Chapter 7: Frequent Itemsets and Association Rules

Information Retrieval & Data Mining
Universität des Saarlandes, Saarbrücken
Winter Semester 2011/12

Chapter VII: Frequent Itemsets and Association Rules*

1. Definitions

- Transaction data, frequent itemsets, closed and maximal itemsets, association rules

2. The Apriori Algorithm

- Monotonicity and candidate pruning, mining closed and maximal itemsets

3. Generating Association Rules

4. Other measures for Association Rules

- Properties of measures

*Zaki & Meira, Chapters 10 and 11; Tan, Steinbach & Kumar, Chapter 6

Chapter VII.1: Definitions

1. The transaction data model
   1.1. Data as subsets
   1.2. Data as binary matrix
2. Itemsets, support, and frequency
3. Closed and maximal itemsets
4. Association rules and confidence
5. Related data mining tasks

The transaction data model

- Data mining considers larger variety of data types than typical IR
- Methods usually work on any data that can be expressed in certain type
  - Graphs, points in metric space, vectors, ...
- The data type used in itemset mining is the **transaction data**
  - Data contains transactions over some set of items

The market basket data

Items are: bread, milk, diapers, beer, and eggs
Transactions are: 1:{bread, milk}, 2:{bread, diapers, beer, eggs}, 3:{milk, diapers, beer}, 4:{bread, milk, diapers, beer}, and 5:{bread, milk, diapers}

The market basket data

Items are: bread, milk, diapers, beer, and eggs
Transactions are: 1:{bread, milk}, 2:{bread, diapers, beer, eggs}, 3:{milk, diapers, beer}, 4:{bread, milk, diapers, beer}, and 5:{bread, milk, diapers}

TID	Bread	Milk	Diapers	Beer	Eggs
1	✔	✔			
2	✔		✔	✔	✔
3		✔	✔	✔	
4	✔	✔	✔	✔	
5	✔	✔	✔		

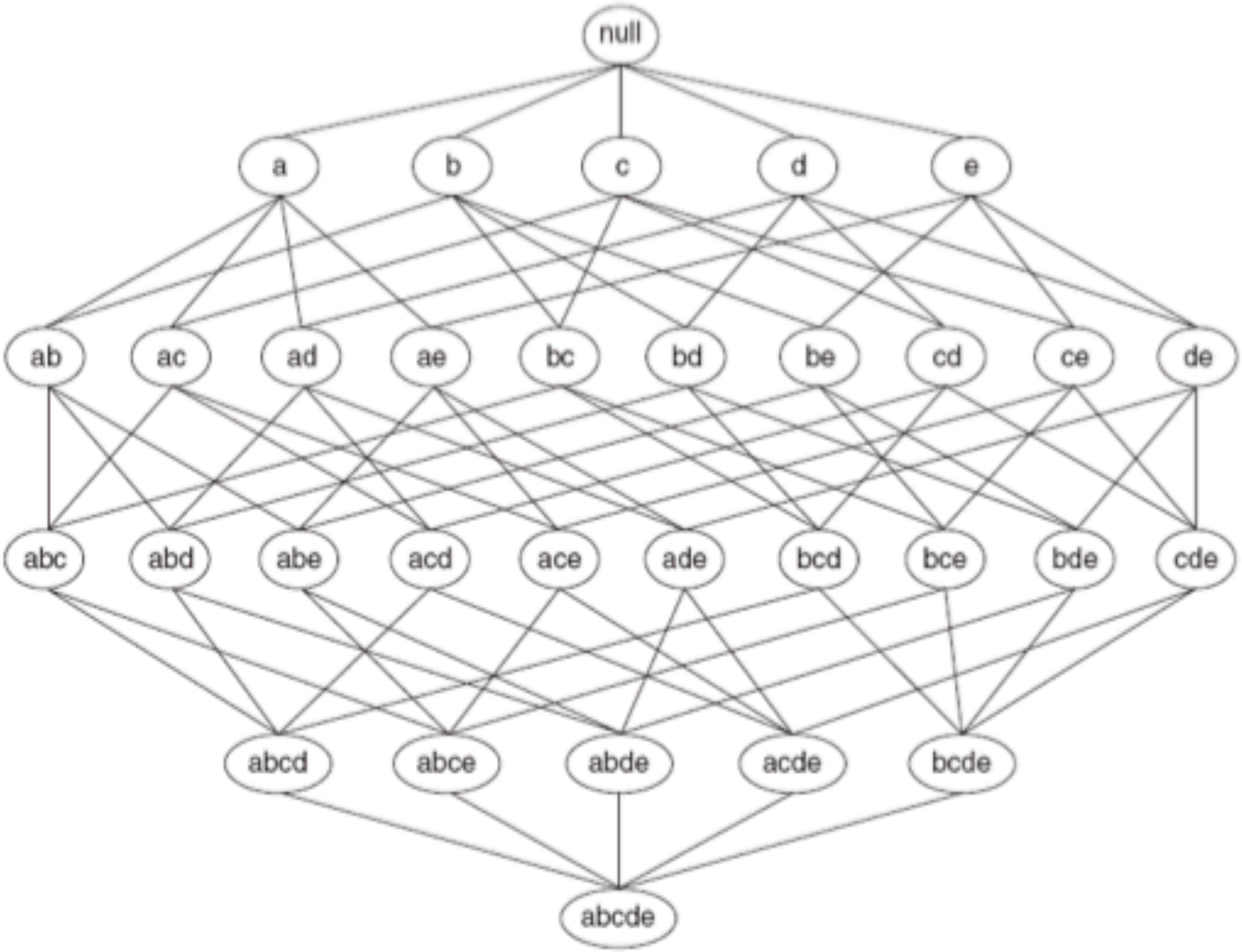
The market basket data

Items are: bread, milk, diapers, beer, and eggs
Transactions are: 1:{bread, milk}, 2:{bread, diapers, beer, eggs}, 3:{milk, diapers, beer}, 4:{bread, milk, diapers, beer}, and 5:{bread, milk, diapers}

Transaction IDs

TID	Bread	Milk	Diapers	Beer	Eggs
1	✔	✔			
2	✔		✔	✔	✔
3		✔	✔	✔	
4	✔	✔	✔	✔	
5	✔	✔	✔		

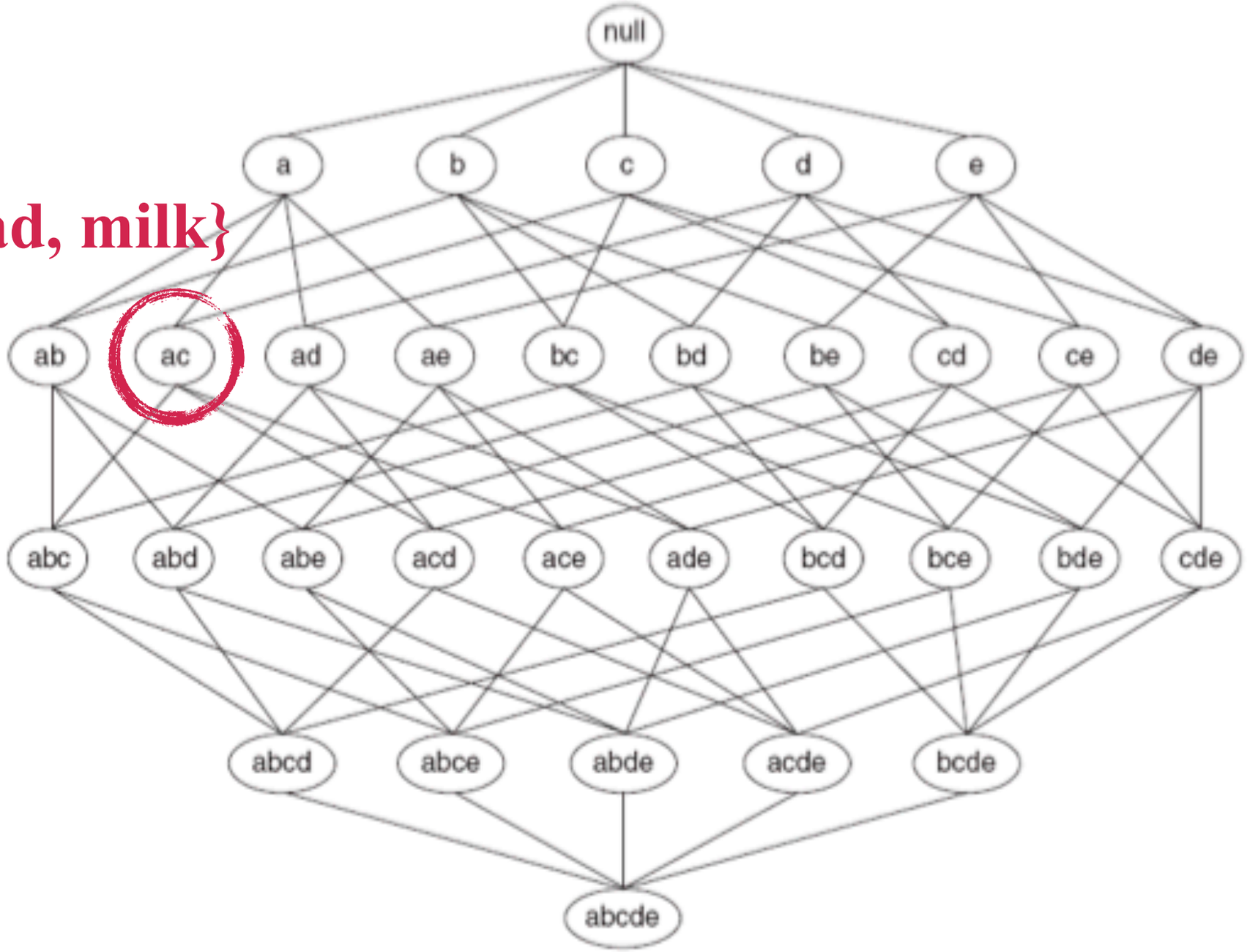
Transaction data as subsets



a: bread
b: beer
c: milk
d: diapers
e: eggs

2^n subsets of n items. Layer k has $\binom{n}{k}$ subsets.

Transaction data as subsets

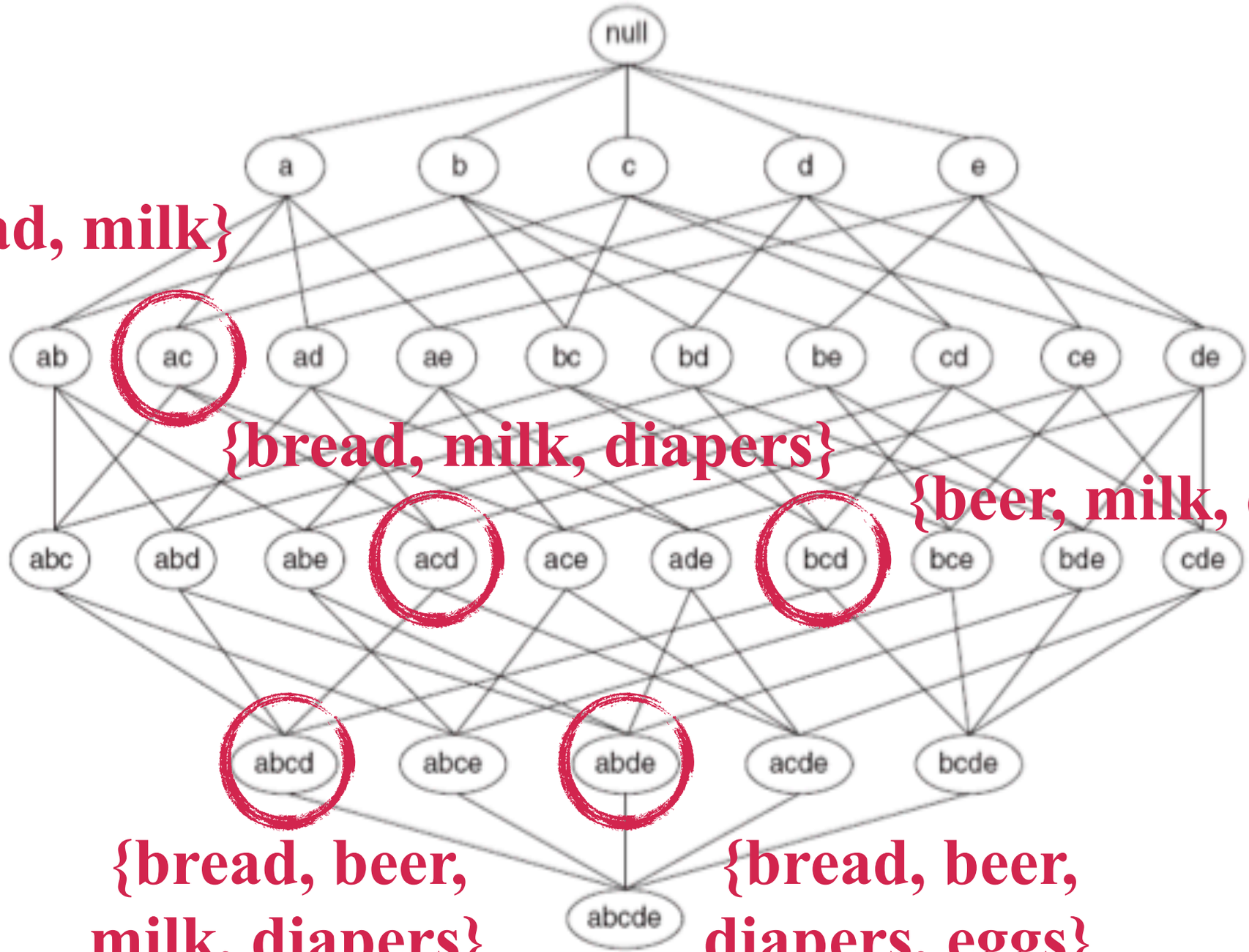


{bread, milk}

a: bread
b: beer
c: milk
d: diapers
e: eggs

2^n subsets of n items. Layer k has $\binom{n}{k}$ subsets.

Transaction data as subsets



{bread, milk}

{bread, milk, diapers}

{beer, milk, diapers}

{bread, beer, milk, diapers}

{bread, beer, diapers, eggs}

a: bread
b: beer
c: milk
d: diapers
e: eggs

2^n subsets of n items. Layer k has $\binom{n}{k}$ subsets.

Transaction data as binary matrix

TID	Bread	Milk	Diapers	Beer	Eggs
1	✔	✔			
2	✔		✔	✔	✔
3		✔	✔	✔	
4	✔	✔	✔	✔	
5	✔	✔	✔		

Transaction data as binary matrix

TID	Bread	Milk	Diapers	Beer	Eggs
1	1	1	0	0	0
2	1	0	1	1	1
3	0	1	1	1	0
4	1	1	1	1	0
5	1	1	1	0	0

Transaction data as binary matrix

TID	Bread	Milk	Diapers	Beer	Eggs
1	1	1	0	0	0
2	1	0	1	1	1
3	0	1	1	1	0
4	1	1	1	1	0
5	1	1	1	0	0

Any data that can be expressed as a binary matrix can be used.

Itemsets, support, and frequency

- An **itemset** is a set of items
  - A transaction t is an itemset with associated transaction ID, $t = (tid, I)$, where I is the set of items of the transaction
- A transaction $t = (tid, I)$ contains itemset X if $X \subseteq I$
- The **support** of itemset X in database D is the number of transactions in D that contain it:

$$supp(X, D) = |\{t \in D : t \text{ contains } X\}|$$

- The **frequency** of itemset X in database D is its support relative to the database size, $supp(X, D) / |D|$
- Itemset is **frequent** if its frequency is above user-defined threshold **minfreq**

Frequent itemset example

TID	Bread	Milk	Diapers	Beer	Eggs
1	1	1	0	0	0
2	1	0	1	1	1
3	0	1	1	1	0
4	1	1	1	1	0
5	1	1	1	0	0

Itemset {Bread, Milk} has support 3 and frequency 3/5
Itemset {Bread, Milk, Eggs} has support and frequency 0
For **minfreq** = 1/2, frequent itemsets are:
{Bread}, {Milk}, {Diapers}, {Beer}, {Bread, Milk}, {Bread, Diapers}, {Milk, Diapers}, and {Diapers, Beer}

Closed and maximal itemsets

- Let F be the set of all frequent itemsets (w.r.t. some **minfreq**) in data D
- Frequent itemset $X \in F$ is **maximal** if it does not have any frequent supersets
  - That is, for all $Y \supset X$, $Y \notin F$
- Frequent itemset $X \in F$ is **closed** if it has no superset with the same frequency
  - That is, for all $Y \supset X$, $supp(Y, D) < supp(X, D)$
    - It can't be that $supp(Y, D) > supp(X, D)$. Why?

Example of maximal frequent itemsets

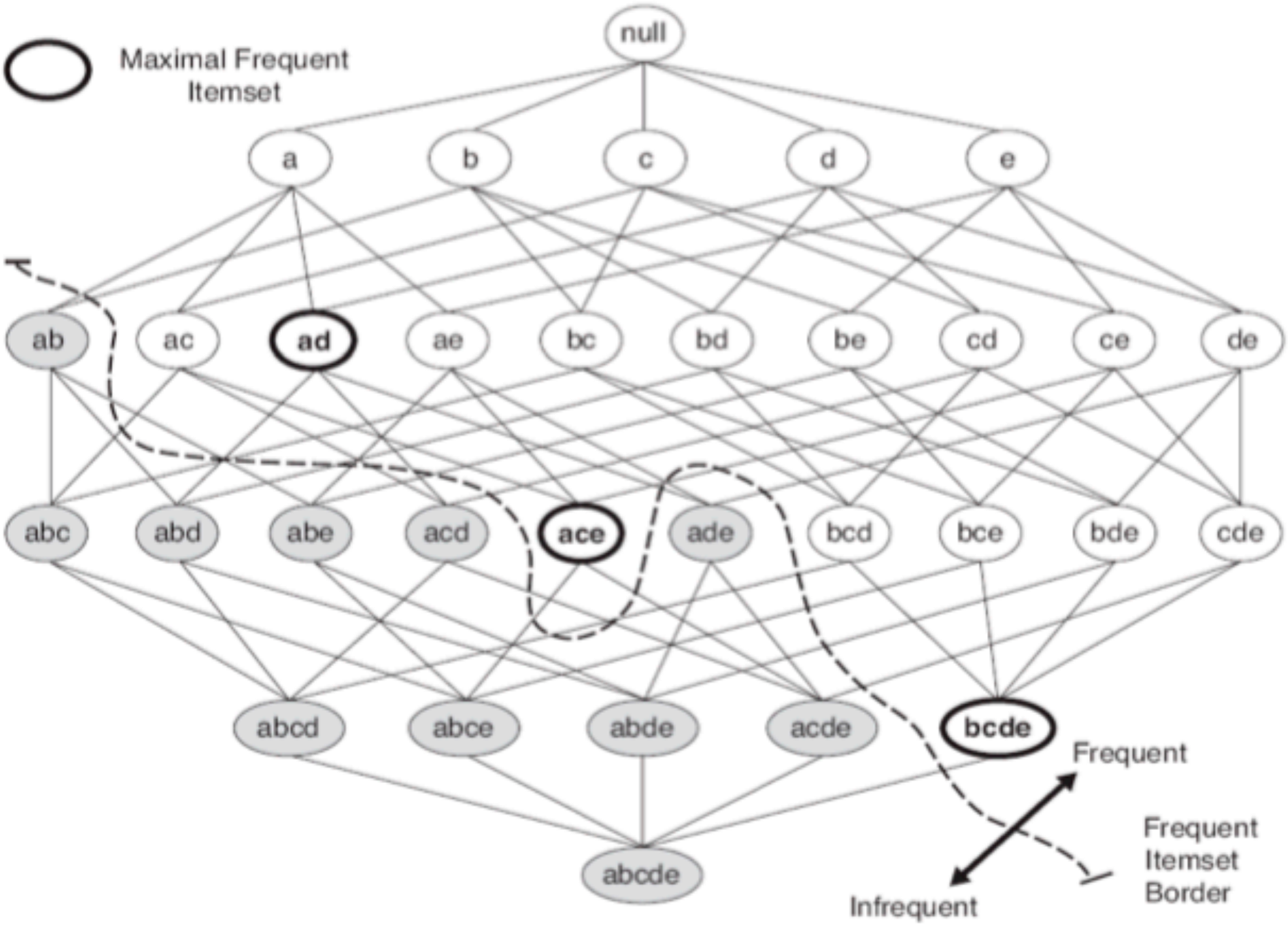


Figure 6.16. Maximal frequent itemset.

Example of maximal frequent itemsets

Not maximal because of {a, c, e}

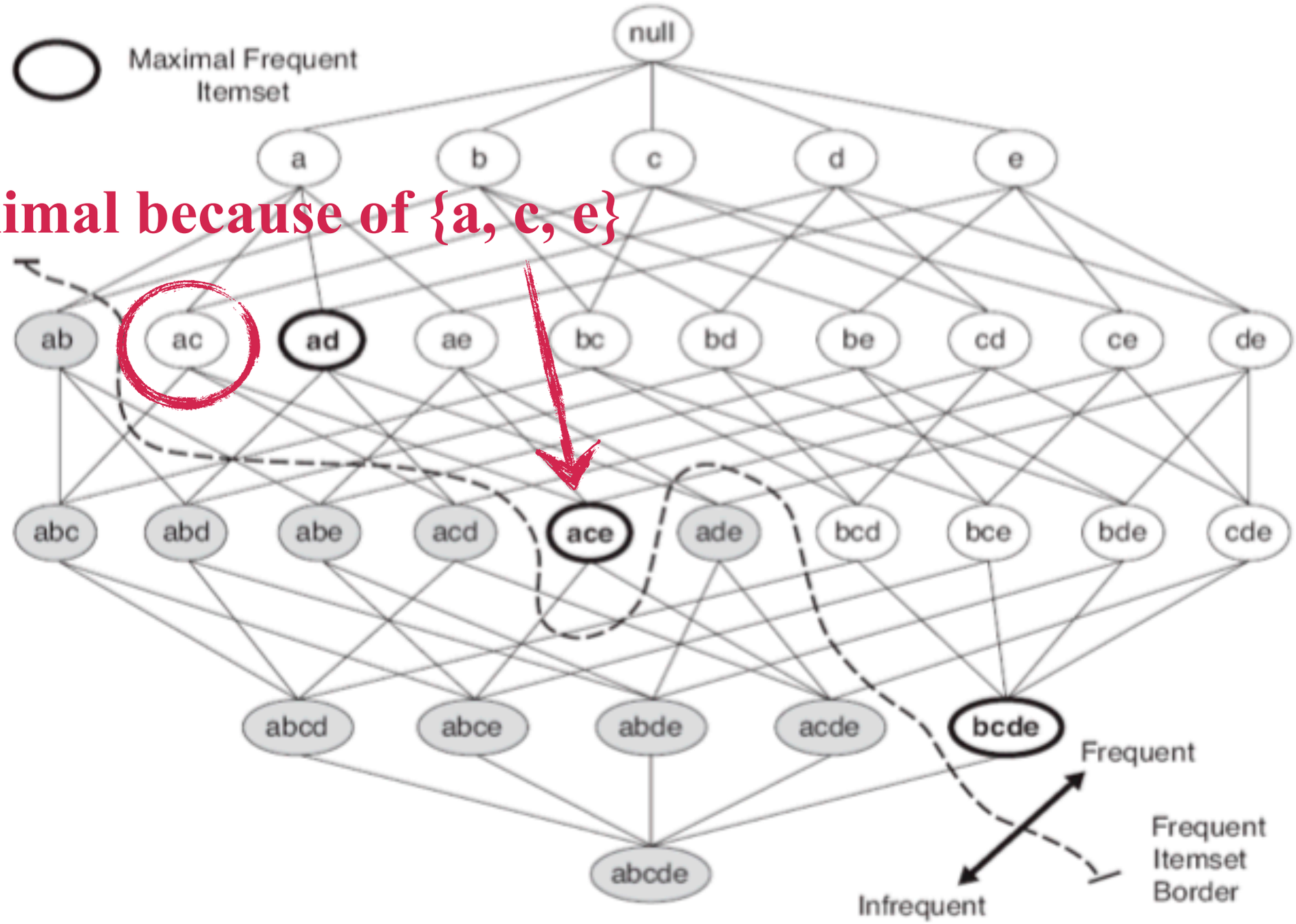


Figure 6.16. Maximal frequent itemset.

Example of maximal frequent itemsets

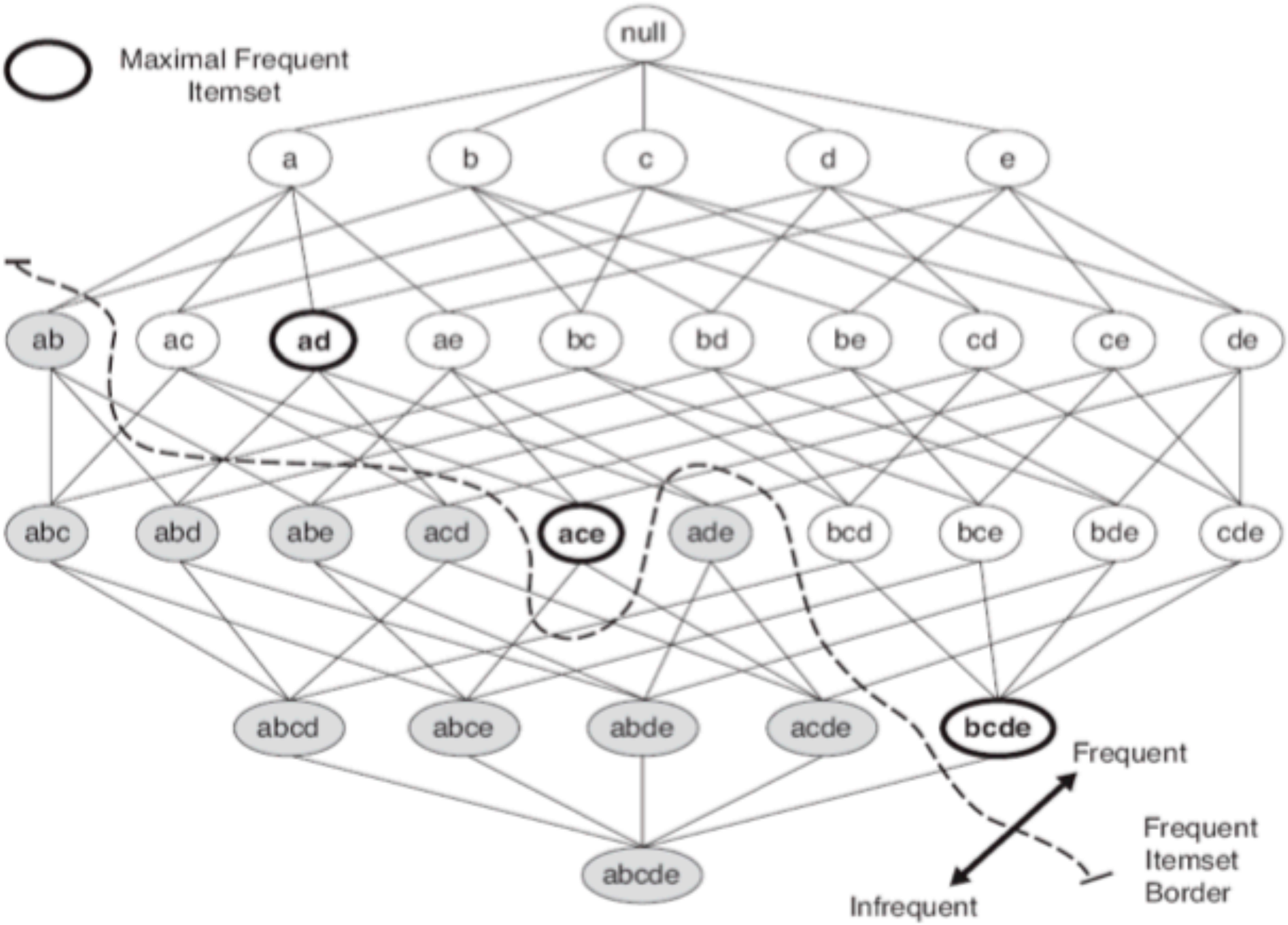


Figure 6.16. Maximal frequent itemset.

Example of closed frequent itemsets

TID	Items
1	abc
2	abcd
3	bce
4	acde
5	de

minsup = 40%

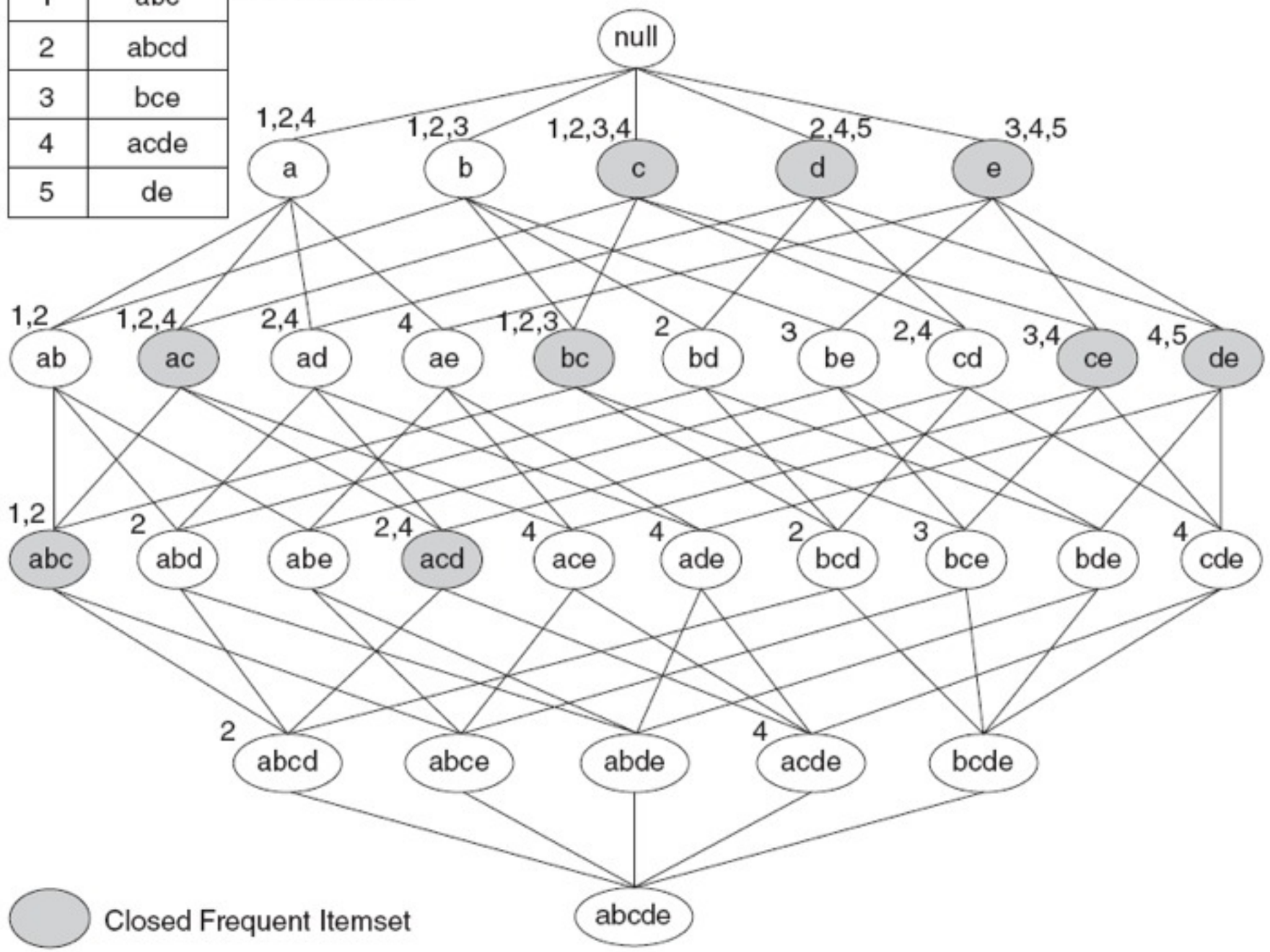
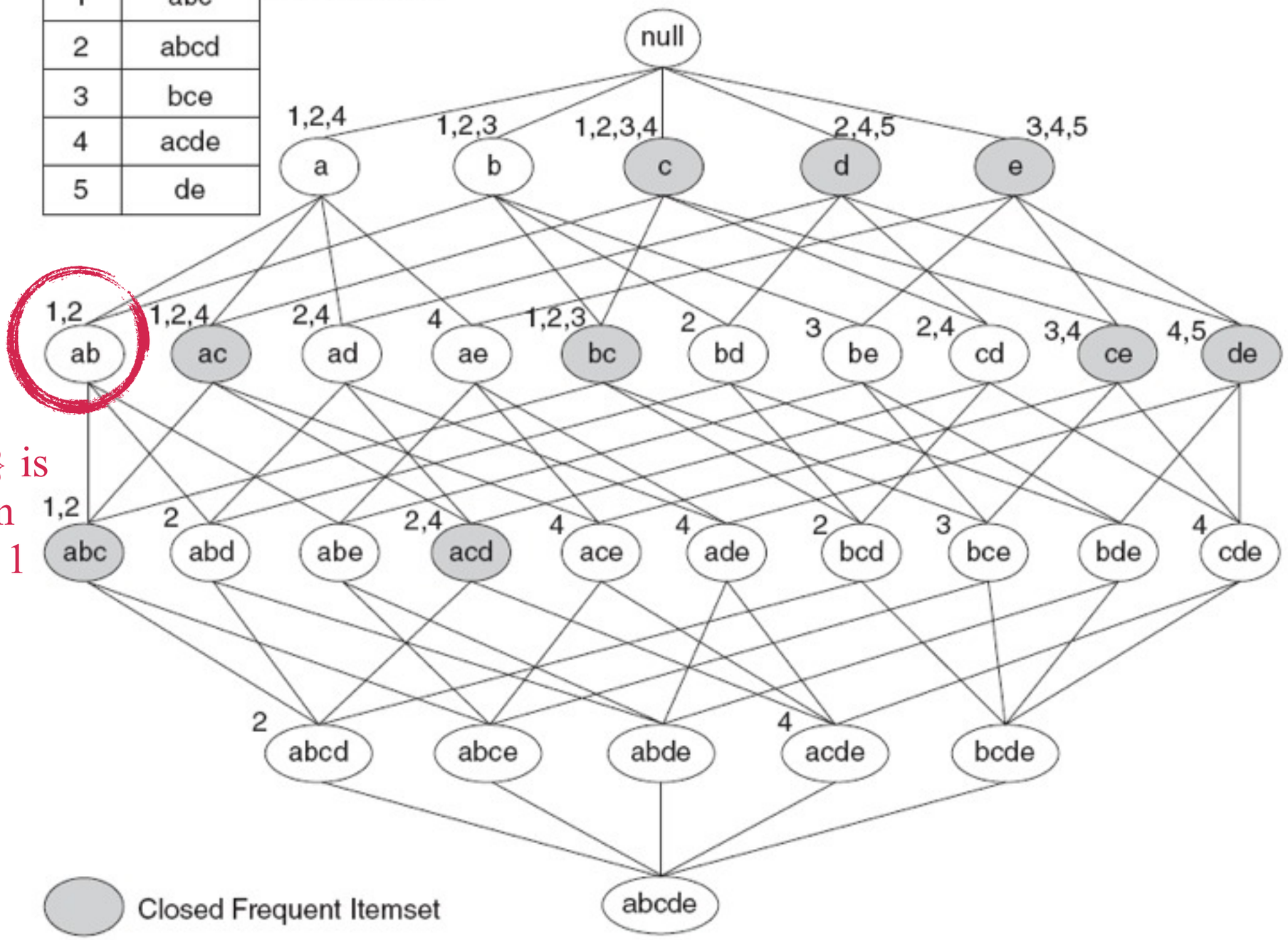


Figure 6.17. An example of the closed frequent itemsets (with minimum support count equal to 40%).

Example of closed frequent itemsets

TID	Items
1	abc
2	abcd
3	bce
4	acde
5	de

minsup = 40%



Itemset {a, b} is contained in transactions 1 and 2

Figure 6.17. An example of the closed frequent itemsets (with minimum support count equal to 40%).

Example of closed frequent itemsets

TID	Items
1	abc
2	abcd
3	bce
4	acde
5	de

minsup = 40%

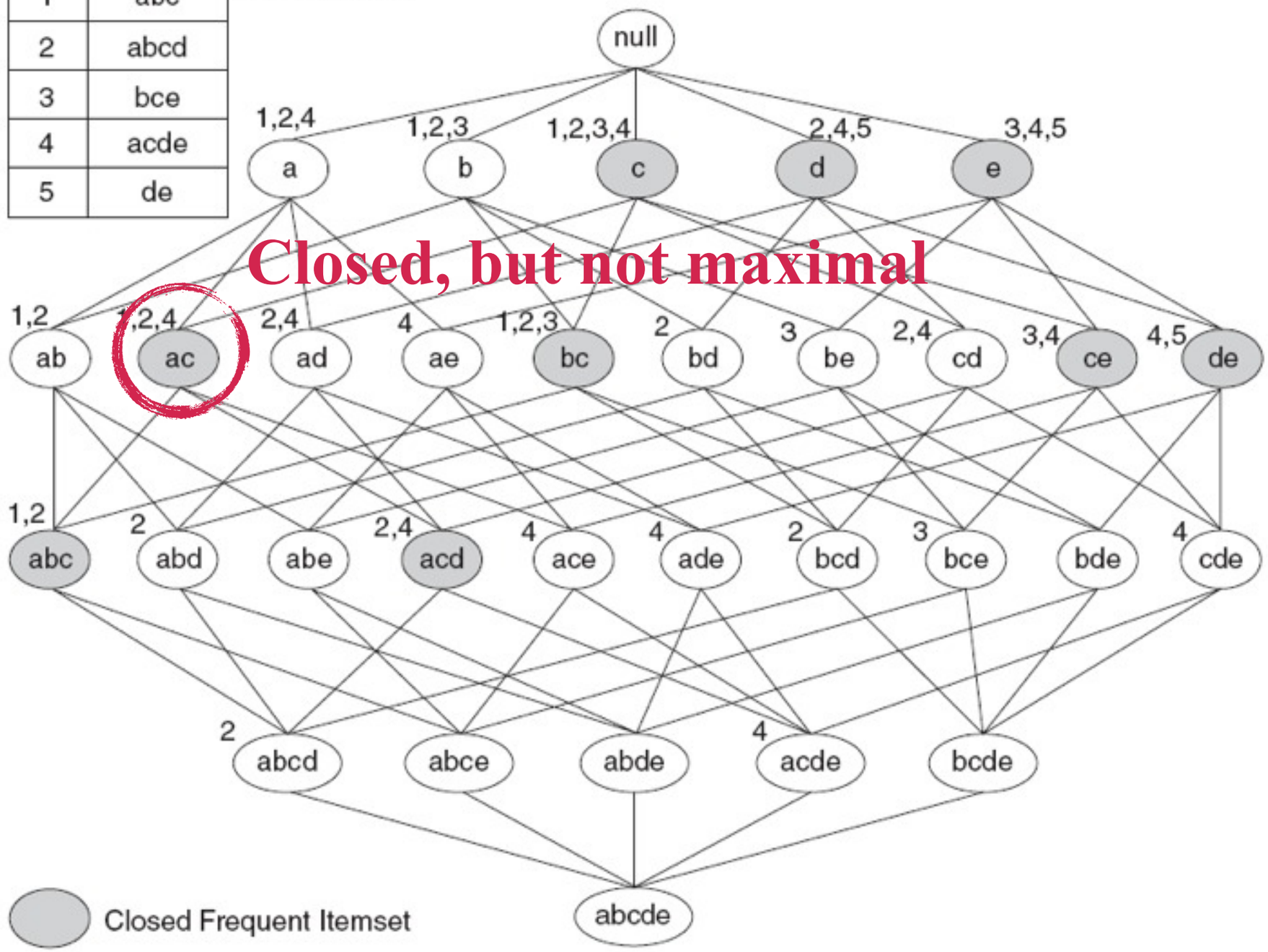


Figure 6.17. An example of the closed frequent itemsets (with minimum support count equal to 40%).

Example of closed frequent itemsets

TID	Items
1	abc
2	abcd
3	bce
4	acde
5	de

minsup = 40%

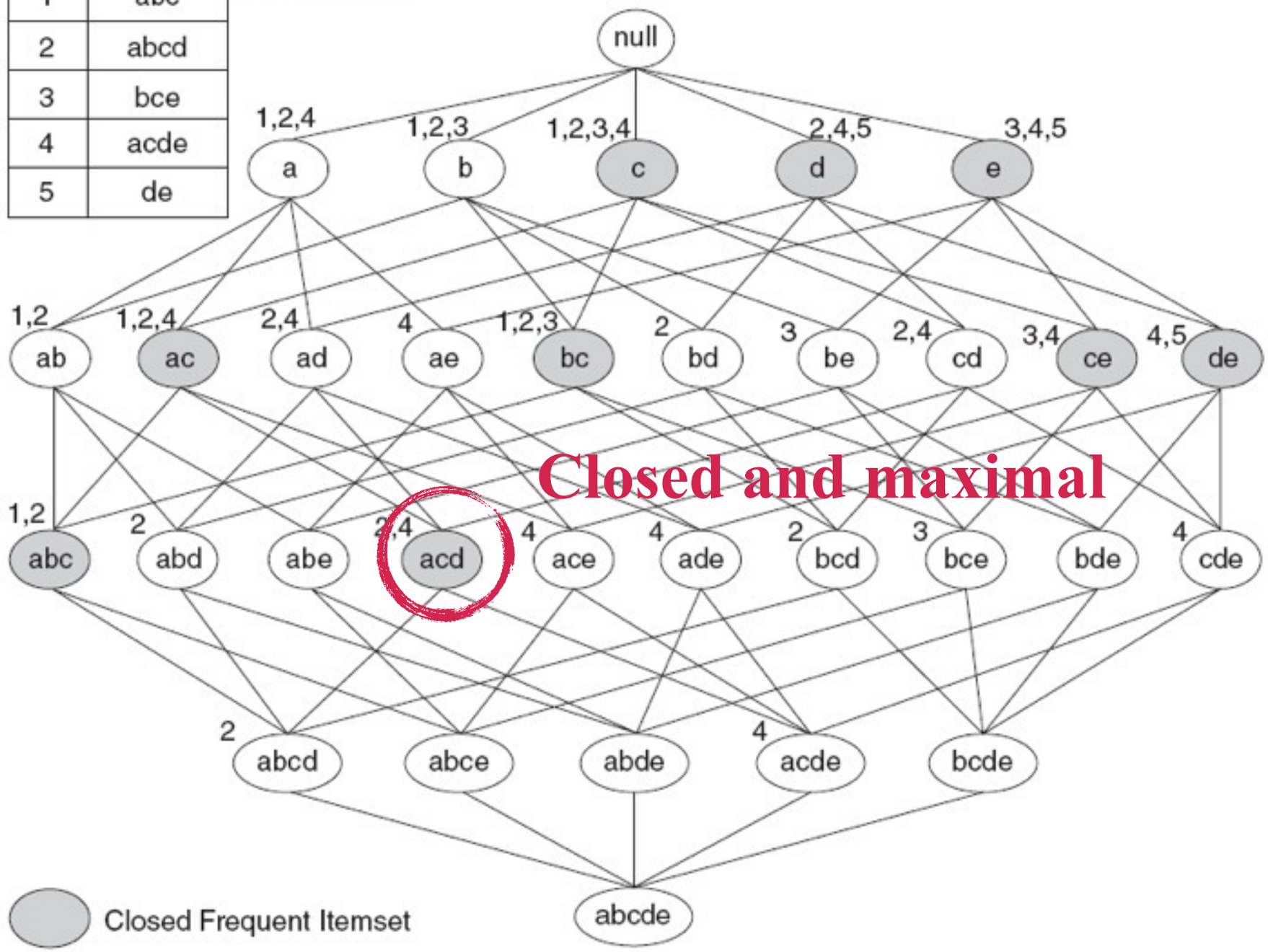


Figure 6.17. An example of the closed frequent itemsets (with minimum support count equal to 40%).

Example of closed frequent itemsets

TID	Items
1	abc
2	abcd
3	bce
4	acde
5	de

minsup = 40%

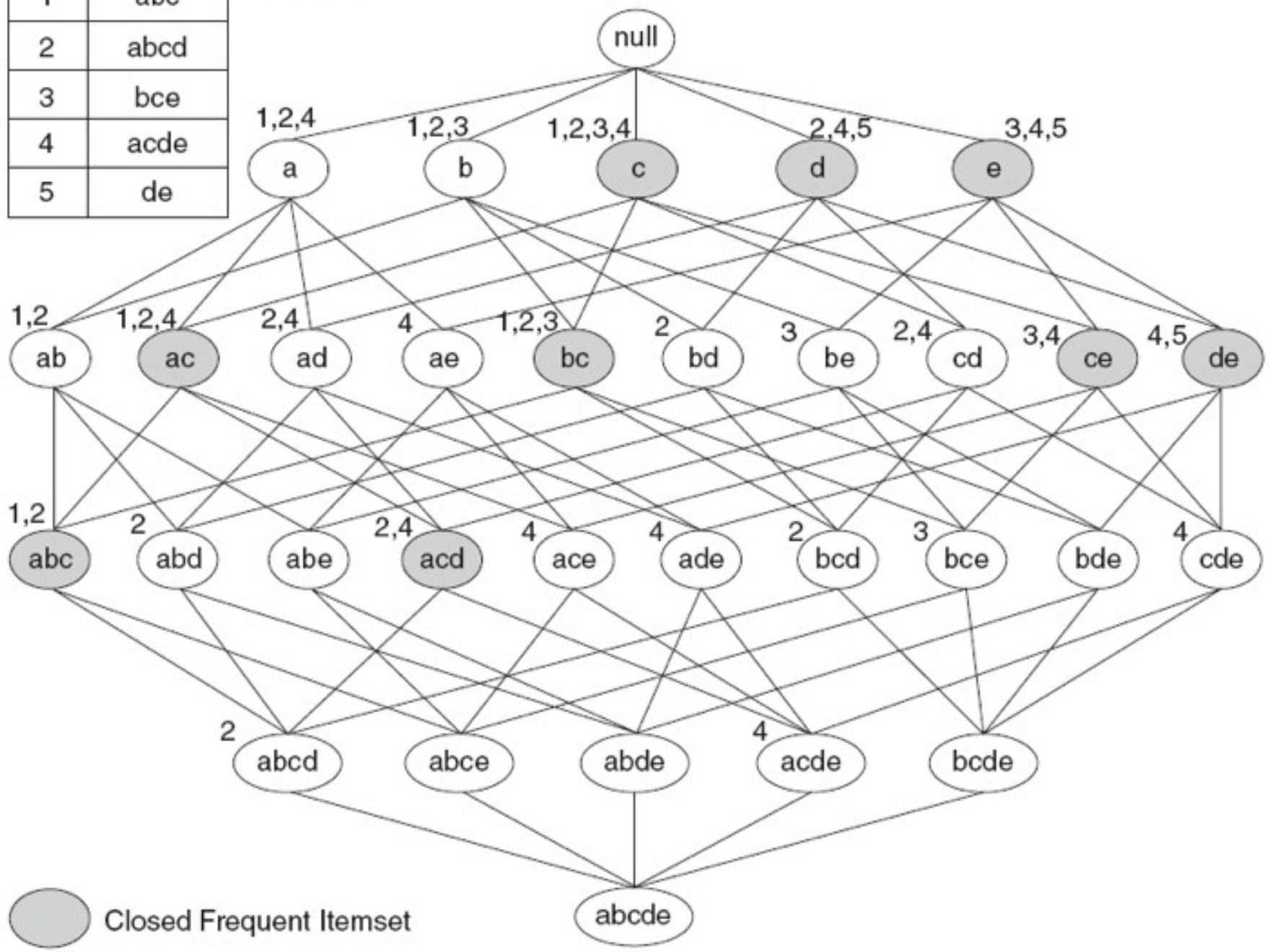


Figure 6.17. An example of the closed frequent itemsets (with minimum support count equal to 40%).

Itemset taxonomy

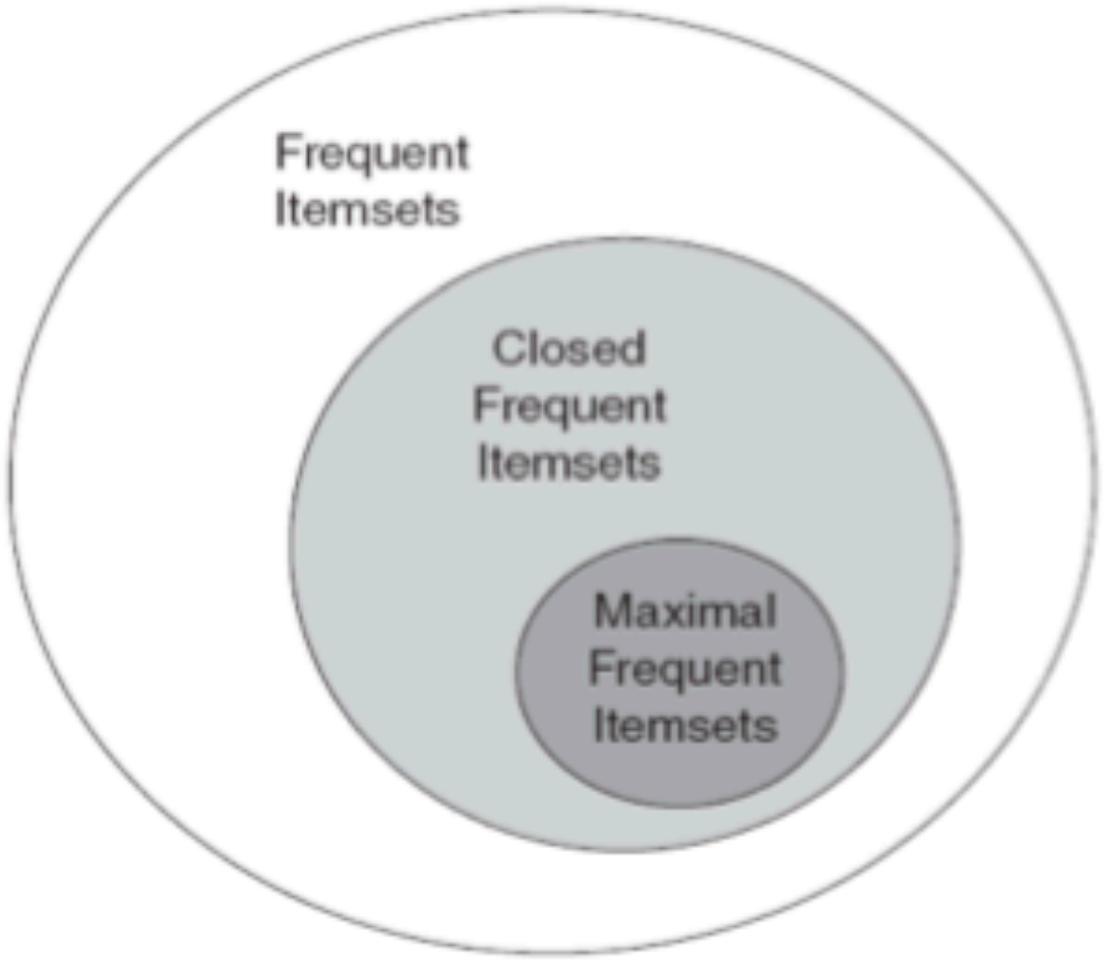


Figure 6.18. Relationships among frequent, maximal frequent, and closed frequent itemsets.

Association rules and confidence

- An **association rule** is a rule of type $X \rightarrow Y$, where X and Y are itemsets
  - If transaction contains itemset X, it (probably) also contains itemset Y
- The **support** of rule $X \rightarrow Y$ in data D is

$$supp(X \rightarrow Y, D) = supp(X \cup Y, D)$$

  - Tan et al. (and other authors) divide this value by $|D|$
- The **confidence** of rule $X \rightarrow Y$ in data D is

$$c(X \rightarrow Y, D) = supp(X \cup Y, D)/supp(X, D)$$

  - The confidence is the empirical conditional probability that transaction contains Y given that it contains X

Association rule examples

TID	Bread	Milk	Diapers	Beer	Eggs
1	1	1	0	0	0
2	1	0	1	1	1
3	0	1	1	1	0
4	1	1	1	1	0
5	1	1	1	0	0

{Bread, Milk} → {Diapers} has support 2 and confidence 2/3
{Diapers} → {Bread, Milk} has support 2 and confidence 1/2
{Eggs} → {Bread, Diapers, Beer} has support 1 and confidence 1

Related data mining tasks

- Frequent itemset mining
  - Given a database and **minfreq**, find all frequent itemsets
  - Often a pre-processing step
- Maximal or closed itemset mining
  - Given a database and **minfreq**, find all maximal or closed itemsets
  - Can provide a succinct presentation of the data
    - Which items appear often together
- Association rule mining
  - Given a database and **minsupp** and **minconf**, find all confident and common association rules
  - Implication analysis: If *X* is bought/observed, what else will probably be bought/observed