



Computational Learning Theory

Formal Concept Analysis and Frequent Item Set Mining

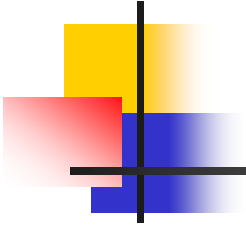
Akihiro Yamamoto 山本 章博

<http://www.iip.ist.i.kyoto-u.ac.jp/member/akihiro/>
akihiro@i.kyoto-u.ac.jp



Contents

- Item Set Mining and the A Priori Algorithm
- Formal Concept Analysis
- Closed Patterns



ITEM SET MINING



A Simple Example

- Set of all items: $X = \{A, B, C, D, E, F\}$

Transaction ID	Item Sets
...	
3256	{A, C, D}
3257	{B, C, E}
3258	{A, B, C, E}
3259	{A, B, E, F}
...

- “Items A and C might be bought together.”



Bit-vector Representation

- Every transaction can be represented as a bit-vector of n dimension, where $n = |X|$.

ID	A	B	C	D	E	F
...						
3256	1	0	1	1	0	0
3257	0	1	1	0	1	0
3258	1	1	1	0	1	0
3259	1	1	0	0	1	1
...						



Bag of Words

- Let $X = \{A_1, A_2, \dots, A_k\}$ be a finite set of words.
- For a sentence s , we define $T(s) = (x_1, x_2, \dots, x_k)$ where
$$x_i = \begin{cases} 1 & \text{if word } A_i \text{ appears in } s \\ 0 & \text{o.w.} \end{cases}$$
for $i = 1, 2, \dots, n$

Example

$W = (\text{arithmetic, book, compute, paper, suppose, square, symbol, write})$

s_1 : **Computing** is normally done by **writing** certain **symbols** on **paper**.

s_2 : We may **suppose** this **paper** is divided into **squares** like a child's **arithmetic book**.

$$T(s_1) = (0, 0, 1, 1, 0, 0, 1, 1)$$

$$T(s_2) = (1, 1, 0, 1, 1, 1, 0, 0)$$



Mathematical Definitions

- Assuming a finite set of all items

$$X = \{A_1, A_2, \dots, A_n\}$$

- A **transaction** is a pair $t = (i, T)$ of an identifier $i \in \mathbf{N}$ and a finite set of items $T \in X$
- A **transaction database** D is a finite set of transactions in which no pair of transactions have a same identifier, that is,
$$t = (i, T) \in D \text{ and } s = (j, S) \in D \text{ imply } i \neq j.$$
- A **pattern** is a finite set of items.
 - Transactions are for training data patterns are rules.



Mathematical Definitions (2)

- For a pattern P and a transaction $t = (i, T)$, we say t satisfies P (or P matches t) iff $P \subset T$.
- Let $D(P) = \{ t \mid P \text{ matches } t \}$.
- The **support** of P in a transaction database D is defined as $\text{supp}(P) = |D(P)| / |D|$.
 - The support is also called the relative frequency.



Definition of Learning Task

- Assuming a set of items X
- For a given transaction database D and a minimal support (threshold) σ s.t. $0 \leq \sigma \leq 1$,
enumerate all patterns P s.t. $\text{supp}(P) \geq \sigma$.



A Very Simple Example

ID	A	B	C	D	E	F
1	1	0	1	1	0	0
2	0	1	1	0	1	0
3	1	1	1	0	1	0
4	1	1	0	0	1	1

$\text{supp}(\{A\}) = \text{supp}(\{B\}) = \text{supp}(\{C\}) = \text{supp}(\{E\}) = 0.75,$

$\text{supp}(\{D\}) = \text{supp}(\{F\}) = 0.25$

$\text{supp}(\{A, B\}) = \text{supp}(\{A, C\}) = 0.5, \text{supp}(\{A, D\}) = 0.25, \dots$

Monotonicity of the Support

Lemma For two patterns P and Q ,

$$P \subseteq Q \Rightarrow \text{supp}(P) \geq \text{supp}(Q)$$

ID	A	B	C	D	E	F
1	1	0	1	1	0	0
2	0	1	1	0	1	0
3	1	1	1	0	1	0
4	1	1	0	0	1	1

$$\text{supp}(\{A\})=0.75 \geq \text{supp}(\{A, B\})=0.25$$

$$\text{supp}(\{B\})=0.5 \geq \text{supp}(\{A, B\})=0.25$$

$$\text{supp}(\{A\})=0.75 \geq \text{supp}(\{A, C\})=0.5$$



Example Run of Apri-Ori Alg.

ID	A	B	C	D	E	F
1	1	0	1	1	0	0
2	0	1	1	0	1	0
3	1	1	1	0	1	0
4	1	1	0	0	1	1

$$\sigma = 0.5$$

$$C_1 = \{\{A\}, \{B\}, \dots, \{F\}\}$$

$$L_1 = \{\{A\}, \{B\}, \{C\}, \{E\}\}$$

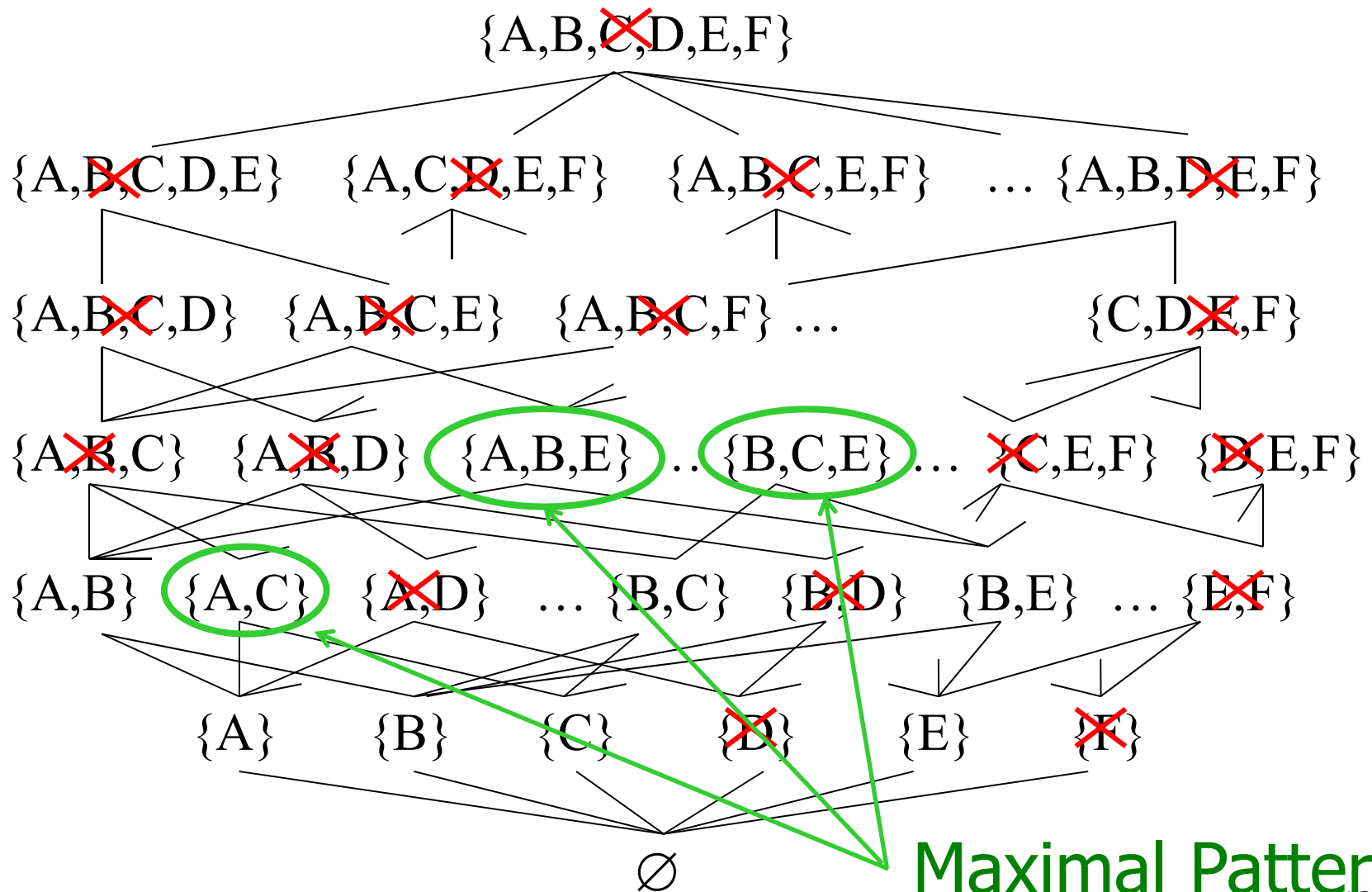
$$C_2 = \{\{A, B\}, \{A, C\}, \\ \{A, E\}, \{B, C\}, \\ \{B, E\}, \{C, E\}\}$$

$$L_2 = \{\{A, B\}, \{A, C\}, \\ \{A, E\}, \{B, C\}, \\ \{B, E\}, \{C, E\}\}$$

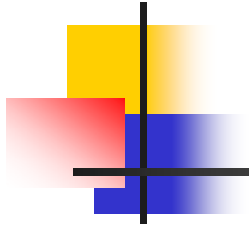
$$C_3 = \{\{A, B, C\}, \{A, B, E\} \\ \{B, C, E\}\}$$

$$L_3 = \{\{A, B, E\}, \{B, C, E\}\}$$

Maximal Patterns in the Hasse Diagram



Maximal Patterns



FORMAL CONCEPT ANALYSIS



Bit-vector Representation

- Every transaction can be represented as a bit-vector of n dimension, where $n = |X|$.

ID	A	B	C	D	E	F
...						
3256	1	0	1	1	0	0
3257	0	1	1	0	1	0
3258	1	1	1	0	1	0
3259	1	1	0	0	1	1
...						



Context Table Representation

- Instead of “1”, we use ●.

ID	A	B	C	D	E	F
...						
3256	●		●	●		
3257		●	●		●	
3258	●	●	●		●	
3259	●	●			●	●
...						

Formal Concepts

- A formal concept is a maximal rectangular filled with ●, without considering the ordering of law and column.

	m ₁	m ₂	m ₃	m ₄	m ₅	m ₆	m ₇	m ₈	m ₉	m ₁₀	m ₁₁	m ₁₂
g ₁	●	●	●	●	●	●	●	●	●	●		
g ₂	●	●	●	●			●	●				
g ₃	●	●	●		●	●					●	●
g ₄	●	●		●	●	●					●	●

	m ₁	m ₂	m ₃	m ₄	m ₇	m ₈	m ₅	m ₆	m ₉	m ₁₀	m ₁₁	m ₁₂
g ₁	●	●	●	●	●	●	●	●	●	●		
g ₂	●	●	●	●	●	●						
g ₃	●	●	●				●	●			●	●
g ₄	●	●		●			●	●			●	●



Intuitive Explanation

In the context of item set mining, a formal concept is a pair of a set A of transaction and a set B of items such that

- every transaction in A contains all items in B ,
- every items in B is contained by all transactions in A ,
- for every item i which is not in B , at least one transaction in A does not contain i , and
- for every transaction t which is not in A , at least one item is not contained by t .



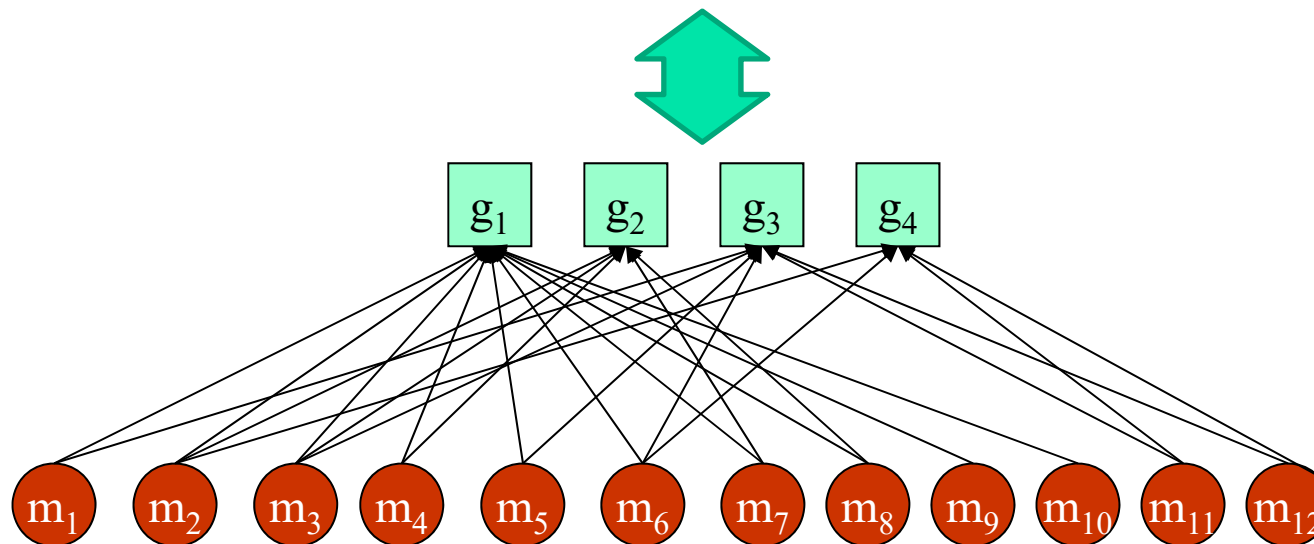
Mathematical Definition

- A formal context $K=(G, M, I)$ consists of two sets G (objects, *Gegenstand*) and M (attributes, *Merkmal*) and a binary relation $I \subseteq G \times M$.
- We define two functions $f: 2^G \rightarrow 2^M$ and $h: 2^M \rightarrow 2^G$
$$f(A) = \{ m \in M \mid (g, m) \in I \text{ for all } g \in A \}$$
$$h(B) = \{ g \in G \mid (g, m) \in I \text{ for all } m \in B \}$$
 - The pair (f, h) is called a Galois connection between 2^G and 2^M .
- A formal concept of K is a pair $C=(A, B)$ with $A \subseteq G$ and $B \subseteq M$ such that $f(A)=B$ and $h(B) = A$, i.e.
$$h(f(A))=A \text{ and } f(h(B))=B.$$
 - A is called the extent of C and B is called the intent of C .

Bipartite Graph Representation

- Every context table can be represented as a bipartite graph.
- Every formal concept is represented as a **bipartite clique**.

	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}	m_{11}	m_{12}
g_1	●	●	●	●	●	●	●	●	●	●		
g_2	●	●	●	●			●	●				
g_3	●	●	●		●	●					●	●
g_4	●	●		●	●	●					●	●





Some Propositions

For a context $K=(G, M, I)$, $A, A_1, A_2 \subseteq G$ and $B, B_1, B_2 \subseteq M$,

- $A_1 \subseteq A_2 \Rightarrow f(A_2) \subseteq f(A_1)$
- $B_1 \subseteq B_2 \Rightarrow h(B_2) \subseteq h(B_1)$
- $A \subseteq h(f(A))$
- $B \subseteq f(h(B))$
- $A \subseteq h(B) \Leftrightarrow B \subseteq f(A) \Leftrightarrow A \times B \subseteq I$

- $h(f(A_1 \cup A_2)) = h(f((h(f(A_1)) \cup h(f(A_2)))))$

- $f(h(B_1 \cup B_2)) = f(h((f(h(B_1)) \cup f(h(B_2)))))$

- $A_1 \subseteq h(f(A_2)) \Rightarrow h(f(A_1)) = h(f(A_2))$

and $h(f(A_1 \cup A)) = h(f(A_2 \cup A))$

- $B_1 \subseteq f(h(B_2)) \Rightarrow f(h(B_1)) = f(h(B_2))$

and $f(h(B_1 \cup B)) = f(h(B_2 \cup B))$



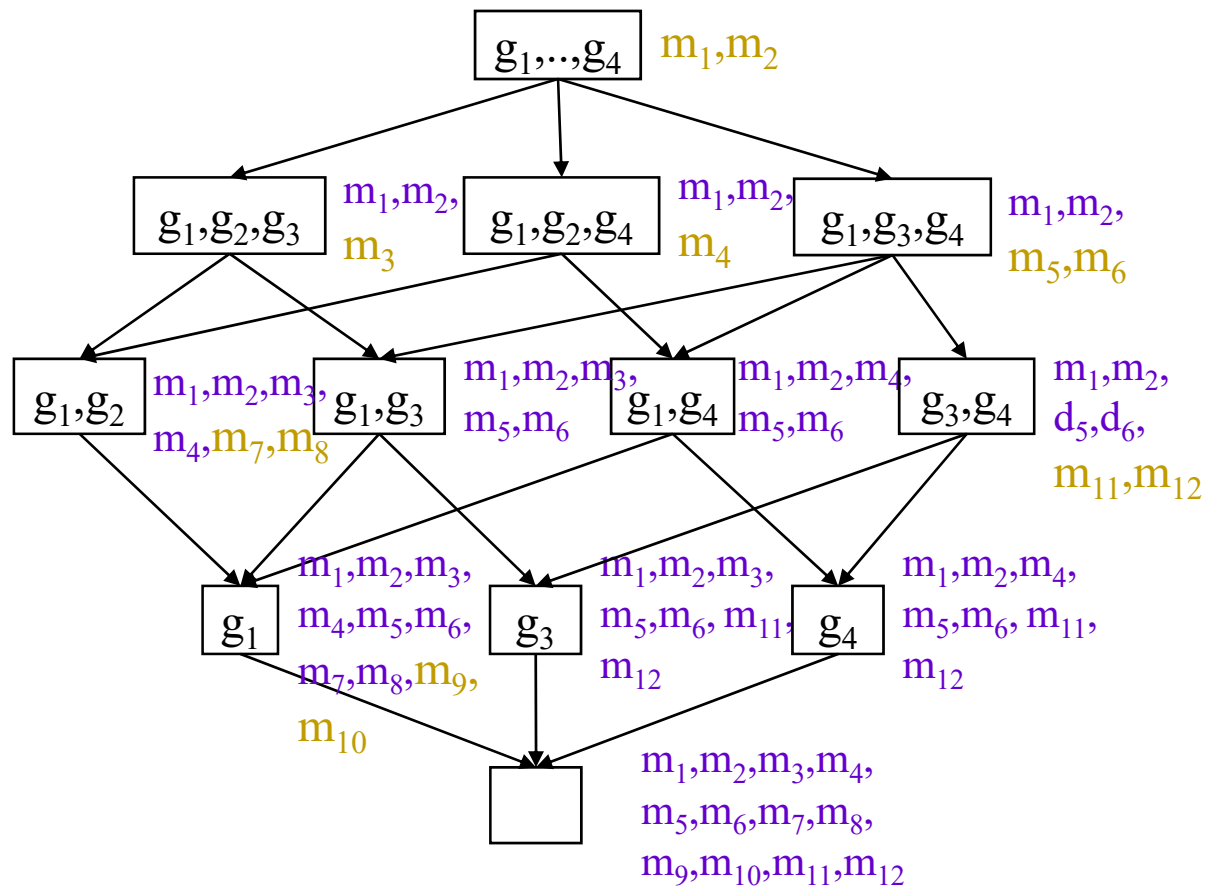
Some Propositions

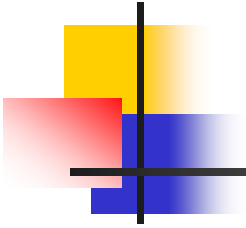
For formal concepts $C_1=(A_1, B_1)$ and $C_2=(A_2, B_2)$,

$$A_1 \subseteq A_2 \Leftrightarrow B_2 \subseteq B_1$$

Hasse Diagram of FCs

- We can draw another Hasse diagram with all of the formal concepts.





CLOSED PATTERNS

Closed Item Sets [Pasquier et al.]

- For a transaction data, we let G is the set of all transaction id and M is the set of all items.
- An pattern B is **closed** iff $B = f(h(B))$, i.e, $(h(B), B)$ is a formal concept.

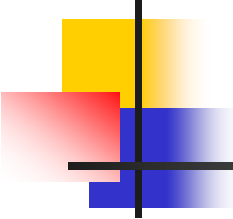
1	a	c	d
2	b	c	e
3	a	b	c
4	b		e
5	a	b	c

$\sigma = 0.5$

Frequent closed pattern: c, ac, be, bce

Frequent but not closed pattern: a, bc, \dots

- For a transaction data, we let G is the set of all transaction ids and M is the set of all items.



Lemmas

Lemma For a context $K=(G, M, I)$, $A \subseteq G$ and $B \subseteq M$

- $h(f(A)) = \bigcap_{g \in G} \{f(\{g\}) \mid A \subseteq f(\{g\})\}$
- $f(h(B)) = \bigcap_{m \in M} \{h(\{m\}) \mid B \subseteq f(\{m\})\}$

Corollary For closed patterns B_2 , if $B_2 \subseteq B_1$ and $B_2 \neq B_1$, then $\text{supp}(B_2) > \text{supp}(B_1)$.

Corollary For two closed patterns B_1 and B_2 , if $B_2 \subseteq B_1$ and $B_2 \neq B_1$, then $\text{supp}(B_2) > \text{supp}(B_1)$.

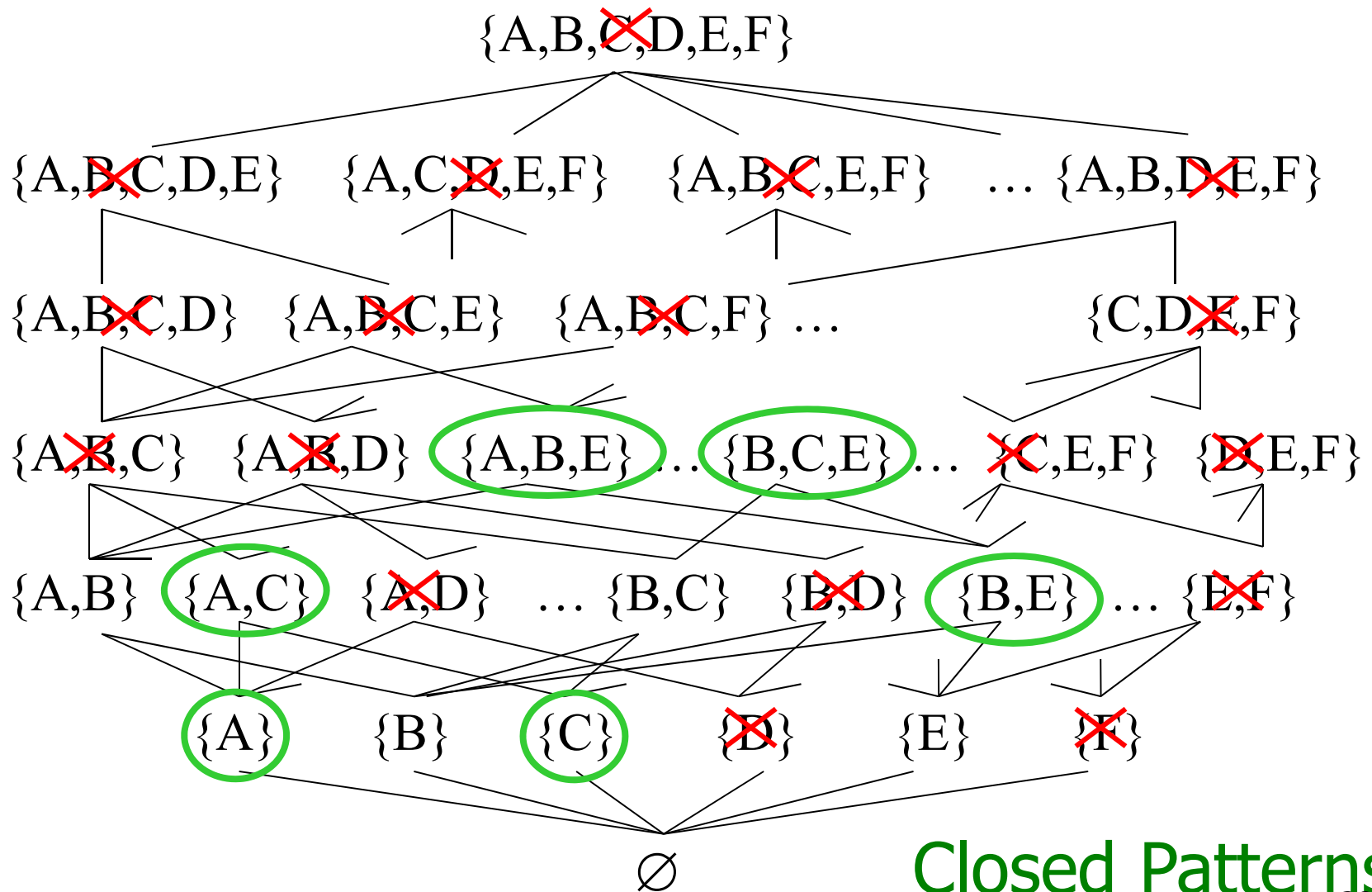
Lemma [Pasquier et al.] Every pattern B_1 of $\text{supp}(B_1) = \sigma$ can be derived from some **closed** pattern B_2 of $\text{supp}(B_2) = \sigma$.



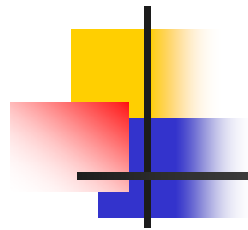
Proposition

Proposition Every **maximally frequent closed** pattern is a **frequent closed** pattern.

Closed Patterns in the Hasse Diagram



Closed Patterns



Frequent Closed ItemSets

$$\sigma = 0.5$$

ID	A	B	C	D	E	F
1	●		●	●		
2		●	●		●	
3	●	●	●		●	
4	●	●			●	●

ID	A	B	C	D	E	F
1	●		●	●		
2		●	●		●	
3	●	●	●		●	
4	●	●			●	●

ID	A	B	C	D	E	F
1	●		●	●		
2		●	●		●	
3	●	●	●		●	
4	●	●			●	●

Frequent Closed ItemSets

$$\sigma = 0.25$$

ID	A	B	C	D	E	F
1	●		●	●		
2		●	●		●	
3	●	●	●		●	
4	●	●			●	●

ID	A	B	C	D	E	F
1	●		●	●		
2		●	●		●	
3	●	●	●		●	
4	●	●			●	●

ID	A	B	C	D	E	F
1	●		●	●		
2		●	●		●	
3	●	●	●		●	
4	●	●			●	●

ID	A	B	C	D	E	F
1	●		●	●		
2		●	●		●	
3	●	●	●		●	
4	●	●			●	●

How to Find Closed Patterns(1)

Start at \emptyset

ID	A	B	C	D	E	F
1	●		●	●		
2		●	●		●	
3	●	●	●		●	
4	●	●			●	●

$\{A,B,C\}$ $\{A,B,D\}$ $\{A,B,E\}$... $\{B,C,E\}$... $\{C,E,F\}$ $\{D,E,F\}$

$\{A,B\}$ $\{A,C\}$ $\{A,D\}$... $\{B,C\}$ $\{B,D\}$ $\{B,E\}$... $\{E,F\}$

$\{A\}$ $\{B\}$ $\{C\}$ $\{D\}$ $\{E\}$ $\{F\}$

\emptyset

Closed Patterns

How to Find Closed Patterns(2)

Add A to \emptyset

ID	A	B	C	D	E	F
1	●		●	●		
2		●	●		●	
3	●	●	●		●	
4	●	●			●	●

{A,B,C} {A,B,D} {A,B,E} ... {B,C,E} ... {C,E,F} {D,E,F}

{A,B} {A,C} {A,D} ... {B,C} {B,D} {B,E} ... {E,F}

{A} {B} {C} {D} {E} {F}

\emptyset

Closed Patterns

How to Find Closed Patterns(3)

Choose transactions containing {A}

ID	A	B	C	D	E	F
1	●		●	●		
2		●	●		●	
3	●	●	●		●	
4	●	●			●	●

{A,B,C} {A,B,D} {A,B,E} ... {B,C,E} ... {C,E,F} {D,E,F}

{A,B} {A,C} {A,D} ... {B,C} {B,D} {B,E} ... {E,F}

{A} {B} {C} {D} {E} {F}

∅

Closed Patterns

How to Find Closed Patterns(4)

Find the pattern common to {1, 3, 4}.
If the pattern {A} is closed, output it.

ID	A	B	C	D	E	F
1	●		●	●		
2		●	●		●	
3	●	●	●		●	
4	●	●			●	●

{A,B,C} {A,B,D} {A,B,E} ... {B,C,E} ... {C,E,F} {D,E,F}

{A,B} {A,C} {A,D} ... {B,C} {B,D} {B,E} ... {E,F}

{A} {B} {C} {D} {E} {F}

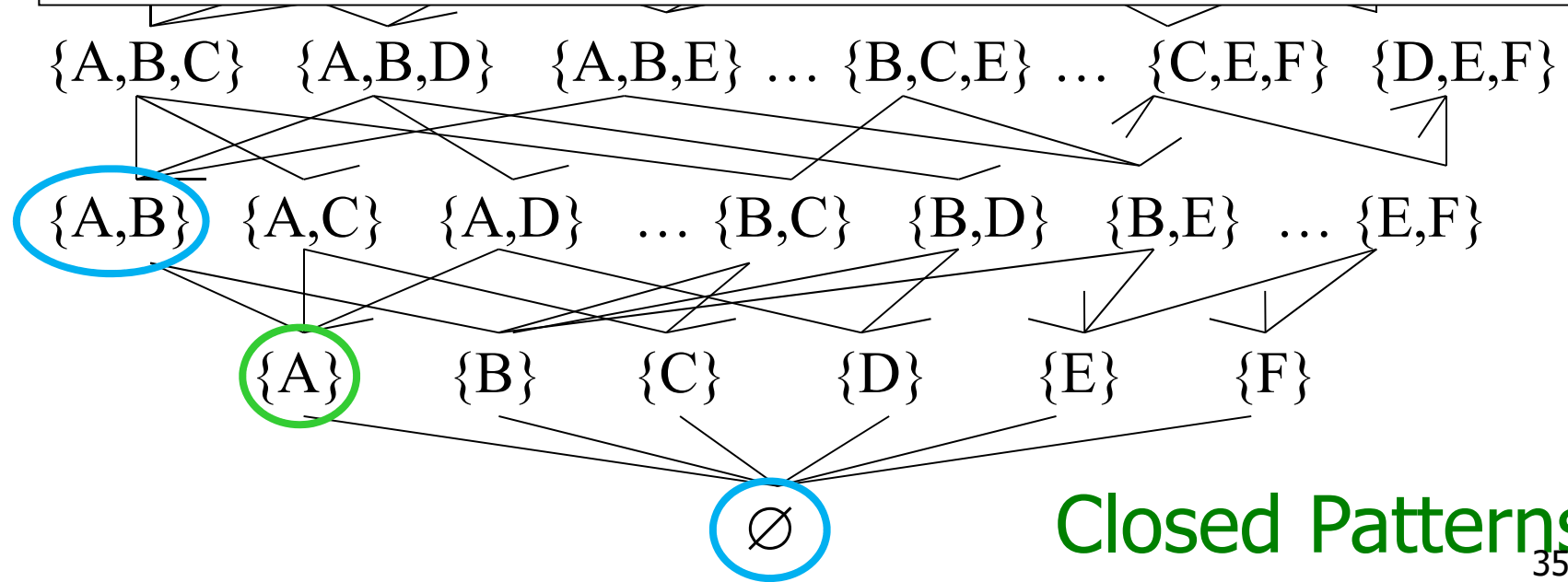
∅

Closed Patterns

How to Find Closed Patterns(5)

Add B to {A}
: B is chosen from the
items $X > A$

ID	A	B	C	D	E	F
1	●		●	●		
2		●	●		●	
3	●	●	●		●	
4	●	●			●	●



How to Find Closed Patterns(6)

Choose transactions containing {B, A}

ID	A	B	C	D	E	F
1	●		●	●		
2		●	●		●	
3	●	●	●		●	
4	●	●			●	●

{A,B,C} {A,B,D} {A,B,E} ... {B,C,E} ... {C,E,F} {D,E,F}

{A,B} {A,C} {A,D} ... {B,C} {B,D} {B,E} ... {E,F}

{A} {B} {C} {D} {E} {F}

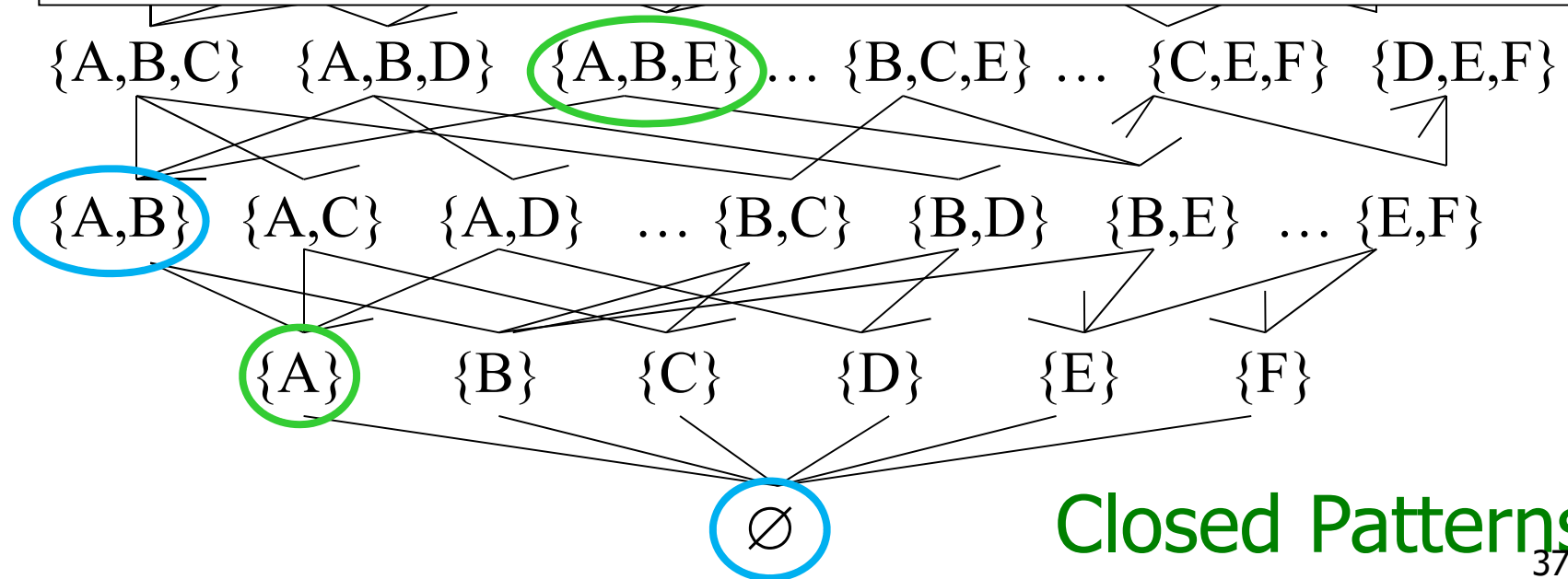
∅

Closed Patterns

How to Find Closed Patterns(7)

Find the pattern common to {3, 4}.
If the pattern {A, B, E} is closed, output it.

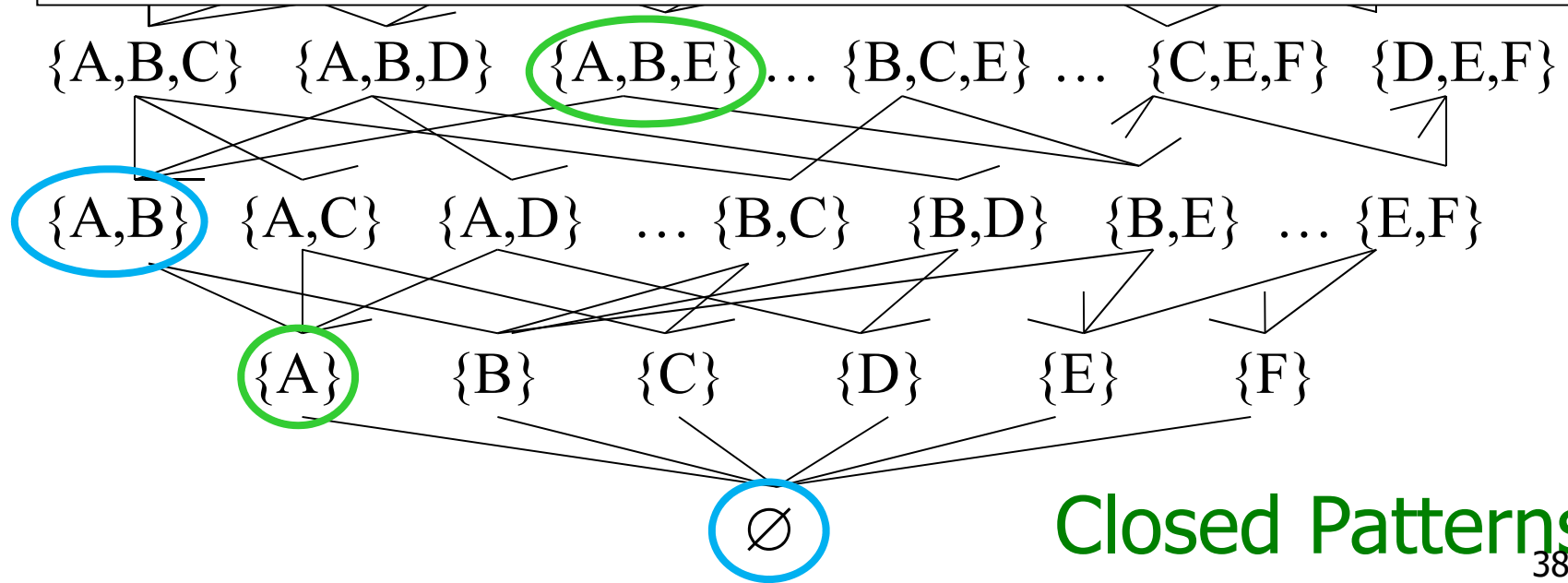
ID	A	B	C	D	E	F
1	●		●	●		
2		●	●		●	
3	●	●	●		●	
4	●	●			●	●



How to Find Closed Patterns(8)

Back track to {A}

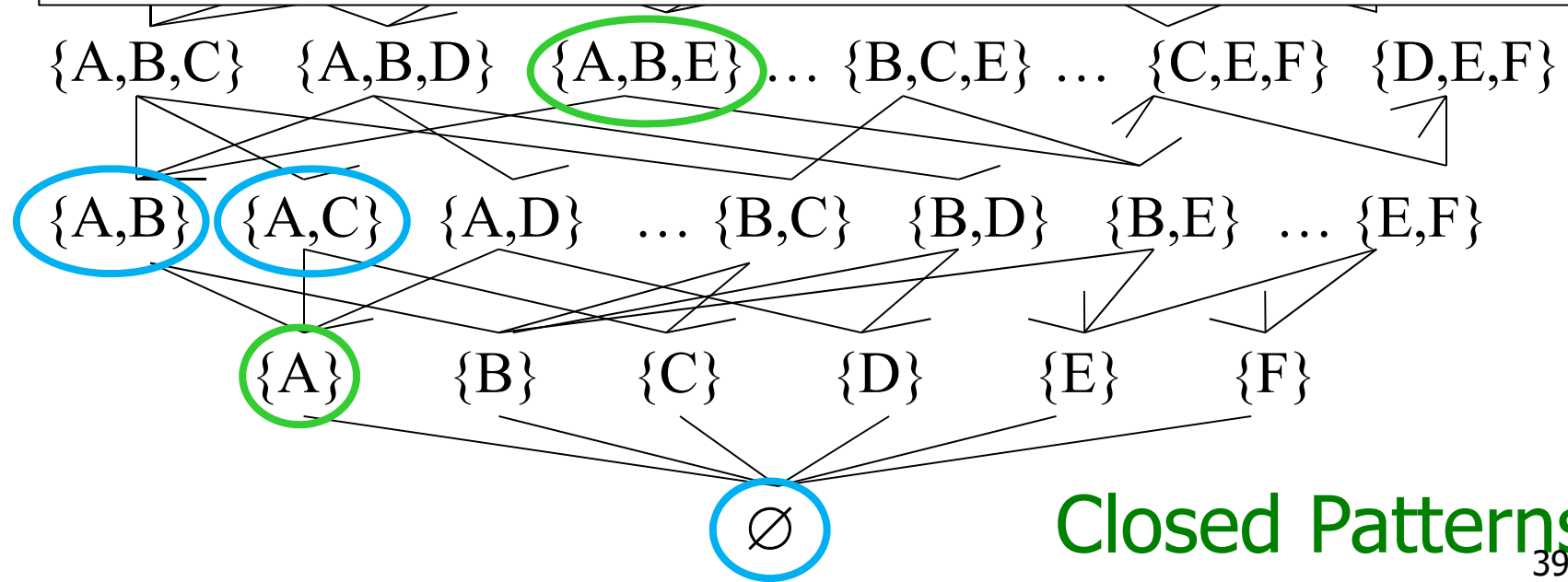
ID	A	B	C	D	E	F
1	●		●	●		
2		●	●		●	
3	●	●	●		●	
4	●	●			●	●



How to Find Closed Patterns(9)

Add C to {A}
 : C is chosen from the
 items $X > A$
 Choose transactions
 containing {C, A}

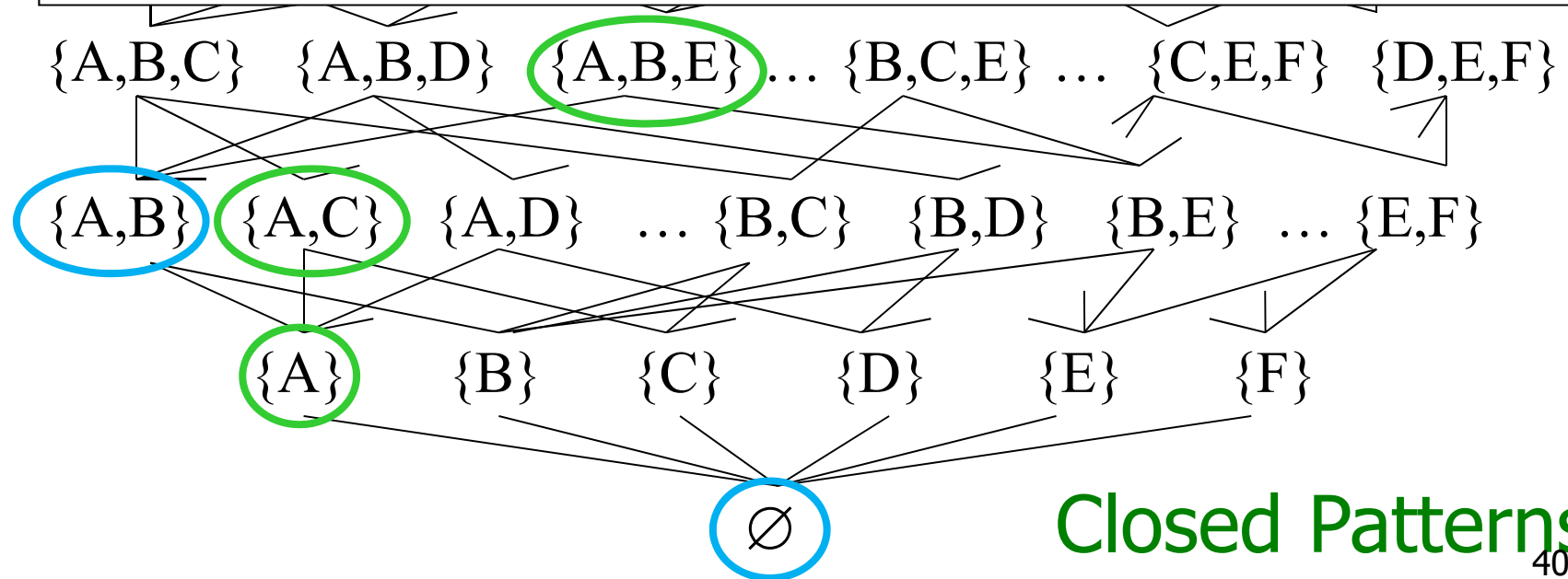
ID	A	B	C	D	E	F
1	●		●	●		
2		●	●		●	
3	●	●	●		●	
4	●	●			●	●



How to Find Closed Patterns(10)

Find the pattern common to {1, 3}.
If the pattern {A, C} is closed, output it.

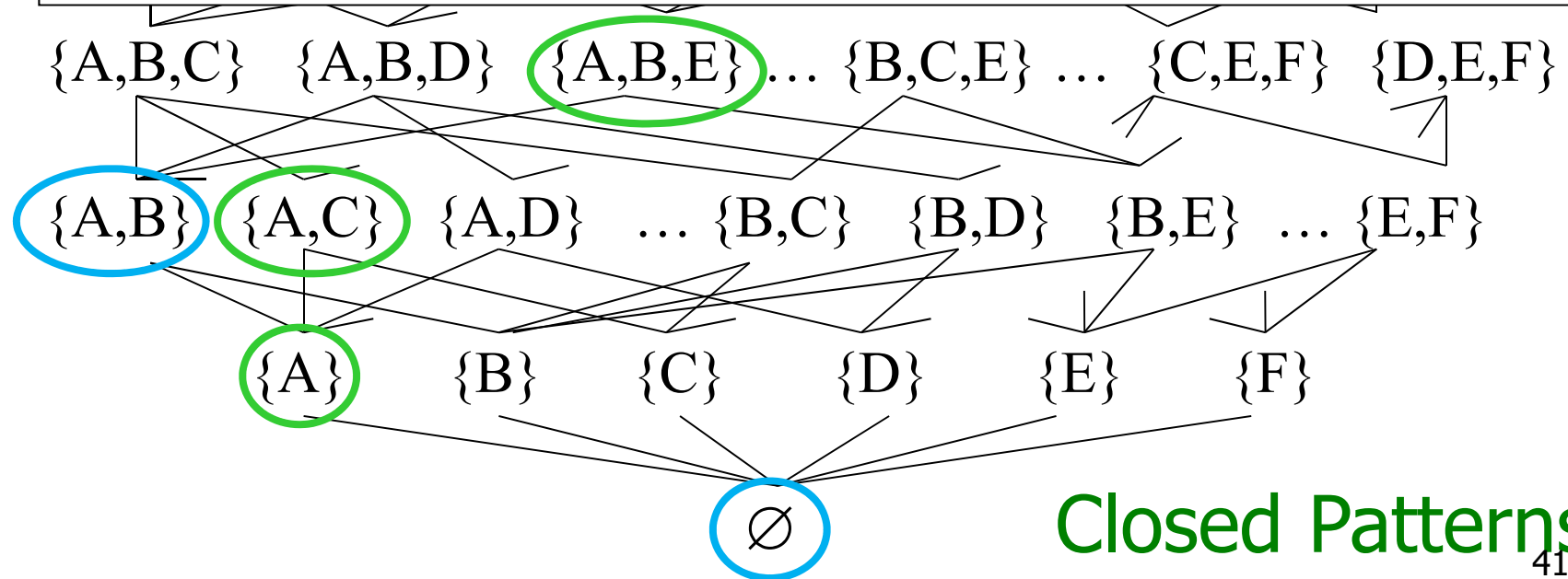
ID	A	B	C	D	E	F
1	●		●	●		
2		●	●		●	
3	●	●	●		●	
4	●	●			●	●



How to Find Closed Patterns(11)

Back track to {A}

ID	A	B	C	D	E	F
1	●		●	●		
2		●	●		●	
3	●	●	●		●	
4	●	●			●	●

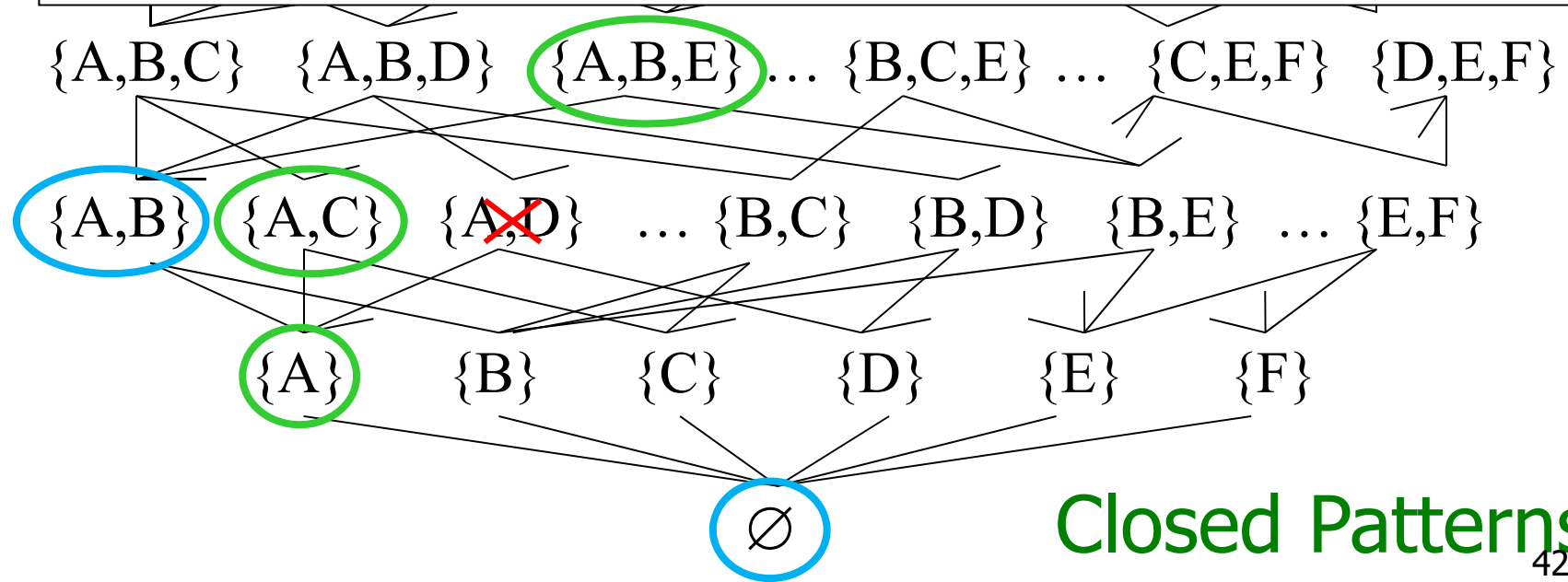


Closed Patterns

How to Find Closed Patterns(12)

Add D to {A}
 : D is chosen from the
 items $X > A$
 But {A, D} is not frequent,
 Again **back track to {A}**

ID	A	B	C	D	E	F
1	●		●	●		
2		●	●		●	
3	●	●	●		●	
4	●	●			●	●



How to Find Closed Patterns(13)

Add E to {A}

: E is chosen from the items $X > A$

Choose transactions

containing {E, A}

Note: We have to check if {A, E} or {A, E, F} is a closed itemset or not.

ID	A	B	C	D	E	F
1	●		●	●		
2		●	●		●	
3	●	●	●		●	
4	●	●			●	●

{A,B,C} {A,B,D} {A,B,E} ... {B,C,E} ... {C,E,F} {D,E,F}

{A,B} {A,C} {A,D} ... {B,C} {B,D} {B,E} ... {E,F}

{A} {B} {C} {D} {E} {F}

∅

Closed Patterns

How to Find Closed Patterns(14)

Add F to {A}

: F is chosen from the items $X > A$

Choose transactions

containing {F, A}

Note: We have to check if {A, F} is a closed itemset or not.

ID	A	B	C	D	E	F
1	●		●	●		
2		●	●		●	
3	●	●	●		●	
4	●	●			●	●

{A,B,C} {A,B,D} {A,B,E} ... {B,C,E} ... {C,E,F} {D,E,F}

{A,B} {A,C} {A,D} ~~{B,C}~~ {B,D} {B,E} ... {E,F}

{A} {B} {C} {D} {E} {F}

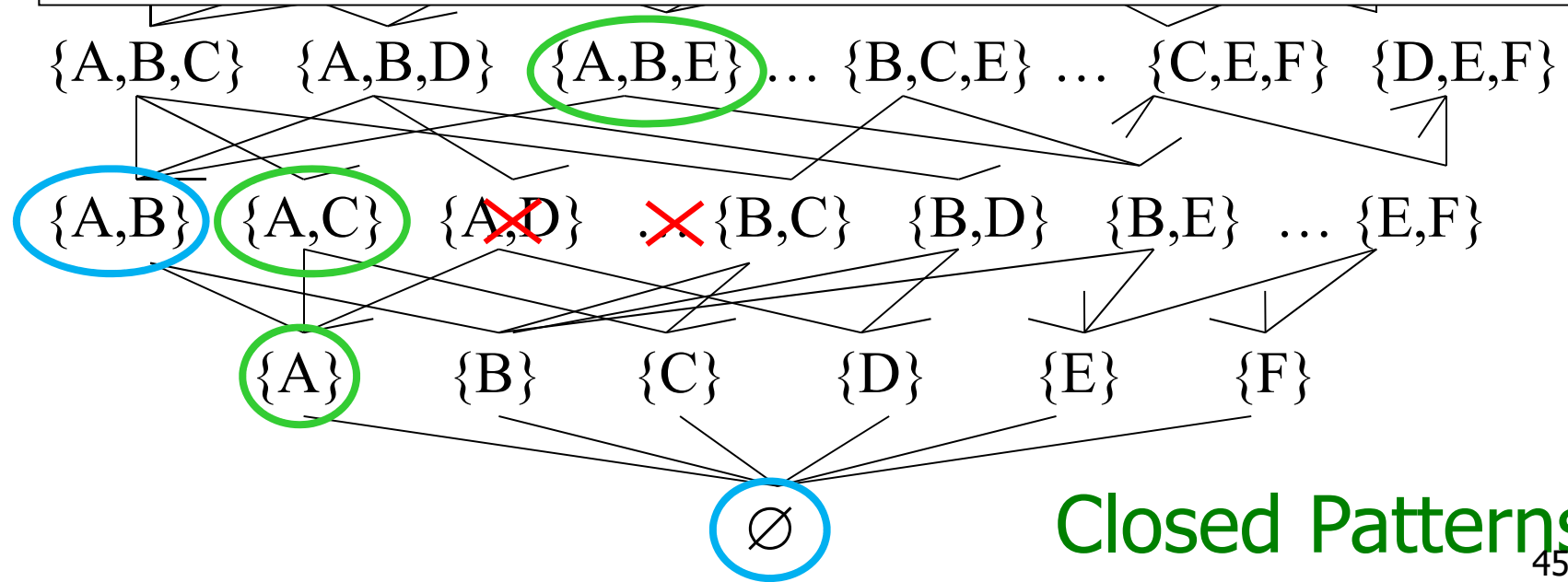
∅

Closed Patterns

How to Find Closed Patterns(15)

Back track to \emptyset

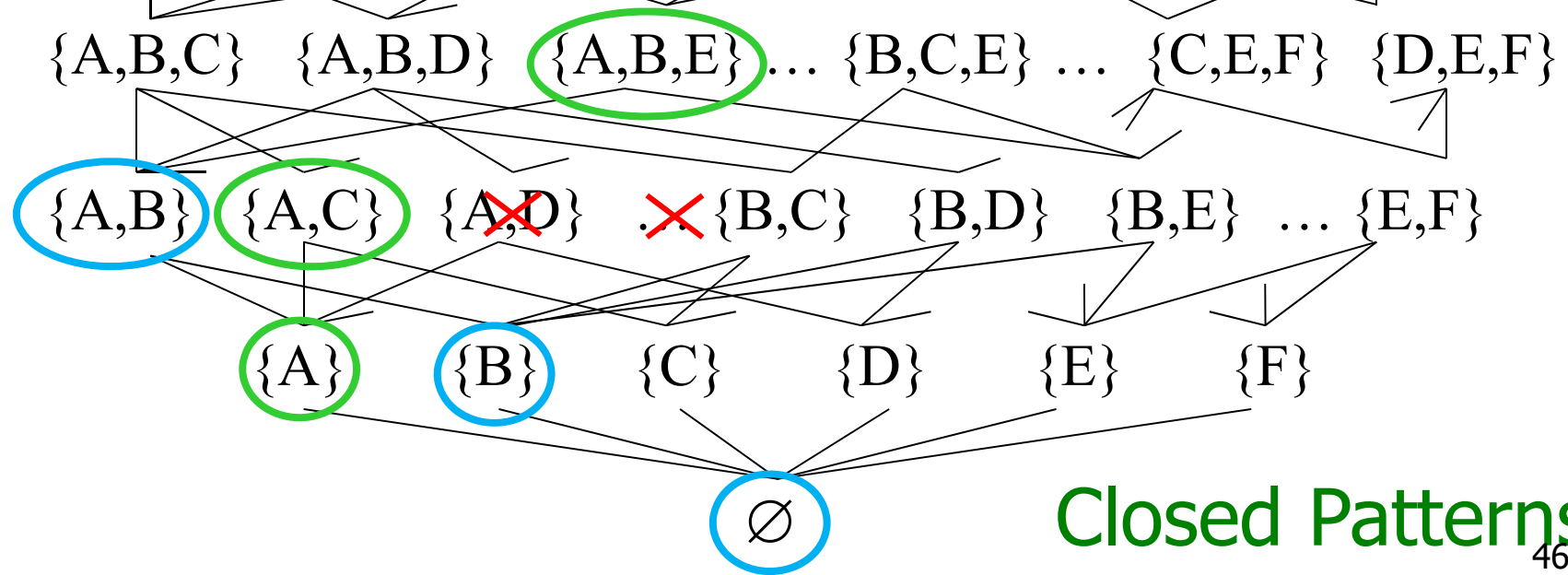
ID	A	B	C	D	E	F
1	●		●	●		
2		●	●		●	
3	●	●	●		●	
4	●	●			●	●



How to Find Closed Patterns(16)

Add B to \emptyset and
Choose transactions
containing {B}

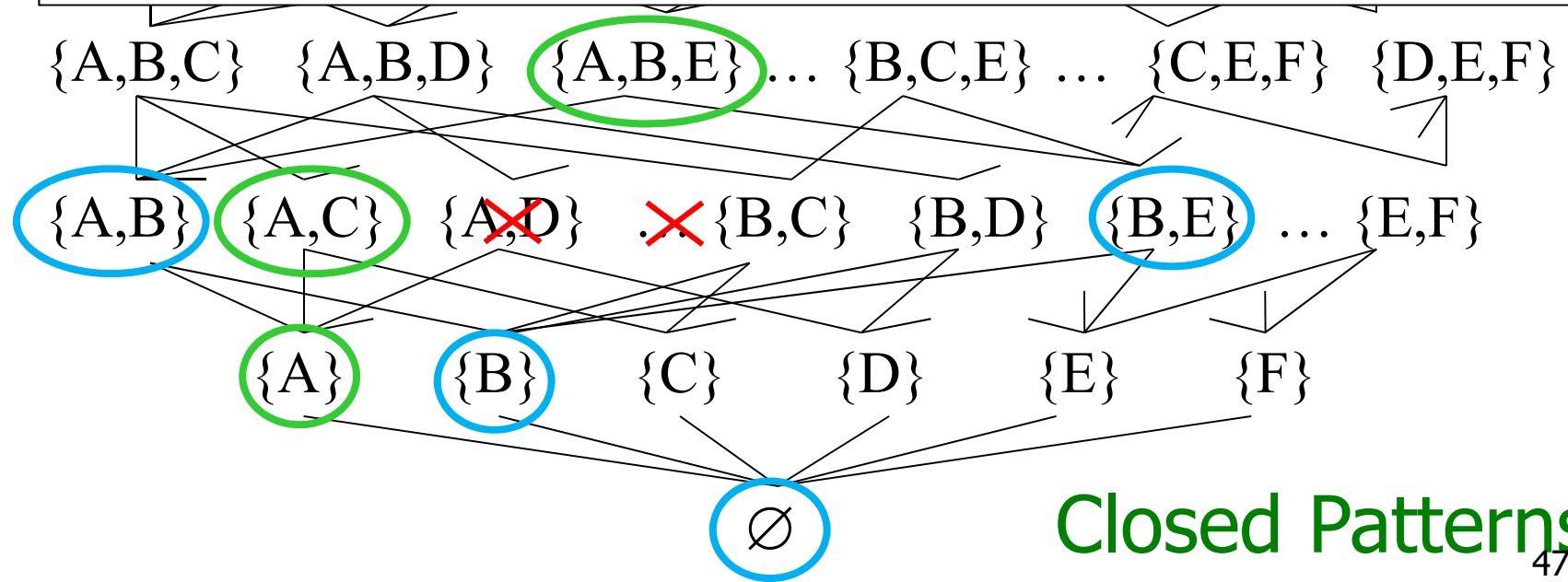
ID	A	B	C	D	E	F
1	●		●	●		
2		●	●		●	
3	●	●	●		●	
4	●	●			●	●



How to Find Closed Patterns(17)

Find the pattern common to {2, 3, 4}.
If the pattern {B, E} is closed, output it.

ID	A	B	C	D	E	F
1	●		●	●		
2		●	●		●	
3	●	●	●		●	
4	●	●			●	●





The LCM Algorithm [Uno et al. 03]

Procedure LCM(P)

output P /* P is a frequent closed itemset */

for each item $X > \text{gap}(P)$ **do**

if P is frequent **and** $\overline{(P \cup \{X\})}$ is closed **then**

 call LCM(P)

end for

where

$\text{gap}(P) = \text{argmin } h(P) = h(P \mid_{\leq X})$ and $h(P) \neq h(P \mid_{< X})$

$\overline{(P \cup \{X\})} = \{ Y \mid Y \in f(h(P \cup \{X\})) \text{ and } Y > X \}$