8TH INTERNATIONAL CONFERENCE EDEMOCRACY 2019

12-13 DEC, 2019 / ATHENS, GREECE

A Constraint-Based Model for the Frequent Itemset Hiding Problem

Vassilios S. Verykios, Elias C. Stavropoulos,

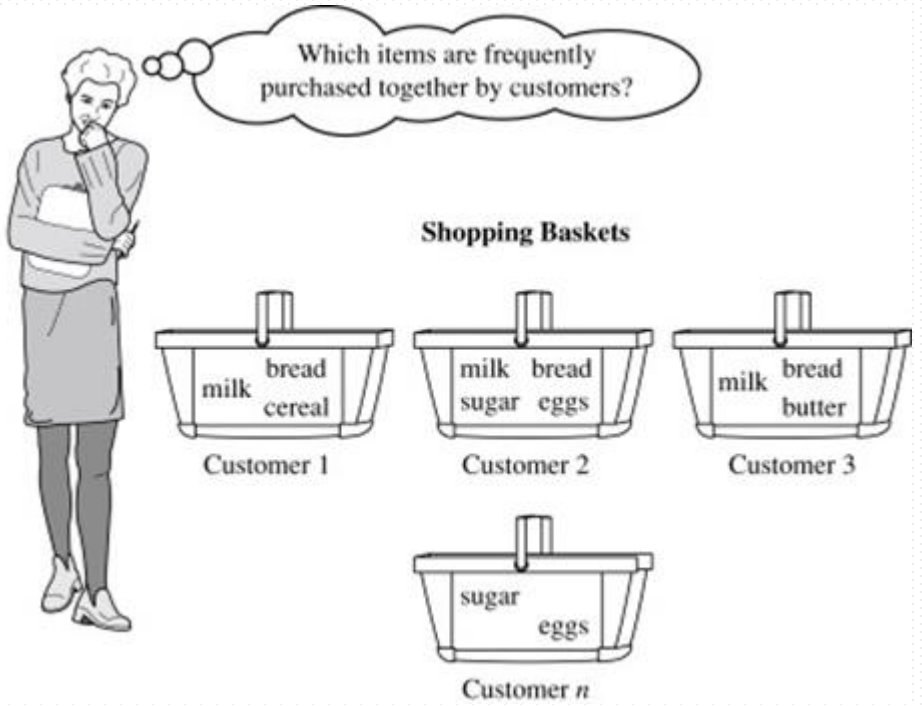
Vasilis Zorkadis, and Ahmed K. Elmagarmid

HELLENIC OPEN UNIVERSITY

BAT LAB
BIG DATA ANALYTICS AND ANONYMIZATION

Frequent Itemset Mining

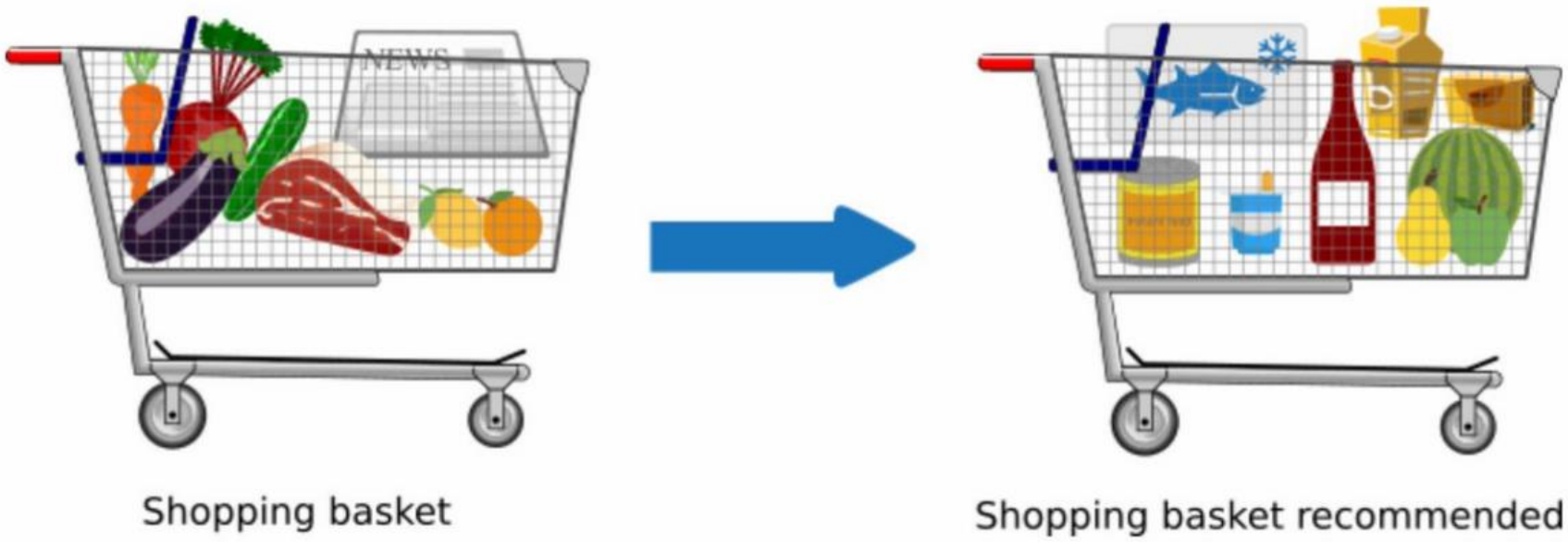
- Discover patters (itemsets) that frequently emerge in a transactional database
- A typical example: Market Basket Analysis
- Analyzes customer buying habits by finding associations between the different items that customers place in their shopping baskets



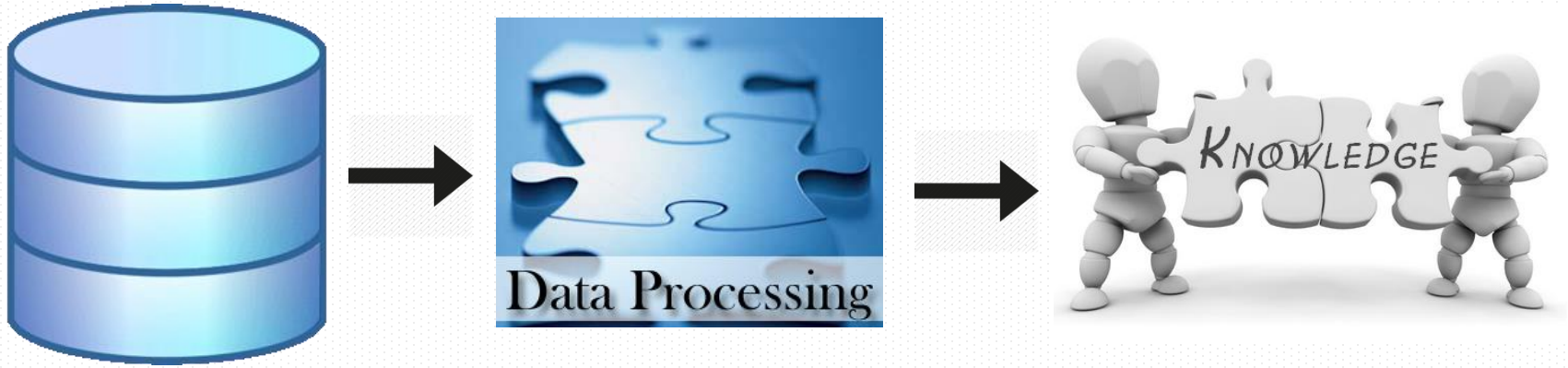
- Next, *association rules* of the form {A}→{B} can be generated, which have a clear antecedent (premise) and consequent (conclusion)

Association rules mining

The discovery of interesting correlation relationships among huge amounts of business transaction records can help in many business decision-making processes such as catalog design, cross-marketing, and customer shopping behavior analysis.



Data Mining & Data Hiding



- Data mining can violate privacy
- Data about individuals may reveal their identity!
- Business related data may reveal trade secrets!

- Privacy Preserving Data Mining Techniques
- Facilitate the mining of data
- Prohibit the leakage of sensitive information (data sanitization)

Problem Definition

Input: A database D, a set of frequent sensitive itemsets S, a support threshold σ_{min}

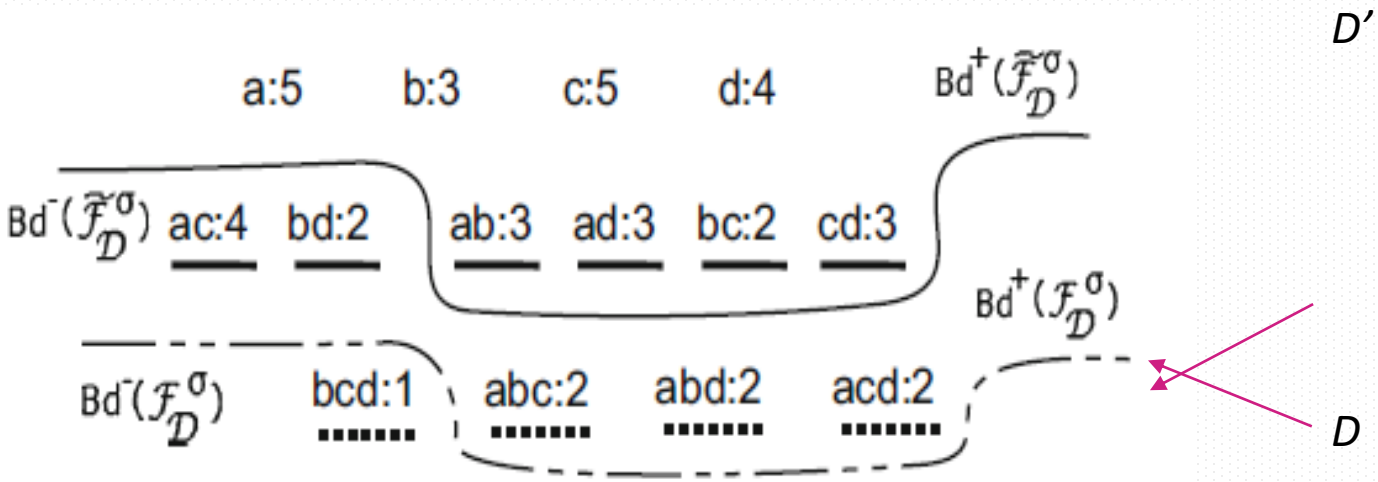
Output: A sanitized database D', where itemsets of S does not appear (are hidden)

- An NP-hard problem (Atallah et al., 1999)
- Approach: Enumerate the ideal *positive border* (i.e. the set of all *maximal itemsets*) of the sanitized database D'
- *Borders form a condensed representation of frequent and infrequent itemsets of D*

An Example

ID	Transaction
1	abcd
2	abc
3	abd
4	acd
5	cd
6	ac

S = {ac, bd, abc, acd}, σ_{min} = 3



- The positive border: $\mathcal{B}d^+(\mathcal{F}_{\mathcal{D}}^{\sigma}) = \left\{X \in \mathcal{F}_{\mathcal{D}}^{\sigma} \mid \forall\, Y, X \subset Y \Rightarrow Y \notin \mathcal{F}_{\mathcal{D}}^{\sigma}\right\}$
- The negative border: $\mathcal{B}d^-(\mathcal{F}_{\mathcal{D}}^{\sigma}) = \left\{X \in \mathcal{P}(\mathcal{I}) \backslash \mathcal{F}_{\mathcal{D}}^{\sigma} \mid \forall\, Y, Y \subset X \Rightarrow Y \in \mathcal{F}_{\mathcal{D}}^{\sigma}\right\}$
- S: The set of sensitive itemsets to be hidden

A Heuristic Coefficient-Based LP method

- Calculate coefficients: $c_i = |A_i| + \left(\log_2 |D|\right)|E_i| + P_i, \forall i \in [1, ..., |D|]$
- How much the frequency of any itemset might be decreased?

$$Thr = \underset{\forall X \in S_i | X \cap j^* \neq \emptyset}{\text{argmax}} \; freq_D(X) - minf$$

- E_i = *endangered* itemsets (*$freq_D$ – Thr < minf*)
- Pricing parameter: $P_i = \sum_{\forall X \in E_i} (1 - freq_D(X))$
- LP Formulation:

$$min \sum_{\forall i:\ T_i \in D} c_i v_i,$$

$$s.t. \sum_{\substack{\forall i,j: \\ X_j \subseteq T_i \in D}} v_i \geq \sigma_D(X_j) - \sigma_{min} + 1, \ \forall X_j \in S_{min}$$

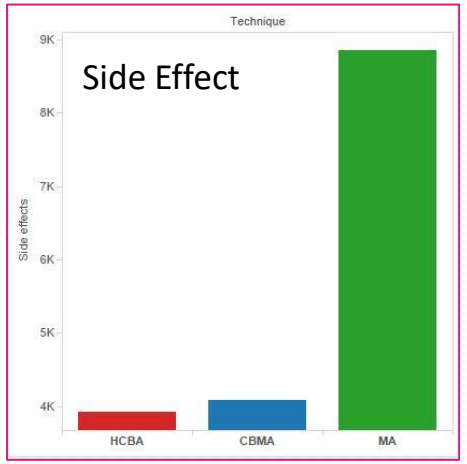
$$v_i \in \{0,1\}, \forall i : T_i \in D$$

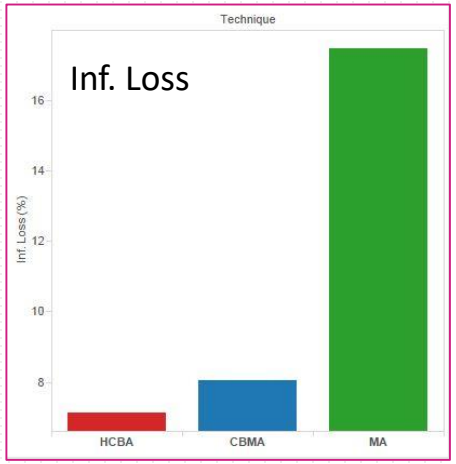
HELLENIC OPEN UNIVERSITY

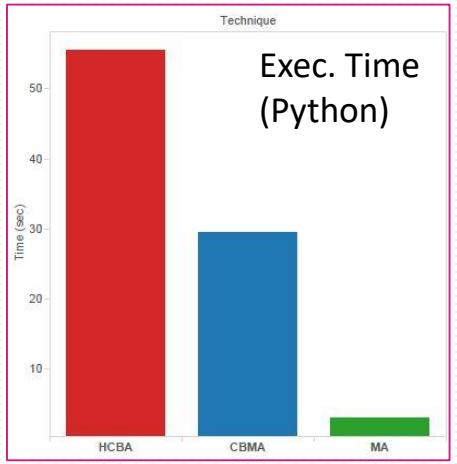
BAT LAB
BIG DATA ANALYTICS AND ANONYMIZATION

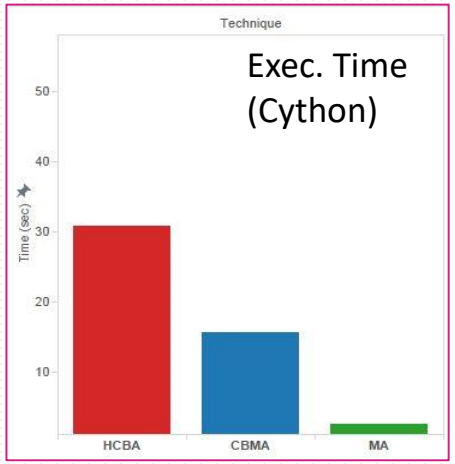
A Heuristic Coefficient-Based LP method

- Sanitization
- Find all sensitive itemsets contained in the first transaction.
- Find the most frequent item in S_i.
- Breaking ties sequence:
  1. Most frequent item in S_i
  2. Least frequent item in $\widetilde{F}$
  3. Item with maximum frequency.
  4. Randomly selected item.
- Repeat for each transaction









HELLENIC OPEN UNIVERSITY

BAT LAB
BIG DATA ANALYTICS AND ANONYMIZATION

A Heuristic Coefficient-Based LP method

Algorithm 1 Intelligent Sanitization Algorithm [12]

1: **for** transactions $T_i \in D : T_i$ is to be sanitized **do**
2: **identify** all sensitive itemsets S_i supported by transaction T_i
3: **while** $S_i \neq \emptyset$ **do**
4: **calculate** $f_j = |\{X \in S_i | j \in X\}|, \forall$ items j in T_i
5: **remove** item $j^* = argmax_j\{f_j\}$
6: **update** $S_i = S_i - \{X \in S_i | j^* \in X\}$
7: **end while**
8: **end for**

The algorithm of Menon et al. (2005)

Algorithm 2 Improved Sanitization Algorithm

1: **for** transactions $T_i \in D$: T_i is to be sanitized **do**
2: **identify** all sensitive itemsets S_i supported by transaction T_i
3: **while** $S_i \neq \emptyset$ **do**
4: **calculate** $f_j = |\{Y \in S_i | j \in Y\}|, \forall$ items j in T_i
5: **calculate** $j^* = argmax_j\{f_j\}$
6: **calculate** $numf = |\{f_j = f_{j^*}\}|, \forall$ items j in T_i
7: **if** $numf = 1$ **then**
8: **remove** item $j^* = argmax_j\{f_j\}$
9: **else**
10: **identify** $C = \{j | f_j = f_{j^*}\}, \forall$ items j in T_i
11: **calculate** $g_j = |\{Y \in \widetilde{F} | j \subseteq Y \subseteq T_i, \forall j \in C\}|$
12: **update** weight $w_j = w_j + g_j$
13: **calculate** $j^* = argmin_{j \in C}\{w_j\}$
14: **if** $|\{j \in C | w_j = w_{j^*}\}| = 1$ **then**
15: **remove** item $j^* = argmin_{j \in C}\{w_j\}$
16: **else**
17: **identify** $M = \{j \in C | w_j = w_{j^*}\}$
18: **remove** item $j^* = argmax_{j \in M}\{freq_D(j)\}$
19: **end if**
20: **end if**
21: **update** $S_i = S_i - \{Y \in S_i | j^* \in Y\}$
22: **end while**
23: **end for**

The algorithm of Kagklis et al. (2014)

HELLENIC OPEN UNIVERSITY

BAT LAB
BIG DATA ANALYTICS AND ANONYMIZATION

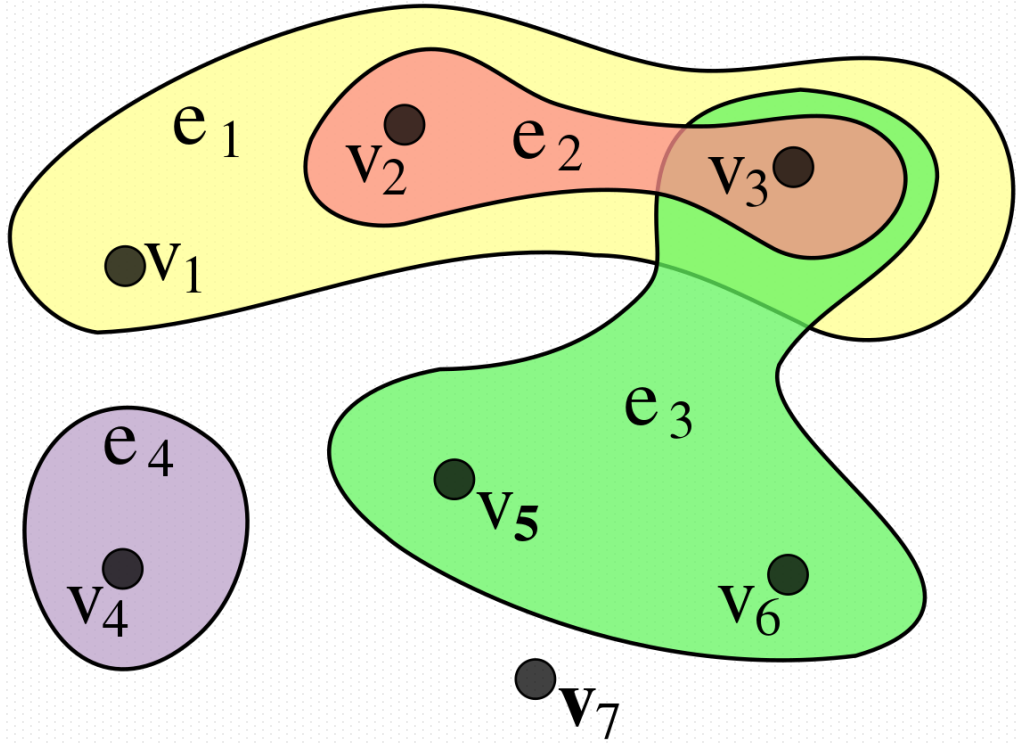
A Heuristic Coefficient-Based LP method

- Experimental Evaluation on real datasets
- Metrics: Side Effects, CPU Time, Information Loss

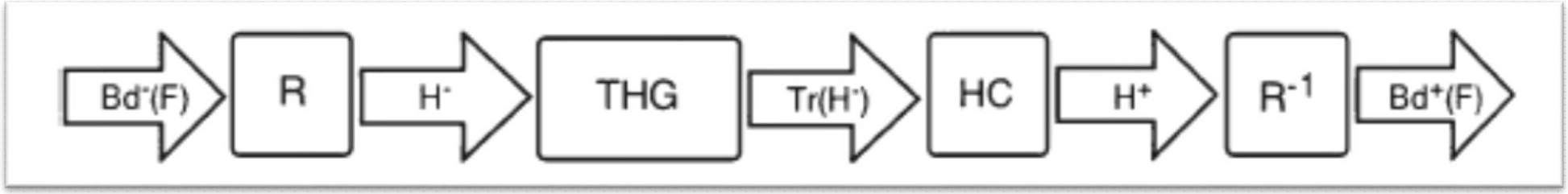
Database (σ_{min})	Sensitive itemsets	Max-Accuracy [12]			Coefficient-Based Max-Accuracy [13]			Heuristic Coefficient-Based Approach		
		SE	Time (sec)	Inf. Loss (%)	SE	Time (sec)	Inf. Loss (%)	SE	Time (sec)	Inf. Loss (%)
BMS1 (48)	10	1426	0.96	9.6	390	1.91	2.5	352	2.78	2.4
	20	3256	0.92	22.1	545	2.38	3.8	444	3.27	3.1
	50	4927	1.47	35.6	1408	4.48	10.3	1286	6.53	9.3
BMS2 (39)	10	12965	5.11	12.5	6403	20.67	7.8	6304	52.03	7.6
	20	49835	6.24	41.9	18084	38.67	19.0	17564	114.42	18.5
	50	57251	11.25	54.8	32975	78.05	35.6	32131	150.15	34.1
retail (44)	10	130	1.73	0.6	53	7.72	0.2	31	12.27	0.2
	20	293	2.1	1.5	150	12.68	0.8	75	23.78	0.5
	50	754	3.1	3.6	331	28.19	1.9	108	51.73	0.9
retai (88)	10	72	1.78	1.1	33	6.0	0.6	5	10.35	0.3
	20	195	1.84	2.6	76	10.05	1.3	21	17.55	0.7
	50	384	2.46	5.3	186	19.92	3.3	108	34.88	2.3
kosarak (4,950)	10	288	0.67	14.3	66	38.92	4.8	43	75.84	3.6
	20	531	1.82	27.2	199	69.97	13.0	133	120.43	9.6
	50	560	4.21	29.3	257	103.97	16.0	211	157.1	13.6

The Transversal Hypergraph model

- Represent transactions of *D* as edges of a hypergraph *H* (a generalized graph)
- Generate the transversal hypergraph *Tr*(*H*) i.e. all minimal hitting sets of *H*
- The output (minimal transversals) may be exponentially large
- The decision problem is in co-NP[$\log^2 n$] (Kavvadias & Stavropoulos, 2003)



The Transversal Hypergraph model



Algorithm 1: An algorithm for enumerating the ideal positive border of the sanitized transaction database.

Input: $\mathcal{S}$, $\mathcal{B}d^{-}(\mathcal{F})$
Set $\mathcal{H}^{-} = \mathcal{R}(\mathcal{B}d^{-}(\mathcal{F}))$ and $\mathcal{H}_{\mathcal{S}} = \mathcal{R}(\mathcal{S})$;
Compute hypergraph $\widetilde{\mathcal{H}}^{-} = Min(\mathcal{H}^{-} \cup \mathcal{H}_{\mathcal{S}})$;
Call THG on input $\widetilde{\mathcal{H}}^{-}$ to generate $Tr(\widetilde{\mathcal{H}}^{-})$;
Compute $\widetilde{\mathcal{H}}^{+} = (Tr(\widetilde{\mathcal{H}}^{-}))^{c}$;
Set $\mathcal{B}d^{+}(\widetilde{\mathcal{F}}) = \mathcal{R}^{-1}(\widetilde{\mathcal{H}}^{+})$;
return $\mathcal{B}d^{+}(\widetilde{\mathcal{F}})$;

The algorithm of Stavropoulos et al. (2015)

THG is any transversal enumeration algorithm, that is time and space efficient i.e. output polynomial (Kavvadias & Stavropoulos, 2005)

HELLENIC OPEN UNIVERSITY

BAT LAB
BIG DATA ANALYTICS AND ANONYMIZATION

A Positive Border-based ILP formulation

- Utilize the ideal positive border to sanitize D with the minimum information loss
- Formulate an ILP and solve it using CLPEX
- Define suitable information loss metrics
- Experimentally evaluate and compare the method vs previous ones

$$min \sum_{\forall i:T_i \in \mathcal{D}} x_i + M \cdot \sum_{\forall j:X_j \in \mathcal{B}d^+(\widetilde{\mathcal{F}})} s_j,$$

$$s.t. \sum_{\forall i:T_i \in \mathcal{D}} a_{ij} x_i \geq \text{supc}_{\mathcal{D}}(X_j) - \sigma + 1, \ \forall X_j \in Min(\mathcal{S}),$$

$$\sum_{\forall i:T_i \in \mathcal{D}} a_{ij} x_i - s_j \leq \text{supc}_{\mathcal{D}}(X_j) - \sigma, \ \forall X_j \in \mathcal{B}d^+(\widetilde{\mathcal{F}}),$$

$$x_i \in \{0, 1\}, \forall i : T_i \in \mathcal{D},$$

$$x_j \in \{0, 1\}, \forall j : X_j \in \mathcal{B}d^+(\widetilde{\mathcal{F}}).$$

Table 9 Information loss on the revised positive border for real and synthetic datasets

Dataset (σ)	Sensitive itemsets	Information loss on positive border (%)					
		MA	EPB-MA	CBMA	MM1	MM2	HCBA
chess (2557)	10	99	94	96	100	100	89
	20	100	98	99	100	100	82
	50	100	99	100	100	100	99
mushroom (1625)	10	85	80	74	77	78	49
	20	69	58	53	65	73	51
	50	98	97	96	92	98	94
BMS1 (51)	10	19	3	4	16	13	3
	20	40	5	6	29	21	4
	50	57	16	15	43	39	15

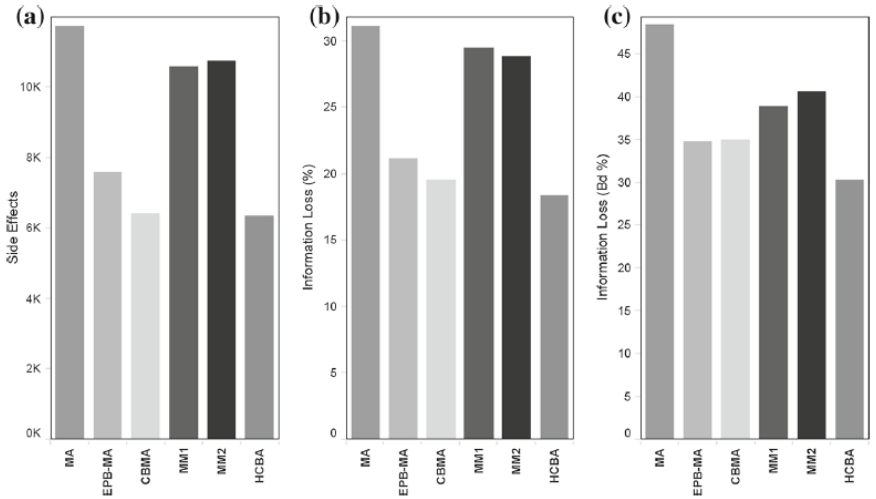


Fig. 3 Mean value of the metrics for the real datasets, **a** side effects, **b** information loss, **c** information loss on $\mathcal{B}d^+(\widetilde{\mathcal{F}})$

HELLENIC OPEN UNIVERSITY

BIG DATA ANALYTICS AND ANONYMIZATION

A Constraint-Based Model for the FIH Problem

Some formal definitions

Definition 1. *[Frequent Itemset Hiding Problem] Given a transaction database* $\mathcal{D}$ *over a set of items* $\mathcal{I} = \{i_1, i_2, i_3, \ldots, i_n\}$*, a support threshold* σ*, and a set of sensitive frequent itemsets* $\mathcal{S}^{\sigma}_{\mathcal{D}}$*, transform* $\mathcal{D}$ *into* $\mathcal{D}'$ *such that:*

1. $sup_{\mathcal{D}'}(X) < \sigma$*, for every* $X \in \mathcal{S}^{\sigma}_{\mathcal{D}}$*,*
2. $|\mathcal{F}^{\sigma}_{\mathcal{D}'} - \widetilde{\mathcal{F}}^{\sigma}_{\mathcal{D}}|$ *is minimized, and*
3. $|\widetilde{\mathcal{F}}^{\sigma}_{\mathcal{D}} - \mathcal{F}^{\sigma}_{\mathcal{D}'}|$ *is minimized.*

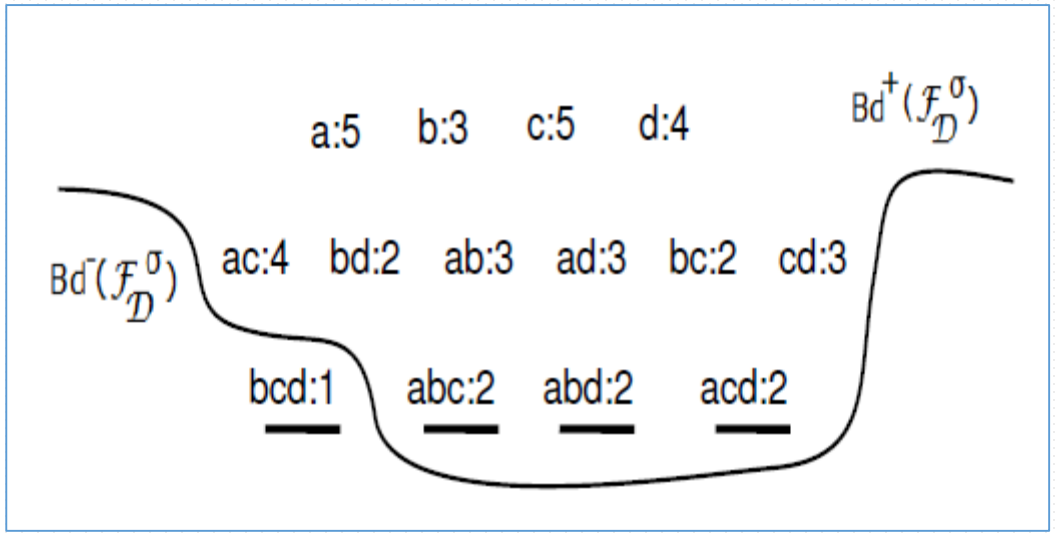
Definition 2. *Given a transaction database* $\mathcal{D}$ *over a set of items* $\mathcal{I}$*, a support threshold* σ*, and a set of sensitive frequent itemsets* $\mathcal{S}^{\sigma}_{\mathcal{D}}$ *of* $\mathcal{D}$*, the negative border* $\mathcal{B}d^{-}(\mathcal{S}^{\sigma}_{\mathcal{D}})$ *of* $\mathcal{S}^{\sigma}_{\mathcal{D}}$ *is the set of minimal itemsets in* $\mathcal{S}^{\sigma}_{\mathcal{D}}$ *with respect to set inclusion.*

A working example

A sample database

Tid	Itemsets
1	$abcd$
2	abc
3	abd
4	acd
5	cd
6	ac

Lattice of itemsets before hiding

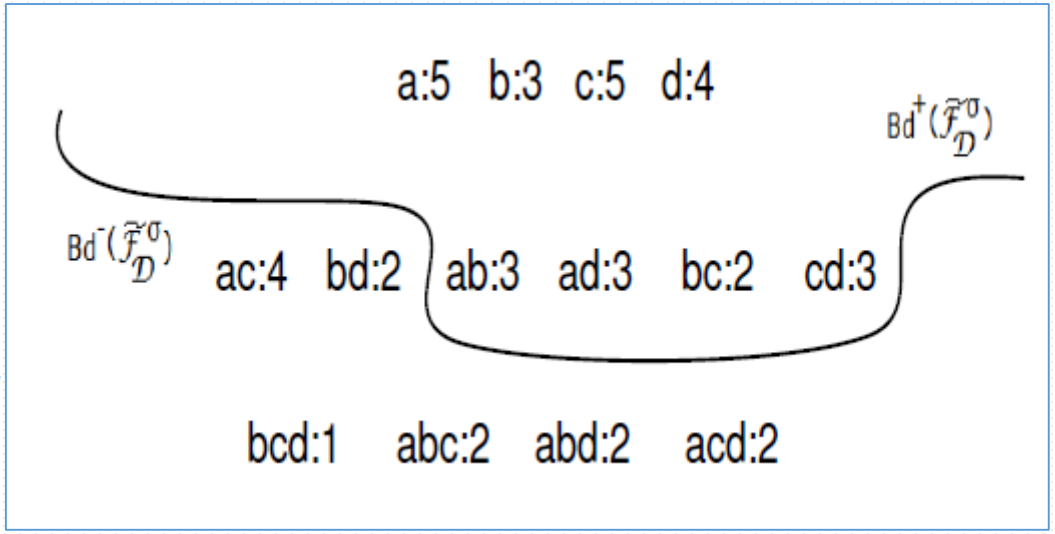


$\sigma = 1/3$

$\mathcal{B}d^+(\mathcal{F}_{\mathcal{D}}^{\sigma}) = \{abc, abd, acd\}$

$\mathcal{B}d^-(\mathcal{F}_{\mathcal{D}}^{\sigma}) = \{bcd\}$

$\mathcal{S}_{\mathcal{D}}^{\sigma} = \{ac, bd, abc, acd\}$



Lattice of itemsets after hiding

$\mathcal{B}d^+(\tilde{\mathcal{F}}_{\mathcal{D}}^{\sigma}) = \{ab, ad, bc, cd\}$ $\mathcal{B}d^-(\mathcal{F}_{\mathcal{D}}^{\sigma}) = \{bcd\}$

HELLENIC OPEN UNIVERSITY

BAT LAB
BIG DATA ANALYTICS AND ANONYMIZATION

A Boolean Formula for Representing Sensitive Itemsets

- Sensitive Itemsets S_D^S are expressed by a Boolean formula B_D^S in Disjunctive Normal Form (DNF) with positive terms
- A DNF is a disjunction of terms (conjunction of literals)

In our example: $\mathcal{S}_{\mathcal{D}}^{\sigma} = \{ac, bd, abc, acd\}$

$$\mathcal{B}_{\mathcal{D}}^{\sigma} = a \cdot c + b \cdot d + a \cdot b \cdot c + a \cdot c \cdot d$$

After removing redundancy

$$\mathcal{B}_{\mathcal{D}}^{\sigma} = a \cdot c + b \cdot d$$

Lemma 1. *The DNF Boolean formula* $\mathcal{B}_{\mathcal{D}}^{\sigma}$ *corresponds to* $\mathcal{B}d^{-}(\mathcal{S}_{\mathcal{D}}^{\sigma})$ *in the border-based theory.*

A Boolean Formula for Representing Sensitive Itemsets

- To satisfy $Bd^{-}(S_D^S)$, it suffices to assign 1 to the variables of a term
- Thus, every term (frequent itemset) defines a pattern of truth assignments (t.a.) that satisfies $Bd^{-}(S_D^S)$
- For the non-frequent itemsets, we want their negative patterns to satisfy $Bd^{-}(S_D^s)$

Theorem 1. *If X is a non-sensitive frequent itemset of the ideal sanitized database, it corresponds to a negated pattern (a truth assignment with zero values) that satisfies $\overline{\mathcal{B}_D^\sigma}$.*

Lemma 2. *The Boolean formula $\overline{\mathcal{B}_D^\sigma}$ is equivalent to an irredundant negative DNF Boolean formula $\mathcal{B}$.*

A Boolean Formula for Representing Sensitive Itemsets

In our example:

- $t_{ac} = 11$ and $t_{bd} = 11$ define the t.a. that satisfies $Bd^{-}(S_D^s)$
- The sensitive itemset abc defines a pattern $t_{abc} = 111$ of t.a. that satisfies B_D^s while its negation 000 does not satisfy the negation of B_D^s
- abc is sensitive and should not be mined from *D'*, while ab is not sensitive (its negation satisfies negation of B_D^s) and it can be mined from *D'*
- $\overline{\mathcal{B}_D^\sigma} = \overline{a \cdot c + b \cdot d}$ should hold for the frequent patterns that can be induced from the sanitized database *D'*

A Constraint-Based Theory for Mining of Borders

- A constraint *C* of itemsets is a function $\mathcal{C} : 2^{\mathcal{I}} \longrightarrow \{\text{true,false}\}$
- An item *X* satisfies a constraint *C* iff *C(X) = true*
- The *theory* of a constraint *C* is the set of itemsets that satisfy *C*

$$\text{Th}(\mathcal{C}) = \{X \in 2^{\mathcal{I}} \mid \mathcal{C}(X) = \text{true}\}$$

- For the itemsets *X* and *Y*, a constraint *C* is *antimonotone* if

$$Y \subseteq X : \mathcal{C}(X) \Rightarrow \mathcal{C}(Y)$$

- The support constraint $\mathcal{C}_{\text{sup}}$ is antimonotone
- The constraint that is related to the non-sensitive itemsets, $\mathcal{C}_{\overline{\text{sen}}}$ is also antimonotone

Proposition 1. *The constraint* $C_{\overline{sen}}$ *holds for an itemset* X *if* X *satisfies* $\mathcal{B}$.

A Constraint-Based Mining Alfgorihtm

Both constraints $\mathcal{C}_{\text{sup}}$ and $\mathcal{C}_{\overline{\text{sen}}}$ are pushed into the frequent itemset mining algorithm that generates

$$\text{Th}_{\mathcal{D}}(\mathcal{C}_{\text{sup}} \wedge \mathcal{C}_{\overline{\text{sen}}})$$

in order to create

$$\mathcal{B}d^{+}(\text{Th}_{\mathcal{D}}(\mathcal{C}_{\text{sup}} \wedge \mathcal{C}_{\overline{\text{sen}}}))$$

Input:
$\mathcal{D}$: transaction database
σ: the minimum support threshold
$\mathcal{B}$: Boolean formula representing $\mathcal{C}_{\overline{\text{sen}}}$
Output:
L: frequent itemsets satisfying $\mathcal{B}$

1: **Description:**
2: $C_1^{\mathcal{B}}$ the candidate items that satisfy $\mathcal{B}$
3: L_1 the frequent items in $C_1^{\mathcal{B}}$
4: **for** $(k = 2; L_{k-1} \neq \emptyset; k{+}{+})$
5: $\quad C_k^B =$ Apriori-genB(L_{k-1});
6: $\quad$ **for each** transaction $t \in \mathcal{D}$
7: $\quad\quad C_t =$**subset**$(C_k^{\mathcal{B}}, t)$;
8: $\quad\quad$ **for each** candidate $c \in C_t$
9: $\quad\quad\quad c.count{+}{+}$;
10: $\quad$ **end**
11: $\quad L_k = \{c \in C_k^{\mathcal{B}} \mid c.count \geq \sigma\}$;
12: **end**
13: return $L = \bigcup_k L_k$;

HELLENIC OPEN UNIVERSITY

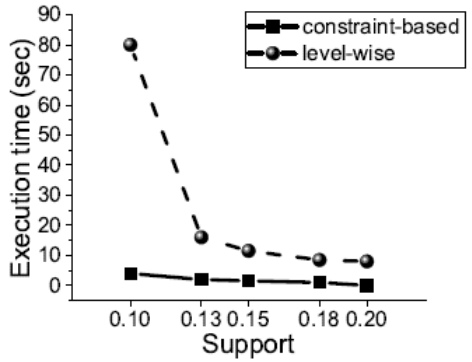
BAT LAB
BIG DATA ANALYTICS AND ANONYMIZATION

Experimental Evaluation

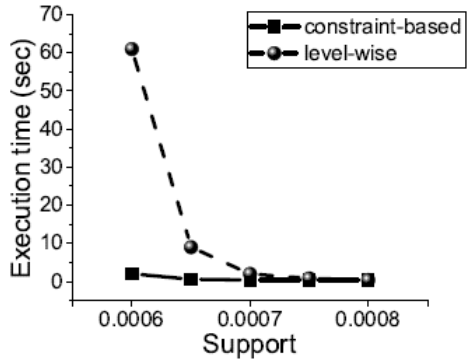
- Comparison with conventional Level-wise Apriori Algorithm
- Metrics: CPU time
- Datasets: real and synthetic

Dataset	Number of Transactions	Number of Items	Aver. Trans. Length
mushroom	8,124	119	23
BMS1	59,602	497	2.5
BMS2	77,512	3,340	5.6
kosarak	990,002	41,270	8.1
T10I4D100K	100,000	870	10.10
T40I10D100K	100,000	942	39.60
1M	1,000,000	2398	4.58
5M	5,000,000	1468	4.07

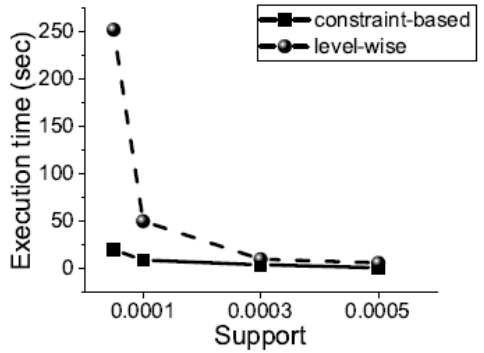
Experimental Evaluation Results



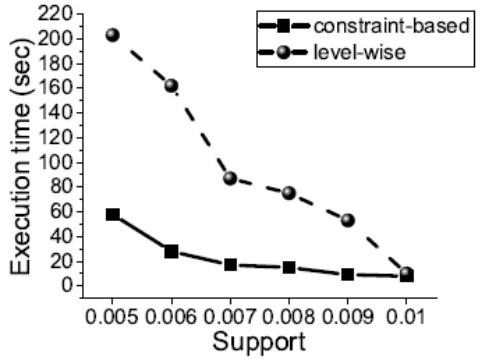
(a) Execution time for *mushroom*



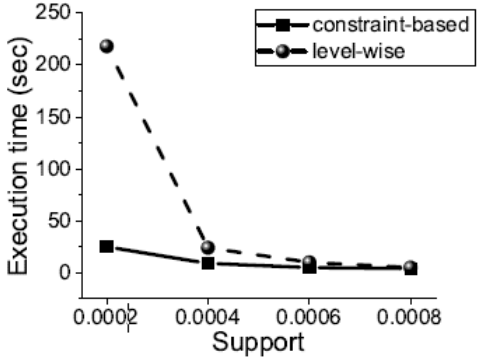
(b) Execution time for *BMS1*



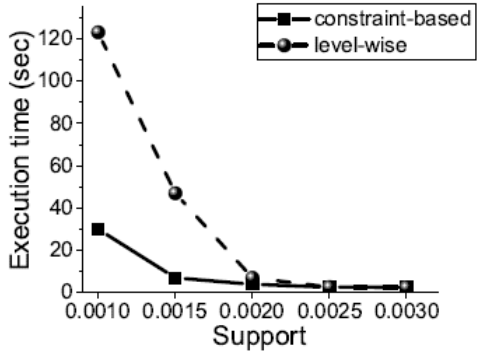
(a) Execution time for *T10I4D100K*



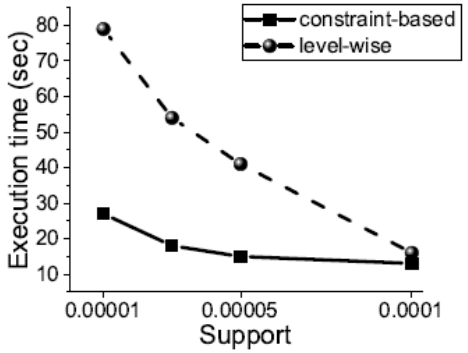
(b) Execution time for *T40I10D100K*



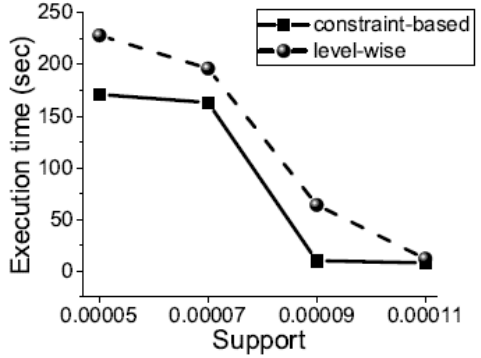
(c) Execution time for *BMS2*



(d) Execution time for *kosarak*



(c) Execution time for *1M*

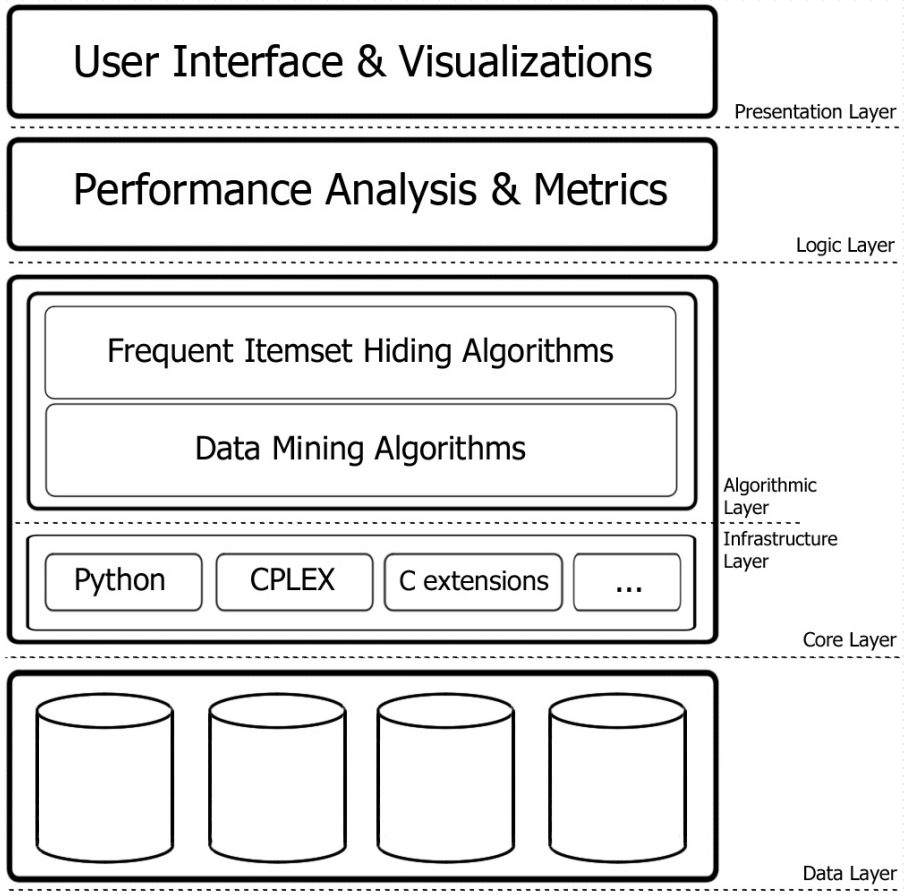


(d) Execution time for *5M*

HELLENIC OPEN UNIVERSITY

BAT LAB
BIG DATA ANALYTICS AND ANONYMIZATION

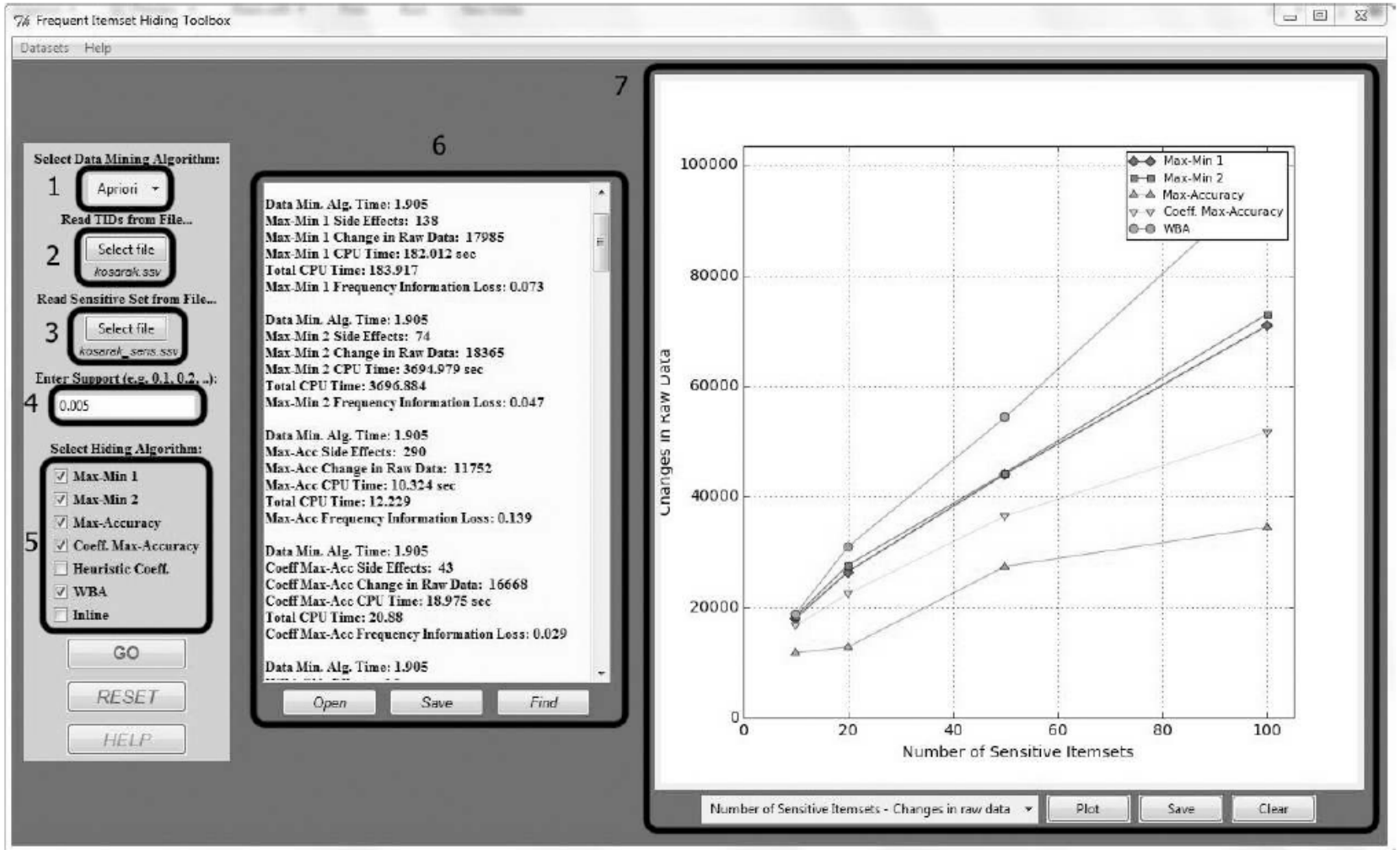
A Frequent Itemset Hiding Toolbox



HELLENIC OPEN UNIVERSITY

BAT LAB
BIG DATA ANALYTICS AND ANONYMIZATION

Toolbox GUI



HELLENIC OPEN UNIVERSITY

BAT LAB
BIG DATA ANALYTICS AND ANONYMIZATION

A Comparison of Border Revision Algorithms

A level-wise algorithm

- Check whether each sensitive itemset is subset of a itemset in the initial border
- Checks which of the subsets of the eliminated itemset should be added to the revised border
- Proceed in a level-wise manner in the lattice of itemsets

A transversal hypergraph-based algorithm

- Compute the hypergraph H that corresponds to the minimal union of the negative border and the sensitive itemsets
- Generate Tr(H)
- Represent Tr(H) as a family of itemsets (the ideal positive border)

A constraint based-algorithm

- Formulate a DNF Boolean formula
- Utilize appropriate constraints related with the minimum support and the non-sensitive itemsets
- Incorporate the above constraints to Apriori algorithm

A Comparison of Border Revision Algorithms

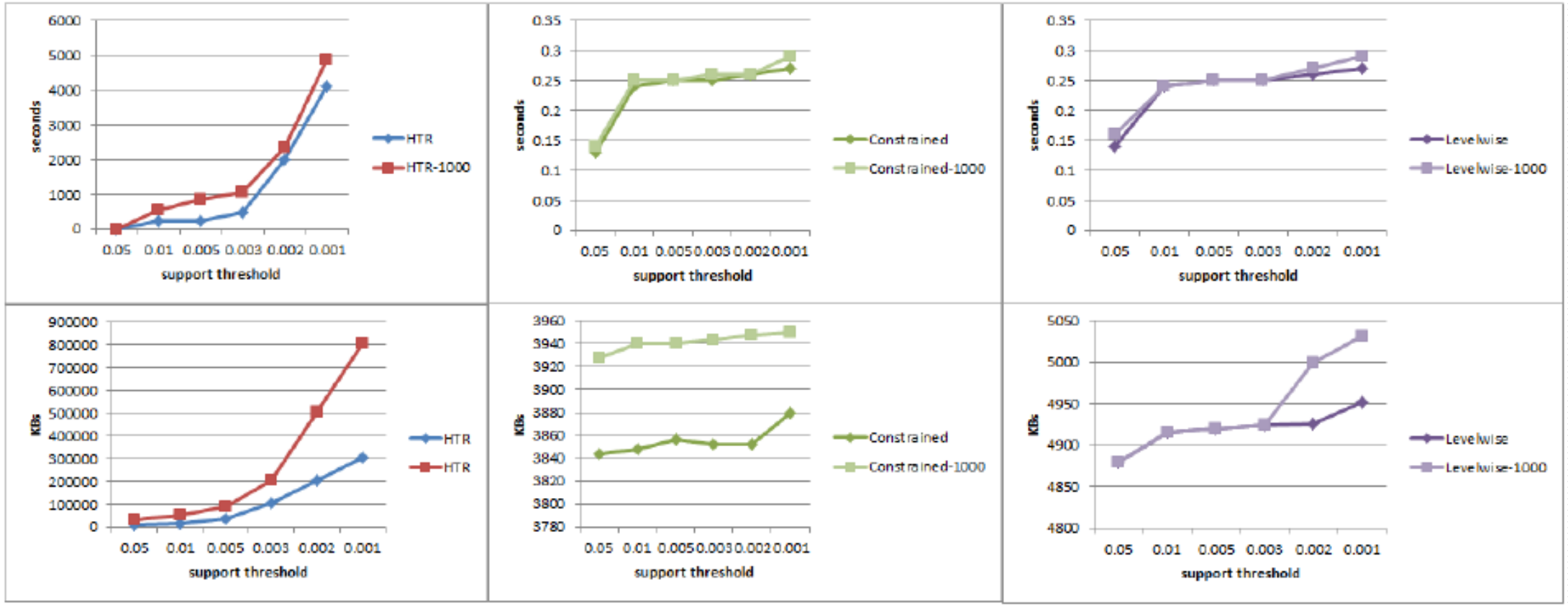


Figure 12: Running time and memory usage using larger $\mathcal{S}_D^g$ for dataset *BMS-WebView-2*.

The Inverse Frequent Itemset Mining Problem

- Given a collection *F* of itemsets, is there a database D where these itemsets are frequent?
- Emerging in privacy preserving data publishing and database generation for benchmarking
- The IFM problem is NP-Hard and in PSPACE (MieliKainen, 2003)
- Strongly related with the *probabilistic satisfiability problem* (Georgakopoulos et al., 1988) and the *transversal duality problem (Fredman & Khachiyan, 1996)*
- Several interesting variants have been proposed (Calders, 2004, Guzzo et al., 2009)
- Recent approaches solve IFM via Large-Scale Linear Programs (Guzzo et al.,2013)
- Open problems: effective column generation techniques to feed Simplex Method, efficient implementation using CPLEX, complexity issues

References

- Vasileios Kagklis, Vassilios S. Verykios, Giannis Tzimas, and Athanasios K. Tsakalidis (2014). An Integer Linear Programming Scheme to Sanitize Sensitive Frequent Itemsets. ICTAI: 771-775.
- Βασίλειος Σ. Βερύκιος, Βασίλειος Καγκλής, Ηλίας Κ. Σταυρόπουλος (2015). Η Επιστήμη των Δεδομένων μέσα από τη Γλώσσα R, Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών.
- Elias C. Stavropoulos, Vassilios S. Verykios, and Vasileios Kagklis (2016). A Transversal Hypergraph Approach for the Frequent Itemset Hiding Problem. Knowledge and Information Systems: Vol. 47, No. 3, pp 625-645.
- Aris Gkoulalas-Divanis, Vasileios Kagklis, and Elias C. Stavropoulos (2018). A Frequent Itemset Hiding Toolbox. ALGOCLOUD: 169-182.
- Β.Σ. Βερύκιος, Σ.Β. Κωτσιαντής, Η.Κ. Σταυρόπουλος και Μ.Μ. Τζαγκαράκης (2019). Η Επιστήμη των Δεδομένων: Βασικές Αρχές, Θεωρία & Εφαρμογές με τη Γλώσσα R, Εκδόσεις Νέων Τεχνολογιών.