# Apports de la PLNE pour la modélisation et l'extraction d'ensembles de motifs

*Séminaire HEUDIASYC*

**Samir Loudni**

15 Mai 2018

# Plan de la présentation

1. La PLNE pour l'extraction de Pattern sets – application au clustering conceptuel

2. Clustering conceptuel équitable

# La PLNE pour l'extraction de pattern sets

# Pattern set mining : Motivations

Une limitation bien connue en fouille de motifs : résultats incluant un grand nombre de motifs non pertinents ou redondants, difficultés à analyser

*Un nouvel intérêt :* vers des approches capables d'extraire des motifs plus pertinents (motifs "actionnables")

➥ Pattern sets

# Pattern set mining

## Définition (Pattern set)

Soit $\mathcal{P}$ un ensemble de motifs locaux. Un pattern set $\Phi \subseteq \mathcal{P}$ est un ensemble de motifs.

**Combinatoire trop importante** : $2^{2^{|\mathcal{I}|}}$

Plusieurs types de pattern sets :

- associative classification (Liu et al. KDD'98), clustering (Wang et al. CIKM'99), top-$k$ patterns (Fu et al. ISMIS'00), tiling (Geerts et al. DS'04), pattern teams (Knobbe et al. ECML/PKDD'06), ...
- approches génériques : constraint-based pattern set mining (De Raedt et al. SDM'07), from local patterns to global models (Knobbe et al. LeGo'08)

# Pourquoi la PLNE ?

- Largement utilisé en recherche opérationnelle
  - planification
  - ordonnancement
  - allocation de fréquence
  - tournées de véhicules
- Cadre générique et flexible
- Étudié depuis plus de 50 ans
  - des solveurs efficace
  - scalable : des centaines de milliers de contraintes/variables

# Contexte – Représentation transactionnelle

- un ensemble de $m$ transactions $\mathcal{T}$, un ensemble de $n$ items $\mathcal{I}$,
- une relation $R \subseteq \mathcal{T} \times \mathcal{I}$ : $(t, i) \in R$ ssi $i \in t$.

| Trans. | Items | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $t_1$ | A | B | | D | | | | |
| $t_2$ | A | | | | E | F | | |
| $t_3$ | A | | | | E | | G | |
| $t_4$ | A | | | | E | | G | |
| $t_5$ | | B | | | E | | G | |
| $t_6$ | | B | | | E | | G | |
| $t_7$ | | | C | | E | | G | |
| $t_8$ | | | C | | E | | G | |
| $t_9$ | | | C | | E | | | H |
| $t_{10}$ | | | C | | E | | | H |
| $t_{11}$ | | | C | | | F | G | H |

# Contexte – Représentation transactionnelle

- un ensemble de $m$ transactions $\mathcal{T}$, un ensemble de $n$ items $\mathcal{I}$,
- une relation $R \subseteq \mathcal{T} \times \mathcal{I}$ : $(t, i) \in R$ ssi $i \in t$.

- **Extension** : soit $I \subseteq \mathcal{I}$, $ext(I) = \{t \in \mathcal{T} | \ \forall i \in I, (t, i) \in R\}$

| Trans. | Items | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $t_1$ | A | B | | D | | | | |
| $t_2$ | A | | | | E | F | | |
| $t_3$ | A | | | | E | | G | |
| $t_4$ | A | | | | E | | G | |
| $t_5$ | | B | | | E | | G | |
| $t_6$ | | B | | | E | | G | |
| $t_7$ | | | C | | E | | G | |
| $t_8$ | | | C | | E | | G | |
| $t_9$ | | | C | | E | | | H |
| $t_{10}$ | | | C | | E | | | H |
| $t_{11}$ | | | C | | | F | G | H |

$$ext(\{B, E, G\}) = \{t_5, t_6\}$$

# Contexte – Représentation transactionnelle

- un ensemble de $m$ transactions $\mathcal{T}$, un ensemble de $n$ items $\mathcal{I}$,
- une relation $R \subseteq \mathcal{T} \times \mathcal{I}$ : $(t, i) \in R$ ssi $i \in t$.

- **Intention** : soit $T \subseteq \mathcal{T}$, $int(T) = \{i \in \mathcal{I} | \ \forall t \in T, (t, i) \in R\}$

| Trans. | Items | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $t_1$ | A | B | | D | | | | |
| $t_2$ | A | | | | E | F | | |
| $t_3$ | A | | | | E | | G | |
| $t_4$ | A | | | | E | | G | |
| $t_5$ | | B | | | E | | G | |
| $t_6$ | | B | | | E | | G | |
| $t_7$ | | | C | | E | | G | |
| $t_8$ | | | C | | E | | G | |
| $t_9$ | | | C | | E | | | H |
| $t_{10}$ | | | C | | E | | | H |
| $t_{11}$ | | | C | | | F | G | H |

$$int(\{t_5, t_6\}) = \{B, E, G\}$$

# Contexte – Représentation transactionnelle

- un ensemble de $m$ transactions $\mathcal{T}$, un ensemble de $n$ items $\mathcal{I}$,
- une relation $R \subseteq \mathcal{T} \times \mathcal{I}$ : $(t, i) \in R$ ssi $i \in t$.

- **Motif** = un sous-ensemble d'items (littéraux)

| Trans. | Items | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $t_1$ | A | B | | D | | | | |
| $t_2$ | A | | | | E | F | | |
| $t_3$ | A | | | | E | | G | |
| $t_4$ | A | | | | E | | G | |
| $t_5$ | | B | | | E | | G | |
| $t_6$ | | B | | | E | | G | |
| $t_7$ | | | C | | E | | G | |
| $t_8$ | | | C | | E | | G | |
| $t_9$ | | | C | | E | | | H |
| $t_{10}$ | | | C | | E | | | H |
| $t_{11}$ | | | C | | | F | G | H |

# Contexte – Représentation transactionnelle

- un ensemble de $m$ transactions $\mathcal{T}$, un ensemble de $n$ items $\mathcal{I}$,
- une relation $R \subseteq \mathcal{T} \times \mathcal{I}$ : $(t, i) \in R$ ssi $i \in t$.

- **Concept formel** $= (T, I) \in \mathcal{T} \times \mathcal{I}$ tel que $I = int(T) \wedge T = ext(I)$
  ➠ **Concept formel** $\equiv$ **Motif fermé**

| Trans. | Items | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $t_1$ | A | B | | D | | | | |
| $t_2$ | A | | | | E | F | | |
| $t_3$ | A | | | | E | | G | |
| $t_4$ | A | | | | E | | G | |
| $t_5$ | | B | | | E | | G | |
| $t_6$ | | B | | | E | | G | |
| $t_7$ | | | C | | E | | G | |
| $t_8$ | | | C | | E | | G | |
| $t_9$ | | | C | | E | | | H |
| $t_{10}$ | | | C | | E | | | H |
| $t_{11}$ | | | C | | | F | G | H |

$\phi = (\{t_5, t_6\}, \{B, E, G\})$

- $ext(\{B, E, G\}) = \{t_5, t_6\}$
- $int(\{t_5, t_6\}) = \{B, E, G\}$
- $\texttt{freq}(\{B, E, G\}, \mathcal{T}) = 2$

# Contexte – Représentation transactionnelle

- un ensemble de $m$ transactions $\mathcal{T}$, un ensemble de $n$ items $\mathcal{I}$,
- une relation $R \subseteq \mathcal{T} \times \mathcal{I}$ : $(t, i) \in R$ ssi $i \in t$.

- **Concept formel** $= (T, I) \in \mathcal{T} \times \mathcal{I}$ tel que $I = int(T) \wedge T = ext(I)$
  ➠ **Concept formel** $\equiv$ **Motif fermé**

| Trans. | Items | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $t_1$ | A | B | | D | | | | |
| $t_2$ | A | | | | E | F | | |
| $t_3$ | A | | | | E | | G | |
| $t_4$ | A | | | | E | | G | |
| $t_5$ | | B | | | E | | G | |
| $t_6$ | | B | | | E | | G | |
| $t_7$ | | | C | | E | | G | |
| $t_8$ | | | C | | E | | G | |
| $t_9$ | | | C | | E | | | H |
| $t_{10}$ | | | C | | E | | | H |
| $t_{11}$ | | | C | | | F | G | H |

**Motif fermé** : Un motif $\phi$ est fermé ssi tous ses sur-ensembles stricts ont une fréquence strictement inférieure.

# Contexte – Clustering conceptuel

**Clustering conceptuel** $\equiv$ ensemble de $k$ **concepts formels** $\Phi = \{\phi_1, \ldots, \phi_k\}$, avec $\phi_j = (T_j, I_j)$, tel que $\{T_1, \ldots, T_k\}$ forme une **partition** de l'ensemble de transactions $\mathcal{T}$.

$$(Q_1) \begin{cases} k_{min} \leq k \leq k_{max} \wedge \\ \texttt{cover}(\Phi, \mathcal{T}) = \mathcal{T} \wedge \\ overlap(\Phi, \mathcal{T}) = 0 \wedge \\ \wedge_{\phi \in \Phi} \textsc{ClosedPattern}(\phi) \end{cases}$$

# Contexte – Clustering conceptuel

**Clustering conceptuel** $\equiv$ ensemble de $k$ **concepts formels** $\Phi = \{\phi_1, \ldots, \phi_k\}$, avec $\phi_j = (T_j, I_j)$, tel que $\{T_1, \ldots, T_k\}$ forme une **partition** de l'ensemble de transactions $\mathcal{T}$.

$$(Q_1) \begin{cases} k_{min} \leq k \leq k_{max} \wedge \\ \texttt{cover}(\Phi, \mathcal{T}) = \mathcal{T} \wedge \\ overlap(\Phi, \mathcal{T}) = 0 \wedge \\ \wedge_{\phi \in \Phi} \textsc{ClosedPattern}(\phi) \end{cases}$$

- $k = \texttt{taille}(\Phi) = |\Phi|$
- $cover(\Phi, \mathcal{T}) = \bigcup_{\phi \in \Phi} cover(\phi, \mathcal{T})$
- $overlap(\Phi, \mathcal{T}) = |\bigcap_{\phi \in \Phi} cover(\phi, \mathcal{T})|$

# Contexte – Clustering conceptuel

**Clustering conceptuel** $\equiv$ ensemble de $k$ **concepts formels** $\Phi = \{\phi_1, \ldots, \phi_k\}$, avec $\phi_j = (T_j, I_j)$, tel que $\{T_1, \ldots, T_k\}$ forme une **partition** de l'ensemble de transactions $\mathcal{T}$.

$$(Q_1) \begin{cases} k_{min} \leq k \leq k_{max} \wedge \\ \texttt{cover}(\Phi, \mathcal{T}) = \mathcal{T} \wedge \\ overlap(\Phi, \mathcal{T}) = 0 \wedge \\ \wedge_{\phi \in \Phi} \textsc{ClosedPattern}(\phi) \end{cases}$$

- $k = \texttt{taille}(\Phi) = |\Phi|$
- $cover(\Phi, \mathcal{T}) = \bigcup_{\phi \in \Phi} cover(\phi, \mathcal{T})$
- $overlap(\Phi, \mathcal{T}) = |\bigcap_{\phi \in \Phi} cover(\phi, \mathcal{T})|$

- $k_{min} = 1$, $k_{max} = 3$

| Trans. | Items | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $t_1$ | A | B | | D | | | | |
| $t_2$ | A | | | | E | F | | |
| $t_3$ | A | | | | E | | G | |
| $t_4$ | A | | | | E | | G | |
| $t_5$ | | B | | | E | | G | |
| $t_6$ | | B | | | E | | G | |
| $t_7$ | | | C | | E | | G | |
| $t_8$ | | | C | | E | | G | |
| $t_9$ | | | C | | E | | | H |
| $t_{10}$ | | | C | | E | | | H |
| $t_{11}$ | | | C | | | F | G | H |

| Sol. | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| $s_1$ | {C, F, G, H} | {E} | {A, B, D} |

# Contexte – Clustering conceptuel

**Clustering conceptuel** ≡ ensemble de ***k* concepts formels** $\Phi = \{\phi_1, \ldots, \phi_k\}$, avec $\phi_j = (T_j, I_j)$, tel que $\{T_1, \ldots, T_k\}$ forme une **partition** de l'ensemble de transactions $\mathcal{T}$.

$$(Q_1) \begin{cases} k_{min} \leq k \leq k_{max} \wedge \\ \texttt{cover}(\Phi, \mathcal{T}) = \mathcal{T} \wedge \\ overlap(\Phi, \mathcal{T}) = 0 \wedge \\ \wedge_{\phi \in \Phi} \textsc{ClosedPattern}(\phi) \end{cases}$$

- $k = \texttt{taille}(\Phi) = |\Phi|$
- $cover(\Phi, \mathcal{T}) = \bigcup_{\phi \in \Phi} cover(\phi, \mathcal{T})$
- $overlap(\Phi, \mathcal{T}) = |\bigcap_{\phi \in \Phi} cover(\phi, \mathcal{T})|$

- $k_{min} = 1$, $k_{max} = 3$

| Trans. | Items | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $t_1$ | A | B | | D | | | | |
| $t_2$ | A | | | | E | F | | |
| $t_3$ | A | | | | E | | G | |
| $t_4$ | A | | | | E | | G | |
| $t_5$ | | B | | | E | | G | |
| $t_6$ | | B | | | E | | G | |
| $t_7$ | | | C | | E | | G | |
| $t_8$ | | | C | | E | | G | |
| $t_9$ | | | C | | E | | | H |
| $t_{10}$ | | | C | | E | | | H |
| $t_{11}$ | | | C | | | F | G | H |

| Sol. | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| $s_1$ | {C, F, G, H} | {E} | {A, B, D} |
| $s_2$ | {B} | {C} | {A, E} |

# Contexte – Clustering conceptuel

**Clustering conceptuel** $\equiv$ ensemble de $k$ **concepts formels** $\Phi = \{\phi_1, \ldots, \phi_k\}$, avec $\phi_j = (T_j, I_j)$, tel que $\{T_1, \ldots, T_k\}$ forme une **partition** de l'ensemble de transactions $\mathcal{T}$.

$$(Q_1) \begin{cases} k_{min} \leq k \leq k_{max} \wedge \\ \texttt{cover}(\Phi, \mathcal{T}) = \mathcal{T} \wedge \\ overlap(\Phi, \mathcal{T}) = 0 \wedge \\ \wedge_{\phi \in \Phi} \textsc{ClosedPattern}(\phi) \end{cases}$$

- $k = \texttt{taille}(\Phi) = |\Phi|$
- $cover(\Phi, \mathcal{T}) = \bigcup_{\phi \in \Phi} cover(\phi, \mathcal{T})$
- $overlap(\Phi, \mathcal{T}) = |\bigcap_{\phi \in \Phi} cover(\phi, \mathcal{T})|$

- $k_{min} = 1$, $k_{max} = 3$

| Trans. | Items | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $t_1$ | A | B | | D | | | | |
| $t_2$ | A | | | | E | F | | |
| $t_3$ | A | | | | E | | G | |
| $t_4$ | A | | | | E | | G | |
| $t_5$ | | B | | | E | | G | |
| $t_6$ | | B | | | E | | G | |
| $t_7$ | | | C | | E | | G | |
| $t_8$ | | | C | | E | | G | |
| $t_9$ | | | C | | E | | | H |
| $t_{10}$ | | | C | | E | | | H |
| $t_{11}$ | | | C | | | F | G | H |

| Sol. | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| $s_1$ | {C, F, G, H} | {E} | {A, B, D} |
| $s_2$ | {B} | {C} | {A, E} |
| $s_3$ | {A} | {C} | {B, E, G} |

# Contexte – Clustering conceptuel

| Trans. | Items | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $t_1$ | A | B | | D | | | | |
| $t_2$ | A | | | | E | F | | |
| $t_3$ | A | | | | E | | G | |
| $t_4$ | A | | | | E | | G | |
| $t_5$ | | B | | | E | | G | |
| $t_6$ | | B | | | E | | G | |
| $t_7$ | | | C | | E | | G | |
| $t_8$ | | | C | | E | | G | |
| $t_9$ | | | C | | E | | | H |
| $t_{10}$ | | | C | | E | | | H |
| $t_{11}$ | | | C | | | F | G | H |

| Sol. | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| $s_1$ | {C, F, G, H} | {E} | {A, B, D} |
| $s_2$ | {B} | {C} | {A, E} |
| $s_3$ | {A} | {C} | {B, E, G} |

**Prise en compte de la dimension optimisation :**

# Contexte – Clustering conceptuel

| Trans. | Items | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $t_1$ | A | B | | D | | | | |
| $t_2$ | A | | | | E | F | | |
| $t_3$ | A | | | | E | | G | |
| $t_4$ | A | | | | E | | G | |
| $t_5$ | | B | | | E | | G | |
| $t_6$ | | B | | | E | | G | |
| $t_7$ | | | C | | E | | G | |
| $t_8$ | | | C | | E | | G | |
| $t_9$ | | | C | | E | | | H |
| $t_{10}$ | | | C | | E | | | H |
| $t_{11}$ | | | C | | | F | G | H |

| Sol. | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| $s_1$ | {C, F, G, H} | {E} | {A, B, D} |

**Prise en compte de la dimension optimisation :**

- **Maximiser la taille** : $\sum_{\phi_i \in \Phi} \texttt{taille}(\phi_i)$
  ➠ solution optimale ($s_1$) de valeur 8

# Contexte – Clustering conceptuel

| Trans. | Items | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $t_1$ | A | B | | D | | | | |
| $t_2$ | A | | | | E | F | | |
| $t_3$ | A | | | | E | | G | |
| $t_4$ | A | | | | E | | G | |
| $t_5$ | | B | | | E | | G | |
| $t_6$ | | B | | | E | | G | |
| $t_7$ | | | C | | E | | G | |
| $t_8$ | | | C | | E | | G | |
| $t_9$ | | | C | | E | | | H |
| $t_{10}$ | | | C | | E | | | H |
| $t_{11}$ | | | C | | | F | G | H |

| Sol. | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| $s_1$ | {C, F, G, H} | {E} | {A, B, D} |

**Prise en compte de la dimension optimisation :**

- **Maximiser la taille** : $\sum_{\phi_i \in \Phi} \texttt{taille}(\phi_i)$
  ➠ solution optimale ($s_1$) de valeur 8
- **Minimiser la diversité** : $\sum_{\phi_i \in \Phi} \texttt{divers}(\phi_i)$
  avec $\phi = (T, I)$ et $\texttt{divers}(\phi) = \sum_{t \in T} |\{i \in \mathcal{I} \mid (i \notin I) \wedge (i \in t)\}|$
  ➠ solution optimale ($s_1$) de valeur 18

# Clustering conceptuel : autres variantes

**co-clustering** $\equiv$ ensemble de **concepts formels** $\Phi = \{\phi_1, \ldots, \phi_k\}$, tel que $\Phi$ forme **une partition** sur l'ensemble de transactions $\mathcal{T}$ et l'ensemble d'items $\mathcal{I}$.

$$(Q_2) \begin{cases} k_{min} \leq k \leq k_{max} \ \wedge \\ \wedge_{\phi \in \Phi} \ \textsc{ClosedPattern}(\phi) \ \wedge \\ \texttt{cover}(\Phi, \mathcal{T}) = |\mathcal{T}| \ \wedge \texttt{cover}(\Phi, \mathcal{I}) = |\mathcal{I}| \ \wedge \\ \mathit{overlap}(\Phi, \mathcal{T}) = 0 \ \wedge \mathit{overlap}(\Phi, \mathcal{I}) = 0 \end{cases}$$

# Clustering conceptuel : autres variantes

**co-clustering** $\equiv$ ensemble de **concepts formels** $\Phi = \{\phi_1, \ldots, \phi_k\}$, tel que $\Phi$ forme **une partition** sur l'ensemble de transactions $\mathcal{T}$ et l'ensemble d'items $\mathcal{I}$.

$$(Q_2) \begin{cases} k_{min} \leq k \leq k_{max} \wedge \\ \wedge_{\phi \in \Phi} \textsc{ClosedPattern}(\phi) \wedge \\ \texttt{cover}(\Phi, \mathcal{T}) = |\mathcal{T}| \ \wedge \texttt{cover}(\Phi, \mathcal{I}) = |\mathcal{I}| \wedge \\ \mathit{overlap}(\Phi, \mathcal{T}) = 0 \ \wedge \mathit{overlap}(\Phi, \mathcal{I}) = 0 \end{cases}$$

**Soft clustering** $\equiv$ consiste à **relâcher** la couverture ($\delta_t \leq \mathcal{T}$) et/ou le non-chevauchement ($\delta_o \geq 1$) sur l'ensemble des transactions.

$$(Q_3) \begin{cases} k_{min} \leq k \leq k_{max} \wedge \\ \wedge_{\phi \in \Phi} \textsc{ClosedPattern}(\phi) \wedge \\ \texttt{cover}(\Phi, \mathcal{T}) \geq \delta_t \ \wedge \mathit{overlap}(\Phi, \mathcal{T}) \leq \delta_o \end{cases}$$

# Modèle PLNE pour le clustering conceptuel [Ouali et al. (IJCAI'16)]

Approche hybride en deux étapes :

1. Extraire l'ensemble $\mathcal{C}$ de tous les concepts formels à l'aide de `LCM` [Uno et al 2004]
2. Rechercher le meilleur clustering qui est une partition de $\mathcal{T}$ avec la `PLNE`

| | |
|---|---|
| `Optimiser` | $z = \sum_{c=1}^{\lvert\mathcal{C}\rvert} v_c \,.\, x_c$ |
| `Sous contraintes` | (1) $\sum_{c=1}^{\lvert\mathcal{C}\rvert} a_{t,c} \,.\, x_c = 1, \quad \forall t \in \mathcal{T}$<br>(2) $\sum_{c=1}^{\lvert\mathcal{C}\rvert} x_c = k$<br>(3) $k_{min} \leq k \leq k_{max}$<br>$k \in \mathbb{N}, \quad x_c \in \{0,1\}, c \in \mathcal{C}$ |

- $(x_c = 1)$ ssi le motif fermé $c$ **appartient** au clustering.
- $v_c$ : valeur exprimant **l'intérêt** du motif fermé $c$ : taille, diversité, aire, ...
- $(a_{t,c})$ matrice booléenne : $(a_{t,c} = 1)$ ssi le motif $c$ **couvre** la transaction $t$.
- $k$ (**nombre de clusters**) est une variable.
- La contrainte (3) permet de contrôler le nombre $k$ de clusters

# Modèle PLNE pour le co-clustering

- $(w_{i,c})$ matrice booléenne : $(w_{i,c} = 1)$ ssi le motif $c$ couvre l'item $i$.

| | |
|---|---|
| `Optimiser` | $z = \sum_{c=1}^{\lvert\mathcal{C}\rvert} v_c \,.\, x_c$ |
| `Sous contraintes` | (1) $\sum_{c=1}^{\lvert\mathcal{C}\rvert} a_{t,c} \,.\, x_c = 1, \quad \forall t \in \mathcal{T}$<br>(2) $k = \sum_{c=1}^{\lvert\mathcal{C}\rvert} x_c$<br>(3) $k_{min} \leq k \leq k_{max}$<br>(4) $\sum_{c=1}^{\lvert\mathcal{C}\rvert} w_{i,c} \,.\, x_c = 1, \quad \forall i \in \mathcal{I}$<br>$k \in \mathbb{N}, \quad x_c \in \{0,1\}, c \in \mathcal{C}$ |

| | $\phi_1$ | $\phi_2$ | $\phi_3$ | $\cdots$ | $\phi_{18}$ |
|---|---|---|---|---|---|
| $t_1$ | 1 | 1 | 0 | $\cdots$ | 0 |
| $t_2$ | 1 | 0 | 1 | $\cdots$ | 0 |
| $t_3$ | 1 | 0 | 1 | $\cdots$ | 1 |
| $t_4$ | 1 | 0 | 1 | $\cdots$ | 1 |
| $t_5$ | 0 | 0 | 0 | $\cdots$ | 1 |
| $t_6$ | 0 | 0 | 0 | $\cdots$ | 1 |
| $t_7$ | 0 | 0 | 0 | $\cdots$ | 1 |
| $t_8$ | 0 | 0 | 0 | $\cdots$ | 1 |
| $t_9$ | 0 | 0 | 0 | $\cdots$ | 0 |
| $t_{10}$ | 0 | 0 | 0 | $\cdots$ | 0 |
| $t_{11}$ | 0 | 0 | 0 | $\cdots$ | 1 |

(a) Matrice $(a_{t,c})$

| | $\phi_1$ | $\phi_2$ | $\phi_3$ | $\cdots$ | $\phi_{18}$ |
|---|---|---|---|---|---|
| A | 0 | 0 | 0 | $\cdots$ | 1 |
| B | 0 | 0 | 0 | $\cdots$ | 1 |
| C | 0 | 0 | 0 | $\cdots$ | 0 |
| D | 0 | 0 | 0 | $\cdots$ | 1 |
| E | 1 | 0 | 1 | $\cdots$ | 0 |
| F | 0 | 0 | 0 | $\cdots$ | 0 |
| G | 0 | 1 | 1 | $\cdots$ | 0 |
| H | 0 | 0 | 0 | $\cdots$ | 0 |

(b) Matrice $(w_{i,c})$

- La contrainte (4) impose le partitionnement des items de $\mathcal{I}$.

# Modèle PLNE pour le soft clustering

- Relâcher :
  - soit **la relation de couverture** : les transactions ne sont pas toutes couvertes, mais au moins $\delta_t$ doivent l'être,
  - soit **la relation de non-chevauchement** : une transaction ne peut apparaître dans plus de $\delta_o$ clusters.

| | |
|---|---|
| Optimiser | $\sum_{c=1}^{\lvert\mathcal{C}\rvert} v_c \cdot x_c$ |
| Sous contraintes | **(5)** $\mathbf{y_t} \leq \sum_{\mathbf{c=1}}^{\lvert\mathcal{C}\rvert} \mathbf{a_{t,c} \cdot x_c} \leq \delta_\mathbf{o} \cdot \mathbf{y_t}, \quad \forall t \in \mathcal{T}$ <br> **(6)** $\sum_{\mathbf{t}\in\mathcal{T}} \mathbf{y_t} \geq \delta_\mathbf{t}$ <br> $k = \sum_{c=1}^{\lvert\mathcal{C}\rvert} x_c$ <br> $k_{min} \leq k \leq k_{max}$ <br> $k \in \mathbb{N},$ <br> $x_c \in \{0,1\}, \quad c \in \mathcal{C}$ <br> $y_t \in \{0,1\}, \quad t \in \mathcal{T}$ |

- $(y_t = 1)$ ssi la transaction $t$ **appartient** à au moins un cluster représenté par le clos $c$.
- La contrainte (6) stipule qu'au moins $\delta_t$ transactions doivent être couvertes.
- La contrainte (5) spécifie que toute transaction $t$ doit **appartenir** à au plus $\delta_o$ clusters.

# Expérimentations : Jeux de données

| datasets | #transactions | #items | densité(%) | #motifs locaux |
|---|---|---|---|---|
| Soybean | 630 | 50 | 32 | 31,759 |
| Primary-tumor | 336 | 31 | 48 | 87,230 |
| Lymph | 148 | 68 | 40 | 154,220 |
| Vote | 435 | 48 | 33 | 227,031 |
| tic-tac-toe | 958 | 27 | 33 | 42,711 |
| Mushroom | **8124** | 119 | 18 | 221,524 |
| Zoo-1 | 101 | 36 | 44 | 4,567 |
| Hepatitis | 137 | 68 | 50 | **3,788,341** |
| Anneal | 812 | 93 | 45 | **1,805,193** |

TABLE – Caractéristiques des différents jeux de données.

# Expérimentations : Temps CPU

- Solutions optimales sur des bases de données avec des millions de clos
- Nombre de clusters $k$ n'est pas fixé : $3 \leq k \leq 10$

| Instance | PSILP-M1-Relâché Maximiser la taille de la description | | | PSILP-M1 Maximiser la taille de la description | | | PSILP-M1-Distance Maximiser la moyenne des similarités internes des clusters | | |
|---|---|---|---|---|---|---|---|---|---|
| | best $k$ | Temps CPU (s.) | | best $k$ | Temps CPU (s.) | | best $k$ | Temps CPU (s.) | |
| | | (1) | (2) | | (1) | (2) | | (1) | (2) |
| Soybean | 10 | 0.74 | **15.52** | 10 | 0.74 | 21.31 | 10 | 0.74 | 8.52 |
| Primary-tumor | 10 | 1.98 | 40.62 | 10 | 1.98 | **35.84** | 8 | 1.98 | 35.4 |
| Lymph | 10 | 3.38 | **25** | 10 | 3.38 | 93.96 | 10 | 3.38 | 149.69 |
| Vote | 10 | 4.42 | **191.12** | 10 | 4.42 | 304.18 | 9 | 4.42 | 66.21 |
| Zoo-1 | 10 | 0.07 | 6.22 | 10 | 0.07 | **2.73** | 9 | 0.07 | 0.35 |
| Mushroom | 10 | 17.1 | **154.75** | 10 | 17.1 | 163.67 | 10 | 17.1 | 169.04 |
| Tic-Tac-Toe | 10 | 0.38 | 420.75 | 10 | 0.38 | **398.86** | 10 | 0.38 | 209.4 |
| Anneal | 10 | 236.85 | **1,493.38** | 10 | 236.85 | 5,744.03 | 10 | 236.85 | 5,744.03 |
| Hepatitis | 10 | 113.84 | **20,825.8** | 10 | 113.84 | 40,014.5 | 10 | 113.84 | 37,745.7 |

# Expérimentations : Qualité des clusterings

- Distance de Jaccard :

$$s : \mathcal{T} \times \mathcal{T} \mapsto [0,1], \quad s(t,t') = \frac{|t \cap t'|}{|t \cup t'|}$$

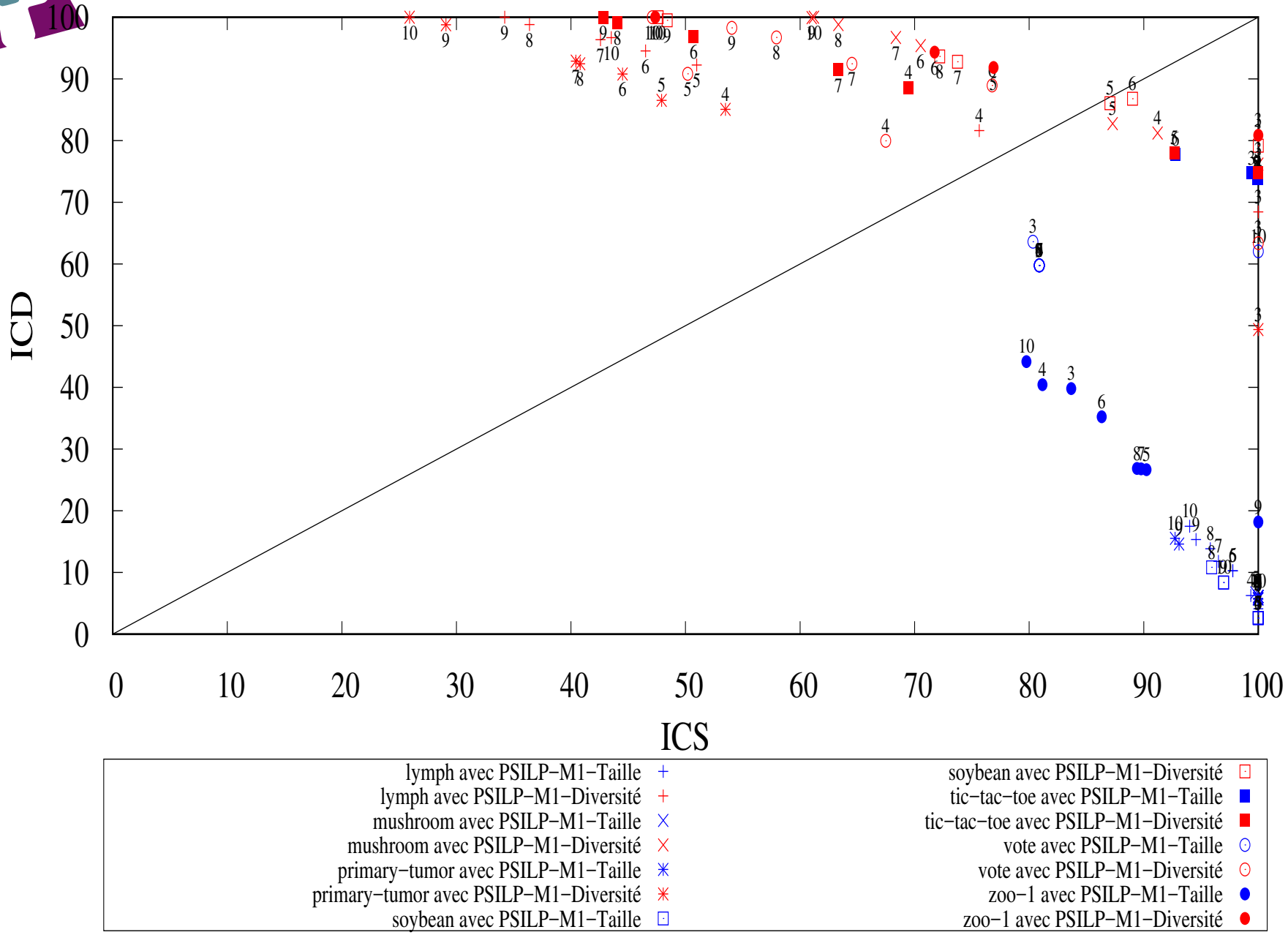
- Similarité intra-cluster :

$$ICS(\phi_1, ..., \phi_k) = \frac{1}{2} \sum_{1 \leq i \leq k} ( \sum_{t,t' \in \phi_i} s(t,t'))$$

- Dissimilarité inter-clusters :

$$ICD(\phi_1, ..., \phi_k) = \sum_{1 \leq i < j \leq k} ( \sum_{t \in \phi_i, t' \in \phi_j} (1 - s(t,t')))$$

# Comparaison entre **différentes mesures**

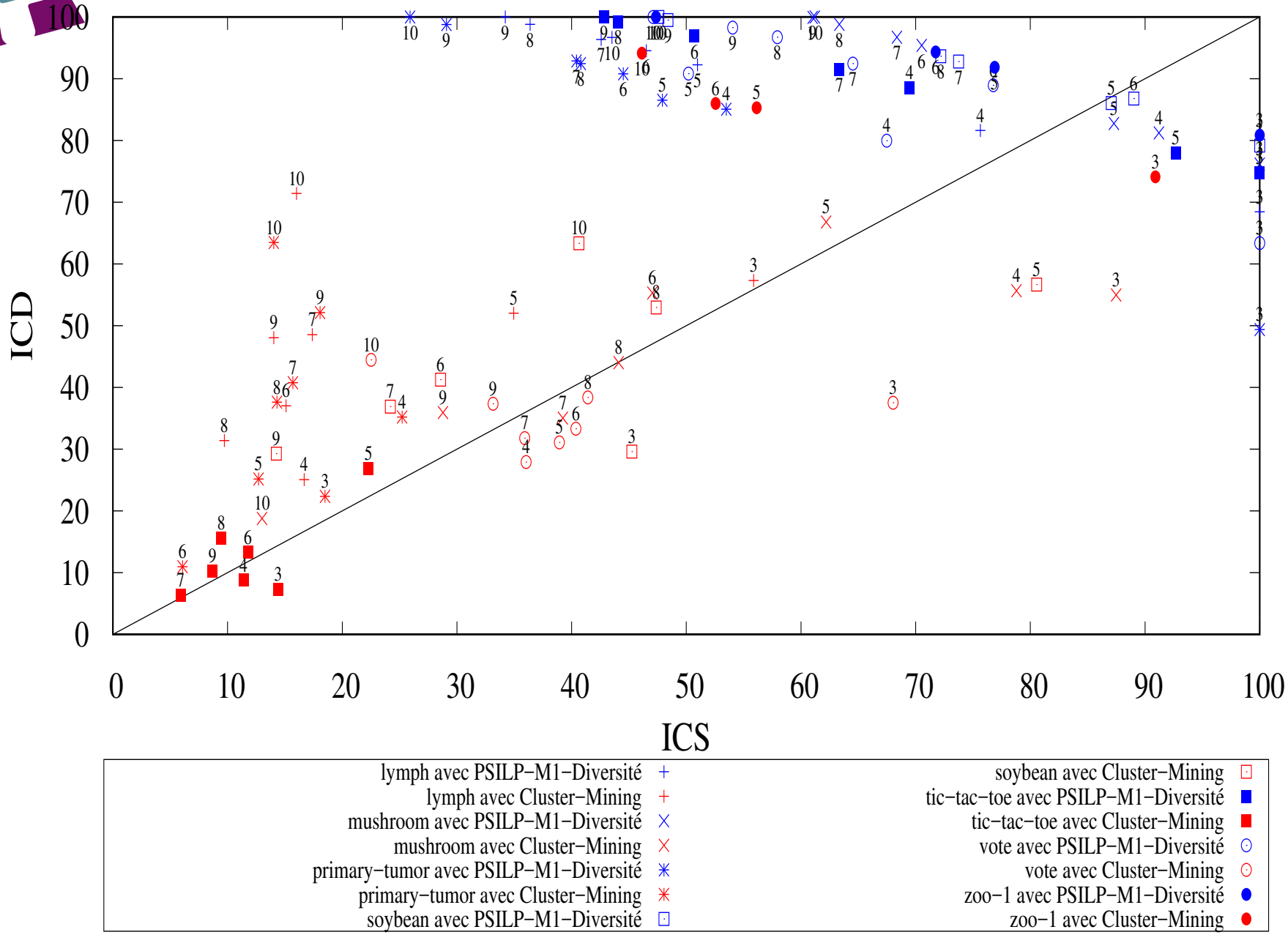


# Comparaison avec **Cluster Mining** (1/3)

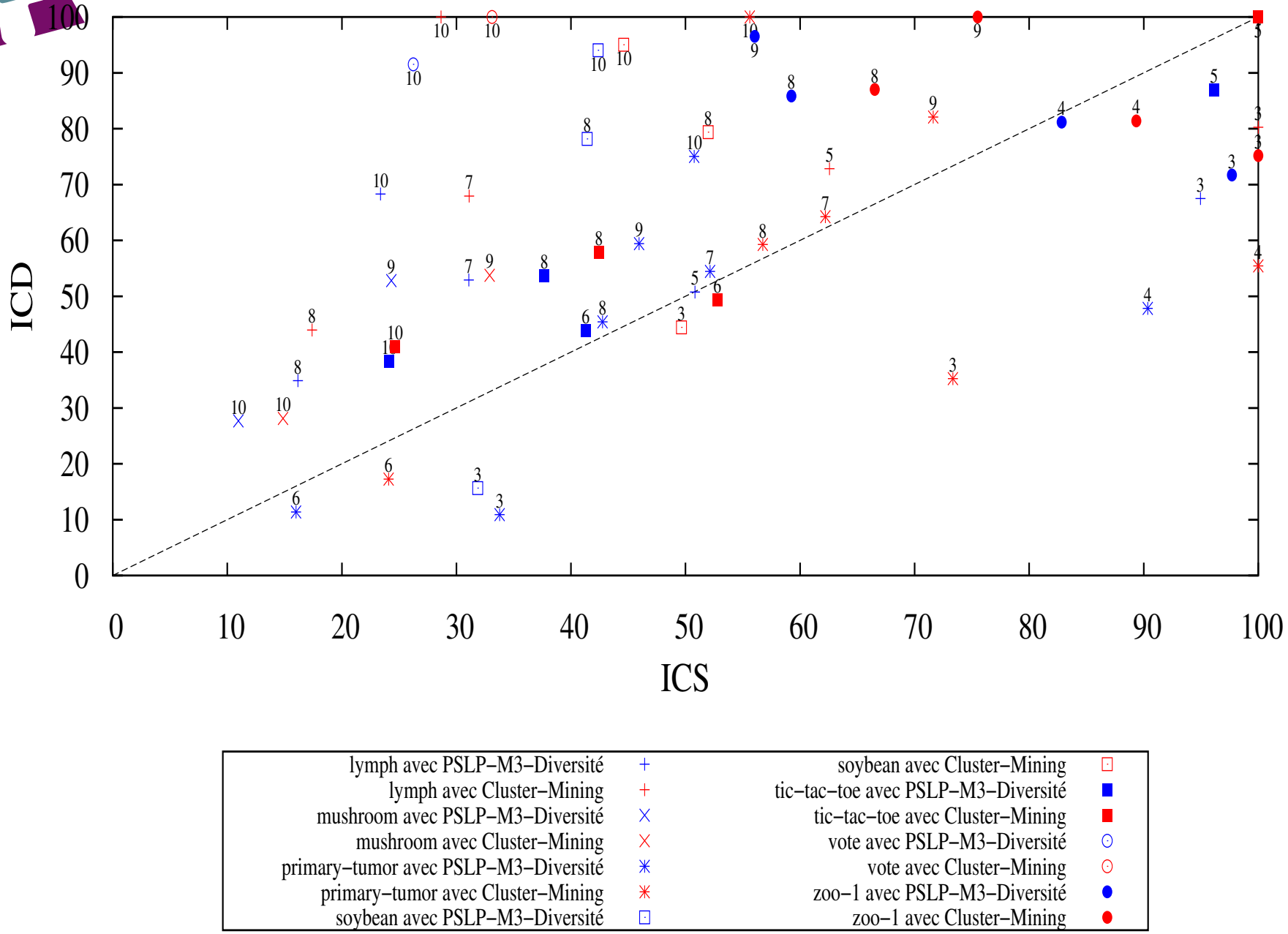
Choix du meilleur clustering pour **Cluster Mining** :

1. Exécuter chaque méthode 100 fois pour lisser les effets de l'initialisation aléatoire de k-Means,
2. Regrouper les résultats en classes d'équivalence dans lesquelles les clusterings sont formés approximativement sur les mêmes clusters,
3. Choisir la plus grande classe d'équivalence et retenir son représentant (un des clusterings de cette classe puisqu'ils ont la même composition).

# Comparaison avec **Cluster Mining** (2/3)



# Comparaison avec **Cluster Mining** (3/3)



# Clustering conceptuel équitable : retour sur l'exemple

| Trans. | Items | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $t_1$ | A | B | | D | | | | |
| $t_2$ | A | | | | E | F | | |
| $t_3$ | A | | | | E | | G | |
| $t_4$ | A | | | | E | | G | |
| $t_5$ | | B | | | E | | G | |
| $t_6$ | | B | | | E | | G | |
| $t_7$ | | | C | | E | | G | |
| $t_8$ | | | C | | E | | G | |
| $t_9$ | | | C | | E | | | H |
| $t_{10}$ | | | C | | E | | | H |
| $t_{11}$ | | | C | | | F | G | H |

| Sol. | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| $s_1$ | {C, F, G, H} | {E} | {A, B, D} |
| $s_2$ | {B} | {C} | {A, E} |
| $s_3$ | {A} | {C} | {B, E, G} |

- Clusterings trouvés souvent favorisent un critère particulier au détriment des autres

# Clustering conceptuel équitable : retour sur l'exemple

| Trans. | Items | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $t_1$ | A | B | | D | | | | |
| $t_2$ | A | | | | E | F | | |
| $t_3$ | A | | | | E | | G | |
| $t_4$ | A | | | | E | | G | |
| $t_5$ | | B | | | E | | G | |
| $t_6$ | | B | | | E | | G | |
| $t_7$ | | | C | | E | | G | |
| $t_8$ | | | C | | E | | G | |
| $t_9$ | | | C | | E | | | H |
| $t_{10}$ | | | C | | E | | | H |
| $t_{11}$ | | | C | | | F | G | H |

| Sol. | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| $s_1$ | {C, F, G, H} | {E} | {A, B, D} |
| $s_2$ | {B} | {C} | {A, E} |
| $s_3$ | {A} | {C} | {B, E, G} |

- Clusterings trouvés souvent favorisent un critère particulier au détriment des autres
  ➠ des clusters ayant une grande description mais de petite fréquence

# Clustering conceptuel équitable : retour sur l'exemple

| Trans. | Items | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $t_1$ | A | B | | D | | | | |
| $t_2$ | A | | | | E | F | | |
| $t_3$ | A | | | | E | | G | |
| $t_4$ | A | | | | E | | G | |
| $t_5$ | | B | | | E | | G | |
| $t_6$ | | B | | | E | | G | |
| $t_7$ | | | C | | E | | G | |
| $t_8$ | | | C | | E | | G | |
| $t_9$ | | | C | | E | | | H |
| $t_{10}$ | | | C | | E | | | H |
| $t_{11}$ | | | C | | | F | G | H |

| Sol. | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| $s_1$ | {C, F, G, H} | {E} | {A, B, D} |
| $s_2$ | {B} | {C} | {A, E} |
| $s_3$ | {A} | {C} | {B, E, G} |

- Clusterings trouvés souvent favorisent un critère particulier au détriment des autres
- **Minimiser la diversité** : une solution optimale $s_1 = (1, 1, 9)$, coût total 18
  ➠ solution non équilibrée

# Clustering conceptuel équitable : retour sur l'exemple

| Trans. | Items | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $t_1$ | A | B | | D | | | | |
| $t_2$ | A | | | | E | F | | |
| $t_3$ | A | | | | E | | G | |
| $t_4$ | A | | | | E | | G | |
| $t_5$ | | B | | | E | | G | |
| $t_6$ | | B | | | E | | G | |
| $t_7$ | | | C | | E | | G | |
| $t_8$ | | | C | | E | | G | |
| $t_9$ | | | C | | E | | | H |
| $t_{10}$ | | | C | | E | | | H |
| $t_{11}$ | | | C | | | F | G | H |

| Sol. | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| $s_1$ | {C, F, G, H} | {E} | {A, B, D} |
| $s_2$ | {B} | {C} | {A, E} |
| $s_3$ | {A} | {C} | {B, E, G} |

- Clusterings trouvés souvent favorisent un critère particulier au détriment des autres
- **Minimiser la diversité** : une solution optimale $s_1 = (1, 1, 9)$, coût total 18
- Une solution plus *équilibrée* : $s_2 = (3, 5, 3)$, coût total 20

# Clustering conceptuel équitable : retour sur l'exemple

| Trans. | Items | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $t_1$ | A | B | | D | | | | |
| $t_2$ | A | | | | E | F | | |
| $t_3$ | A | | | | E | | G | |
| $t_4$ | A | | | | E | | G | |
| $t_5$ | | B | | | E | | G | |
| $t_6$ | | B | | | E | | G | |
| $t_7$ | | | C | | E | | G | |
| $t_8$ | | | C | | E | | G | |
| $t_9$ | | | C | | E | | | H |
| $t_{10}$ | | | C | | E | | | H |
| $t_{11}$ | | | C | | | F | G | H |

| Sol. | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| $s_1$ | {C, F, G, H} | {E} | {A, B, D} |
| $s_2$ | {B} | {C} | {A, E} |
| $s_3$ | {A} | {C} | {B, E, G} |

- Clusterings trouvés souvent favorisent un critère particulier au détriment des autres
- **Minimiser la diversité** : une solution optimale $s_1 = (1, 1, 9)$, coût total 18
- Une solution plus *équilibrée* : $s_2 = (3, 5, 3)$, coût total 20
- Considérer des **critères d'optimisation dédiés** pour trouver des clusterings équilibrés
  - **Maxmin** (sur la fréquence) ➠ plusieurs défauts : $(1, 1, 1, 100)$ et $(100, 100, 100, 1)$

# Équité et optimisation multi-agents

P un problème d'optimisation mono-critère multi-agents

- $n$ agents $N = \{1, .., n\}$ à considérer pour optimiser P
- solution de P $\Leftrightarrow$ vecteur d'utilité $x = (x_1, ..., x_n) \in R^n_+$ ($x_i$ = utilité de l'agent $i$)
- comparer les solutions = comparer les vecteurs d'utilité

Optimisation équitable d'un tel problème, équilibre entre :

- juste répartition des utilités entre les agents
- mais également performance globale (i.e. solution efficace)

Largement étudié en optimisation

- Allocation équitable de ressources indivisibles (Bouveret et al. AAMAS'05)
- Paper assignment problems (Goldsmith et al. wrshAAAI'07, Lian et al. AAAI'18)
- ...

# Formalisation de l'équité

Le principe d'équité traduit l'idée de favoriser des solutions Pareto-optimales qui équilibrent au mieux la satisfaction ou la dissatisfaction des agents :

- **Principe de transfert** pour capturer la notion de "juste répartition" des utilités [Shorrocks, Economica 1983] :

Soit $x = (x_1, x_2, \ldots, x_i, \ldots, x_j, \ldots, x_n) \in R^n_+$ tel que $x_i > x_j$. Alors pour $\epsilon$ tel que $0 < \epsilon \leq \frac{x_i - x_j}{2}$, $x - \epsilon e_i + \epsilon e_j \succ x$ où $e_i$ (resp. $e_j$) est le vecteur dont la *ieme* (resp. *jeme*) composante vaut 1, toutes les autres étant nulles.

➠ La solution $y = (\mathbf{9}, 10, \mathbf{9}, 10)$ devrait être préférée à $x = (\mathbf{11}, 10, \mathbf{7}, 10)$ car il y a un transfert de $\epsilon = 2$ (i.e. $\frac{11-7}{2}$), qui permet d'obtenir $y$ à partir de $x$.

- **Symétrie** pour formaliser le fait que tous les agents sont traités de manière équivalente :

Pour tout $x \in R^n_+$, pour toute permutation $\sigma$ de $\{1, \ldots, n\}$, $(x_{\sigma(1)}, ..., x_{\sigma(n)}) \sim (x_1, ..., x_n)$.

➠ les deux vecteurs d'utilité (5,3,0) et (0,3,5) sont considérés équivalents.

- **P-Monotonie** pour garantir la compatibilité avec la dominance de Pareto :

Pour tout $x, y \in R^n_+$, $x \succsim_P y \Rightarrow x \succsim_{\|} y$ et $x \succ_P y \Rightarrow x \succ_{\|} y$.

# Dominance de Lorenz généralisée
## [Marshall et Olkin, 1979 ; Shorrocks, 1983]

- Principe de transfert ne permet pas d'établir de préférence entre deux vecteurs ayant des utilités moyennes différentes.
- Combiner la P-monotonie avec le principe de transfert pour choisir les vecteurs les plus équilibrés et les plus efficaces.

  Soit $w = (9, 10, 11, 10)$ et $z = (11, 10, 7, 8)$, lequel le plus équitable ?
  On va utiliser les vecteurs $x = (11, 10, 7, 10)$ et $y = (9, 10, 9, 10)$.
  - $w \succ_P y$ (par P-monotonie)
  - $y \succ_{\|} x$ (par le principe de transfert)
  - $x \succ_P z$ par P-monotonie

  ➠ $w \succ_{\|} z$ (par *transitivité*)
- Une relation de dominance équitable $\rightsquigarrow$ dominance de Lorenz généralisée :

$$\forall x, y \in R^n_+, x \succsim_L y \Longleftrightarrow L(x) \succsim_P L(y)$$

avec $L_i(x) = \sum_{j=1}^{i} x_{\sigma(j)}$ et $\sigma$ une permutation sur $\{1, \ldots, n\}$ telle que $x_{\sigma(1)} \leq x_{\sigma(2)} \leq \ldots \leq x_{\sigma(n)}$

# L'ordre infini de la dominance de Lorenz généralisée

- L-dominance d'ordres supérieurs plus **discriminantes** en **réitérant** la transformation **L(.)** : $L^2(x) = L(L(x))$

  Exemple : $x = (9, 6, 8, 12)$ et $y = (10, 8, 6, 10)$ :

  - $L(x) = (6, 14, 23, 35)$ et $L(y) = (6, 14, 24, 34) \rightarrow$ pas de dominance
  - $L^2(x) = (6, 20, 43, 78)$ et $L^2(y) = (6, 20, 44, 78) \rightarrow x \succ_L^2 y$

- L'ordre **infini** de la dominance de Lorenz généralisée : $\succ_L^\infty = \bigcup_{k \geq 1} \succ_L^k$

- Comment savoir quand terminer ?
  ➥ exprimer $L^\infty$ directement sur les vecteurs

# Somme ordonnée pondérée (OWA)

- OWA est une **généralisation** de certaines fonctions d'agrégation comme la **moyenne** et **Maximin**.
- Permet de **pondérer** les critères relativement à leur **rang** [Yager 88] :

$$W(x) = \sum_{k=1}^{n} w_i x_{\sigma(i)}$$

avec $w = (w_1, \ldots, w_n) \in [0,1]^n$ et $x_{\sigma(1)} \leq x_{\sigma(2)} \leq \cdots \leq x_{\sigma(n)}$

- Golden et Perny (2010) ont permis d'établir un lien entre l'agrégateur OWA et l'ordre infini de la dominance généralisée de Lorenz :

$$\forall x, y \in R_+^n, x \succeq_L^\infty y \Leftrightarrow W(x) > W(y)$$

avec $W(x) = \sum_{i=1}^{n} sin(\frac{(n+1-i)\pi}{2n+1}) x_{\sigma(i)}$

# Modèle PLNE pour le clustering conceptuel équilibré [Ouali et al. (PAKDD'18)]

- Chaque agent représente un concept et sa propre utilité correspondant à la fréquence du concept

$$\texttt{Max} \sum_{c=1}^{|\mathcal{C}|} \omega_c \,.\, r_c$$

$$\text{s.t.} \begin{cases} \text{Clustering.} \begin{cases} \text{(C1)} \quad \sum_{c=1}^{|\mathcal{C}|} a_{t,c} \,.\, x_c = 1, & \forall t \in \mathcal{T} \\ \text{(C2)} \quad k_{min} \leq \sum_{c=1}^{|\mathcal{C}|} x_c \leq k_{max} \end{cases} \\ \texttt{OWA} \text{ sorting.} \begin{cases} \text{(O1)} \quad r_c - (v_i \,.\, x_i) \leq M \times z_{c,i}, & \forall\, i, c = 1, ..., |\mathcal{C}| \\ \text{(O2)} \quad \sum_{i=1}^{|\mathcal{C}|} z_{c,i} \leq c - 1, & \forall\, c = 1, ..., |\mathcal{C}| \end{cases} \\ x_c \in \{0,1\},\ r_c \in R^+, \quad \forall\, c = 1, ..., |\mathcal{C}| \\ z_{c,i} \in \{0,1\}, \quad \forall\, i, c = 1, ..., |\mathcal{C}| \end{cases}$$

# Modèle PLNE pour le clustering conceptuel équilibré [Ouali et al. (PAKDD'18)]

- Chaque agent représente un concept et sa propre utilité correspondant à la fréquence du concept

$$\texttt{Max} \sum_{c=1}^{|\mathcal{C}|} \omega_c \,.\, r_c$$

$$\text{s.t.} \begin{cases} \text{Clustering.} \begin{cases} \text{(C1)} & \sum_{c=1}^{|\mathcal{C}|} a_{t,c} \,.\, x_c = 1, & \forall t \in \mathcal{T} \\ \text{(C2)} & k_{min} \leq \sum_{c=1}^{|\mathcal{C}|} x_c \leq k_{max} \end{cases} \\ \texttt{OWA} \text{ sorting.} \begin{cases} \text{(O1)} & r_c - (v_i \,.\, x_i) \leq M \times z_{c,i}, & \forall\, i, c = 1, ..., |\mathcal{C}| \\ \text{(O2)} & \sum_{i=1}^{|\mathcal{C}|} z_{c,i} \leq c - 1, & \forall\, c = 1, ..., |\mathcal{C}| \end{cases} \\ x_c \in \{0,1\},\ r_c \in R^+, \quad \forall\, c = 1, ..., |\mathcal{C}| \\ z_{c,i} \in \{0,1\}, \quad \forall\, i, c = 1, ..., |\mathcal{C}| \end{cases}$$

➠ requière $(|\mathcal{T}| + |\mathcal{C}|^2 + 1)$ contraintes et $(|\mathcal{C}| + |\mathcal{C}|^2 + 1)$ variables

# Modèle PLNE simplifié pour le clustering conceptuel équilibré

$$\texttt{Max} \sum_{c=1}^{|\mathcal{C}|} \omega_c \,.\, (v_c^{\uparrow} \,.\, x_c^{\uparrow})$$

$$\text{s.t.} \begin{cases} (\text{C1}), \quad (\text{C2}) \\ x_c \in \{0, 1\}, \\ \quad \forall\ c = 1, ..., |\mathcal{C}| \end{cases}$$

- **Contraintes de tri.** Les valeurs d'utilité sont connues à l'avance. Le tri est **effectué immédiatement après** la recherche de motifs fermés.

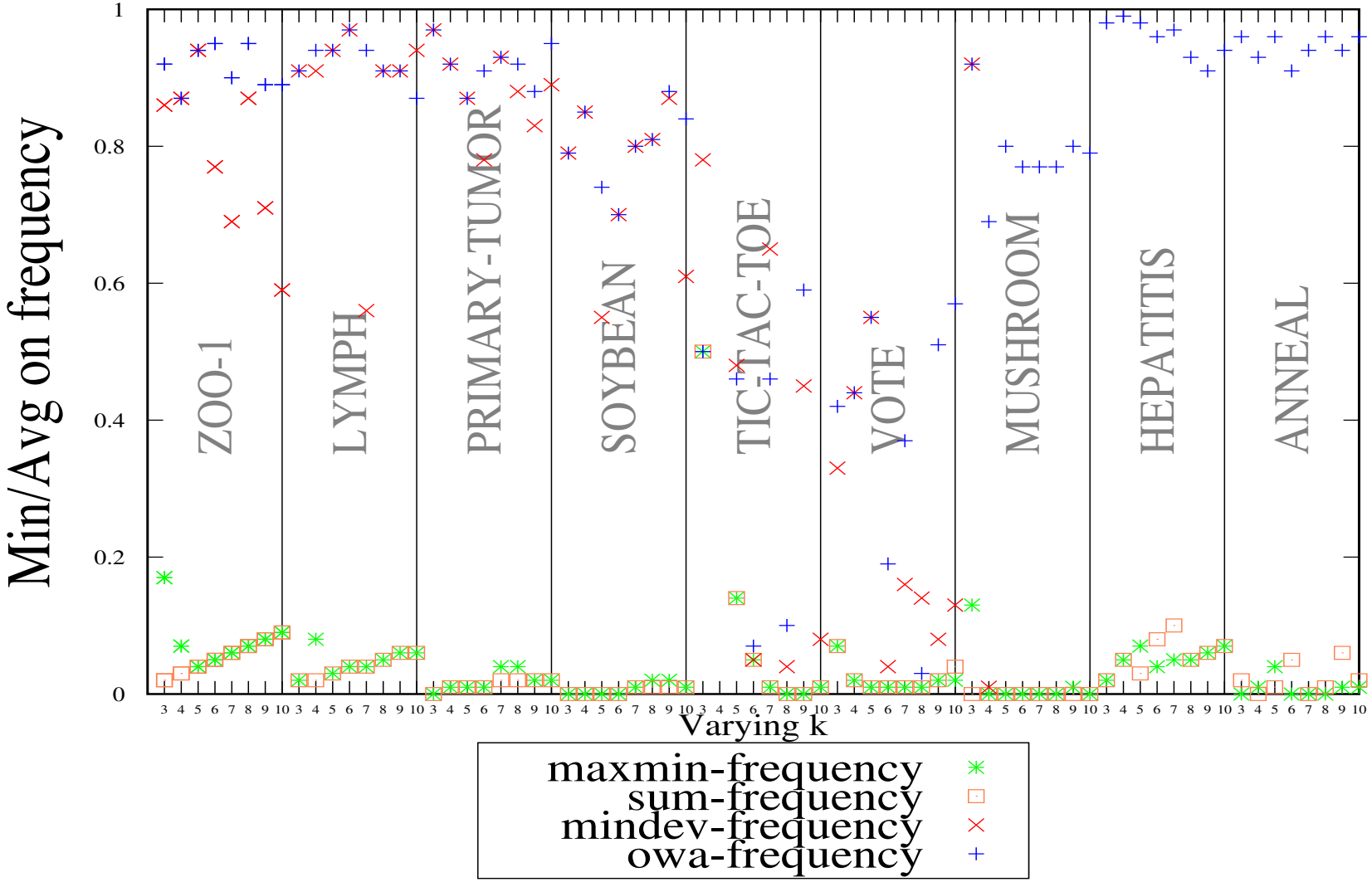
# Expérimentations : Qualité des clusterings (1/4)

- Ratio entre la fréquence Min et Avg (Min/Avg)
  ➥ Pour $m$ transactions, Avg $\equiv (m/k)$.

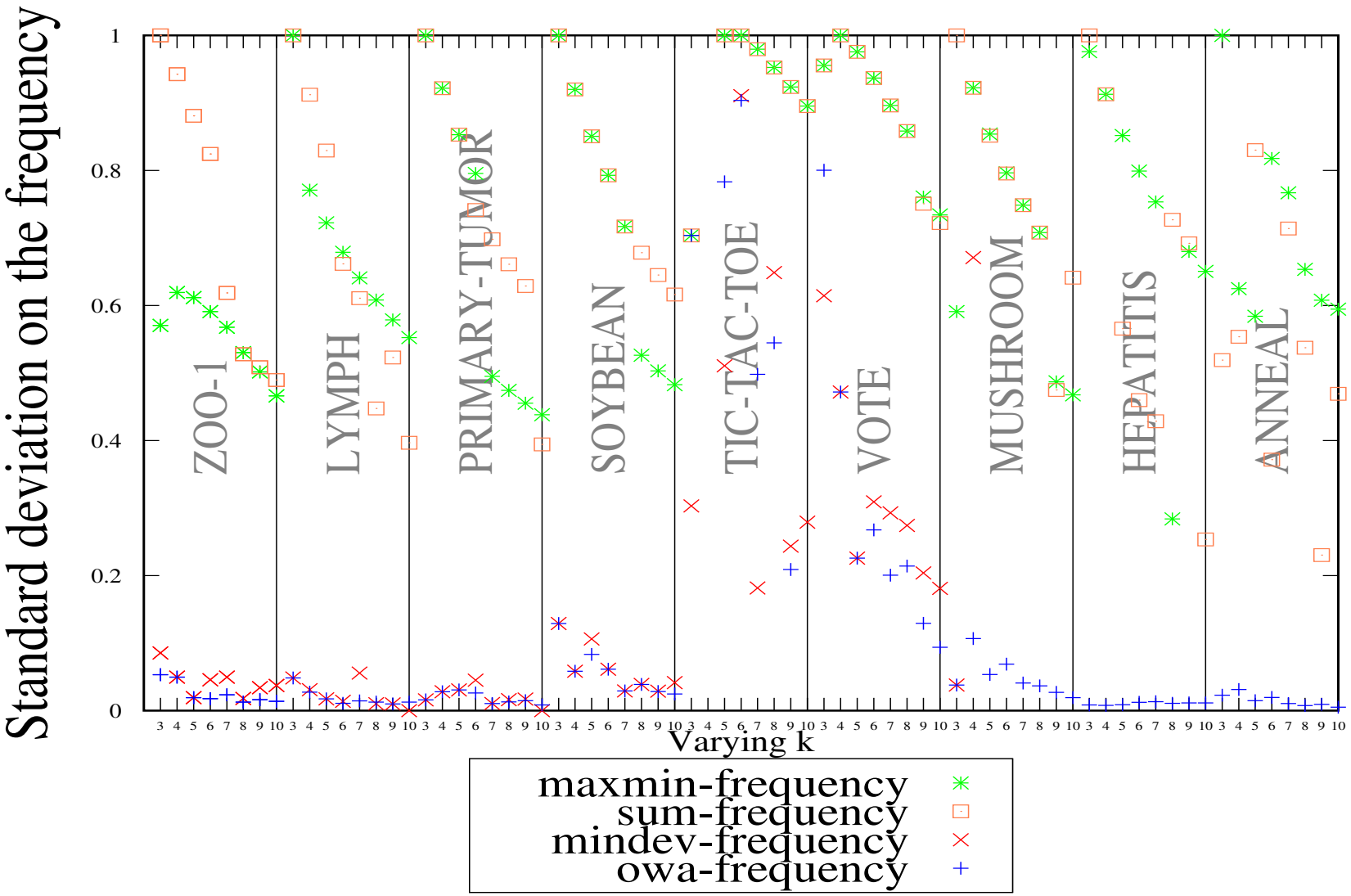
- Déviation Standard (i.e. StdDev)

- Déviation entre la plus petite et la plus grande description (devSize)

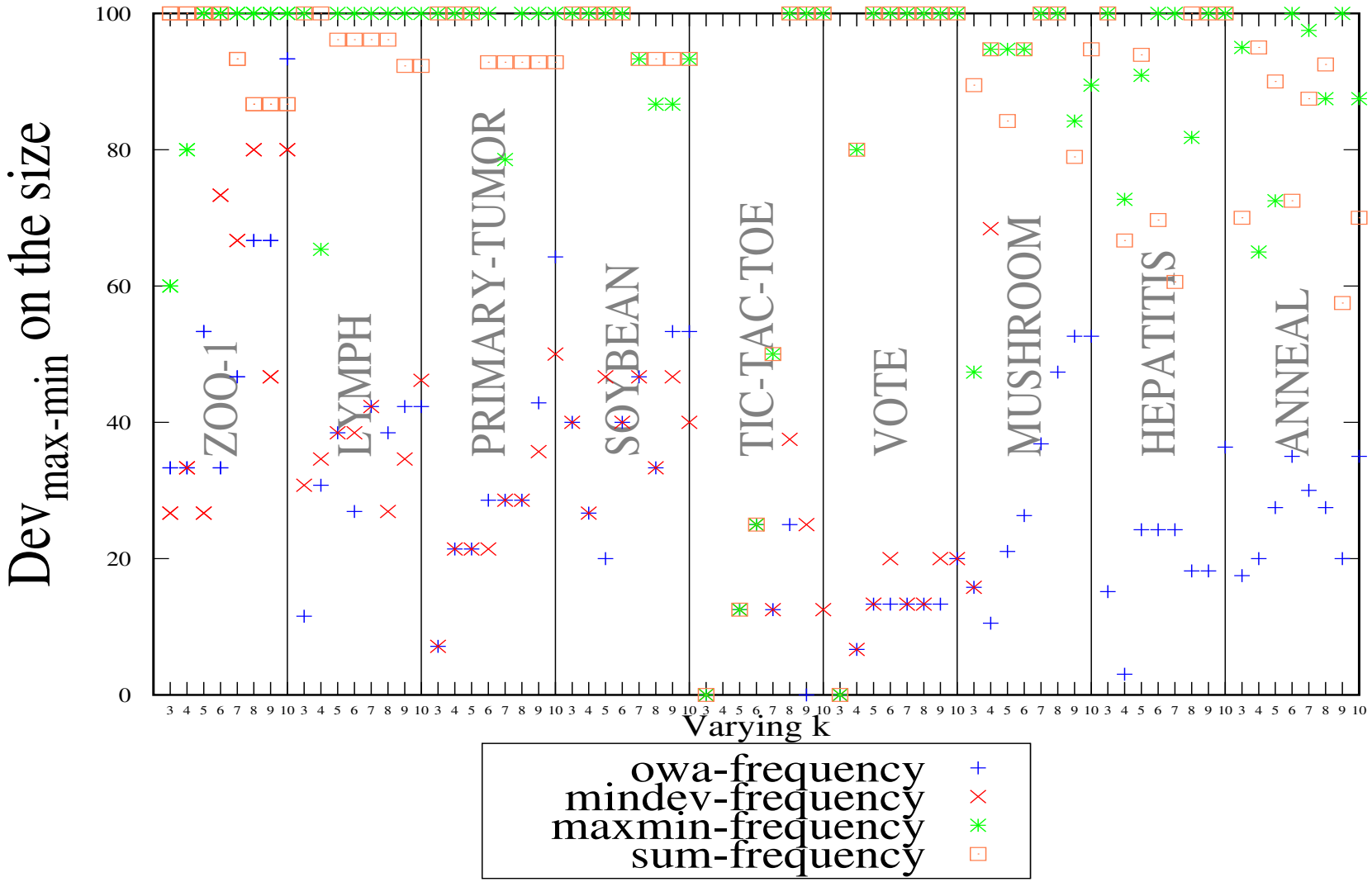
# Expérimentations : Qualité des clusterings (2/4)



# Expérimentations : Qualité des clusterings (3/4)



# Expérimentations : Qualité des clusterings (4/4)



Merci !