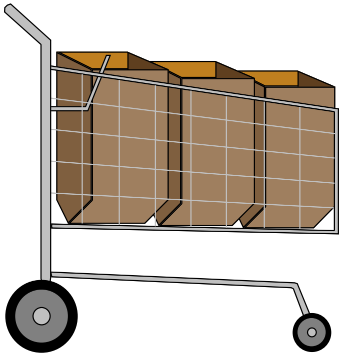
Association Rule Mining

Association Rule Mining

- Consider shopping cart filled with several items
- Market basket analysis tries to answer the following questions:
  - Who makes purchases?
  - What do customers buy together?
  - In what order do customers purchase items?

Market Basket Analysis



- INPUT: list of purchases by purchaser
  - do not have names
- ***identify purchase patterns***
  - what items tend to be purchased together
    - obvious: steak-potatoes; beer-pretzels
  - what items are purchased sequentially
    - obvious: house-furniture; car-tires
  - what items tend to be purchased by season

Association Rules

- Categorize customer purchase behavior
- identify ***actionable*** information
  - purchase profiles
  - profitability of each purchase profile
  - use for marketing
    - layout or catalogs
    - select products for promotion
    - space allocation, product placement

Association Rules

- ***Benefits***
  - selection of promotions, merchandising strategy
    - sensitive to price: Italian entrees, pizza, pies, Oriental entrees, orange juice
  - uncover consumer spending patterns
    - correlations: orange juice & waffles
  - joint promotional opportunities

Applications

- Retail outlets
- Telecommunications
- Banks
- Insurance
  - link analysis for fraud
- Medical
  - symptom analysis

Purchase Profiles

- Each profile has an average profit per basket
  - Kids' fashion $15.24 **push these**
  - Men's fashion $13.41
  - ….
  - Smoker $2.88 *don't push*
  - Student/home office $2.55 *these*

Market Basket Analysis

- ***Affinity Positioning***
  - coffee, coffee makers in close proximity
- ***Cross-Selling***
  - cold medicines, kleenex, orange juice

Association Rules

- Wal-Mart customers who purchase Barbie dolls have a 60% likelihood of also purchasing one of three types of candy bars [*Forbes*, Sept 8, 1997]

- Customers who purchase maintenance agreements are very likely to purchase large appliances (Linoff and Berry experience)

- When a new hardware store opens, one of the most commonly sold items is toilet bowl cleaners (Linoff and Berry experience)

How can Association Rules be used?

- Let the rule discovered be

$\{\text{Bagels},...\} \rightarrow \{\text{Potato Chips}\}$

- Potato chips as consequent => Can be used to determine what should be done to boost its sales

- Bagels in the antecedent => Can be used to see which products would be affected if the store discontinues selling bagels

- Bagels in antecedent and Potato chips in the consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato Chips

What Is Association Rule Mining

- Rule form

  Antecedent $\rightarrow$ Consequent

- Examples
  - buys(x, “computer”) $\rightarrow$ buys(x, “financial management software”)
  - age(x, “30..39”) ^ income(x, “42..48K”) → buys(x, “car”)

Definition: Frequent Itemset

- **Itemset**
  - A collection of one or more items
    - Example: {Milk, Bread, Diaper}
  - k-itemset
    - An itemset that contains k items
- **Support count (σ)**
  - Frequency of occurrence of an itemset
  - E.g. σ({Milk, Bread,Diaper}) = 2
- **Support**
  - Fraction of transactions that contain an itemset
  - E.g. s({Milk, Bread, Diaper}) = 2/5
- **Frequent Itemset**
  - An itemset whose support is greater than or equal to a *minsup* threshold

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

Definition: Association Rule

- **Association Rule**
  - An implication expression of the form $X \rightarrow Y$, where X and Y are itemsets
  - Example: $\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\}$
- **Rule Evaluation Metrics**
  - Support (s)
    - Fraction of transactions that contain both X and Y
  - Confidence (c)
    - Measures how often items in Y appear in transactions that contain X

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**

Example:

$$\{\text{Milk}, \text{Diaper}\} \Rightarrow \text{Beer}$$

$$s = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{|\text{T}|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{\sigma(\text{Milk}, \text{Diaper})} = \frac{2}{3} = 0.67$$

Association Rule

Association rules are rules presenting association or correlation between itemsets.

$$\text{support}(A \Rightarrow B) = P(A \cup B)$$

$$\text{confidence}(A \Rightarrow B) = P(B/A)$$

$$= \frac{P(A \cup B)}{P(A)}$$

$$\text{lift}(A \Rightarrow B) = \frac{\text{confidence}(A \Rightarrow B)}{P(B)}$$

$$= \frac{P(A \cup B)}{P(A)P(B)}$$

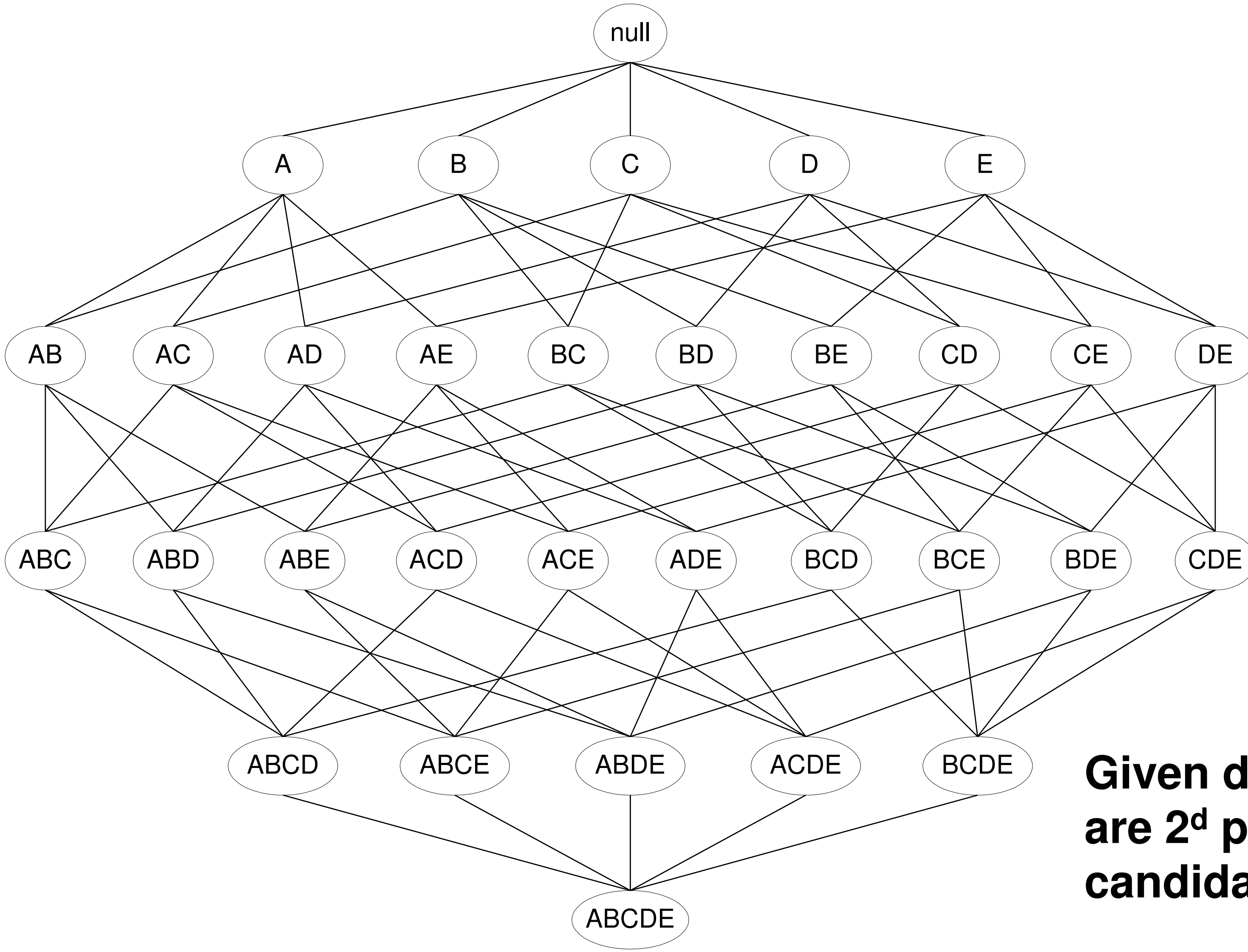
where $P(A)$ is the percentage (or probability) of cases containing A.

Frequent Itemsets Mining

TID	Transactions
100	{ A, B, E }
200	{ B, D }
300	{ A, B, E }
400	{ A, C }
500	{ B, C }
600	{ A, C }
700	{ A, B }
800	{ A, B, C, E }
900	{ A, B, C }
1000	{ A, C, E }

- Minimum support level 50%
  - {A},{B},{C},{A,B}, {A,C}

Frequent Itemset Generation



Given d items, there are 2^d possible candidate itemsets

Frequent Itemset Generation

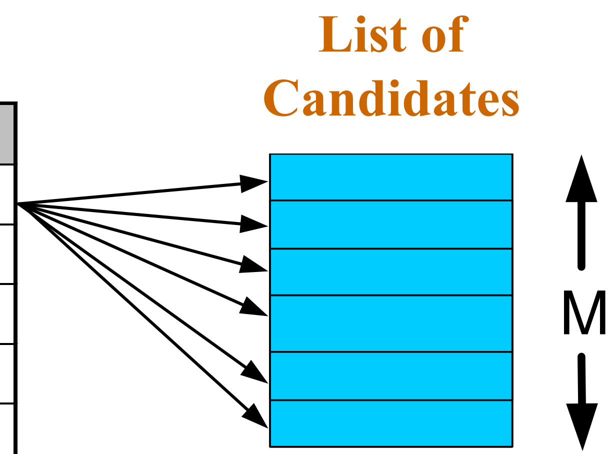
- Brute-force approach:
  - Each itemset in the lattice is a candidate frequent itemset
  - Count the support of each candidate by scanning the database

Transactions

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

N

w



  - Match each transaction against every candidate
  - Complexity ~ O(NMw) => Expensive since $M = 2^d$!!!

Reducing Number of Candidates

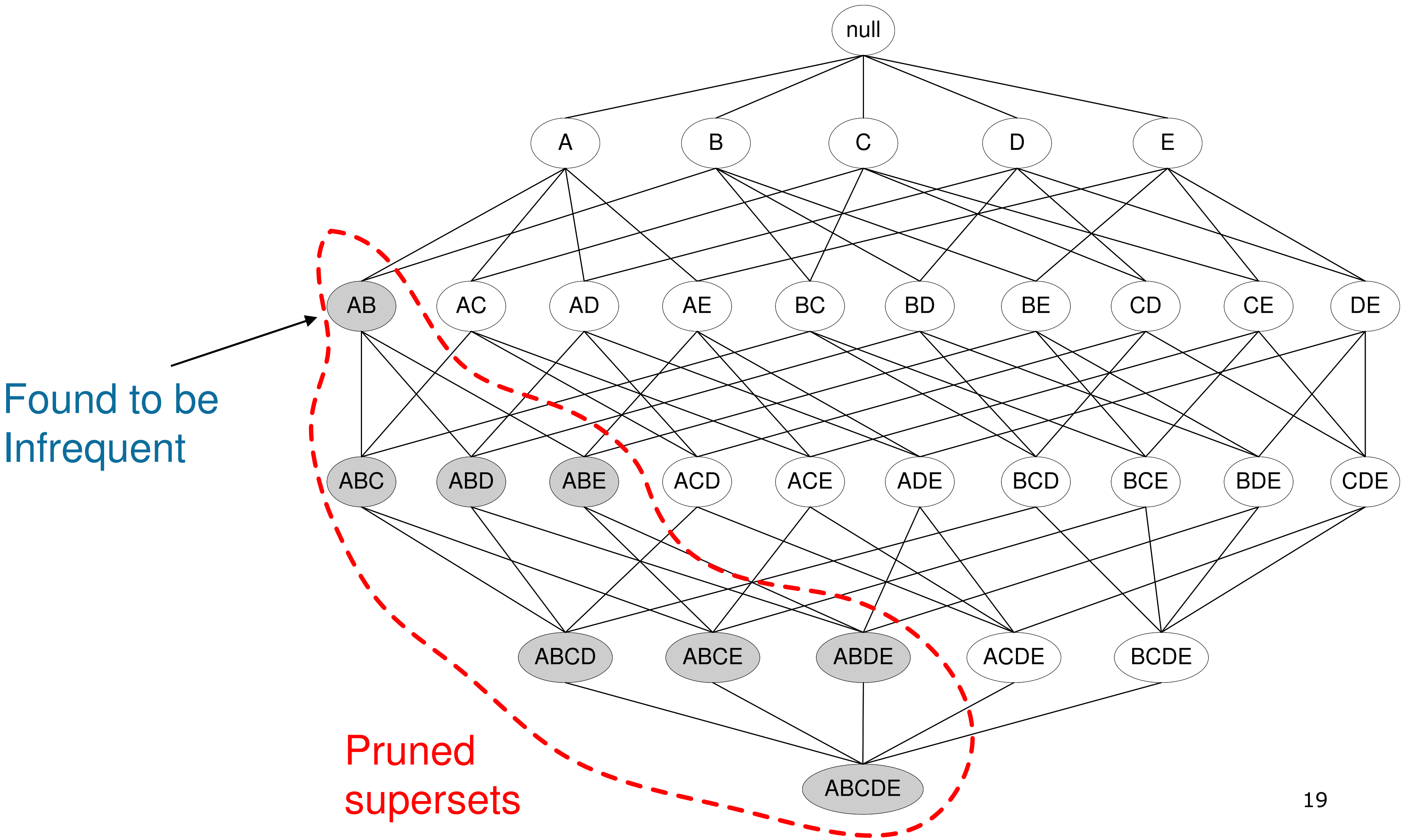
- Apriori principle:
  - If an itemset is frequent, then all of its subsets must also be frequent

- Apriori principle holds due to the following property of the support measure:

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

  - Support of an itemset never exceeds the support of its subsets
  - This is known as the anti-monotone property of support

Illustrating Apriori Principle



Illustrating Apriori Principle

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

Items (1-itemsets)

Itemset	Count
{Bread,Milk}	**3**
{Bread,Beer}	**2**
{Bread,Diaper}	**3**
{Milk,Beer}	**2**
{Milk,Diaper}	**3**
{Beer,Diaper}	**3**

Pairs (2-itemsets)

(No need to generate candidates involving Coke or Eggs)

Minimum Support = 3

Triplets (3-itemsets)

Itemset	Count
{Bread,Milk,Diaper}	**3**

If every subset is considered,

$$^6C_1 + {}^6C_2 + {}^6C_3 = 41$$

With support-based pruning,

$$6 + 6 + 1 = 13$$

Challenges of Frequent Itemset Mining

- Challenges
  - Multiple scans of transaction database
  - Huge number of candidates
  - Tedious workload of support counting for candidates
- Improving Apriori: general ideas
  - Reduce passes of transaction database scans
  - Shrink number of candidates
  - Facilitate support counting of candidates

Alternative Methods for Frequent Itemset Generation

- Representation of Database
  - horizontal vs vertical data layout

Horizontal Data Layout

TID	Items
1	A,B,E
2	B,C,D
3	C,E
4	A,C,D
5	A,B,C,D
6	A,E
7	A,B
8	A,B,C
9	A,C,D
10	B

Vertical Data Layout

A	B	C	D	E
1	1	2	2	1
4	2	3	4	3
5	5	4	5	6
6	7	8	9	
7	8	9		
8	10			
9				

ECLAT

- For each item, store a list of transaction ids (tids)

Horizontal Data Layout

TID	Items
1	A,B,E
2	B,C,D
3	C,E
4	A,C,D
5	A,B,C,D
6	A,E
7	A,B
8	A,B,C
9	A,C,D
10	B

Vertical Data Layout

A	B	C	D	E
1	1	2	2	1
4	2	3	4	3
5	5	4	5	6
6	7	8	9	
7	8	9		
8	10			
9				

TID-list

ECLAT

- Determine support of any k-itemset by intersecting tid-lists of two of its (k-1) subsets.

A
1
4
5
6
7
8
9

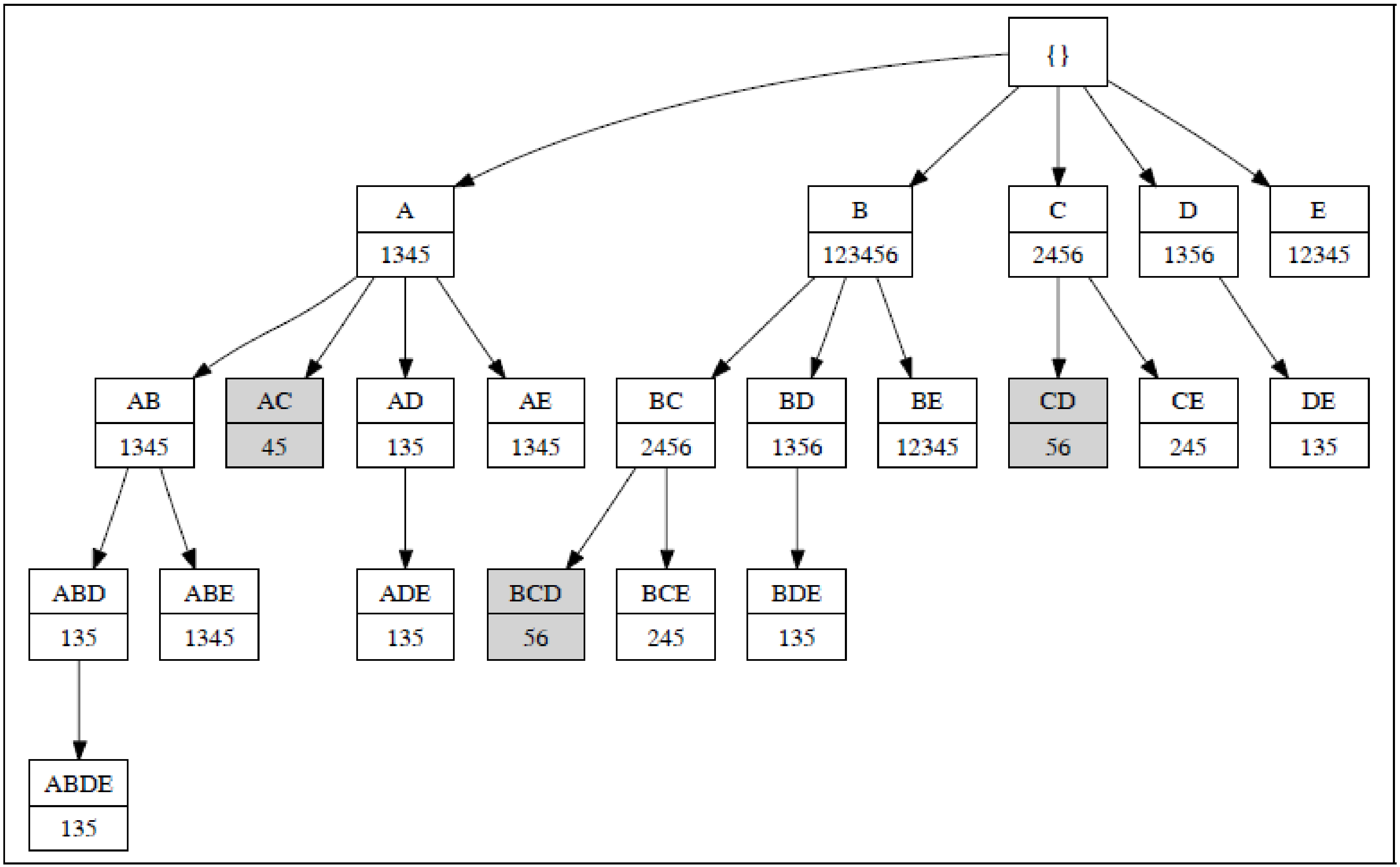
$\wedge$

B
1
2
5
7
8
10

$\rightarrow$

AB
1
5
7
8

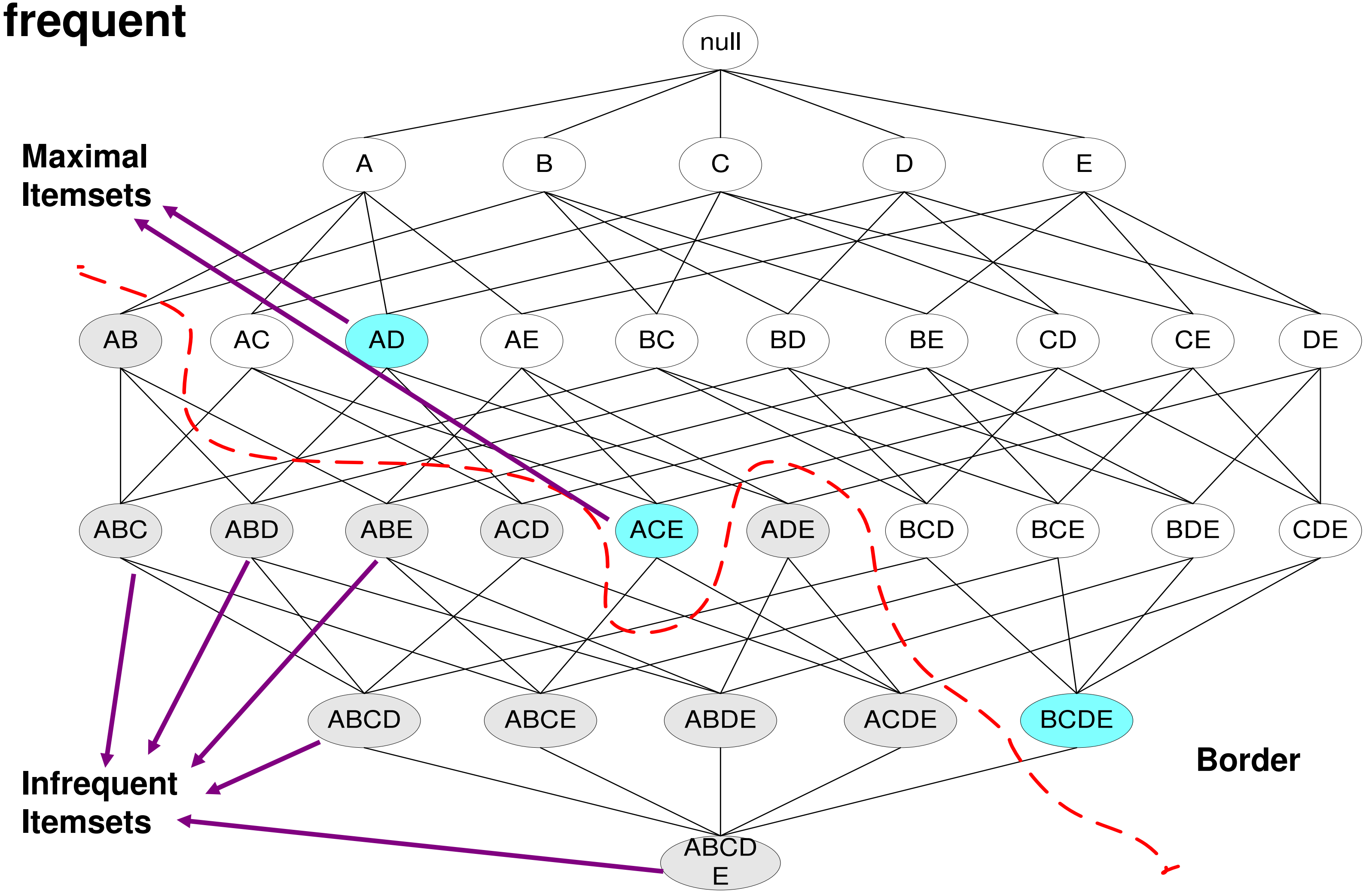
- 3 traversal approaches:
  - top-down, bottom-up and hybrid
- Advantage: very fast support counting
- Disadvantage: intermediate tid-lists may become too large for memory



{ }
A
1345
B
123456
C
2456
D
1356
E
12345
AB
1345
AC
45
AD
135
AE
1345
BC
2456
BD
1356
BE
12345
CD
56
CE
245
DE
135
ABD
135
ABE
1345
ADE
135
BCD
56
BCE
245
BDE
135
ABDE
135

Maximal Frequent Itemset

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



Closed Itemset

- An itemset is closed if none of its immediate supersets has the same support as the itemset

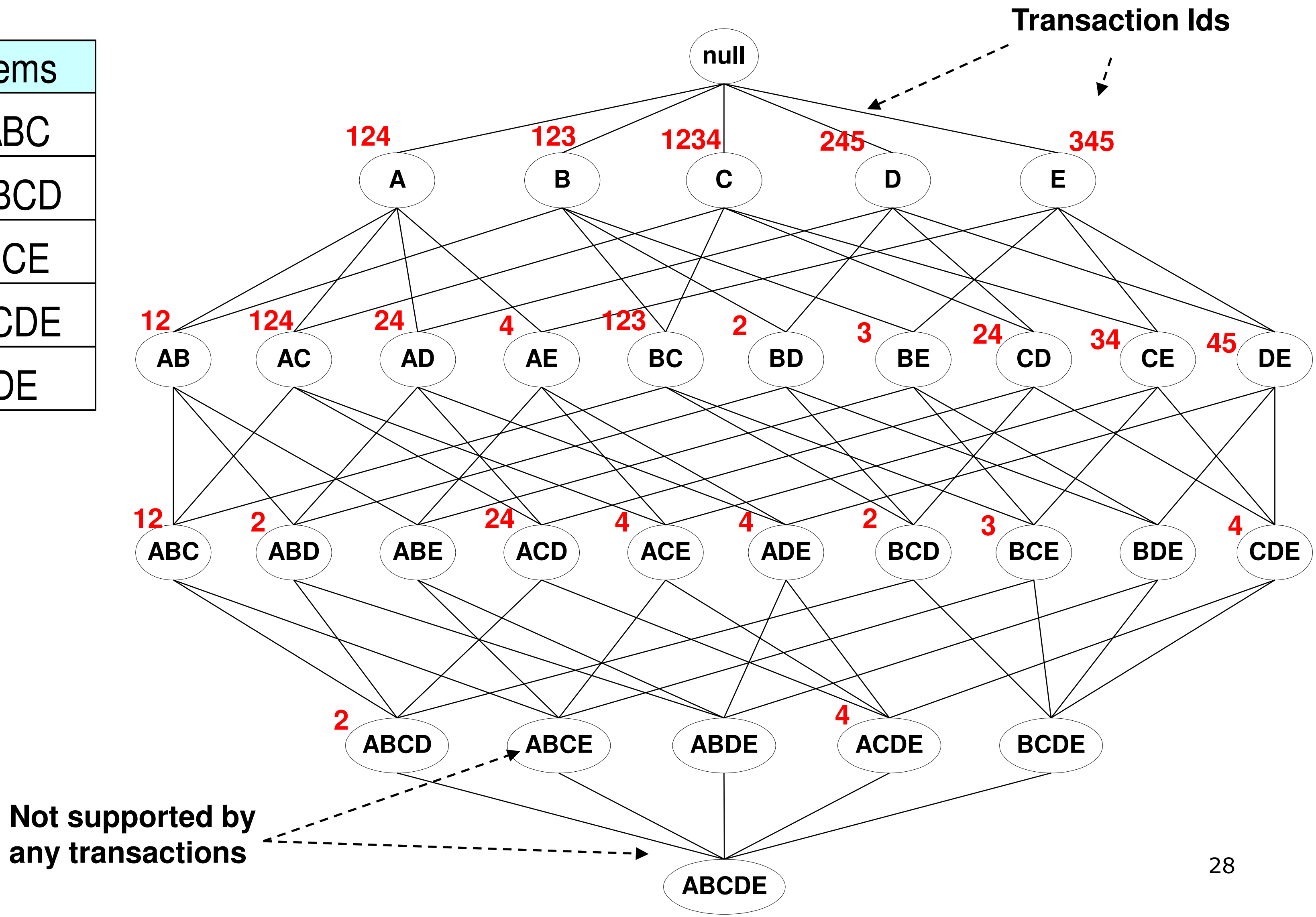
TID	Items
1	{A,B}
2	{B,C,D}
3	{A,B,C,D}
4	{A,B,D}
5	{A,B,C,D}

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

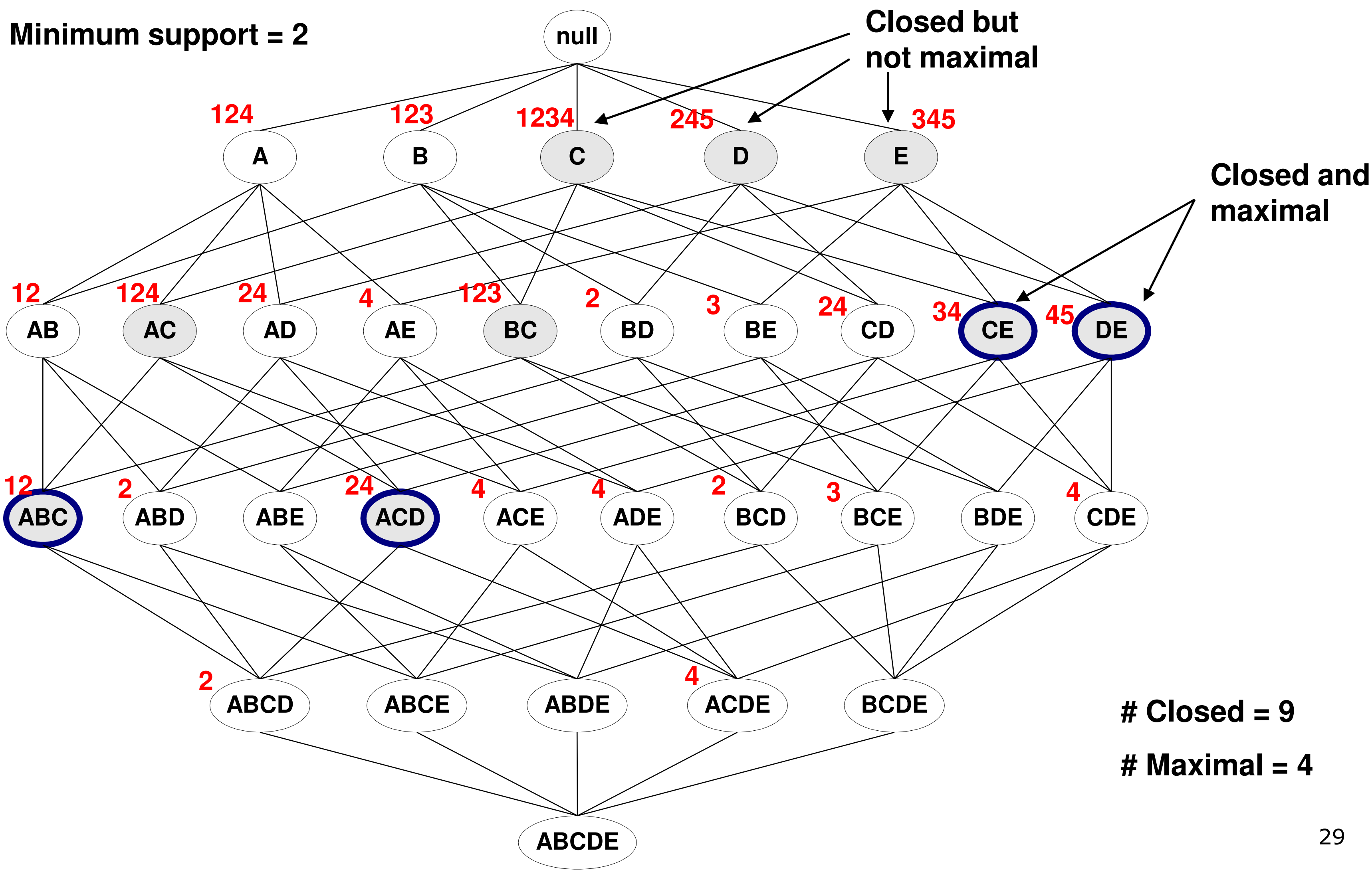
Itemset	Support
{A,B,C}	2
{A,B,D}	3
{A,C,D}	2
{B,C,D}	3
{A,B,C,D}	2

Maximal vs Closed Itemsets

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



Maximal vs Closed Frequent Itemsets



Maximal vs Closed Itemsets