Block Interaction: A Generative Summarization Scheme for Frequent Patterns

Ruoming Jin
Kent State University

Joint work with Yang Xiang (OSU), Hui Hong (KSU) and Kun Huang (OSU)

Frequent Pattern Mining

- Summarizing the underlying datasets, providing key insights
- Key building block for data mining toolbox
  - Association rule mining
  - Classification
  - Clustering
  - Change Detection
  - etc...
- Application Domains
  - Business, biology, chemistry, WWW, computer/networing security, software engineering, ...

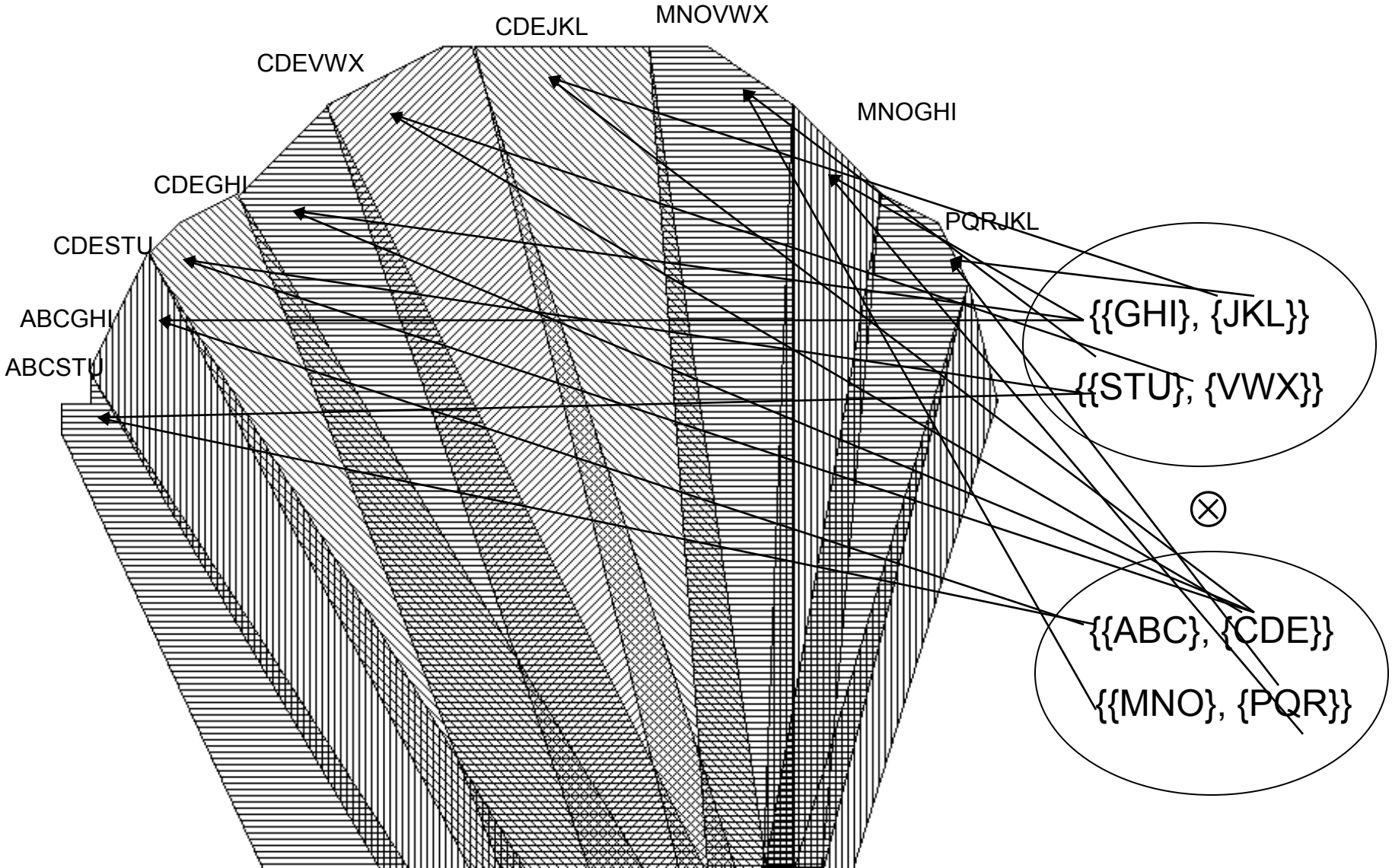
The Problem

- The number of patterns is too large
- Attempts
  - Maximal Frequent Itemsets
  - Closed Frequent Itemsets
  - Non-Derivable Itemsets
  - Compressed or Top-k Patterns
  - …
- Issues
  - Significant Information Loss
  - Large Size

Pattern Summarization

- Using a small number of itemsets to best represent the entire collection of frequent itemsets
  - The Spanning Set Approach [Afrati-Gionis-Mannila, KDD04]
  - Exact Description = Maximal Frequent Itemsets
  - No support information
- The problem: Can we summarize a collection of frequent itemsets and provide accurate support information using only a small number of frequent itemsets?

Itemset Contour (KDD'09)



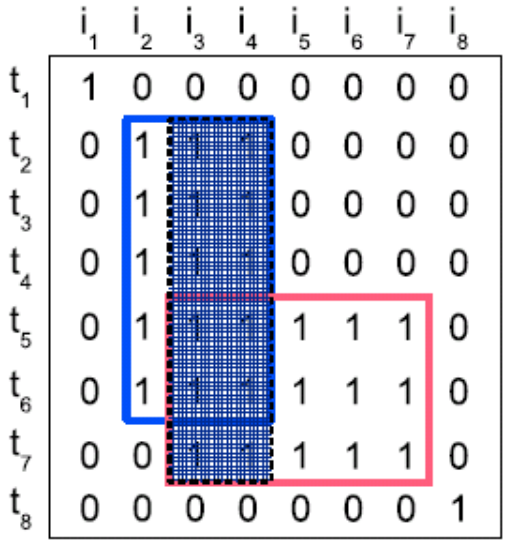
Generative Block-Interaction Model

- Core blocks (hyper-rectangles, tiles, etc)
  - Cartesian products of itemsets and its support transactions
- Core blocks interact with each other through two operators
  - Vertical Union, Horizontal Union
- Each itemset and its frequency can be accurately recovered through the combination of the core blocks

Vertical Operator

Block Vertical Union ($\oplus$)**:** Given two blocks $B_1 = T_1 \times I_1$ and $B_2 = T_2 \times I_2$, the block vertical union operator generates a new block with the itemset being the *intersection* of two itemsets $I_1 \cap I_2$ and the transaction set being the *union* of two transaction sets $T_1 \cup T_2$ (Figure 1(a)):

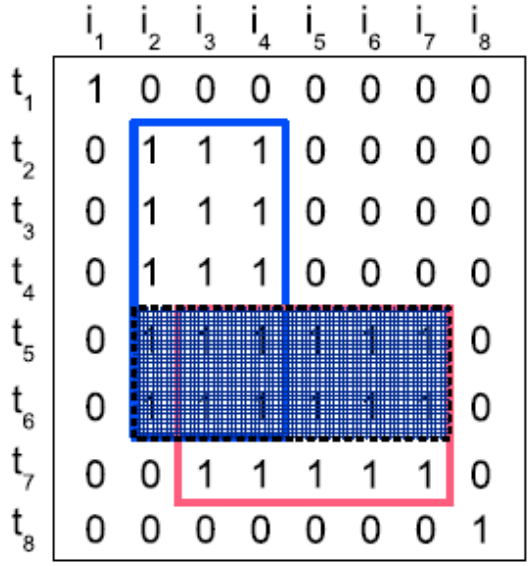
$$B_1 \oplus B_2 = T_1 \times I_1 \oplus T_2 \times I_2 = (T_1 \cup T_2) \times (I_1 \cap I_2)$$



Horizontal Operator

Block Horizontal Union ($\ominus$)**:** Given two blocks $B_1 = T_1 \times I_1$ and $B_2 = T_2 \times I_2$, the block horizontal union operator generates a new block with the itemset being the *union* of two itemsets $I_1 \cup I_2$ and the transaction set being the *intersection* of two transaction sets $T_1 \cap T_2$ (Figure 1(b)):

$$B_1 \ominus B_2 = T_1 \times I_1 \ominus T_2 \times I_2 = (T_1 \cap T_2) \times (I_1 \cup I_2)$$



Block Support

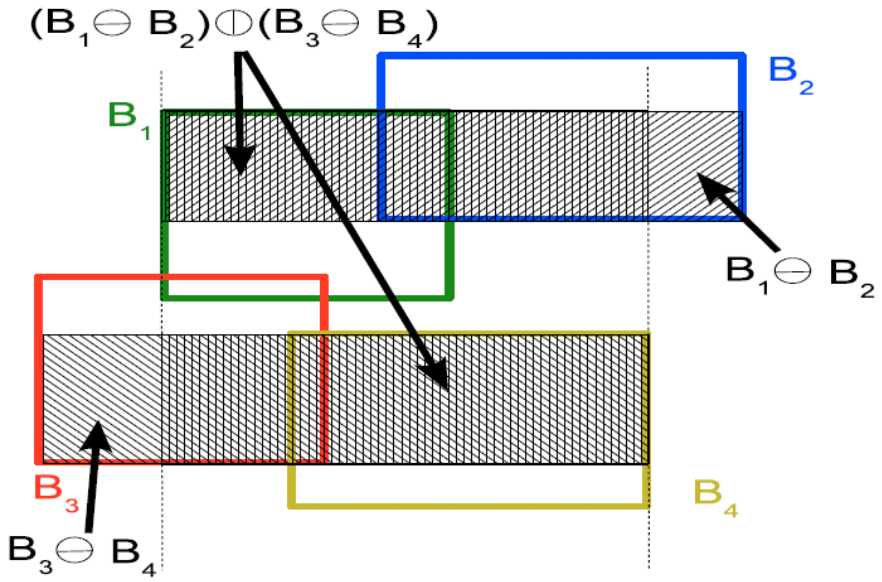
DEFINITION 1. (**Block Support**) *Given an itemset I, if a block B in $\mathcal{P}(\mathcal{B})$* subsumes *I, i.e., $I \subseteq I(B)$, and if $|T(B)| \geq (1-\epsilon)supp(I)$, where ϵ is a user-preferred accuracy level for support recovery, then we say I is* **supported** *or* **explained** *by block B, denoted as $B \models I$. For a given set of frequent itemset F_α, and a set of core blocks $\mathcal{B}$, if any itemset I in F_α is supported by at least one block B in the closure $\mathcal{P}$ of $\mathcal{B}$, then we say that F_α is supported or explained by $\mathcal{B}$ or $\mathcal{P}$, denoted as $\mathcal{B} \models F_\alpha$ or $\mathcal{P} \models F_\alpha$.*

(2X2) Block-Interaction Model

DEFINITION 2. *((2×2)-**Block Support**) Given an itemset I, if a block B in the block closure $\mathcal{P}$ supports I, i,e., $B \models I$, and if this block can be expressed in the following format,*

$$B = (B_1 \ominus B_2) \oplus (B_3 \ominus B_4), \tag{1}$$

where $B_1, B_2, B_3, B_4 \in \mathcal{B}$ are core blocks, then we say I is (2×2)-block supported by $\mathcal{B}$. If each itemset I in F_α is (2×2)-block supported by $\mathcal{B}$, then we say $\mathcal{B}$ is a (2×2)-block interaction model for F_α.



Minimal 2X2 Block Model Problem

- Given the (2×2) block interaction model, our goal is to provide a generative view of an entire collection of itemsets Fα using only a small set of core blocks B.

NP-Hardness

THEOREM 1. *Given a transaction database DB and a collection of frequent itemsets F_α, it is NP-hard to find a minimal (2×2)-block interaction model.*

Proof sketch of Theorem 1 can be found in our technical report [17].

NP-Hardness

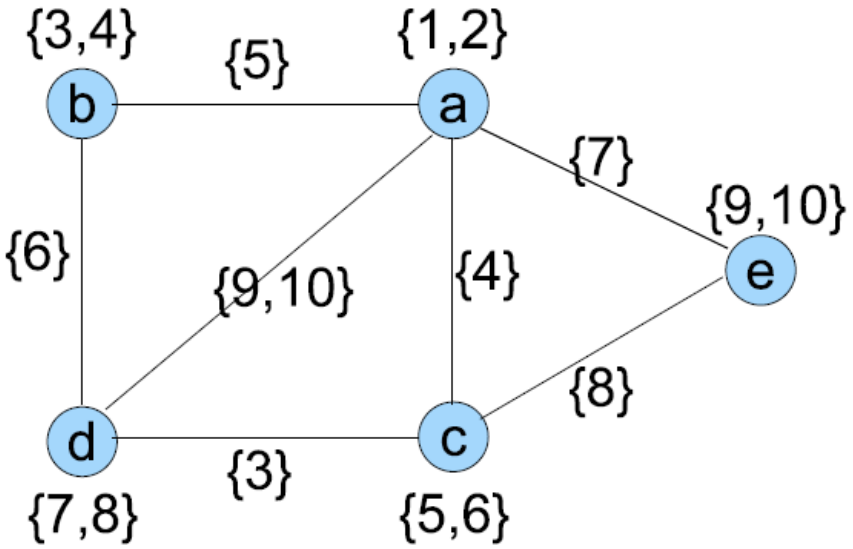
This problem is more closely related to the recently proposed *set-cover-with-pairs problem* [15].

DEFINITION 5. **(Set-Cover-with-Pairs Problem)** *Let U be the ground set and let $S = \{1, \ldots, M\}$ be a set of objects. For every $\{i, j\} \subseteq S$, let $\mathcal{C}(i, j)$ be the collection of elements in U covered by the pair $\{i, j\}$. The objective of the set cover with pairs (SCP) problem is to find a subset $S' \subseteq S$ such that*

$$\mathcal{C}(S') = \bigcup_{\{i,j\} \subseteq S'} \mathcal{C}(i, j) = U$$

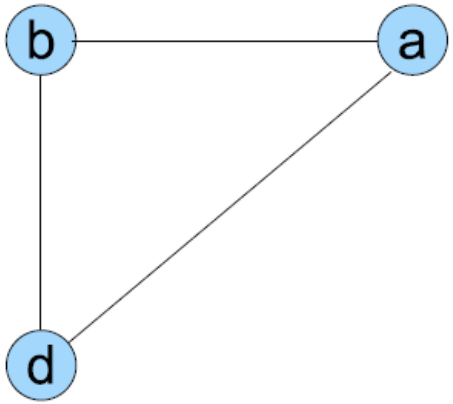
with a minimum number of objects.

Example



Ground set: {1,2,3,4,5,6,7,8,9,10}

(1)



An induced subgraph which is an optima solution for the graph set cover problem

(2)

Two Stage Approach

Stage 1 (Minimizing Vertical Union Decomposition): In the first stage, we seek a minimal number of blocks ($\mathcal{C}$) which use only the $\oplus$ operator to support the entire collection of closed frequent itemsets (CF_α). Those blocks $\mathcal{C}$ discovered in the first stage then will be decomposed using $\ominus$ operator in the second stage. Specifically, the goal of the first stage is as follows:

DEFINITION 6. *(***Subproblem 1: Minimal Vertical Union Decomposition Problem***)* *Given a collection of closed frequent itemset CF_α, we seek a small set of blocks, $\mathcal{C} = \{C_1, \cdots, C_m\}$, where $C_i = I_i \times T(I_i)$ and $I_i \in CF_\alpha^2$, such that each itemset $I \in CF_\alpha$ can be supported or explained by at most two blocks C_i and C_j in $\mathcal{C}$, $C_i \oplus C_j \models I$ with respect to accuracy level ϵ_1 ($\epsilon_1 \leq \epsilon$): $I \subseteq I(C_i \oplus C_j)$ and*

$$|T(C_i \oplus C_j)| \geq (1 - \epsilon_1) supp(I).$$

Two Stage Approach

Stage 2 (Minimizing Horizontal Union Decomposition): In the second stage, we will seek a minimal number of blocks ($\mathcal{B}$) to support the blocks ($\mathcal{C}$) discovered in the first stage. Formally, the goal of this stage is formally described as follows.

DEFINITION 7. *(***Subproblem 2: Minimal Horizontal Union Decomposition Problem***)* *Let $\mathcal{C}$ be the set of blocks discovered in the first stage, we seek a minimal number of closed supporting blocks, $\mathcal{B} = \{B_1, \cdots, B_k\}$, where $B_i = I_i \times T(I_i), I_i \in CF_\alpha$, such that the itemset $I(C)$ of each block $C \in \mathcal{C}$ is supported or explained by at most two blocks B_i and B_j in $\mathcal{B}$, $B_i \ominus B_j \models I(C)$ with respect to accuracy level $\epsilon_2 = (\epsilon - \epsilon_1)/2$, i.e., $I(C) \subseteq I(B_i \ominus B_j)$ and*

$$|T(B_i \ominus B_j)| \geq (1 - \epsilon_2) supp(I(C)).$$

Algorithm

Stage1:
Block Vertical Union

Stage2:
Block Horizontal Union

{Stage 1: Block Vertical Union ($\oplus$) Decomposition}
1: $CF_\alpha^2 \leftarrow \{I_1 \cup I_2 | I_1, I_2 \in CF_\alpha\}$;
2: $CF \leftarrow CF_{(1-\epsilon)\alpha} \cap CF_\alpha^2$ {reducing the candidate blocks};
3: $CF^2 \leftarrow \{I_1 \cup I_2 | I_1, I_2 \in CF\}$;
4: ComputeSupport($CF^2 \backslash CF_{(1-\epsilon)\alpha}$); {Compute support for each itemset in CF}
{Vertex Set Construction:}
5: **for all** $I_v \in CF$ **do**
6: $V_1 \leftarrow V_1 \cup \{(I_v, supp(I_v)\}$;
7: $S(v) \leftarrow \{I \in CF_\alpha | I \subseteq I_v \wedge supp(I_v) \geq (1-\epsilon_1) supp(I)\}$;
8: **end for**
{Edge Set Construction:}
9: **for all** $(I_1, I_2) \in CF \times CF$ **do**
10: $S(v_1, v_2) \leftarrow \{I \in CF_\alpha | I_1 \cap I_2 \supseteq I \wedge supp(I_1) + supp(I_2) - supp(I_1 \cup I_2) \geq (1-\epsilon_1) supp(I)\} \backslash (S(v_1) \cup S(v_2))$;
11: **if** $S(v_1, v2) \neq \emptyset$ **then**
12: $E_1 \leftarrow E_1 \cup \{(v_1, v_2)\}$;
13: **end if**
14: **end for**
15: $\mathcal{C} \leftarrow$ GraphSetCover($G_1(V_1, E_1), CF_\alpha$);

{Stage 2: Block Horizontal Union ($\ominus$) Decomposition}
16: $U \leftarrow \{I(C) | C \in \mathcal{C}\}$;
{Vertex Set Construction:}
17: **for all** $I_v \in CF_\alpha$ **do**
18: $V_2 \leftarrow V_2 \cup \{(I_v, supp(I_v)\}$;
19: $S(v) \leftarrow \{I \in U | I \subseteq I_v \wedge supp(I_v) \geq (1-\epsilon_2) supp(I)\}$;
20: **end for**
{Edge Set Construction:}
21: **for all** $(I_1, I_2) \in CF_\alpha \times CF_\alpha$ **do**
22: $S(v_1, v_2) \leftarrow \{I \in U | I_1 \cup I_2 \supseteq I \wedge supp(I_1 \cup I_2) \geq (1-\epsilon_2) supp(I)\} \backslash (S(v_1) \cup S(v_2))$;
23: **if** $S(v_1, v2) \neq \emptyset$ **then**
24: $E_2 \leftarrow E_2 \cup \{(v_1, v_2)\}$;
25: **end if**
26: **end for**
27: $\mathcal{B} \leftarrow$ GraphSetCover($G_2(V_2, E_2), U$);

Experiment

- How does our block interaction model(**B.I.) compare with** the state-of-art summarization schemes, including Maximal Frequent Itemsets (**MFI), Close Frequent Itemsets (CFI),** Non-Derivable Frequent Itemsets (**NDI), and Representative** pattern (δ**-Cluster).**

- How do different parameters, including α and ϵ, affect the conciseness of the block modeling, i.e., the number of core blocks?

Experiment Setup

- **Group 1: In the first group of experiments, we vary the support** level α for each dataset with a fixed user-preferred accuracy level ϵ (either 5% or 10%) and fix ϵ1 = ϵ/2 .

- **Group 2: In the second group of experiments, we study how userpreferred** accuracy level ϵ would affect the model conciseness (the number of core blocks). Here, we vary ϵ generally in the range from 0.1 to 0.2 with a fixed support level α and ϵ1 = ϵ/2 .

- **Group 3: In the third group of experiments, we study how the** distribution of accuracy level ϵ1 in the two stages would affect the model conciseness. We vary ϵ1 between 0.1ϵ and 0.9ϵ with fixed support level α and the overall accuracy level ϵ.

Data Description

Datasets	$\mathcal{I}$	$\mathcal{T}$	density
connect	129	67557	dense
pumsb	7116	49046	dense
chess	75	3,196	dense
retail	16469	88162	sparse
T40I10D100K	1000	100000	sparse

Table 1: Datasets Characters. $\mathcal{I}$ is the total number of items and $\mathcal{T}$ is the total number of transactions

Group1 Results (varying support)

α	MFI	CFI	NDI	δ-Cluster	B.I.
0.92	175	2212	168	178	56
0.91	192	2819	184	196	56
0.90	222	3486	199	222	72
0.89	261	4218	223	279	85
0.88	313	5106	240	332	89

Table 2: Group1.Connect: $\epsilon = 0.05$

α	MFI	CFI	NDI	δ-Cluster	B.I.
0.90	259	1465	585	259	48
0.89	348	2186	763	348	82
0.88	500	3160	501	988	88
0.87	633	4508	1200	634	99
0.86	825	6245	1470	826	262

Table 3: Group1.Pumsb: $\epsilon = 0.1$

α	MFI	CFI	NDI	δ-Cluster	B.I.
0.875	74	1059	133	83	81
0.850	119	1885	172	137	82
0.825	176	3189	218	209	126
0.800	226	5083	281	288	109
0.775	325	7679	352	426	266

Table 4: Group1.Chess: $\epsilon = 0.05$

α	MFI	CFI	NDI	δ-Cluster	B.I.
0.007	167	315	317	294	136
0.006	219	417	418	391	176
0.005	284	580	582	545	241
0.004	424	831	838	783	335
0.003	692	1393	1410	1325	538

Table 5: Group1.Retail: $\epsilon = 0.05$

α	MFI	CFI	NDI	δ-Cluster	B.I.
0.032	608	685	686	685	458
0.031	645	730	731	730	472
0.030	700	793	794	793	486
0.029	741	842	843	842	495
0.028	812	924	925	924	506

Table 6: Group1.T40I10D100K: $\epsilon = 0.1$

Group2 Results (varying accuracy)

ϵ	MFI	CFI	NDI	δ-Cluster	B.I.
0.06	222	3486	199	225	104
0.08	222	3486	199	223	50
0.1	222	3486	199	222	40
0.12	222	3486	199	222	27
0.14	222	3486	199	222	19

Table 7: Group2.Connect: $\alpha = 0.9$

ϵ	MFI	CFI	NDI	δ-Cluster	B.I.
0.06	219	417	418	390	176
0.07	219	417	418	389	175
0.08	219	417	418	389	175
0.09	219	417	418	220	233
0.1	219	417	418	389	203

Table 8: Group2.Retail: $\alpha = 0.006$

Group3 Results

ϵ_1	MFI	CFI	NDI	δ-Cluster	B.I.
0.01	259	1465	585	259	28
0.03	259	1465	585	259	39
0.05	259	1465	585	259	48
0.07	259	1465	585	259	87
0.09	259	1465	585	259	258

Table 9: Group3.Pumsb: $\alpha = 0.9$, $\epsilon = 0.1$

ϵ_1	MFI	CFI	NDI	δ-Cluster	B.I.
0.005	219	417	418	391	216
0.015	219	417	418	391	401
0.025	219	417	418	391	176
0.035	219	417	418	391	175
0.045	219	417	418	391	175

Table 10: Group3.Retail: $\alpha = 0.006$, $\epsilon = 0.05$

Case Study

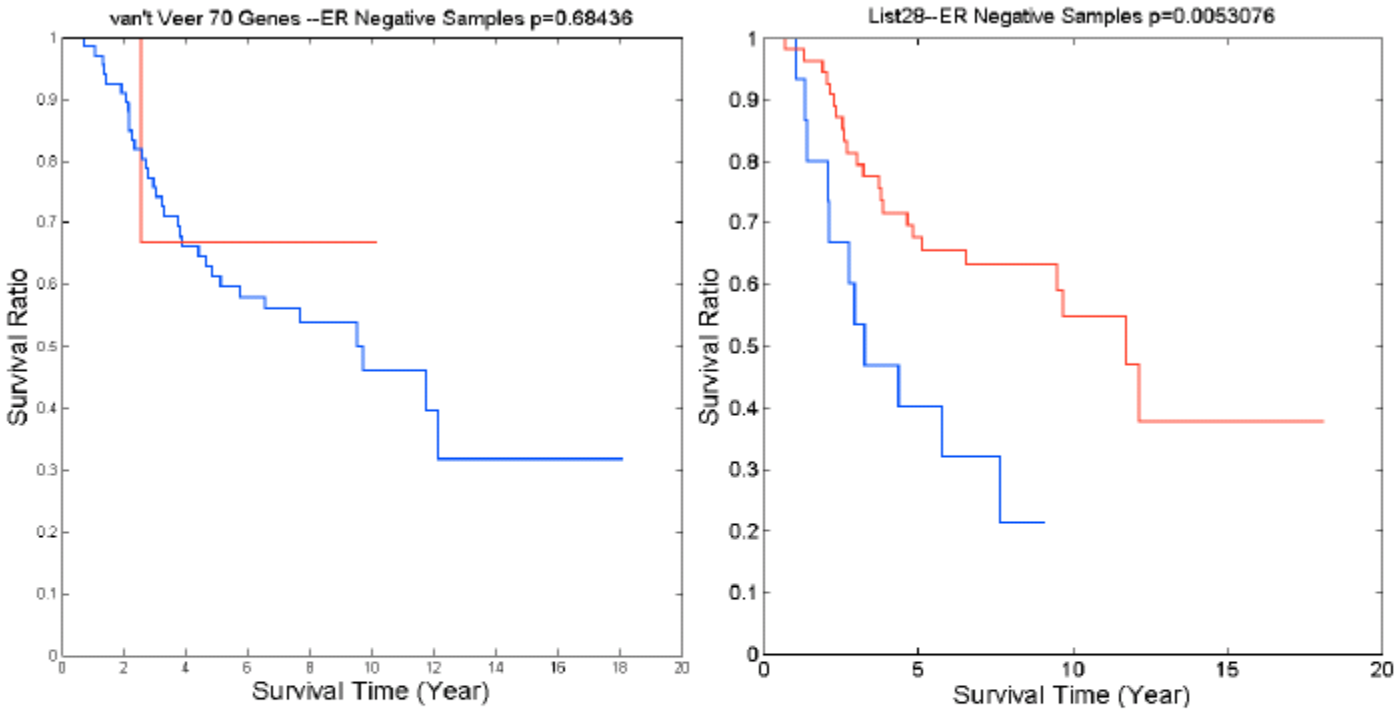


Figure 4: Left: The Kaplan-Meier curves for the two groups from the ER-negative patients separated using the cluster generated from by the 70-gene signature. Right: The Kaplan-Meier curves for the same group of patients obtained using the BI algorithm.

Questions

- How does the complexity of frequent itemsets arise?
- Can the large number of frequent itemsets be generated from a small number of patterns through their interactions?
- Can we summarize a collection of frequent itemsets and provide support information using only a small number of frequent itemsets?
- How can we evaluate the usefulness of concise patterns?

Thanks!!!

Questions?