Association Rule Mining (Market Basket Analysis)

Data Mining

Prof. Sujee Lee

Department of Systems Management Engineering

Sungkyunkwan University

Association Rule Mining

Association Rule Mining

- Association Rule Mining is a data mining technique used to find patterns and relationships in large datasets.
- Often used in market basket analysis
- Given a set of **transactions**, find **rules** that will predict the occurrence of an item based on the occurrences of other items in the transaction

- Market-Basket (POS) 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

- Examples of Association Rules

$$\{\text{Diaper}\} \rightarrow \{\text{Beer}\}$$
$$\{\text{Milk, Bread}\} \rightarrow \{\text{Eggs, Coke}\}$$
$$\{\text{Beer, Bread}\} \rightarrow \{\text{Milk}\}$$

Key Concepts

- **Itemset**
  - A collection of one or more items
    - e.g. {Milk, Bread, Diaper}
  - k-itemset: An itemset that contains k items
- **Support Count (sc)**
  - Number of occurrence of an itemset
  - e.g. sc({Milk, Bread, Diaper})=2
- **Support (supp)**
  - Fraction of transactions that contain an itemset
  - e.g. supp({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

Key Concepts

- **Association Rule**
  - **If X, then Y. (X → Y)**
    - X : Antecedent
    - Y: Consequent
  - Many rules are possible
    - For the itemset {Bread, Milk}:
      - Bread → Milk
      - Milk → Bread
    - For the itemset {Bread, Milk, Diaper}:
      - Bread, Milk → Diaper
      - Bread, Diaper → Milk
      - …

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

Rule Evaluation Metrics

- **Support (supp)**
  - Fraction of transactions that contain both X and Y

$$\mathrm{supp}(X \rightarrow Y) = \frac{\mathrm{sc}(\{X, Y\})}{N}, \ N = \text{number of total transactions}$$

- **Confidence (conf)**
  - Measures how often items in Y appear in transactions that contain X.

$$\mathrm{conf}(X \rightarrow Y) = P(Y \mid X) = \frac{\mathrm{sc}(X,Y)/N}{\mathrm{sc}(X)/N} = \frac{\mathrm{sc}(X, Y)}{\mathrm{sc}(X)}$$

- **Lift (lift) {=1, independent; >1, positive relationship; <1, negative relationship}**
  - Measure of how much more likely items X and Y are to occur together than expected by chance.
  - Ratio of observed support to that expected if X and Y were independent.

$$\mathrm{lift}(X \rightarrow Y) = \frac{\mathrm{conf}(X \rightarrow Y)}{P(Y)} = \frac{\mathrm{supp}(X, Y)}{\mathrm{supp}(X) \times \mathrm{supp}(Y)}$$

Rule Evaluation Metrics

- **Example**
  - Rule: {Milk, Diaper} → {Beer}

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

$$\text{supp}(\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\})$$
$$= \frac{\text{sc}(\{\text{Milk, Diaper, Beer}\})}{5} = \frac{2}{5} = 0.4$$

$$\text{conf}(\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\})$$
$$= \frac{\text{sc}(\{\text{Milk, Diaper, Beer}\})}{\text{sc}(\{\text{Milk, Diaper}\})} = \frac{2}{3} = 0.67$$

$$\text{lift}(\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\})$$
$$= \frac{\text{supp}(\{\text{Milk, Diaper, Beer}\})}{\text{supp}(\{\text{Milk, Diaper}\}) \times \text{supp}(\{\text{Beer}\})} = \frac{2/5}{3/5 \times 3/5} = 1.11$$

Rule Evaluation Metrics

- **Support**
  - A measure of significance (importance) of an itemset.
  - "Larger is better" does not hold always: *Rare item problem*

- **Confidence**
  - Different values for the rules X→ Y and Y → X
  - Sensitive to the frequency of Y
  - Caused by the way confidence is calculated, Y with a high support will automatically produce a high confidence value even if there exist no association between X and Y.

- **Lift**
  - Measures how many times more often X and Y occur together.
  - Useful rules have the lift values greater than 1.

Association Rule Mining Task

- **Association Rule Mining Task**
  - Given a set of transactions, the goal of association rule mining is to find all rules having
    - supp ≥ *minsup* threshold
    - conf ≥ *minconf* threshold

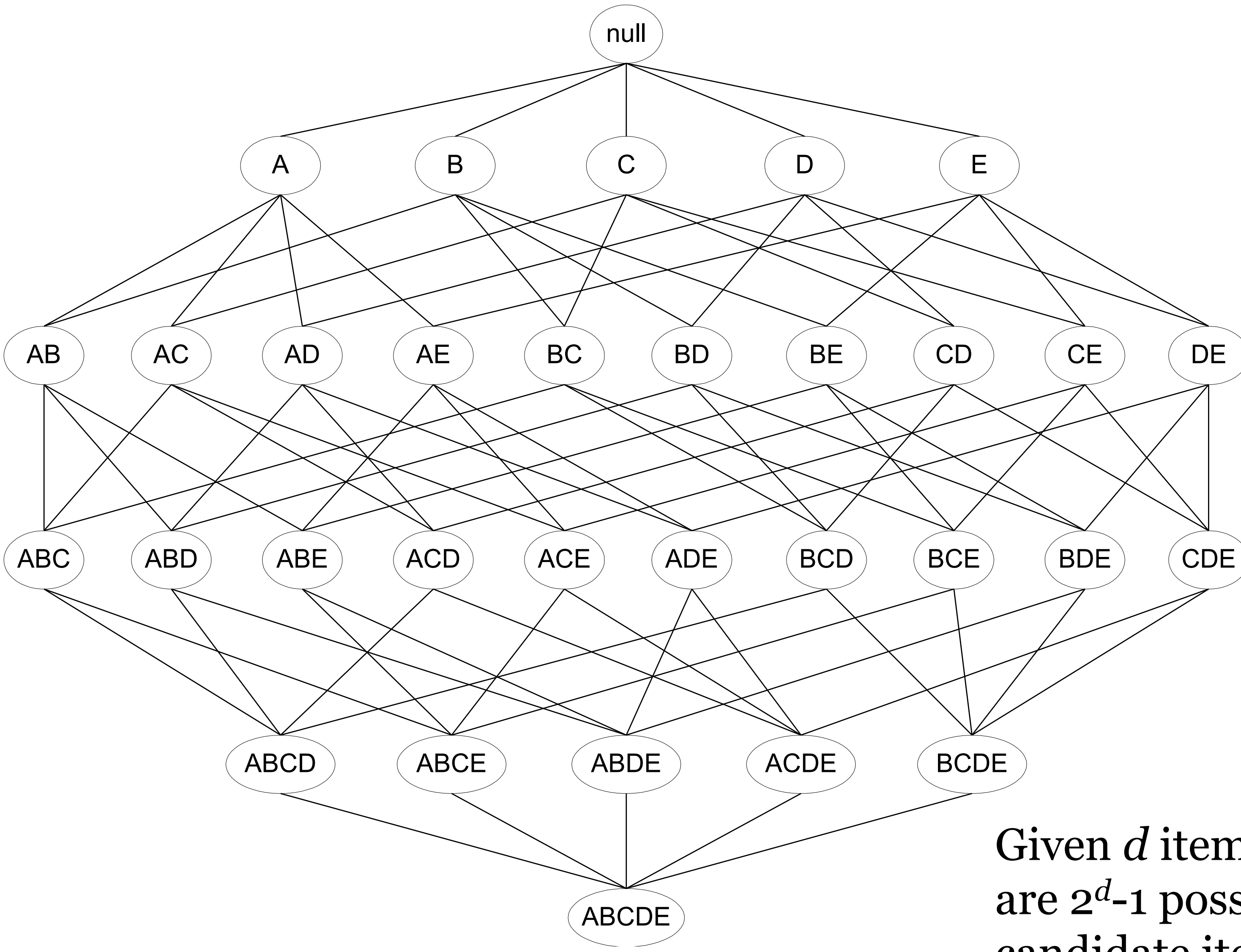
- **Brute-force approach**
  - List all possible association rules
  - Compute the support and confidence for each rule
  - Prune rules that fail the *minsup* and *minconf* thresholds
  - Computationally expensive!

Mining Association Rules

- **Two-step approach**
  1. **Frequent Itemset Generation**
     - Generate all itemsets whose supp ≥ *minsup*
  2. **Rule Generation**
     - Generate high confidence rules (conf ≥ *minconf*) from each frequent itemset, where each rule is a binary partitioning of a frequent itemset

- Frequent Itemset Generation is computationally expensive.

Possible Candidate Itemsets



Given d items, there are 2^d-1 possible candidate itemsets

Apriori Algorithm

- **Apriori principle**
  - **If an itemset is frequent, then all of its subsets must also be frequent. (Apriori property)**
  - Apriori principle holds due to the following property of the support measure

$$\forall X, Y: \text{if } X \subset Y, \text{then } \mathrm{supp}(Y) \leq \mathrm{supp}(X)$$

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

Apriori Algorithm

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

* Revised example

- **Illustrating apriori algorithm**

Items (1-itemsets)

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

Pairs (2-itemsets)

(No need to generate candidates involving Coke or Eggs)

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

Itemset	Count
{Bread,Milk,Diaper}	3

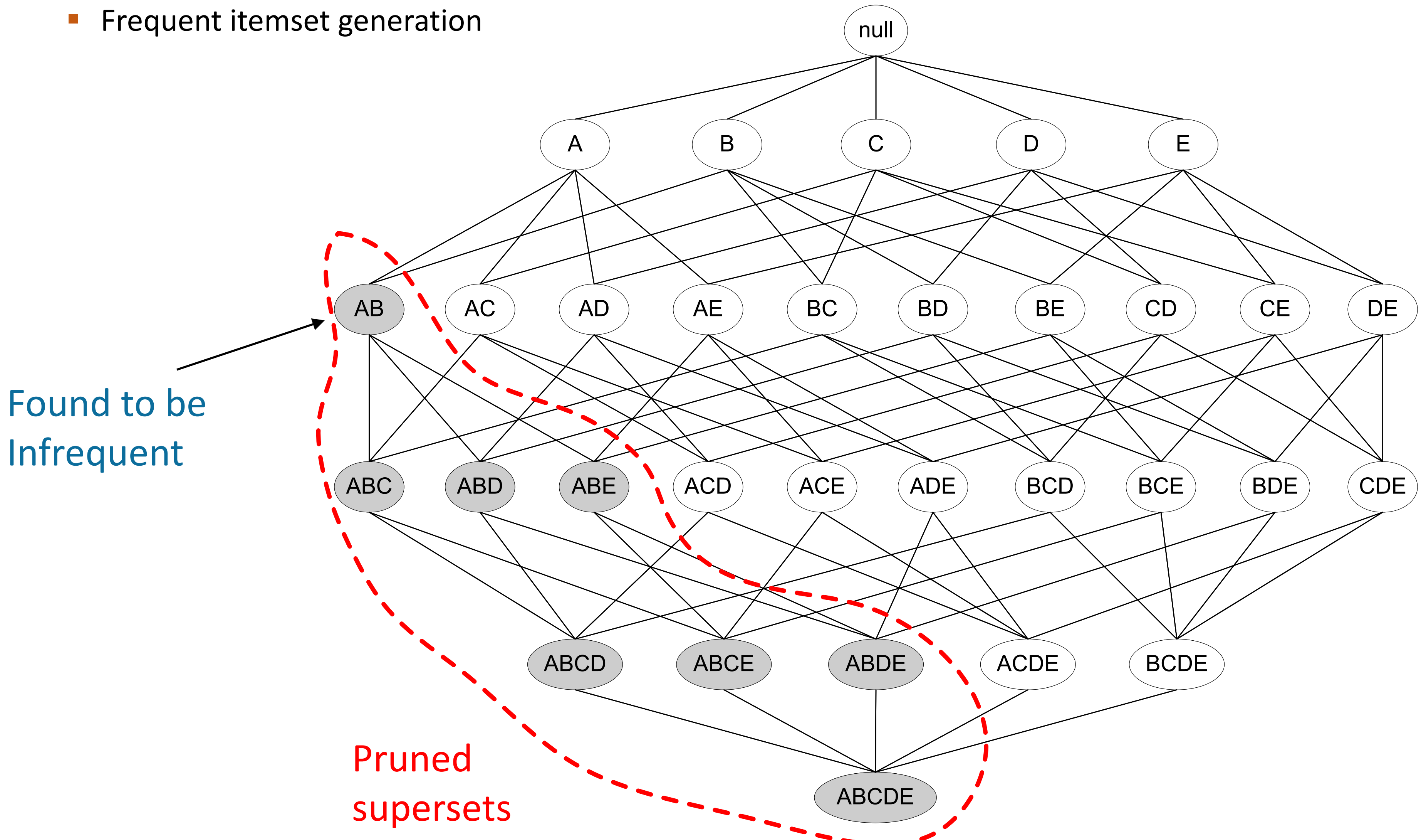
Minimum support count = 3

Apriori Algorithm

- **Frequent Itemset Generation**
  - Let k=1
  - Generate frequent itemsets of length 1
  - Repeat until no new frequent itemsets are identified
    - Generate length (k+1) candidate itemsets from length k frequent itemsets
    - Prune candidate itemsets containing subsets of length k that are infrequent
    - Count the support of each candidate by scanning the DB
    - Eliminate candidates that are infrequent, leaving only those that are frequent

Apriori Algorithm

- **Illustrating apriori algorithm**
  - Frequent itemset generation



Apriori Algorithm

- **Rule Generation**

  1. All the high-confidence rules, satisfying the greater-than-or-equal-to-*minconf*, that have only one item in the rule consequent are extracted. (1-item consequent rules)
  2. These rules are then used to generate new candidate rules. (2-item consequent rules)
     - The new candidate rule is generated by merging the consequents of two rules.
  3. Repeat the procedure until (*g*-1)-item consequent rules are generated, where g is the number of items in a frequent itemset.

  - Example

a frequent itemset: $\{A,B,C\}$

$\{B,C\} \rightarrow \{A\} \qquad \{A,C\} \rightarrow \{B\} \qquad \{A,B\} \rightarrow \{C\}$

$\{C\} \rightarrow \{A,B\} \qquad \{B\} \rightarrow \{A,C\} \qquad \{A\} \rightarrow \{B,C\}$

Apriori Algorithm

Rule Generation

- For the rules generated from the same frequent itemset Y, the following theorem holds.
- *If a rule* $X \rightarrow Y - X$ *does not satisfy the confidence threshold, then any rule* $X' \rightarrow Y - X'$ *, where* X' *is a subset of* X*, must not satisfy the confidence threshold as well.*

$$\mathrm{conf}(X \rightarrow Y - X) = \frac{\mathrm{supp}(Y)}{\mathrm{supp}(X)} = \frac{\mathrm{sc}(Y)}{\mathrm{sc}(X)}$$

$$\mathrm{conf}(X' \rightarrow Y - X') = \frac{\mathrm{supp}(Y)}{\mathrm{supp}(X')} = \frac{\mathrm{sc}(Y)}{\mathrm{sc}(X')}$$

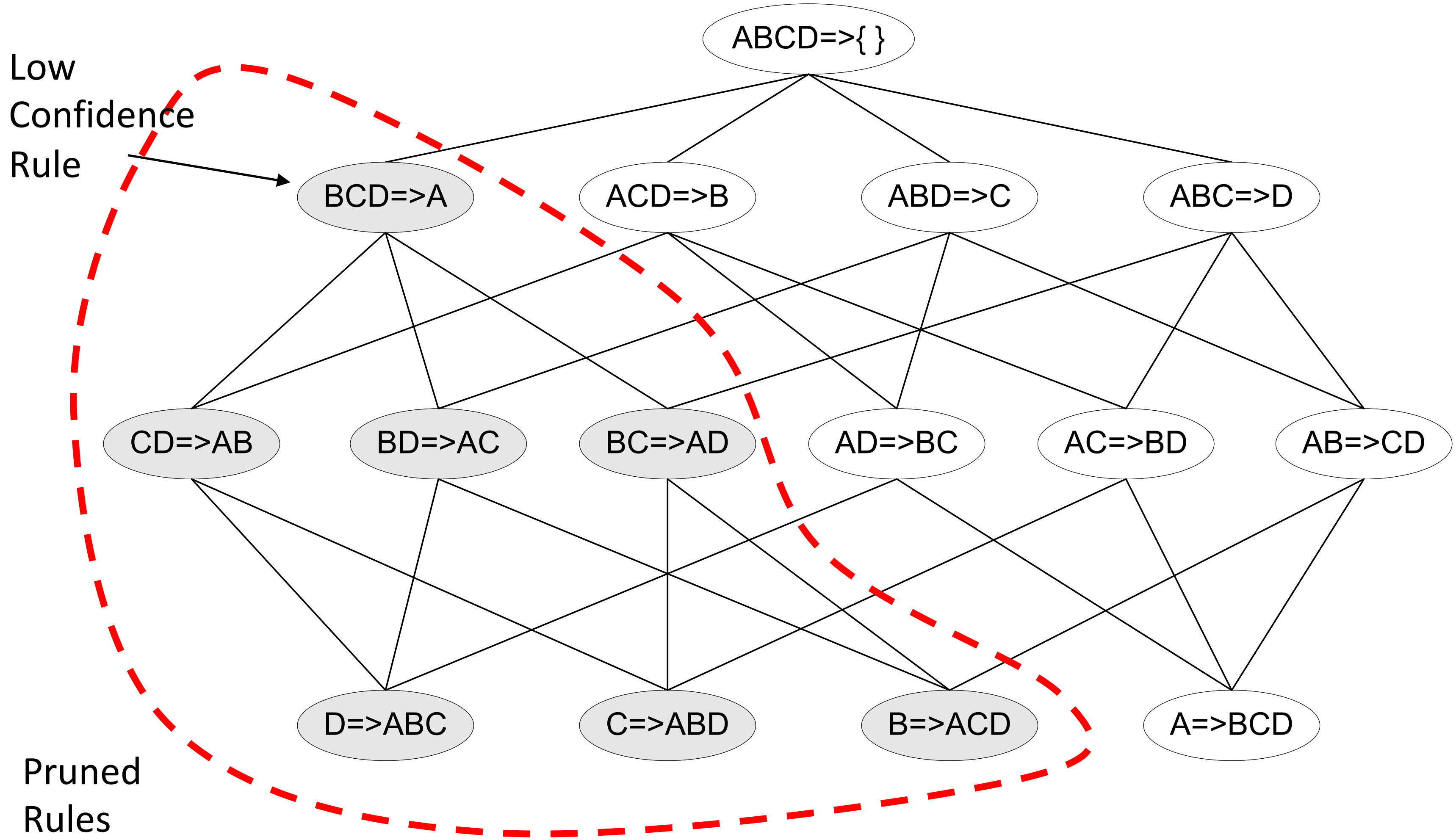
- By anti-monotone property

$$X' \subset X \Rightarrow \mathrm{supp}(X') \geq \mathrm{supp}(X) \text{ or } \mathrm{sc}(X') \geq \mathrm{sc}(X)$$

$$\therefore \mathrm{conf}(X \rightarrow Y - X) \geq \mathrm{conf}(X' \rightarrow Y - X')$$

Apriori Algorithm

- **Illustrating apriori algorithm**
  - Rule generation



Apriori Algorithm

- **Rule Generation**
  - From the previous example

- Frequent itemsets -

item	SC
Bread	4
Milk	4
Beer	3
Diaper	4

item	SC
Bread, Milk	3
Bread, Diaper	3
Milk, Diaper	3
Beer, Diaper	3

item	SC
Bread, Milk, Diaper	3

- Example rules (minconf=1) -

$$\text{Bread} \rightarrow \text{Milk}, \quad \text{conf}(\text{Bread} \rightarrow \text{Milk}) = \frac{\text{supp}(\text{Bread, Milk})}{\text{supp}(\text{Bread})} = \frac{3}{4} \quad \times$$

$$\text{Milk} \rightarrow \text{Bread}, \quad \text{conf}(\text{Milk} \rightarrow \text{Bread}) = \frac{\text{supp}(\text{Bread, Milk})}{\text{supp}(\text{Milk})} = \frac{3}{4} \quad \times$$

$$\text{Beer} \rightarrow \text{Diaper}, \quad \text{conf}(\text{Beer} \rightarrow \text{Diaper}) = \frac{\text{supp}(\text{Beer, Diaper})}{\text{supp}(\text{Beer})} = \frac{3}{3} \quad \bigcirc$$

$$\text{Bread, Milk} \rightarrow \text{Diaper}, \quad \text{conf}(\text{Bread, Milk} \rightarrow \text{Diaper}) = \frac{\text{supp}(\text{Bread, Milk, Diaper})}{\text{supp}(\text{Bread, Milk})} = \frac{3}{3} \quad \bigcirc$$