This thesis is funded by the French National Research Agency (ANR) through the FIDD project, grant ANR-24-CE23-0711.

HETEROGENEOUS PATTERN SAMPLING ACCORDING TO FREQUENCY

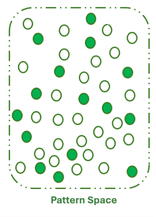
Rayane Lachache[1], Djawad Bekkoucha[1,3], Abdelkader Ouali[1], Bruno Crémilleux[1], Thi-Bich-Hanh Dao[2], and Christel Vrain[2]

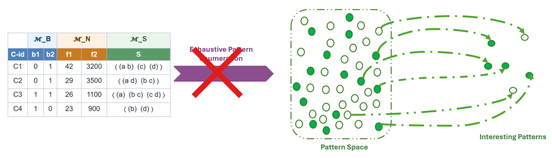
[1]University of Caen Normandy, CNRS, UMR 6072 GREYC, 14032 Caen, France
[2]University of Orléans, INSA Centre Val de Loire, LIFO EA 4022, Orléans, France
[3]LISN, CNRS (UMR 9015), Paris Saclay University, Gif-sur-Yvette, France
rayane.lachache@unicaen.fr

1. Motivation

	$\mathcal{M}_B$		$\mathcal{M}_N$		$\mathcal{M}_S$
C-id	b1	b2	f1	f2	S
C1	0	1	42	3200	⟨(a b) (c) (d)⟩
C2	0	1	29	3500	⟨(a d) (b c)⟩
C3	1	1	26	1100	⟨(a) (b c) (c d)⟩
C4	1	0	23	900	⟨(b) (d)⟩

Exhaustive Pattern enumeration





Drawbacks :
- high computational cost;
- huge pattern spaces;
- redundant outputs.

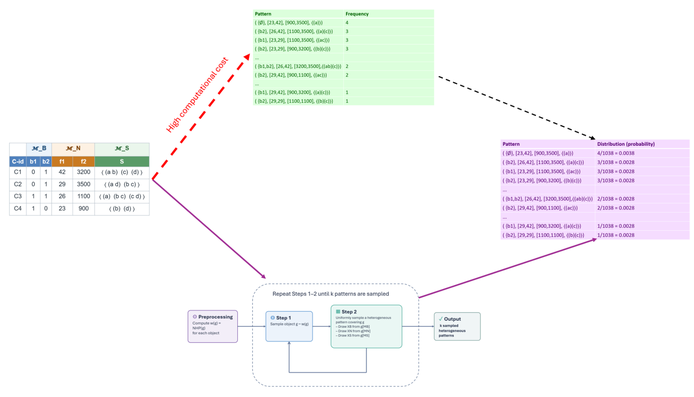
Advantages :
- low computational cost;
- user controls sample size;
- controlled sampling distribution.

2. Problem Statement

Sampling goal :
draw a heterogeneous pattern proportionally to its frequency.

$$P(X_H) = \frac{\text{freq}(X_H, H)}{\sum_{X'_H \in \mathcal{L}_H} \text{freq}(X'_H, H)}$$

$$X_H = \langle X_B, X_N, X_S \rangle$$



3. State of the Art

Stochastic methods :
- **Graphs :** [El Hasan et al., VLDB 2009], [Bendimerad et al., ICDM 2016]
- **Formal concepts :** [Boley et al., DSM 2010]
- **Tiles :** [Bendimerad et al., IDA 2020]

Declarative method :
- **Itemsets :** [Dzyuba et al., DAMI 2017]

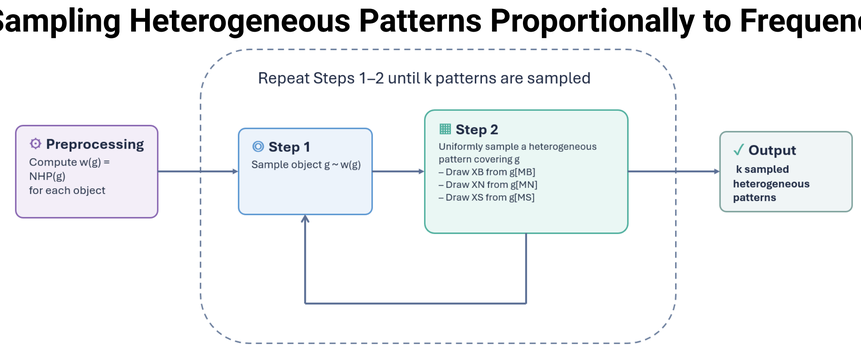
Multi-step methods :
- **Itemsets :** [Boley et al., KDD 2011]
- **Sequences :** [Diop et al., ICDM 2018]
- **Numerical data :** [Giacometti et al., SDM 2018], [Bekkoucha et al., DKE 2026]

Our contribution : SEHP is the first frequency-based sampling method for heterogeneous tabular data.

Published work : [Lachache et al., IDA 2026] *Heterogeneous Pattern Sampling according to Frequency*.

4. SEHP Method

Sampling Heterogeneous Patterns Proportionally to Frequency



Preprocessing : compute one weight per object

$$w(g) = \text{NHP}(g) \qquad \text{NHP}(g) = \text{NI}(g) \times \text{NIP}(g) \times \text{NSP}(g)$$

	M_B		M_N		M_S	$W(g)$
C-id	b1	b2	f1	f2	S	NHP
C1	0	1	42	3200	⟨(a b) (c) (d)⟩	12
C2	0	1	29	3500	⟨(a d) (b c)⟩	18
C3	1	1	26	1100	⟨(a) (b c) (c d)⟩	24
C4	1	0	23	900	⟨(b) (d)⟩	46

Step 1

Draw an object g :

$$P(g) = \frac{w(g)}{\sum_{g' \in G} w(g')}$$

	M_B		M_N		M_S	$W(g)$
C-id	b1	b2	f1	f2	S	NHP
C1	0	1	42	3200	⟨(a b) (c) (d)⟩	12
C2	0	1	29	3500	⟨(a d) (b c)⟩	18
C3	**1**	**1**	**26**	**1100**	**⟨(a) (b c) (c d)⟩**	**24**
C4	1	0	23	900	⟨(b) (d)⟩	46

Step 1 draw one object proportionally to its weight

Step 2

Uniformly draw a pattern covering g :

$$X_H = \langle X_B, X_N, X_S \rangle$$

	M_B		M_N		M_S	$W(g)$
C-id	b1	b2	f1	f2	S	NHP
C1	0	1	42	3200	⟨(a b) (c) (d)⟩	12
C2	0	1	29	3500	⟨(a d) (b c)⟩	18
C3	**1**	**1**	**26**	**1100**	**⟨(a) (b c) (c d)⟩**	**24**
C4	1	0	23	900	⟨(b) (d)⟩	46

Step 2: uniform draw of a pattern covering g
X_H = ⟨{b}, ⟨[23,29],[900,1100]⟩,⟨(b)(d)⟩)

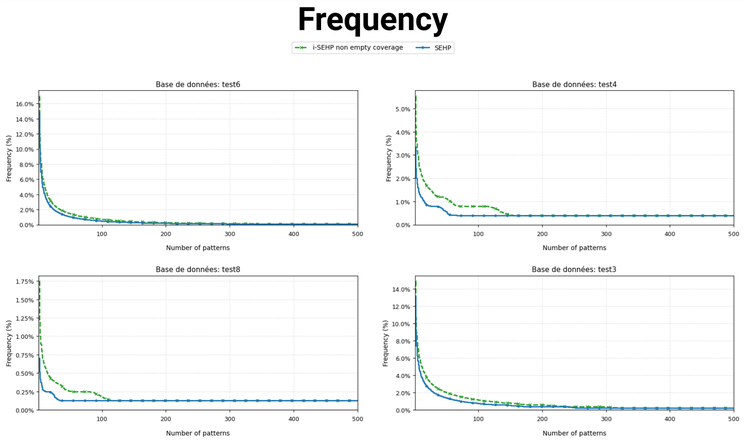
5. Theoretical Guarantee

$$Z = \sum_{g \in G} \text{NHP}(g) \implies P(X_H) = \sum_{g \in \text{cover}(X_H, H)} \frac{\text{NHP}(g)}{Z} \frac{1}{\text{NHP}(g)} = \frac{\text{freq}(X_H, H)}{Z}.$$

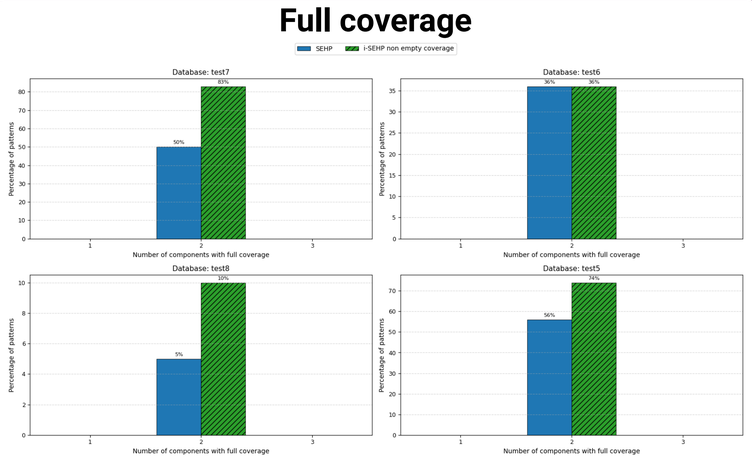
✓ **SEHP samples heterogeneous patterns proportionally to their frequency.**

6. Experimental Results

Frequency



Full coverage



Baseline : I-SEHP

I-SEHP independently samples :

$$X_B \sim \text{freq}(X_B), \qquad X_N \sim \text{freq}(X_N), \qquad X_S \sim \text{freq}(X_S)$$

Then it aggregates them :

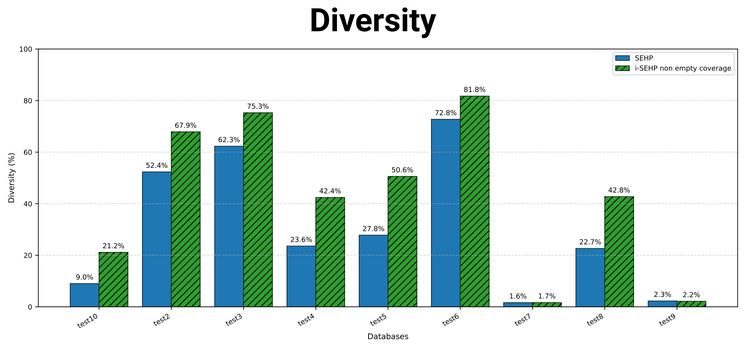
$$X_H = (X_B, X_N, X_S)$$

with coverage

$$\text{cover}(X_H, H) = \text{cov}_B \cap \text{cov}_N \cap \text{cov}_S$$

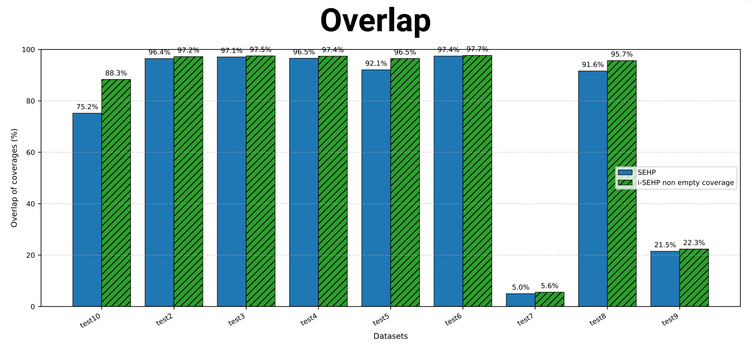
Independent draws may produce **empty-cover** patterns, requiring rejection.

Diversity



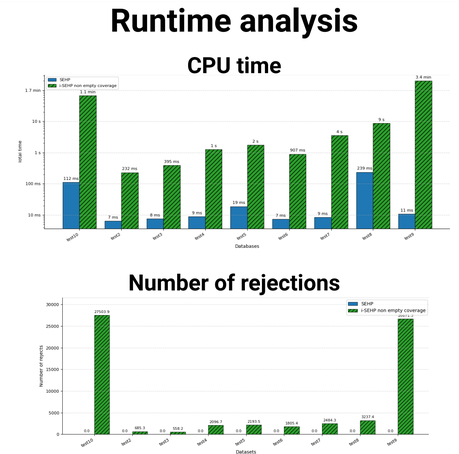
$$\text{Diversity}(K, H) = \frac{\left|\{\text{cover}(X_H^1, H), \ldots, \text{cover}(X_H^{|K|}, H)\}\right|}{|K|}$$

Overlap



$$\text{Overlap}(K, H) = \frac{2}{|K|(|K|-1)} \sum_{1 \le i < j \le |K|} \left(1 - \frac{|\text{cover}(X_H^i, H) \cap \text{cover}(X_H^j, H)|}{|\text{cover}(X_H^i, H) \cup \text{cover}(X_H^j, H)|}\right).$$

Runtime analysis



Main messages

- ✓ SEHP avoids empty-cover patterns by construction.
- ✓ SEHP requires no rejection.
- ✓ SEHP is faster on most datasets.
- ✗ I-SEHP may be slowed down by repeated rejection.
- ✗ Both methods are affected by the long-tail phenomenon.
- ✗ I-SEHP produces more full-coverage sub-patterns than SEHP.
- ✓ SEHP provides a formal frequency-based guarantee.
- ✓ Diversity and overlap characterize exploration of the pattern space.

7. Future Directions

- Extending SEHP to richer interestingness measures beyond frequency;
- Integrating user-defined constraints into the heterogeneous sampling process;
- Leveraging sampled heterogeneous patterns as features for supervised and unsupervised learning;
- Embedding SEHP within interactive pattern mining workflows.