowledge and Data Engineering*, 27(12) :3417–3431.
- Shi, Q., Xiao, Y., Bessis, N., Lu, Y., Chen, Y., and Hill, R. (2012). Optimizing k2 trees : A case for validating the maturity of network of practices. *Computers & Mathematics with Applications*, 63(2) :427–436.
- Uthayakumar, J., Vengattaraman, T., and Dhavachelvan, P. (2018). A survey on data compression techniques : From the perspective of data quality, coding schemes, data type and applications. *Journal of King Saud University-Computer and Information Sciences*.
- W. GUERMAH, T. B. (2018). Compression de graphes : étude et classification. Master’s thesis, Esi.
- Wang, H., Yang, Y., Liu, B., and Fujita, H. (2019). A study of graph-based system for multi-view clustering. *Knowledge-Based Systems*, 163 :1009–1019.
- Yang, J. and Leskovec, J. (2013). Overlapping community detection at scale : a nonnegative matrix factorization approach. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 587–596. ACM.
- Yıldırım, H., Chaoji, V., and Zaki, M. J. (2012). Grail : a scalable index for reachability queries in very large graphs. *The VLDB Journal—The International Journal on Very Large Data Bases*, 21(4) :509–534.
- Zhang, H., Duan, Y., Yuan, X., and Zhang, Y. (2014a). Assg : Adaptive structural summary for rdf graph data. In *International Semantic Web Conference (Posters & Demos)*, pages 233–236. Citeseer.
- Zhang, Y., Xiong, G., Liu, Y., Liu, M., Liu, P., and Guo, L. (2014b). Delta-k 2-tree for compact representation of web graphs. In *Asia-Pacific Web Conference*, pages 270–281. Springer.