Traverses minimales d'un hypergraphe : Applications et Analyse

Céline Hébert Alain Bretto Loïck Lhote

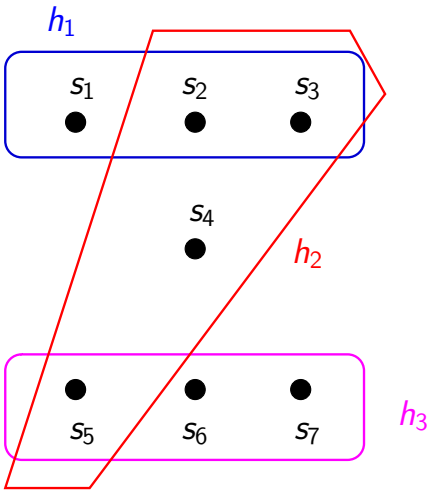
GREYC, Université de Caen Basse-Normandie

ALEA 2007, Marseille

Hypergraphe

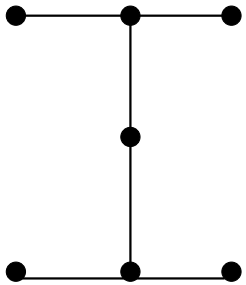
Hypergraphe $\mathcal{H} = (V, E)$ où

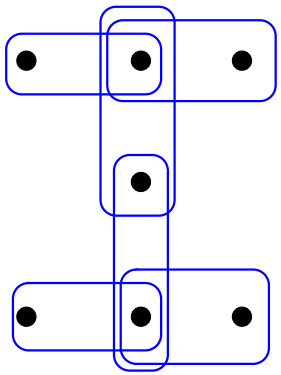
- V est l'ensemble des sommets
- et $E \subset 2^{|V|}$ est l'ensemble des hyperarêtes



	s_1	s_2	s_3	s_4	s_5	s_6	s_7
h_1	1	1	1	0	0	0	0
h_2	0	1	1	1	1	0	0
h_3	0	0	0	0	1	1	1

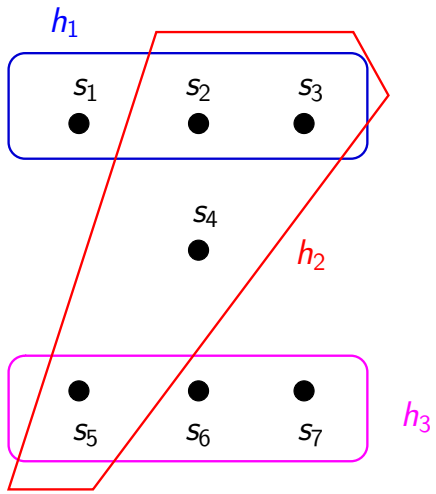
Les hypergraphes généralisent les graphes





Traverses minimales

Traverse = ensemble de sommets qui intersecte toutes les hyperarêtes
Traverse minimale= traverse minimale au sens de l'inclusion



Exemples :

- $\{s_1, s_7\}$ n'est pas une traverse
- $\{s_1, s_3, s_5\}$ est une traverse non-minimale
- $\{s_1, s_5\}$ est une traverse minimale

Problèmes :

- Etant donné un hypergraphe $\mathcal{H}$, peut-on générer toutes ses traverses minimales $\mathcal{T}(\mathcal{H})$?
- Si oui, comment et à quel coût ?

Applications

Application en Intelligence Artificielle

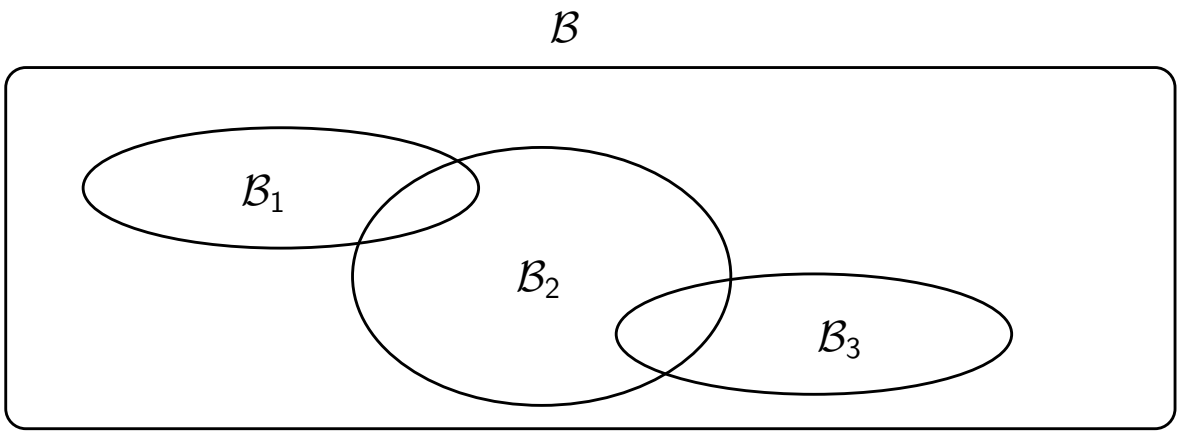
Contexte : ajouter une nouvelle connaissance p à un base de connaissances $\mathcal{B}$

- premier cas : p n'est pas en contradiction avec $\mathcal{B}$
  ⇝ on ajoute p à $\mathcal{B}$
- deuxième cas : p est en contradiction avec $\mathcal{B}$
  ⇝ on enlève un ensemble $\mathcal{R} \subset \mathcal{B}$ de connaissances tel que $\neg p$ ne peut plus être inférée de $\mathcal{B} \backslash \mathcal{R}$

Application en Intelligence Artificielle

Contexte : ajouter une nouvelle connaissance p à un base de connaissances $\mathcal{B}$

- premier cas : p n'est pas en contradiction avec $\mathcal{B}$
  ➟ on ajoute p à $\mathcal{B}$
- deuxième cas : p est en contradiction avec $\mathcal{B}$
  ➟ on enlève un ensemble $\mathcal{R} \subset \mathcal{B}$ de connaissances tel que $\neg p$ ne peut plus être inférée de $\mathcal{B} \backslash \mathcal{R}$



Propriété 1. au moins une connaisance de chaque ensemble $\mathcal{B}_i$ doit être supprimée
Propriété 2. on veut faire le minimum de changements

Application en Logique

Problème MSAT. Etant donnée une CNF $F = C_1 \wedge \ldots \wedge C_p$ telle que toutes les clauses sont soit positives, soit négatives. F est-elle satisfaisable ?

➠ C'est un problème NP-complet.

Exemple sur les variables $X = \{x_1, x_2, x_3, x_4\}$:

$$x_1 \quad \wedge \quad (x_2 \vee x_4) \quad \wedge \quad (\neg x_2 \vee \neg x_4) \quad \wedge \quad \neg x_3$$

Application en Logique

Problème MSAT. Etant donnée une CNF $F = C_1 \wedge \ldots \wedge C_p$ telle que toutes les clauses sont soit positives, soit négatives. F est-elle satisfaisable ?

➠ C'est un problème NP-complet.

Exemple sur les variables $X = \{x_1, x_2, x_3, x_4\}$:

$$x_1 \quad \wedge \quad (x_2 \vee x_4) \quad \wedge \quad (\neg x_2 \vee \neg x_4) \quad \wedge \quad \neg x_3$$

On définit deux hypergraphes $\mathcal{H}^+ = (X, E^+)$ et $\mathcal{H}^- = (X, E^-)$ avec

$$E^+ = \{\{x_1\}, \{x_2, x_4\}\}, \qquad E^- = \{\{x_2, x_4\}, \{x_3\}\}$$

Pour deux hypergraphes $\mathcal{H}$ et $\mathcal{H}'$, on écrit $\mathcal{H} \succeq \mathcal{H}'$ si toute hyperarête de $\mathcal{H}$ contient une hyperarête de $\mathcal{H}'$: $\forall e \in E, \quad \exists e' \in E' \quad \text{t.q.} \quad e' \subseteq e$.

Application en Logique

Problème MSAT. Etant donnée une CNF $F = C_1 \wedge \ldots \wedge C_p$ telle que toutes les clauses sont soit positives, soit négatives. F est-elle satisfaisable ?

➠ C'est un problème NP-complet.

Exemple sur les variables $X = \{x_1, x_2, x_3, x_4\}$:

$$x_1 \quad \wedge \quad (x_2 \vee x_4) \quad \wedge \quad (\neg x_2 \vee \neg x_4) \quad \wedge \quad \neg x_3$$

On définit deux hypergraphes $\mathcal{H}^+ = (X, E^+)$ et $\mathcal{H}^- = (X, E^-)$ avec

$$E^+ = \{\{x_1\}, \{x_2, x_4\}\}, \qquad E^- = \{\{x_2, x_4\}, \{x_3\}\}$$

Pour deux hypergraphes $\mathcal{H}$ et $\mathcal{H}'$, on écrit $\mathcal{H} \succeq \mathcal{H}'$ si toute hyperarête de $\mathcal{H}$ contient une hyperarête de $\mathcal{H}'$: $\forall e \in E, \quad \exists e' \in E' \quad \text{t.q.} \quad e' \subseteq e$.

Propriété : Une instance F de MSAT n'est pas satisfaisable ssi $\mathcal{T}(\mathcal{H}^+) \succeq \mathcal{H}^-$

Exemple : $\mathcal{T}(\mathcal{H}^+) = \{\{x_1, x_2\}, \{x_1, x_4\}\} \not\succeq \mathcal{H}^-$

Application en Fouille de données

s_1	s_2	s_3	s_4	s_5	s_6	s_7
1	1	1	0	0	0	0
0	1	1	1	1	0	0
0	0	0	0	1	1	1

Motif fréquent = ensemble de colonnes qui sont présentes en même temps dans plusieurs lignes.

Exemple

- $\{s_1, s_4\}$ n'est pas un motif fréquent
- $\{s_2, s_3\}$ est un motif fréquent

Application en Fouille de données

s_1	s_2	s_3	s_4	s_5	s_6	s_7
1	1	1	0	0	0	0
0	1	1	1	1	0	0
0	0	0	0	1	1	1

Motif fréquent = ensemble de colonnes qui sont présentes en même temps dans plusieurs lignes.

Exemple

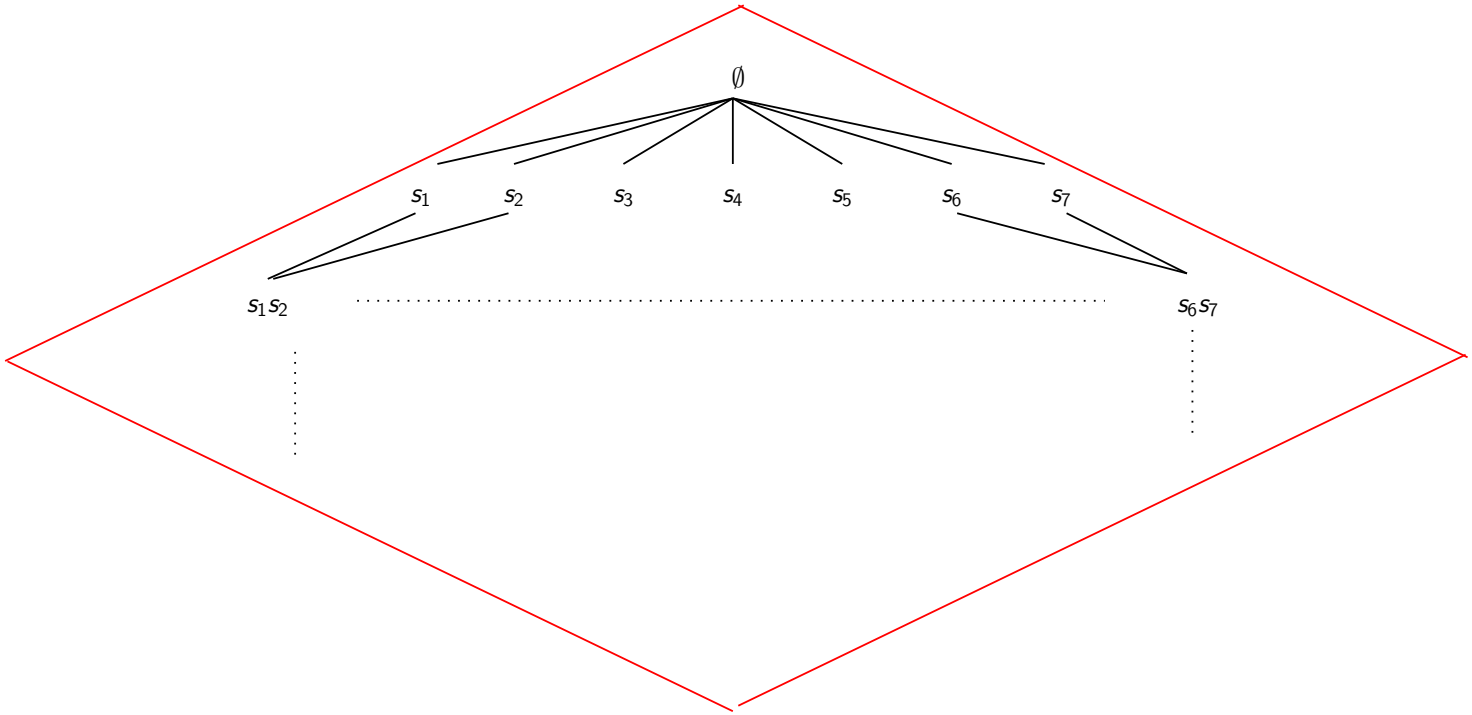
- $\{s_1, s_4\}$ n'est pas un motif fréquent
- $\{s_2, s_3\}$ est un motif fréquent

Bordure négative = ensemble des motifs non-fréquents dont tous les sous-motifs sont fréquents.

Exemple

- $\{s_1, s_4\}$ est un motif de la bordure négative
- $\{s_4, s_5, s_6\}$ n'est pas dans la bordure négative car $\{s_4, s_6\}$ n'est pas un motif fréquent

Treillis des motifs

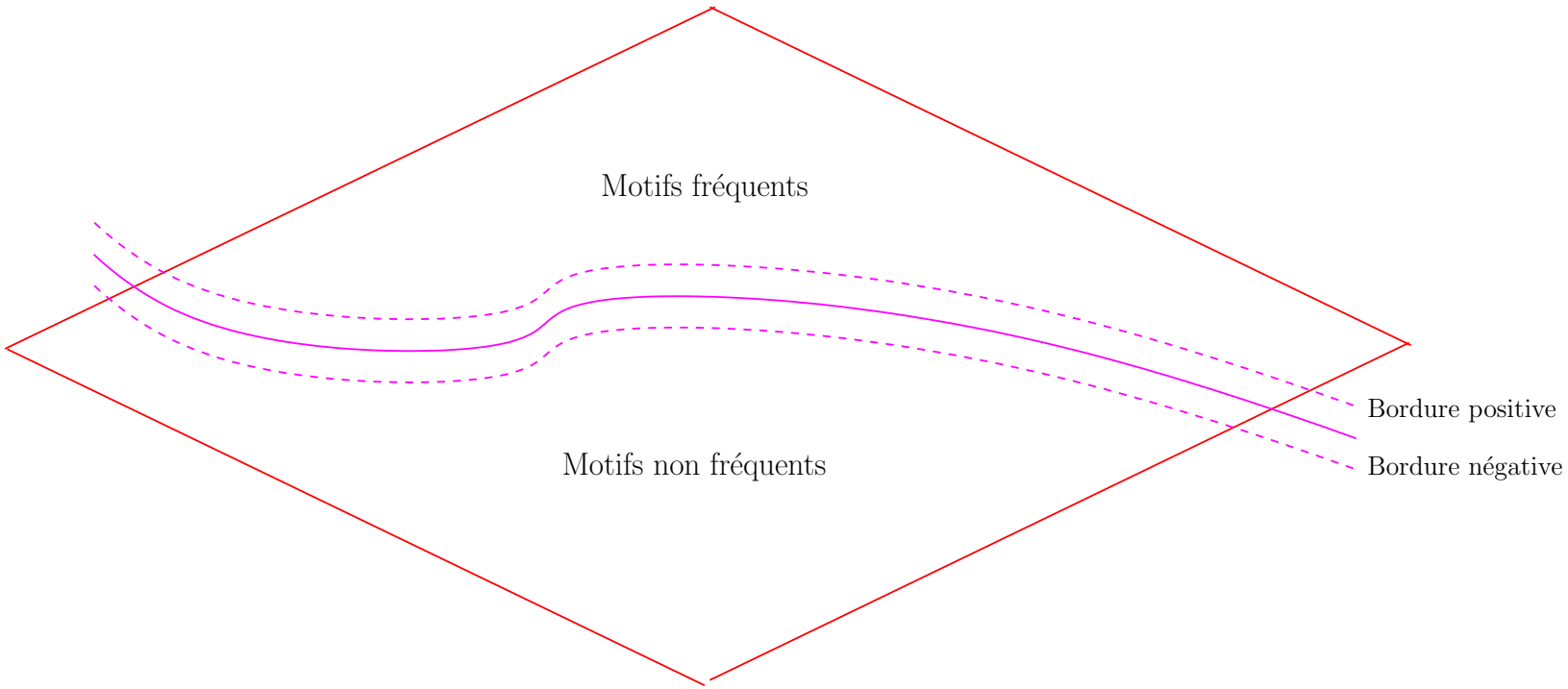


Intérêts de la bordure négative :

- représentation condensée des motifs fréquents
- Complexité algorithmes par niveaux

  Nb de motifs fréquents + taille de la bordure négative

Treillis des motifs

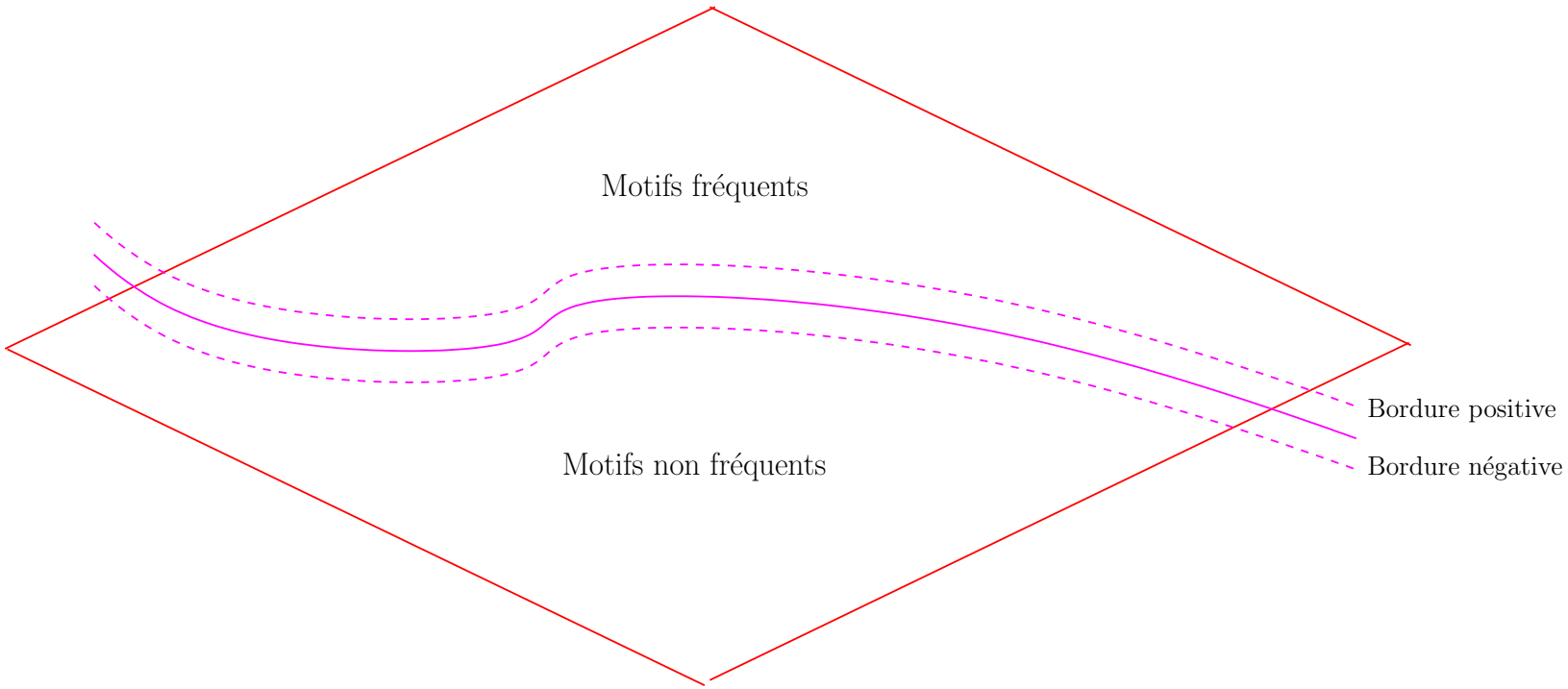


Intérêts de la bordure négative :

- représentation condensée des motifs fréquents
- Complexité algorithmes par niveaux

  Nb de motifs fréquents + taille de la bordure négative

Treillis des motifs



Intérêts de la bordure négative :

- représentation condensée des motifs fréquents
- Complexité algorithmes par niveaux

  Nb de motifs fréquents + taille de la bordure négative

Bordure négative et traverses minimales

Base

s_1	s_2	s_3	s_4	s_5	s_6	s_7
1	1	1	0	0	0	0
0	1	1	1	1	0	0
0	0	0	0	1	1	1

$-->$

Base opposée

s_1	s_2	s_3	s_4	s_5	s_6	s_7
0	0	0	1	1	1	1
1	0	0	0	0	1	1
1	1	1	1	0	0	0

$=$

$\mathcal{T}(\mathcal{H})$
est en bijection avec
la bordure négative

$<--$

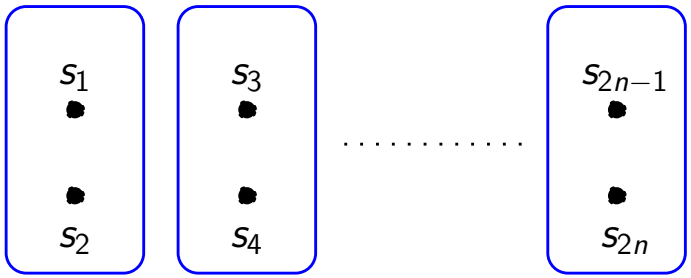
Hypergraphe

s_1	s_2	s_3	s_4	s_5	s_6	s_7
0	0	0	1	1	1	1
1	0	0	0	0	1	1
1	1	1	1	0	0	0

Analyse en moyenne

Difficulté théorique

Nombre de traverses minimales

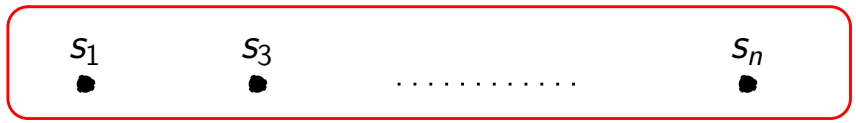


$|\mathcal{T}(\mathcal{H})| = 2^n$

$|\mathcal{T}(\mathcal{H})| = 1$

Taille de la plus grande traverse

$\max = n$



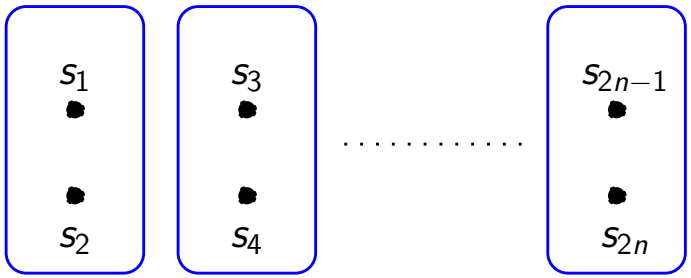
$\max = 1$

➠ Bonne notion de complexité : fait intervenir la taille de la sortie

Question ouverte : générer les traverses minimales est-il un problème de complexité polynomiale en la taille de la sortie (output-polynomial) ?

Difficulté théorique

Nombre de traverses minimales

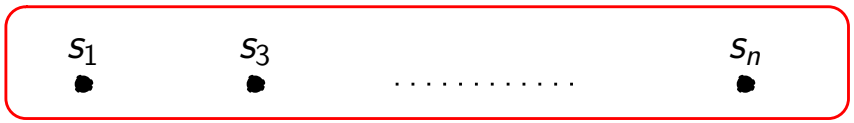


$|\mathcal{T}(\mathcal{H})| = 2^n$

$|\mathcal{T}(\mathcal{H})| = 1$

Taille de la plus grande traverse

$\max = n$



$\max = 1$

➟ Bonne notion de complexité : fait intervenir la taille de la sortie

Question ouverte : générer les traverses minimales est-il un problème de complexité polynomiale en la taille de la sortie (output-polynomial) ?

Algorithmes

Fredman et Khachiyan (Dual, 1996)

- pour étudier la complexité théorique du problème de dualisation d'une DNF
- complexité :

$$O(T^{\log T}) \quad \text{si } T \text{ est la taille combinée de l'entrée et de la sortie}$$

C. Hébert et A. Bretto (MT-Miner, 2005)

- algorithme par niveaux
- utilise des règles d'élagages
  - Règle 1 : tout sur-ensemble d'une traverse ne peut être une traverse minimale.
  - Règle 2 : si un ensemble de sommets E contient un sous-ensemble qui intersecte le même nombre d'hyperarêtes, alors E et tout ses sur-ensembles ne peuvent être des traverses minimales.
- complexité :

$$O(T \cdot 2^{\max})$$

avec T la taille combinée de l'entrée et de la sortie et max la taille de la plus grande traverse minimale

Question : quel est le meilleur algorithme ?

Algorithmes

Fredman et Khachiyan (Dual, 1996)

- pour étudier la complexité théorique du problème de dualisation d'une DNF
- complexité :

$$O(T^{\log T}) \quad \text{si } T \text{ est la taille combinée de l'entrée et de la sortie}$$

C. Hébert et A. Bretto (MT-Miner, 2005)

- algorithme par niveaux
- utilise des règles d'élagages
  - Règle 1 : tout sur-ensemble d'une traverse ne peut être une traverse minimale.
  - Règle 2 : si un ensemble de sommets E contient un sous-ensemble qui intersecte le même nombre d'hyperarêtes, alors E et tout ses sur-ensembles ne peuvent être des traverses minimales.
- complexité :

$$O(T \cdot 2^{\text{max}})$$

avec T la taille combinée de l'entrée et de la sortie et max la taille de la plus grande traverse minimale

Question : quel est le meilleur algorithme ?

Expériences

Modèle aléatoire de Erdös-Rényi :

- nombre de sommets : 50
- nombre d'hyperarêtes : 1000
- probabilité d'appartenir à une hyperarête : p

p	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
Dual	326	fail	fail	fail	fail	fail	fail	fail	59
MT-Miner	0.25	4	48	530	fail	fail	fail	fail	fail
$\lvert\mathcal{T}(\mathcal{H})\rvert$	26939	339372	2634205	16237137	?	?	?	?	4396
max	3	5	7	8	?	?	?	?	?

Modèle aléatoire pour les analyses :

- nombre de sommets : n
- nombre d'hyperarêtes : $m = \alpha \cdot n$, $\alpha \in \mathbb{R}^+$
- probabilité d'appartenir à une hyperarête : p

Question : est-ce un bon modèle ?

Expériences

Modèle aléatoire de Erdös-Rényi :

- nombre de sommets : 50
- nombre d'hyperarêtes : 1000
- probabilité d'appartenir à une hyperarête : p

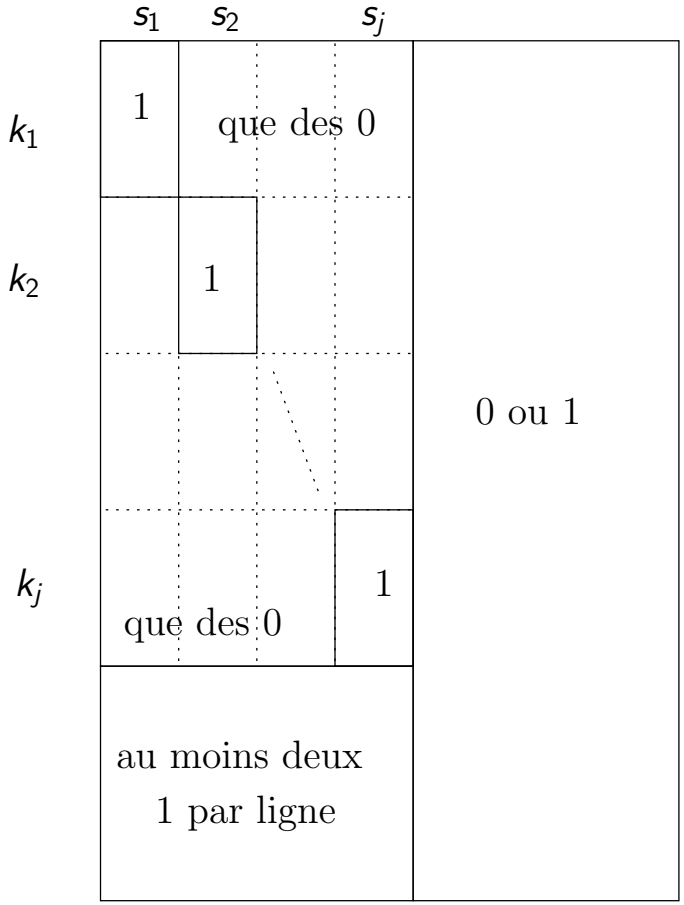
p	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
Dual	326	fail	fail	fail	fail	fail	fail	fail	59
MT-Miner	0.25	4	48	530	fail	fail	fail	fail	fail
$\lvert\mathcal{T}(\mathcal{H})\rvert$	26939	339372	2634205	16237137	?	?	?	?	4396
max	3	5	7	8	?	?	?	?	?

Modèle aléatoire pour les analyses :

- nombre de sommets : n
- nombre d'hyperarêtes : $m = \alpha \cdot n$, $\alpha \in \mathbb{R}^+$
- probabilité d'appartenir à une hyperarête : p

Question : est-ce un bon modèle ?

Première formule



Nombre moyen de traverses minimales de longueur j

$$N_j = \binom{m}{j} \sum_{\substack{k_1+\ldots+k_j=j \\ k_i \geq 1}}^{n} \binom{n}{k_1,\ldots,k_j} (pq^{j-1})^{k_1+\ldots+k_j} (1-q^j-jpq^{j-1})^{n-(k_1+\ldots+k_j)}$$

Formule intégrale

On peut écrire

$$N_j = \binom{m}{j} F(pq^{j-1}, 1 - q^j - jpq^{j-1}) \qquad \text{avec}$$

$$F(X,Y) = \sum_{\substack{k_1 + \ldots + k_j = j \\ k_i \geq 1}} \binom{n}{k_1, \ldots, k_j} X^{k_1+\ldots+k_j} Y^{n-(k_1+\ldots+k_j)}$$

Lemme

La fonction F vérifie la formule intégrale

$$F(X,Y) = \frac{n!}{(n-j)!} \int_{[0,X]^j} (Y + t_1 + \ldots + t_j)^{n-j} dt_1 \ldots dt_j$$

Nombre moyen de traverses minimales

$$\mathrm{E}[|\mathcal{T}(\mathcal{H})|] = \sum_{j=1}^{m} \binom{m}{j} \frac{n!}{(n-j)!} \int_{[0,pq^{j-1}]^j} (1 - q^j - t_1 - \ldots - t_j)^{n-j} dt_1 \ldots dt_j$$

On sépare la somme en deux parties : $j|\log q| < 0.75 \log n$ et $j|\log q| \geq 0.75 \log n$

Théorème

Le nombre moyen de traverses minimales dans un hypergraphe à m sommets, n hyperarêtes et sous le modèle d'Erdös et Rényi de paramètre $p = 1 - q$ est asymptotiquement

$$\mathrm{E}[|\mathcal{T}(\mathcal{H})|] \sim w_{j_0} + w_{j_0+1} \quad \text{avec} \quad w_j = \binom{m}{j} (1 - e^{npq^{j-1}})^j.$$

$$\text{et} \qquad j_0 = \frac{1}{|\log q|} (\log n + \log\log n - \log\log\log n + O(1))$$

Plus simplement

$$\mathrm{E}[|\mathcal{T}(\mathcal{H})|] = O\left(n^{\log n/|\log q|}\right)$$

Taille moyenne de la plus grande traverse

Théorème

La taille moyenne de la plus grande traverse minimale dans un hypergraphe à m sommets, n hyperarêtes et sous le modèle d'Erdös et Rényi de paramètre $p = 1 - q$ est asymptotiquement

$$\mathrm{E}[\max] = \frac{1}{|\log q|}(2\log n - \log\log n) + O(\log\log\log n)$$

Comparaison des algorithmes

Fredman et Khachiyan (1996)

- Complexité :

$$O(T^{\log T}) \quad \text{si } T \text{ est la taille combinée de l'entrée et de la sortie}$$

- Asymptotique :

$$O\left(\exp \frac{\log^4 n}{|\log q|^2}\right)$$

C. Hébert et A. Bretto

- Complexité :

$$O(T \cdot 2^{\max})$$

- Asymptotique :

$$O\left(\exp \frac{\log^2 n}{|\log q|}\right)$$

Conclusion

Résultats

- la taille de la plus grande traverse est logarithmique en le nombre d'hyperarêtes
- le nombre de traverses minimales est sous exponentiel
- on peut faire des comparaisons d'algorithmes

Améliorations

- autres modèles probabilistes
- améliorer les complexités données
- résultats qui mélangent pire des cas et en moyenne