ASSOCIATION RULE MINING

BEIYU LIN

RULE GENERATION

Anti-monotone

$A \subseteq B$ (F) $\sigma(B) \geq \min\sigma$

$\sigma(A) \geq \sigma(B) \geq \min\sigma$

$\frac{1}{\sigma(A)} \leq \frac{1}{\sigma(B)}$

{ABC}, {BCD} F $\Rightarrow$ {AB} {AC} {BC} {BD} {CD}

Find F w/ k=2 $\Leftrightarrow$ subsets w k=2 {AD} $\rightarrow$? F

- Given a frequent itemset L, find all non-empty subsets $f \subset L$ such that $f \rightarrow L - f$ satisfies the minimum confidence requirement

F {ABCD} $\Rightarrow$ all F, All subset

  - If {A,B,C,D} is a frequent itemset, candidate rules:

    ABC $\rightarrow$D, ABD $\rightarrow$C, ACD $\rightarrow$B, BCD $\rightarrow$A,
    A $\rightarrow$BCD, B $\rightarrow$ACD, C $\rightarrow$ABD, D $\rightarrow$ABC
    AB $\rightarrow$CD, AC $\rightarrow$ BD, AD $\rightarrow$ BC, BC $\rightarrow$AD,
    BD $\rightarrow$AC, CD $\rightarrow$AB,

$2^4 - 1$

{ABCD}

X $\rightarrow$ Y

k=1 {A} $\rightarrow$ {BCD}

⋮

- If $|L| = k$, then there are $2^k - 2$ candidate association rules (ignoring $L \rightarrow \varnothing$ and $\varnothing \rightarrow L$)

{D}

k=2 {AB}

{AC}

⋮

RULE GENERATION

$\{A\} \subseteq \{AB\} \Rightarrow \sigma(\{A\}) \geq \sigma(\{AB\})$

- In general, confidence does not have an anti-monotone property

  c(ABC →D) can be larger or smaller than c(AB →D)

  $c(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$

  $c(ABC \to D) = \frac{\sigma(ABCD)}{\sigma(ABC)}$; $c(AB \to D) = \frac{\sigma(ABD)}{\sigma(AB)}$

- But confidence of rules generated from the same itemset has an anti-monotone property
  - E.g., Suppose {A,B,C,D} is a frequent 4-itemset:

    c(ABC → D) ≥ c(AB → CD) ≥ c(A → BCD)

    $c(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$

    $c(ABC \to D) = \frac{\sigma(ABCD)}{\sigma(ABC)}$

    $c(AB \to CD) = \frac{\sigma(ABCD)}{\sigma(AB)}$

    $\frac{1}{\sigma(AB)} \leq \frac{1}{\sigma(ABC)}$

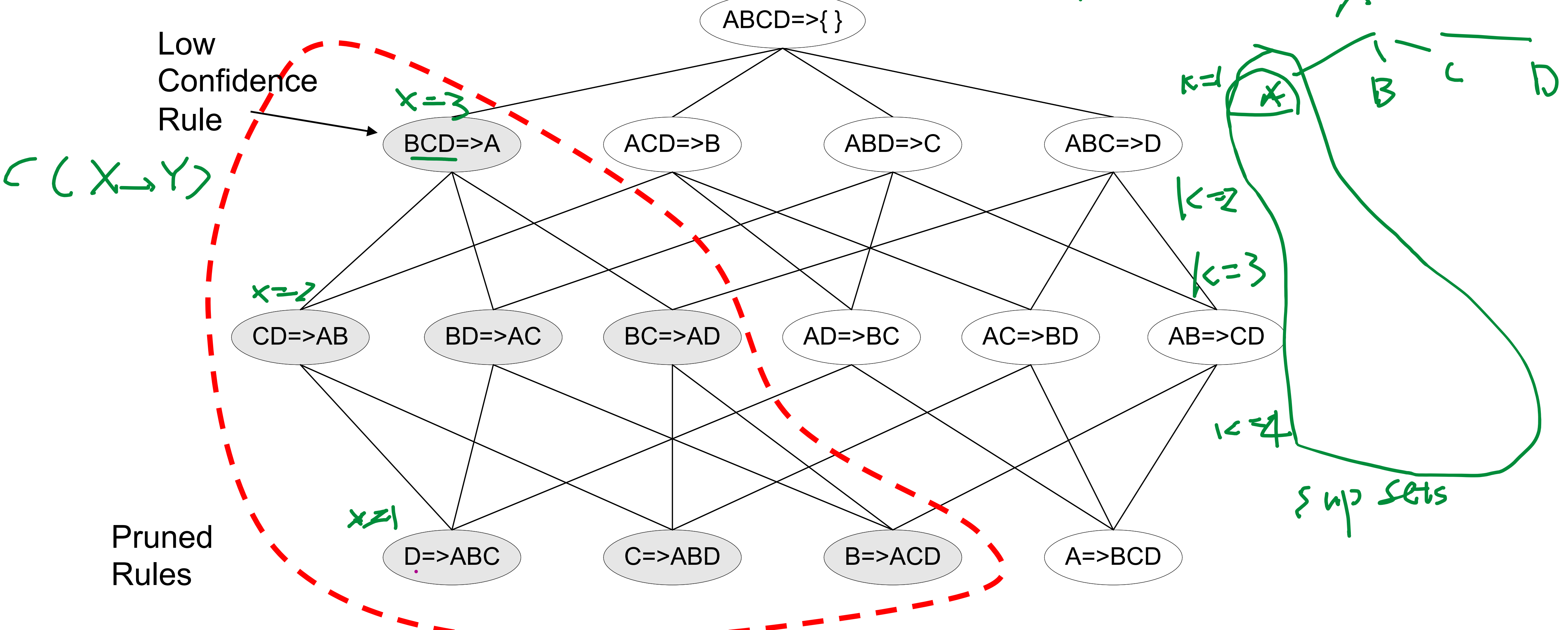
    $AB \subseteq ABC \Rightarrow \sigma(AB) \geq \sigma(ABC)$

  - Confidence is anti-monotone w.r.t. number of items on the RHS of the rule

RULE GENERATION FOR APRIORI ALGORITHM

Lattice of rules

σ | S



ASSOCIATION ANALYSIS: BASIC CONCEPTS

Algorithms and Complexity

FACTORS AFFECTING COMPLEXITY OF APRIORI

- Choice of minimum support threshold
- Dimensionality (number of items) of the data set
- Size of database
- Average transaction width

FACTORS AFFECTING COMPLEXITY OF APRIORI

- Choice of minimum support threshold
  - lowering support threshold results in more frequent itemsets
  - this may increase number of candidates and max length of frequent itemsets
- Dimensionality (number of items) of the data set 6
- Size of database
- Average transaction width 4

TID	*Items*	
1	**Bread, Milk**	2
2	**Beer, Bread, Diaper, Eggs**	4
3	**Beer, Coke, Diaper, Milk**	4
4	**Beer, Bread, Diaper, Milk**	4
5	**Bread, Coke, Diaper, Milk**	4

IMPACT OF SUPPORT BASED PRUNING

TID	Items
1	Bread, Milk
2	Beer, Bread, Diaper, Eggs
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Bread, Coke, Diaper, Milk

Items (1-itemsets)

Item	Count 6
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Minimum Support = 3

If every subset is considered,

$^6C_1 + {}^6C_2 + {}^6C_3$

$6 + 15 + 20 = 41$

With support-based pruning,

$6 + 6 + 4 = 16$

Minimum Support = 2

If every subset is considered,

$^6C_1 + {}^6C_2 + {}^6C_3 + {}^6C_4$

$6 + 15 + 20 + 15 = 56$

FACTORS AFFECTING COMPLEXITY OF APRIORI

- Choice of minimum support threshold
  - lowering support threshold results in more frequent itemsets
  - this may increase number of candidates and max length of frequent itemsets
- Dimensionality (number of items) of the data set
  - More space is needed to store support count of itemsets
  - if number of frequent itemsets also increases, both computation and I/O costs may also increase
- Size of database
- Average transaction width

TID	*Items*
1	**Bread, Milk**
2	**Beer, Bread, Diaper, Eggs**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Bread, Coke, Diaper, Milk**

FACTORS AFFECTING COMPLEXITY OF APRIORI

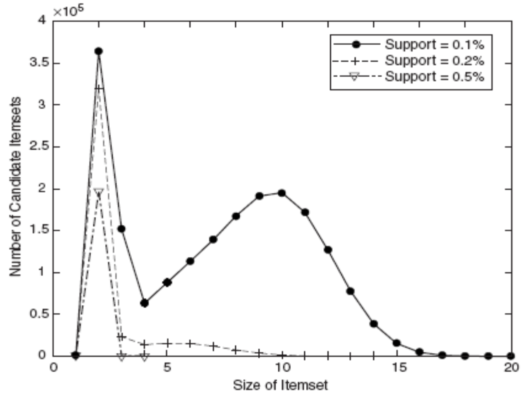
- Choice of minimum support threshold
  - lowering support threshold results in more frequent itemsets
  - this may increase number of candidates and max length of frequent itemsets
- Dimensionality (number of items) of the data set
  - More space is needed to store support count of itemsets
  - if number of frequent itemsets also increases, both computation and I/O costs may also increase
- Size of database
  - run time of algorithm increases with number of transactions
- Average transaction width

TID	*Items*
1	**Bread, Milk**
2	**Beer, Bread, Diaper, Eggs**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Bread, Coke, Diaper, Milk**

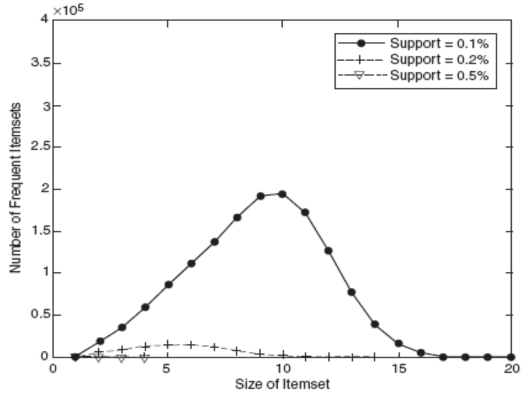
FACTORS AFFECTING COMPLEXITY OF APRIORI

- Choice of minimum support threshold
  - lowering support threshold results in more frequent itemsets
  - this may increase number of candidates and max length of frequent itemsets
- Dimensionality (number of items) of the data set
  - More space is needed to store support count of itemsets
  - if number of frequent itemsets also increases, both computation and I/O costs may also increase
- Size of database
  - run time of algorithm increases with number of transactions
- Average transaction width
  - transaction width increases the max length of frequent itemsets
  - number of subsets in a transaction increases with its width, increasing computation time for support counting

FACTORS AFFECTING COMPLEXITY OF APRIORI

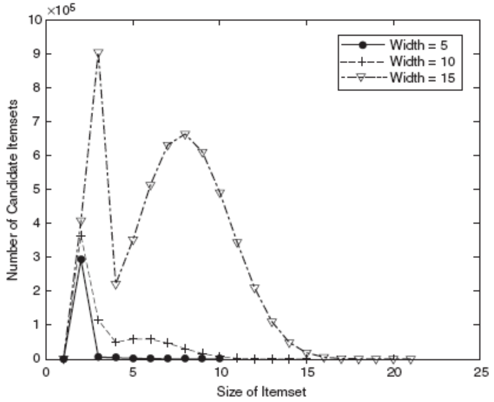


(a) Number of candidate itemsets.

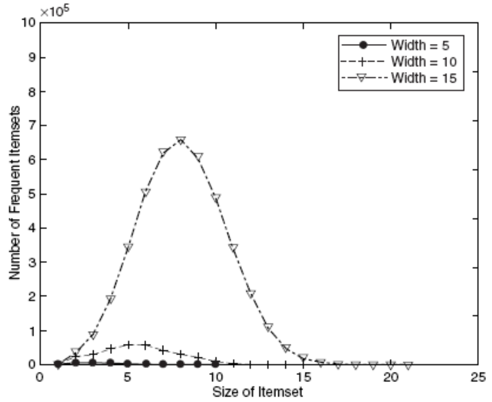


(b) Number of frequent itemsets.

Figure 6.13. Effect of support threshold on the number of candidate and frequent itemsets.



(a) Number of candidate itemsets.



(b) Number of Frequent Itemsets.

Figure 6.14. Effect of average transaction width on the number of candidate and frequent itemsets.

COMPACT REPRESENTATION OF FREQUENT ITEMSETS

- Some frequent itemsets are redundant because their supersets are also frequent

Consider the following data set. Assume support threshold =5

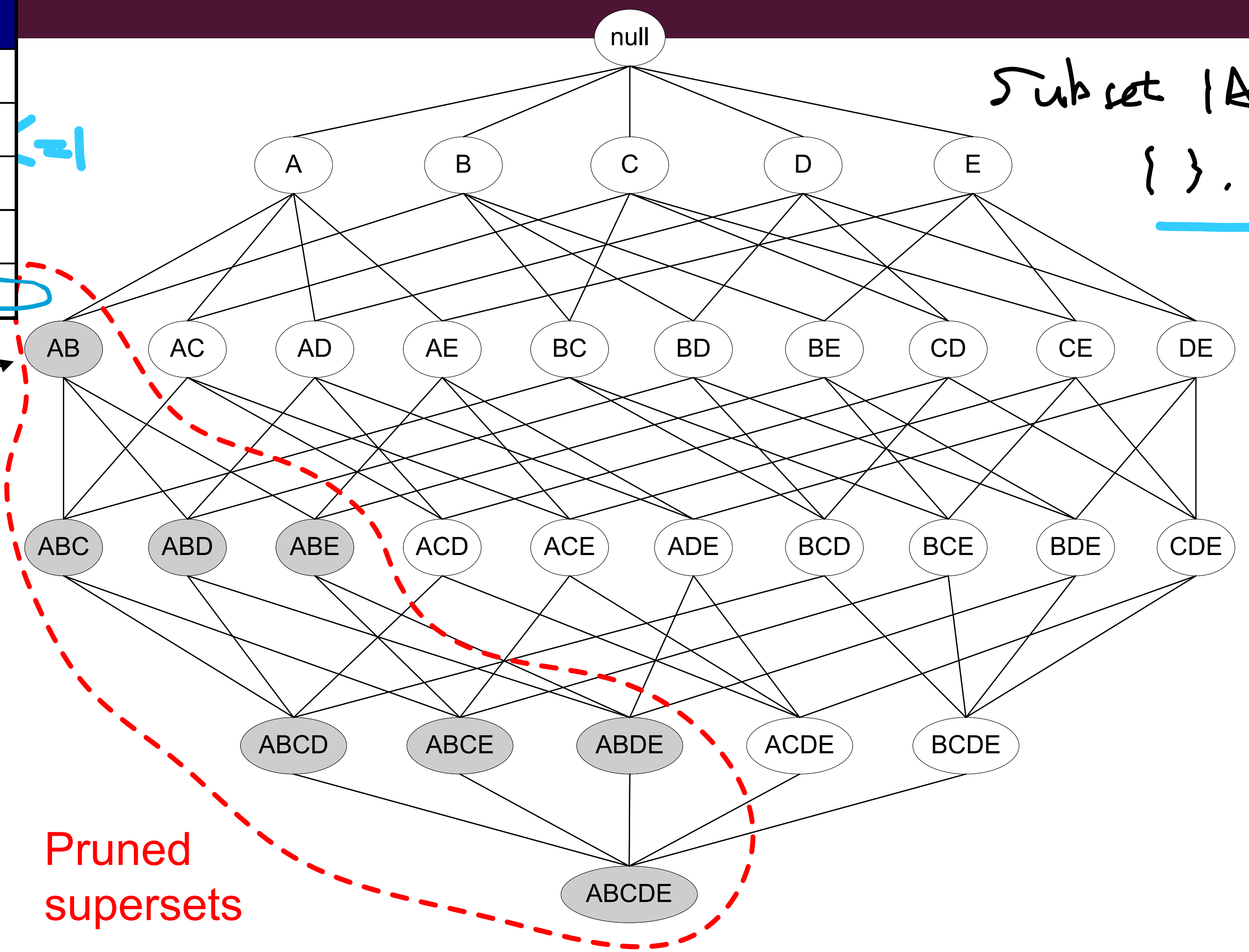
TID	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1

Number of frequent itemsets $= 3 \times \sum_{k=1}^{10} \binom{10}{k}$

- Need a compact representation

ILLUSTRATING APRIORI PRINCIPLE

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke



Found to be Infrequent

Pruned supersets

Subset {ABC} 2^3

{ }, {A}, ----

supsets

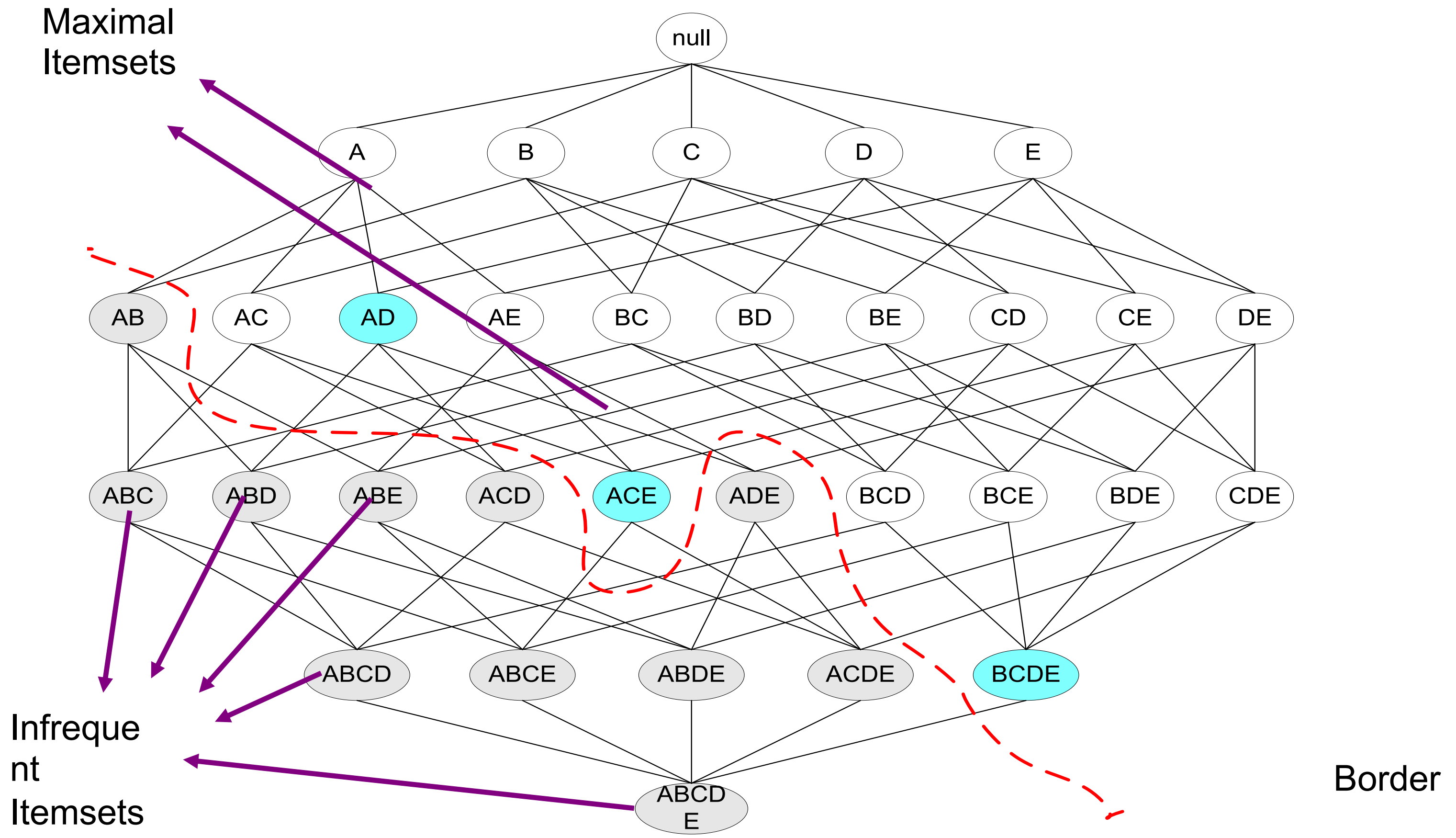
{ABC}

{ABCD}

{ABCE}

MAXIMAL FREQUENT ITEMSET

An itemset is maximal frequent if it is frequent and none of its immediate supersets is frequent



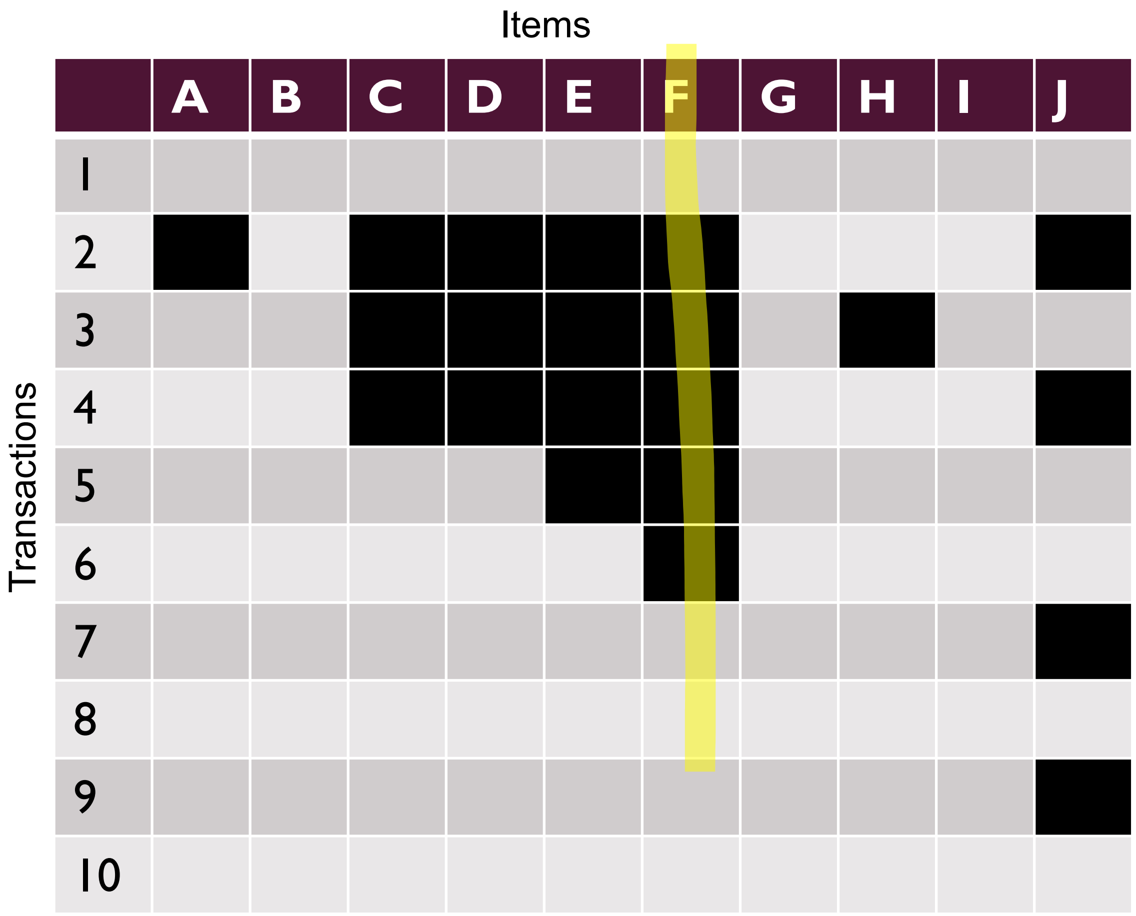
WHAT ARE THE MAXIMAL FREQUENT ITEMSETS IN THIS DATA?

TID	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1

Minimum support threshold = 5

(A1-A10)
(B1-B10)
(C1-C10)

AN ILLUSTRATIVE EXAMPLE



Support threshold (by count) : 5

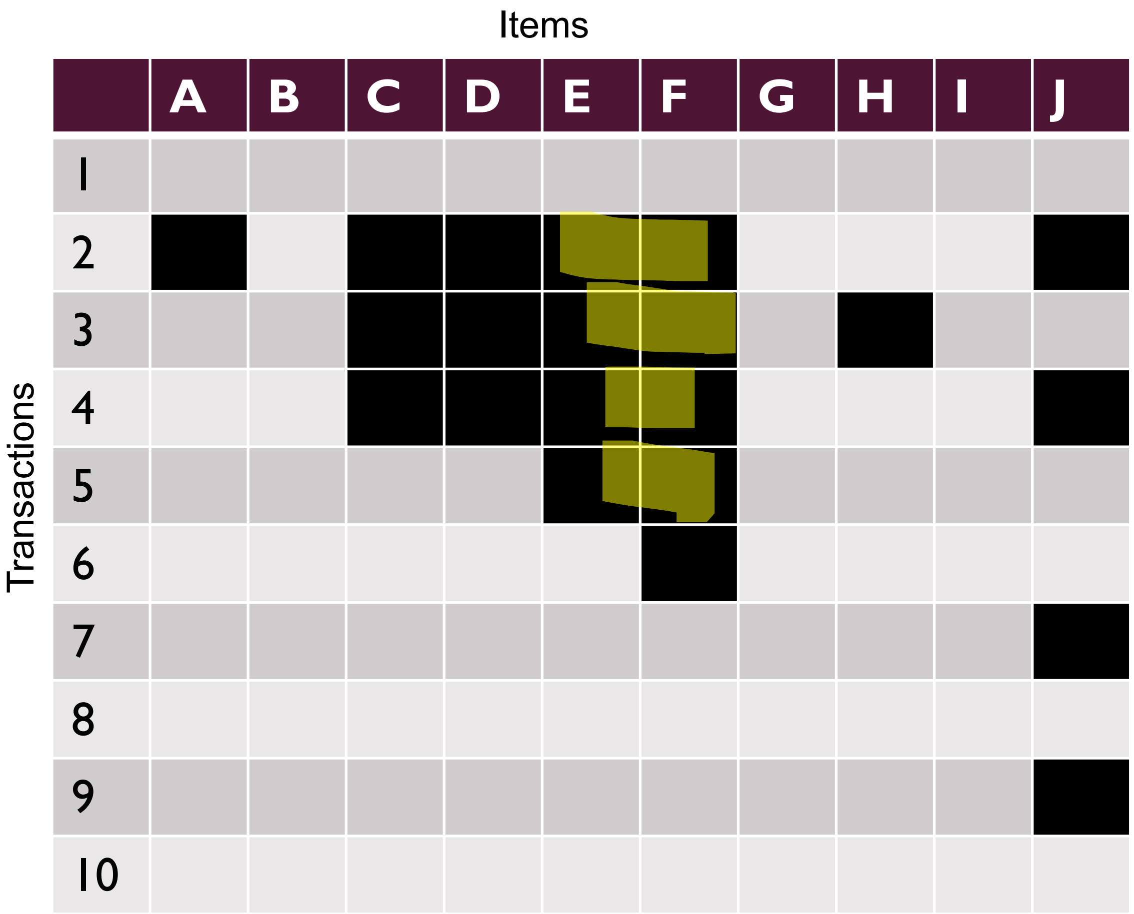
Frequent itemsets: ?

Maximal itemsets: ?

Freq : {F}

M : {F}

AN ILLUSTRATIVE EXAMPLE



Support threshold (by count) : 5

Frequent itemsets: {F}

Maximal itemsets: {F}

Support threshold (by count): 4

Frequent itemsets: ?

Maximal itemsets: ?

Frq: {E} {F} {J} {EF}

M: {J}, {EF}

{E} $\subseteq$ {EF}

i.e. {EF} superset of {E}

AN ILLUSTRATIVE EXAMPLE

Items

Transactions

	A	B	C	D	E	F	G	H	I	J
1										
2	■		■	■	■	■				■
3			■	■	■	■		■		
4			■	■	■	■				■
5					■	■				
6						■				
7										■
8										
9										■
10										

Support threshold (by count) : 5

Frequent itemsets: {F}

Maximal itemsets: {F}

Support threshold (by count): 4

Frequent itemsets: {E}, {F}, {E,F}, {J}

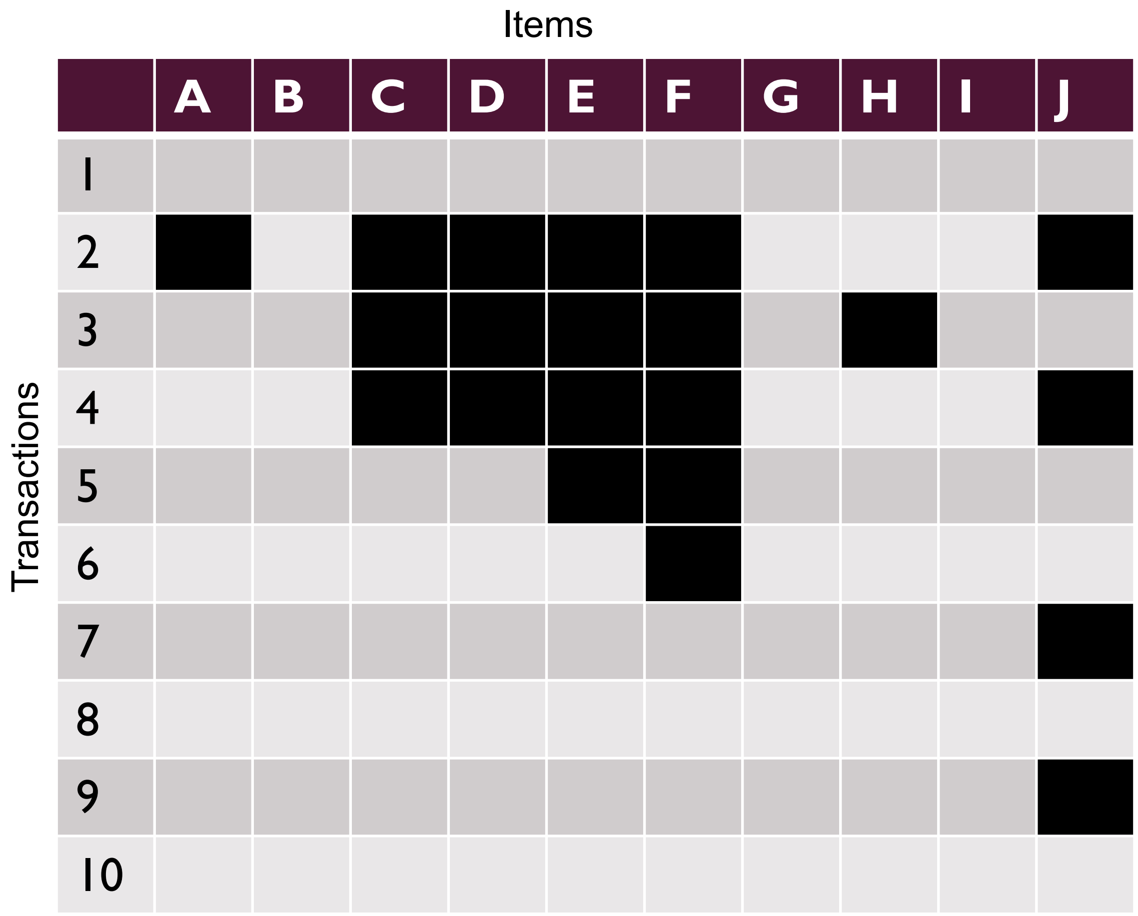
Maximal itemsets: {E,F}, {J}

Support threshold (by count): 3

Frequent itemsets: ?

Maximal itemsets: ?

AN ILLUSTRATIVE EXAMPLE



Support threshold (by count) : 5

Frequent itemsets: {F}

Maximal itemsets: {F}

Support threshold (by count): 4

Frequent itemsets: {E}, {F}, {E,F}, {J}

Maximal itemsets: {E,F}, {J}

Support threshold (by count): 3

Frequent itemsets:

All subsets of {C,D,E,F} + {J}

Maximal itemsets:

{C,D,E,F}, {J}

CLOSED ITEMSET

- An itemset X is closed if none of its immediate supersets has the same support as the itemset X.
- X is not closed if at least one of its immediate supersets has support count as X.

WEKA – ASSOCIATE RULE

CLOSED ITEMSET

- An itemset X is closed if none of its immediate supersets has the same support as the itemset X.
- X is not closed if at least one of its immediate supersets has support count as X.

$S(\{A\}) = 4$ ✓

$S(\{A\}) = \frac{5}{5} = \frac{4}{5}$

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,B,C,D}
4	{A,B,D}
5	{A,B,C,D}

NC

k=1

k=2

5

Itemset	Support
{A}	4
{B}	5
{C}	3
{D}	4
{A,B}	4
{A,C}	2
{A,D}	3
{B,C}	3
{B,D}	4
{C,D}	3

{B} closed {A,B}

Itemset	Support
{A,B,C}	2
{A,B,D}	3
{A,C,D}	2
{B,C,D}	2
{A,B,C,D}	2

k=3

k=4

MAXIMAL VS CLOSED ITEMSETS

F itemset

σ/s

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

T = 3

TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE

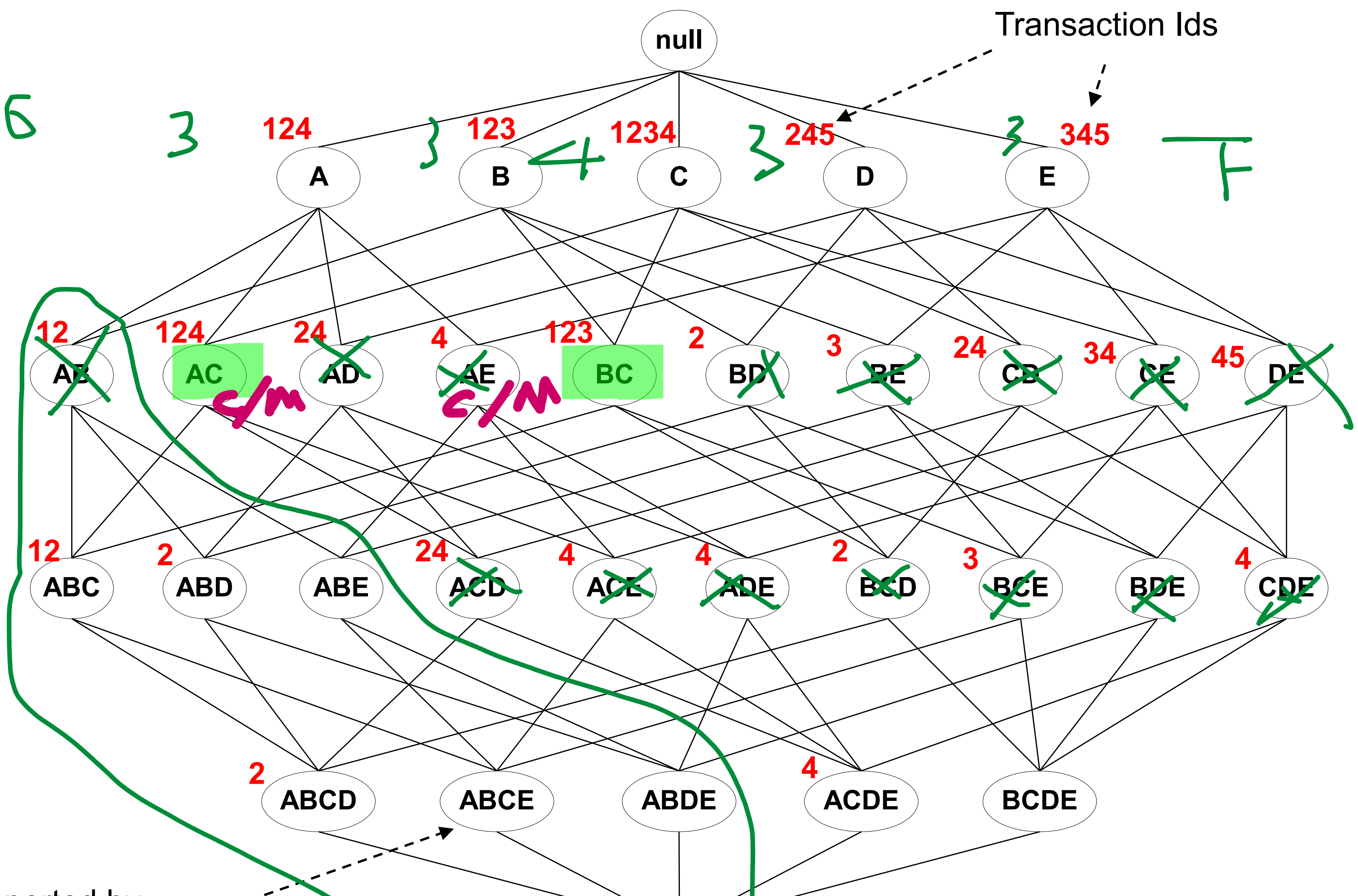
6

F:

{A} -- {E}

{AC} {BC}

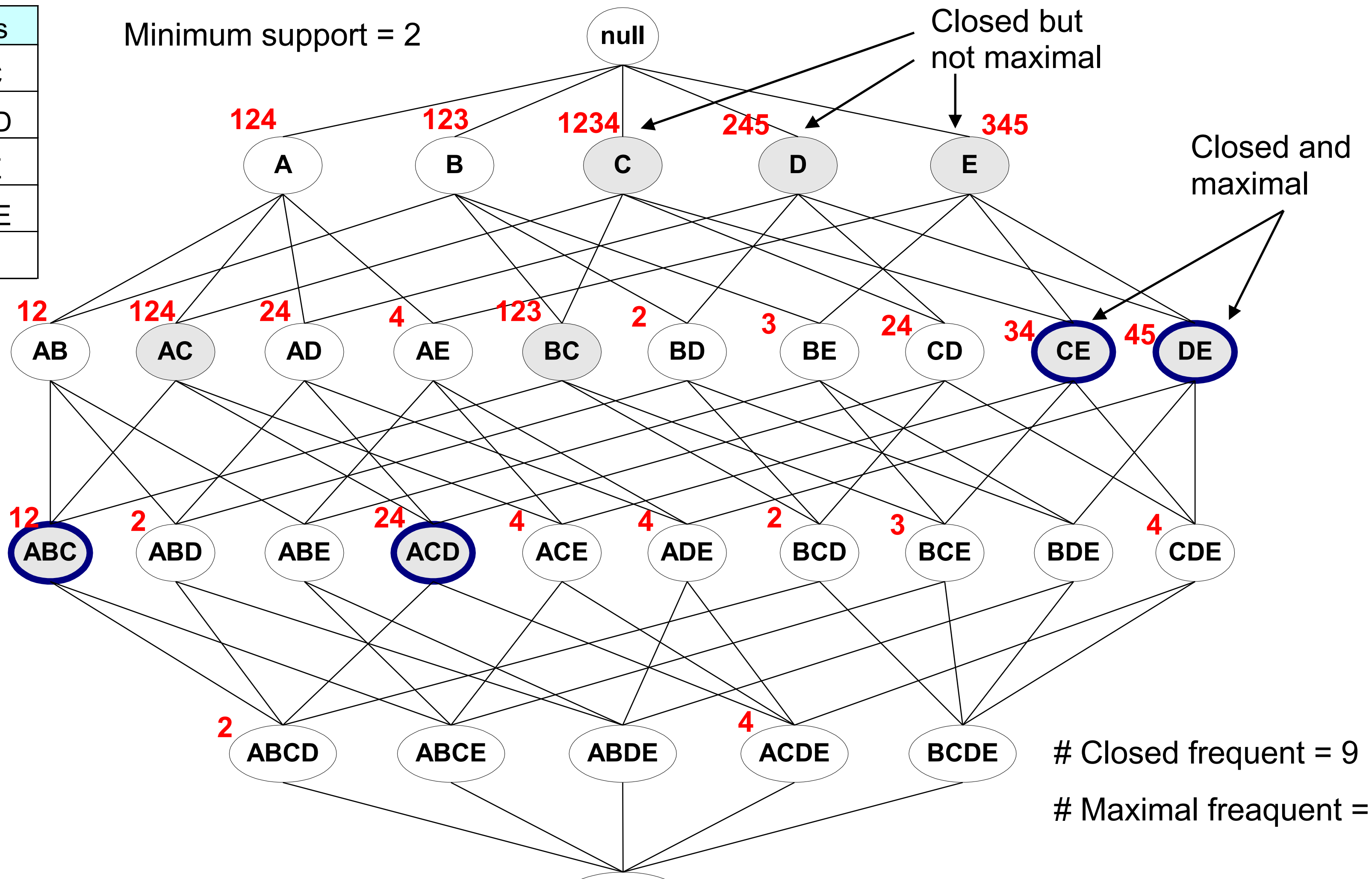
M: {AC} {BC}



MAXIMAL FREQUENT VS CLOSED FREQUENT ITEMSETS

TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE

Minimum support = 2



Closed frequent = 9

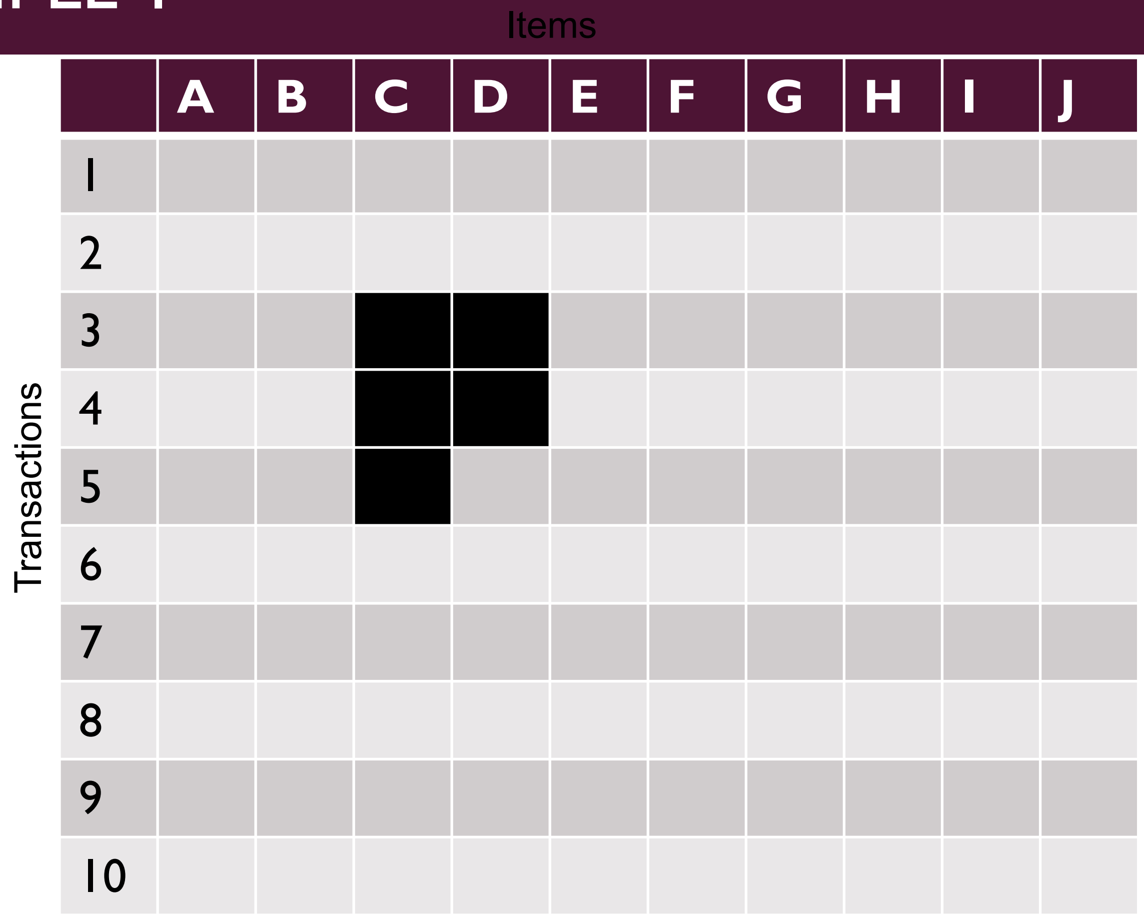
Maximal freaquent = 4

WHAT ARE THE CLOSED ITEMSETS IN THIS DATA?

TID	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1

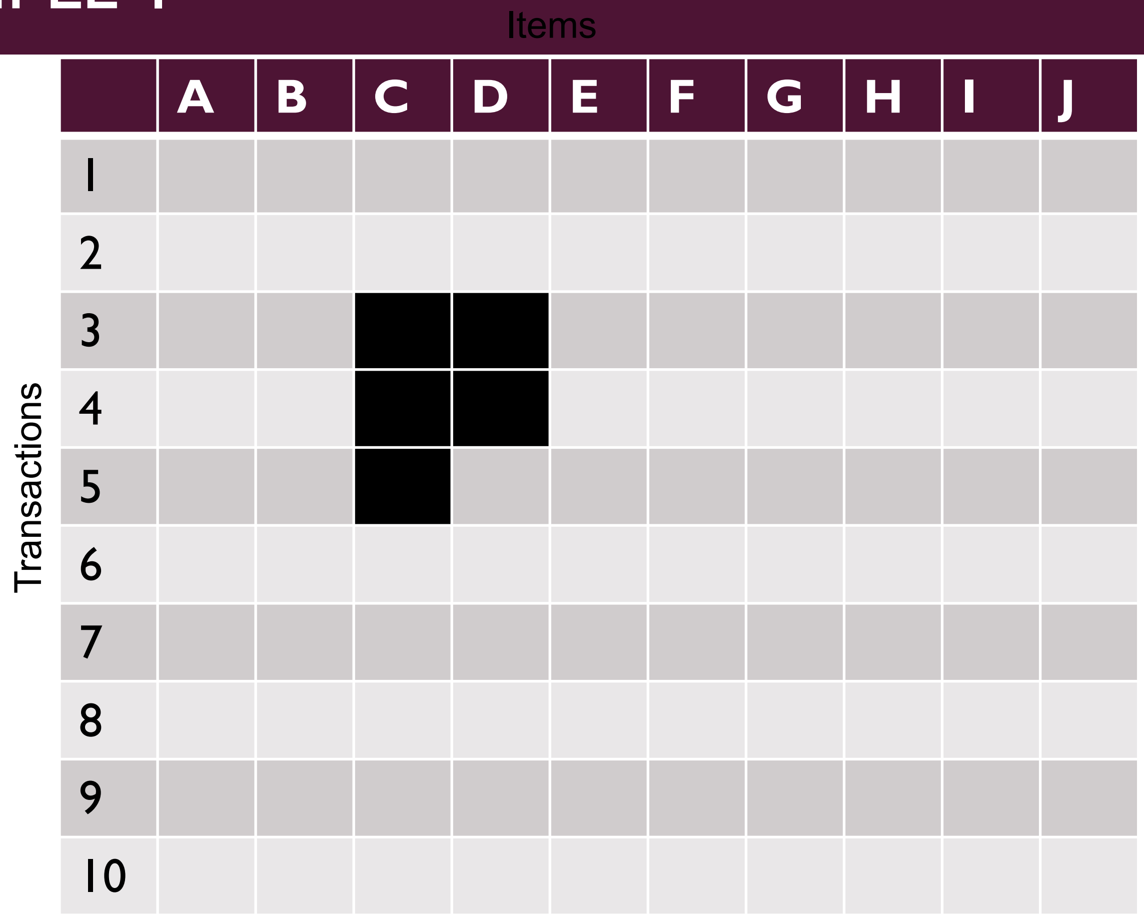
(A1-A10)
(B1-B10)
(C1-C10)

EXAMPLE 1



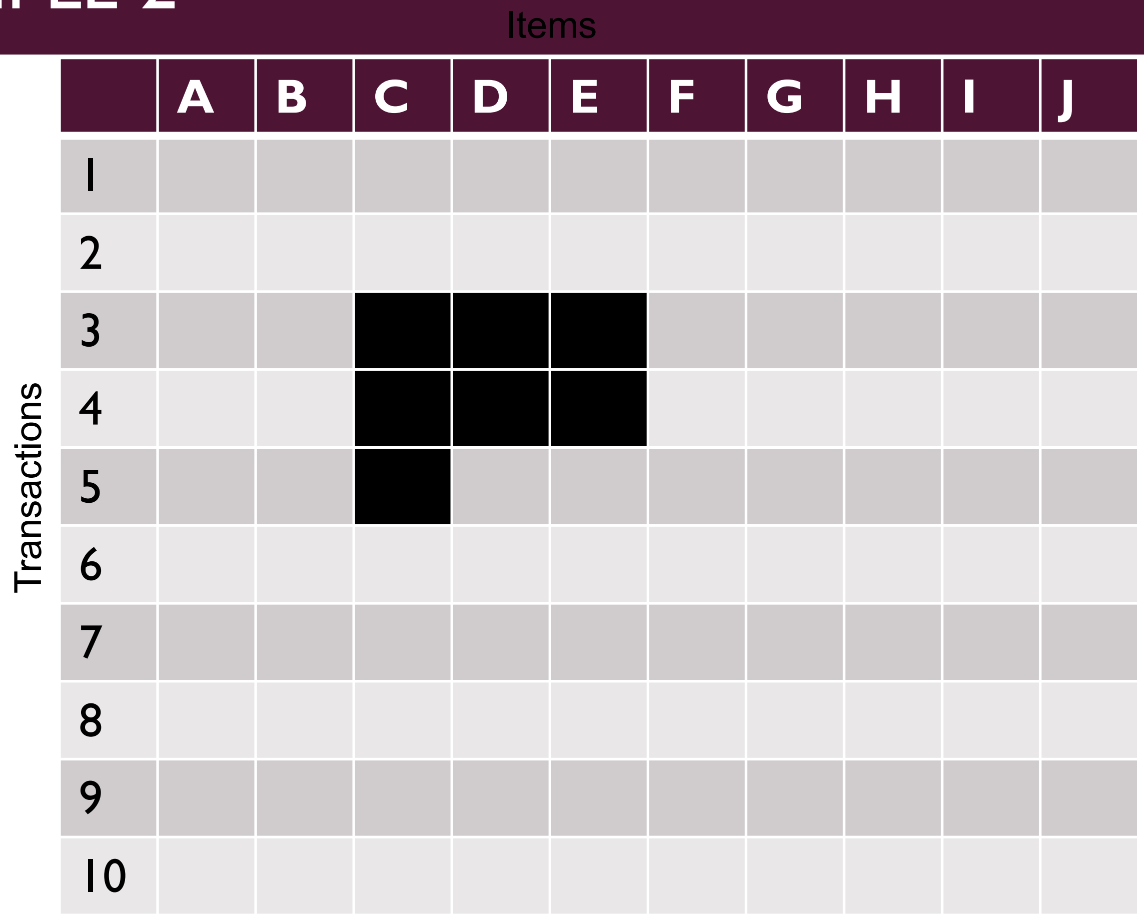
Itemsets	Support (counts)	Closed itemsets
{C}	3	✓
{D}	2	✗
{C,D}	2	✓

EXAMPLE 1



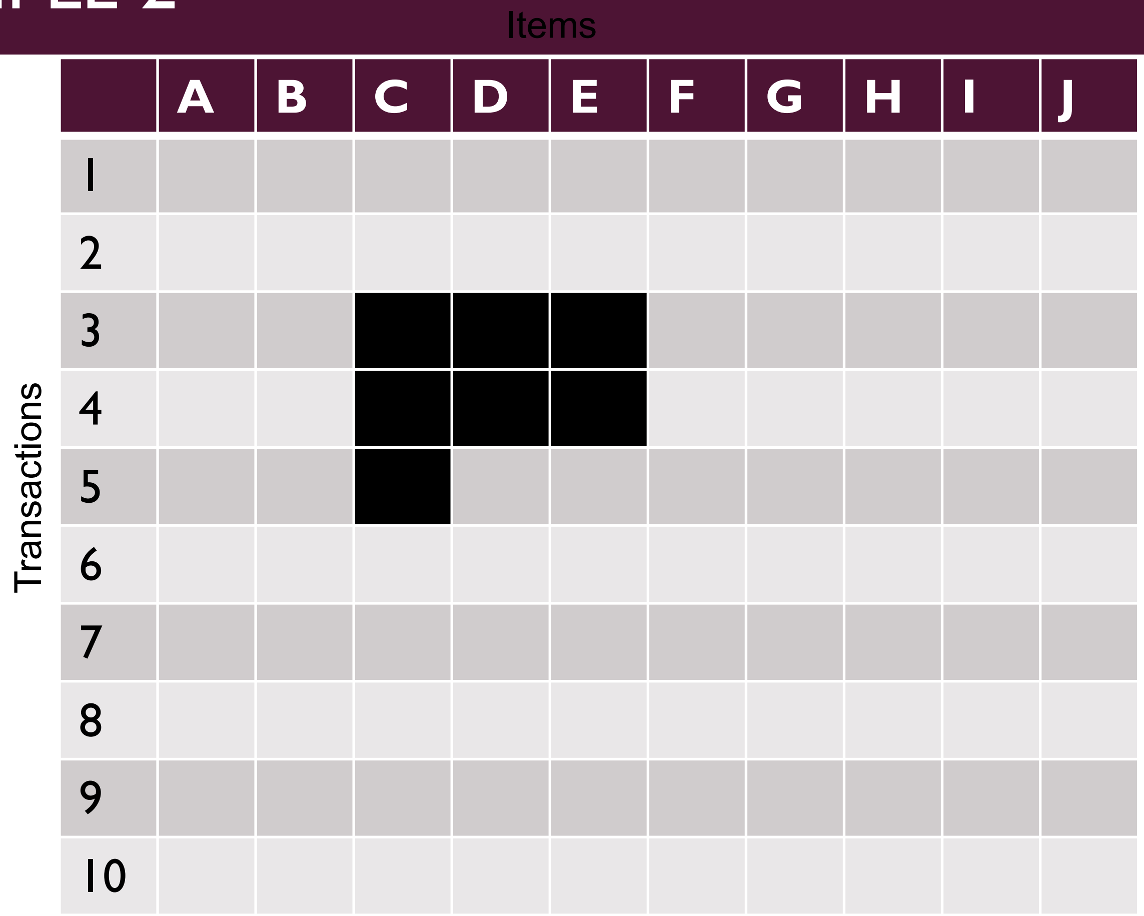
Itemsets	Support (counts)	Closed itemsets
{C}	**3**	✓
{D}	2	
{C,D}	**2**	✓

EXAMPLE 2



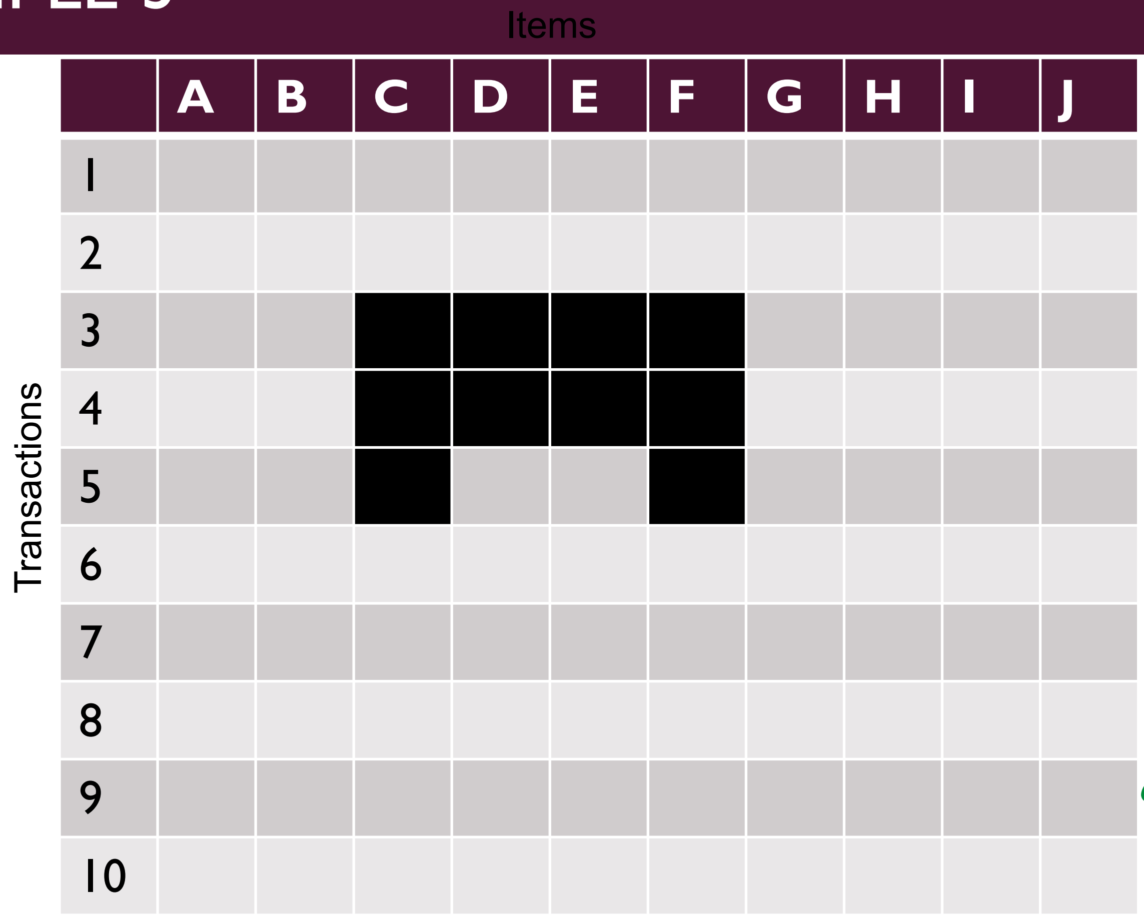
Itemsets	Support (counts)	Closed itemsets
{C}	3	
{D}	2	
{E}	2	✗
{C,D}	2	✗
✓ {C,E}	2	
✓ {D,E}	2	
✓ {C,D,E}	2	

EXAMPLE 2



Itemsets	Support (counts)	Closed itemsets
{C}	**3**	✓
{D}	2	
{E}	2	
{C,D}	2	
{C,E}	2	
{D,E}	2	
{C,D,E}	**2**	✓

EXAMPLE 3



Closed itemsets: {C,D,E,F}, {C,F}

What are closed?

MF ⇒ closed

⇍

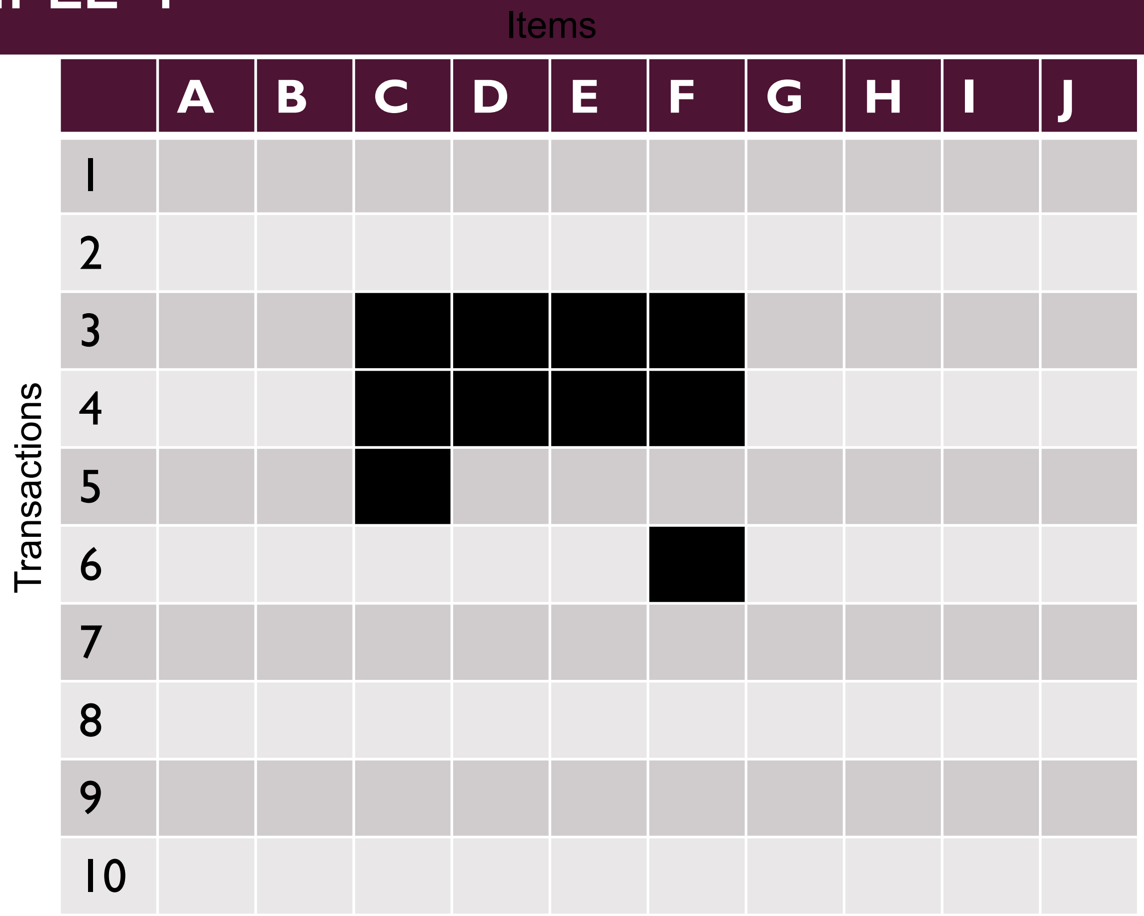
① MF ⇒ closed ⇒ σ(sup) < σ(IF)

$\boxed{F}$ + all sup IF

⇕

σ(all sup) < σ(IT)

EXAMPLE 4



Closed itemsets: {C,D,E,F}, {C}, {F}

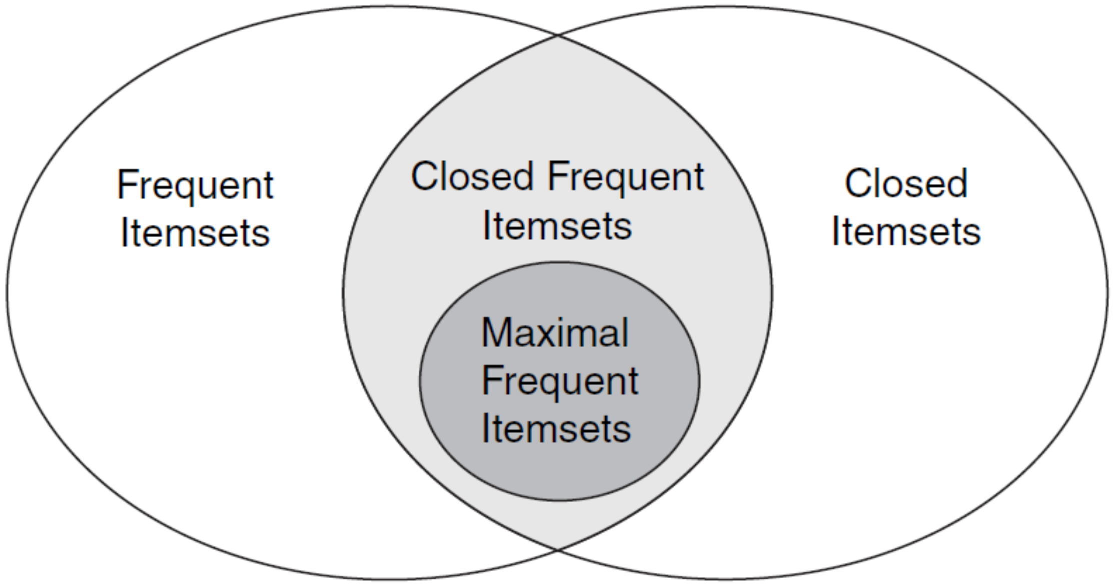
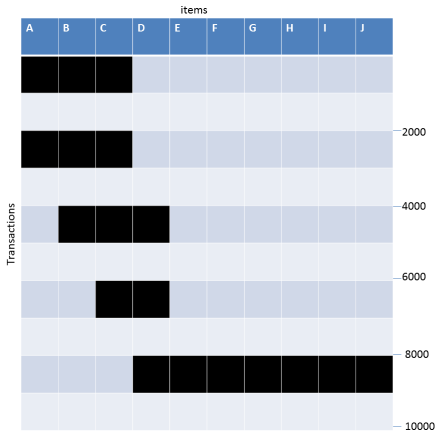


Figure 5.18. Relationships among frequent, closed, closed frequent, and maximal frequent itemsets.

EXAMPLE QUESTION



Data Set: C

F (min s = 3)

s ≥ 3

6 5 C

F MF closed

{C}, {D}, {B} {BC}

MF: {D}, {BC}

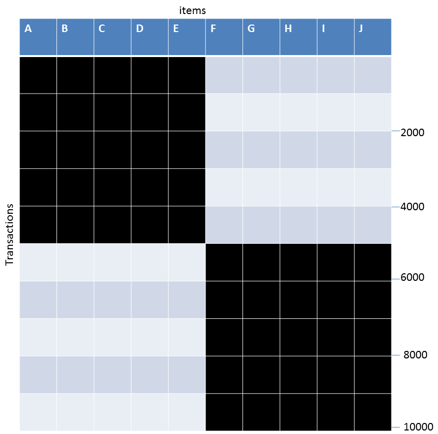
Closed: {D}, {BC}, {C}, {CD}, {ABC}, {BCD}, {DEFGHIJ}

a. What is the number of frequent itemsets for each dataset? Which dataset will produce the most number of frequent itemsets?
b. Which dataset will produce the longest frequent itemset?
c. Which dataset will produce frequent itemsets with highest maximum support?
d. Which dataset will produce frequent itemsets containing items with widely varying support levels (i.e., itemsets containing items with mixed support, ranging from 20% to more than 70%)?
e. What is the number of maximal frequent itemsets for each dataset? Which dataset will produce the most number of maximal frequent itemsets?
f. What is the number of closed frequent itemsets for each dataset? Which dataset will produce the most number of closed frequent itemsets?

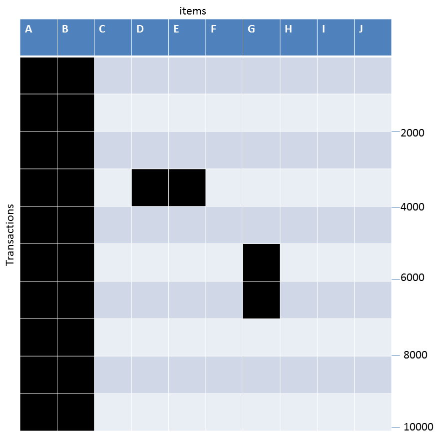
EXAMPLE QUESTION

6 5 C

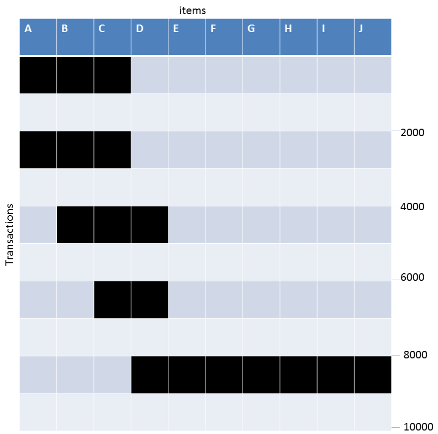
F MF closed



DataSet: A



Data Set: B



Data Set: C

- Given the following transaction data sets (dark cells indicate presence of an item in a transaction) and a support threshold of 20%, answer the following questions
  a. What is the number of frequent itemsets for each dataset? Which dataset will produce the most number of frequent itemsets?
  b. Which dataset will produce the longest frequent itemset?
  c. Which dataset will produce frequent itemsets with highest maximum support?
  d. Which dataset will produce frequent itemsets containing items with widely varying support levels (i.e., itemsets containing items with mixed support, ranging from 20% to more than 70%)?
  e. What is the number of maximal frequent itemsets for each dataset? Which dataset will produce the most number of maximal frequent itemsets?
  f. What is the number of closed frequent itemsets for each dataset? Which dataset will produce the most number of closed frequent itemsets?