Week 11
Mining Association Rules (Part I)

Seokho Chi
Assistant Professor I Ph.D.
SNU Construction Innovation Lab

Source: Tan, Kumar, Steinback (2006)

Association Rule Mining

- 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 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

Example of Association Rules

{Diaper} → {Beer},
{Milk, Bread} → {Eggs,Coke},
{Beer, Bread} → {Milk},

Implication means co-occurrence, not causality!

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Association Rule Mining

- Itemset: a collection of one or more items
  - 1 item set: {milk}, 3 item set: {milk, bread, diaper}
- Support count (θ): frequency of occurrence of an itemset
  - θ({milk, bread, diaper}) = 2
- Support (S): fraction of transactions that contain an itemset
  - S({milk, bread, diaper}) = 2/5
- Frequent itemset: an item set whose support is greater than or equal to a minimum support threshold(*minsup*)

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Association Rule Mining

- Association rule: an implication expression of the form X → Y where X and Y are item sets
  - {milk, diaper} → {bread}, {milk} → {diaper, bread}

- Rule evaluation metrics
  - Support (S): fraction of transactions that contain both X and Y
  - Confidence (C): measure how often items in Y appear transactions that contain X

  - {Milk, Diaper} → {Beer}
  - S = 2/5 : milk, diaper & beer among total
  - C = 2/3: beer among milk, diaper

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Association Rule Discovery: Application 1

- Marketing and Sales Promotion:
  - Let the rule discovered be

    {Bagels, … } --> {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 consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Association Rule Discovery: Application 2

- Inventory Management:
  - Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
  - Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

Association Rule Mining Task

- Given a set of transactions T, the goal of association rule mining is to find all rules having
  - support ≥ *minsup* threshold
  - confidence ≥ *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 prohibitive!

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Mining Association Rules

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 of Rules:

{Milk,Diaper} → {Beer} (s=0.4, c=0.67)
{Milk,Beer} → {Diaper} (s=0.4, c=1.0)
{Diaper,Beer} → {Milk} (s=0.4, c=0.67)
{Beer} → {Milk,Diaper} (s=0.4, c=0.67)
{Diaper} → {Milk,Beer} (s=0.4, c=0.5)
{Milk} → {Diaper,Beer} (s=0.4, c=0.5)

Observations:

- All the above rules are binary partitions of the same itemset:
  {Milk, Diaper, Beer}
- Rules originating from the same itemset have identical support but can have different confidence
- Thus, we may decouple the support and confidence requirements

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

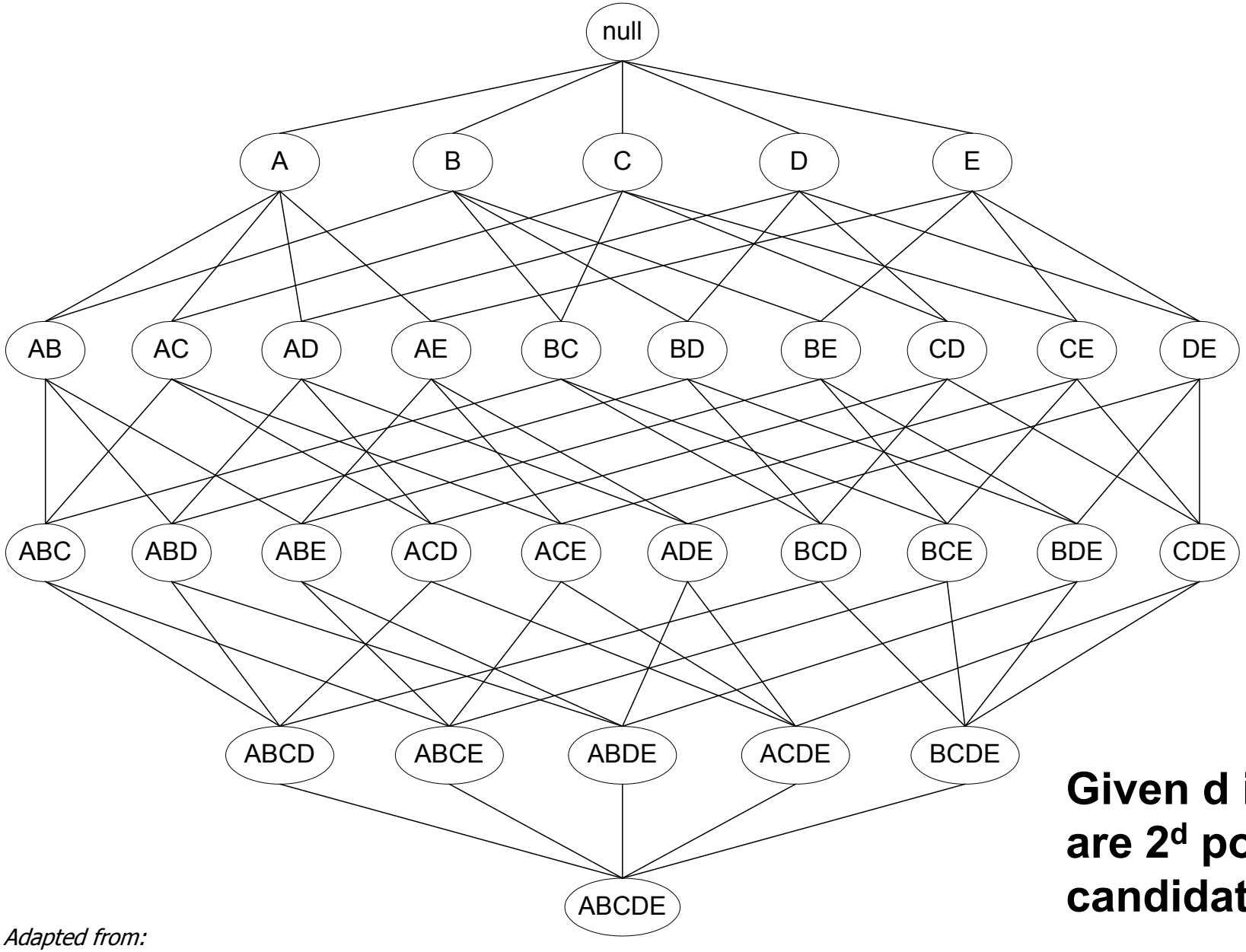
Mining Association Rules

- Two-step approach:
  1. Frequent Itemset Generation
     - Generate all itemsets whose support $\geq$ minsup
  2. Rule Generation
     - Generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset

- Frequent itemset generation is still computationally expensive

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Frequent Itemset Generation

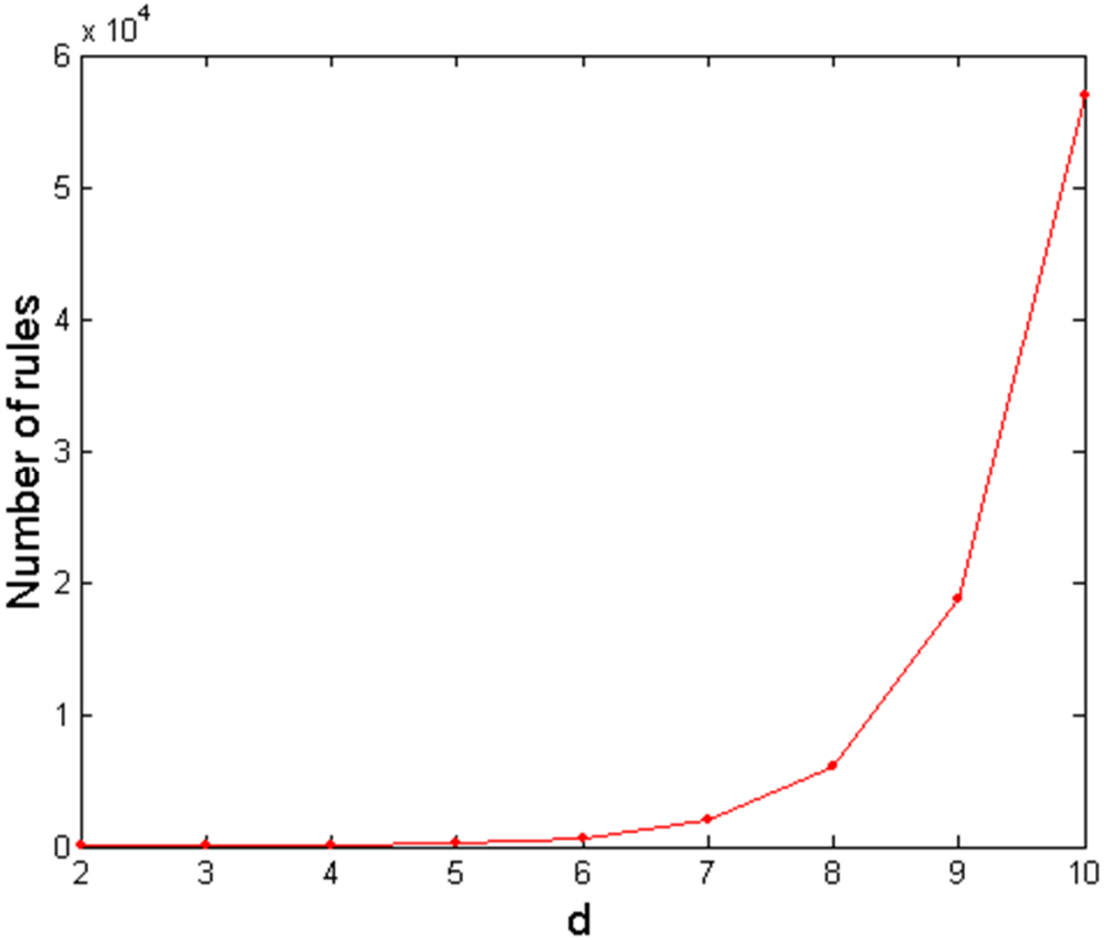


Given d items, there are 2^d possible candidate itemsets

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Computational Complexity

- Given d unique items:
  - Total number of itemsets = 2^d
  - Total number of possible association rules:



$$R = \sum_{k=1}^{d-1}\left[\binom{d}{k} \times \sum_{j=1}^{d-k}\binom{d-k}{j}\right]$$

$$= 3^d - 2^{d+1} + 1$$

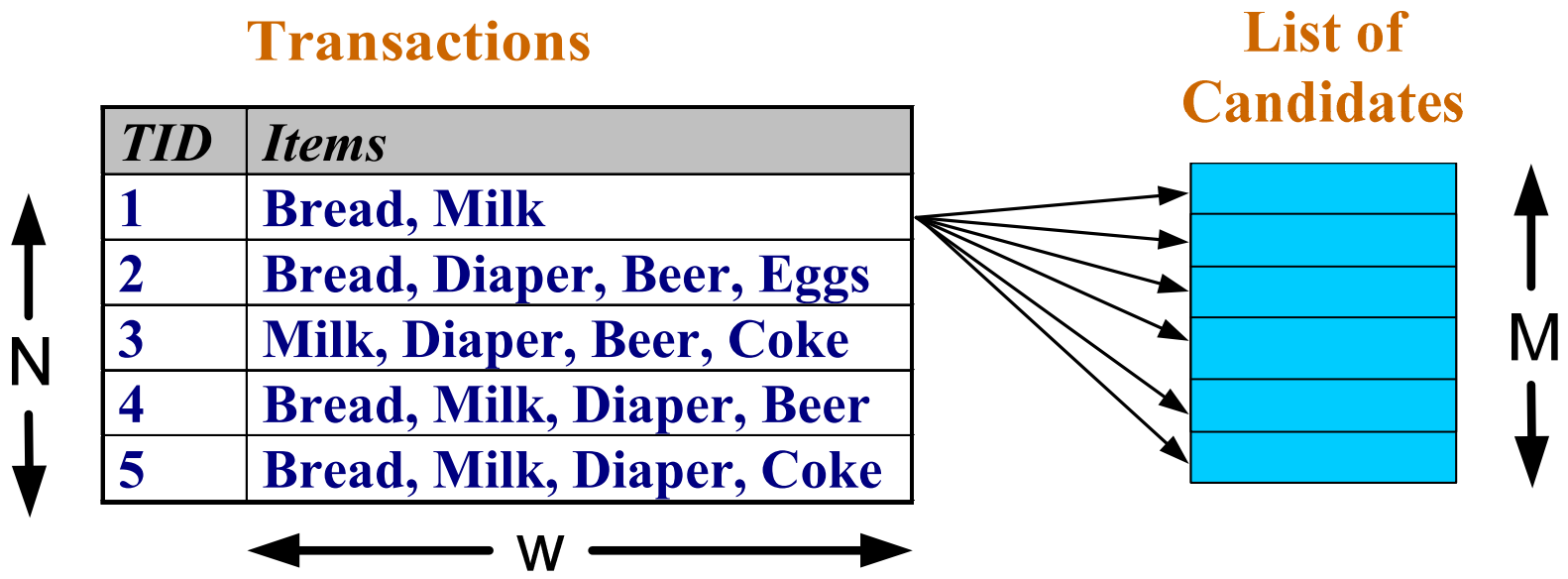
If d=6, R = 602 rules

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

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



  - Match each transaction against every candidate

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Frequent Itemset Generation Strategies

- Reduce the number of candidates (M)
  - Complete search: $M=2^d$
  - Use pruning techniques to reduce M
- Reduce the number of transactions (N)
  - Reduce size of N as the size of itemset increases
  - Used by DHP and vertical-based mining algorithms
- Reduce the number of comparisons (NM)
  - Use efficient data structures to store the candidates or transactions
  - No need to match every candidate against every transaction

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Reducing Number of Candidates

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

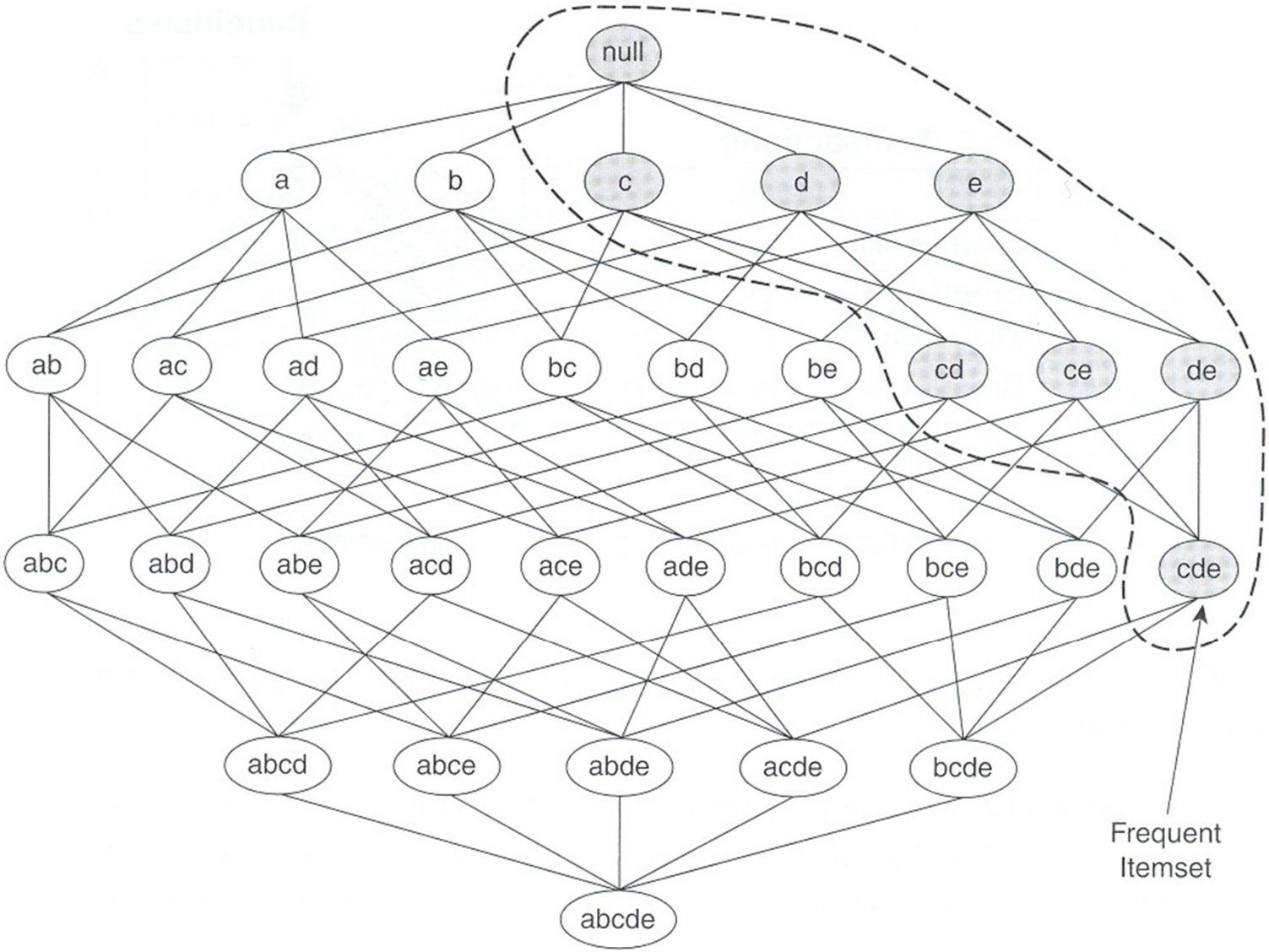
- Apriori principle holds due to the following property of the support measure:

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

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

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

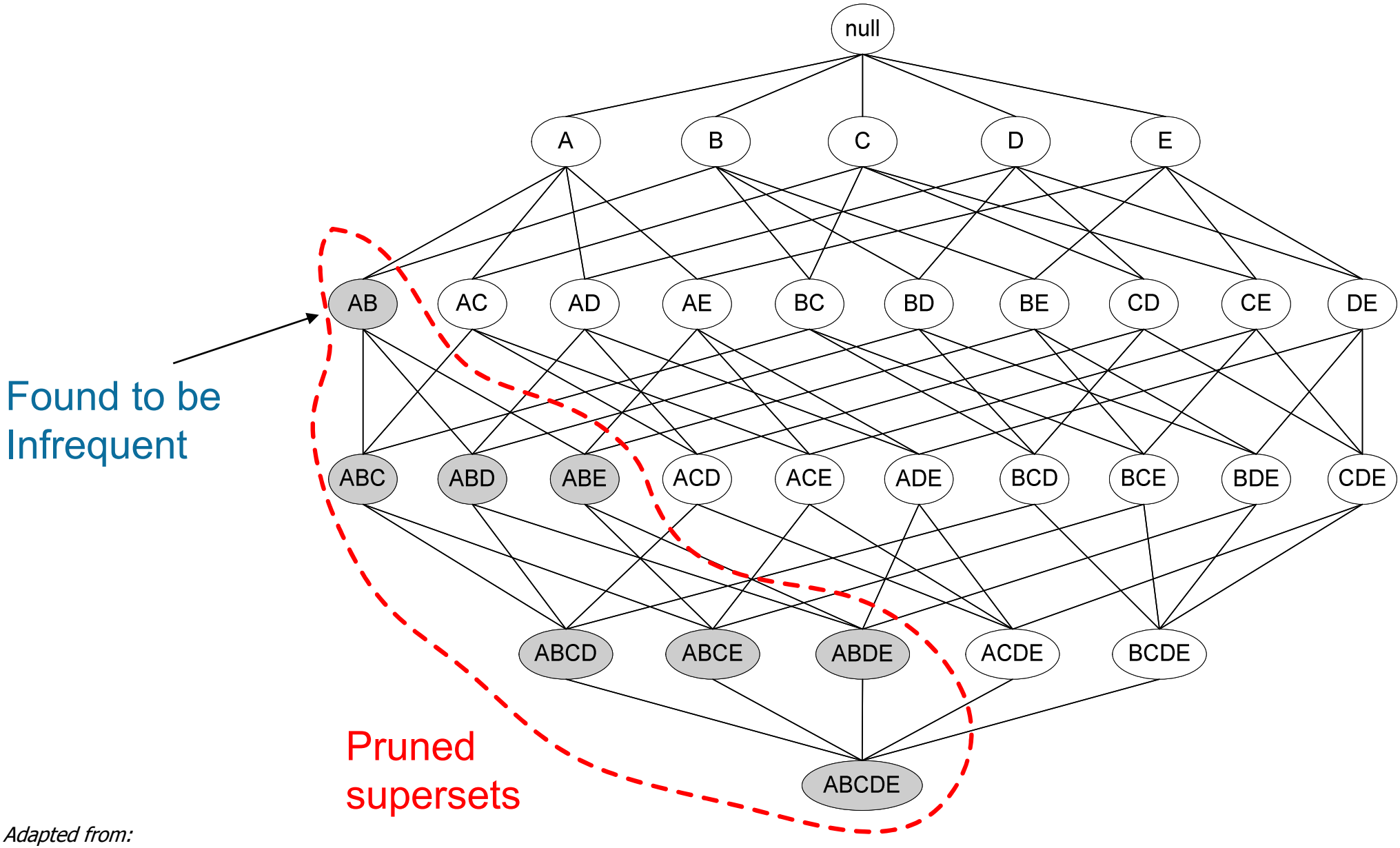
Illustrating Apriori Principle



If "cde" is frequent, all things are also frequent!

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Illustrating Apriori Principle



Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

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,

$$6 + 6 + 1 = 13$$

Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Apriori Algorithm

- Method:
  - 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

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Candidate Generation

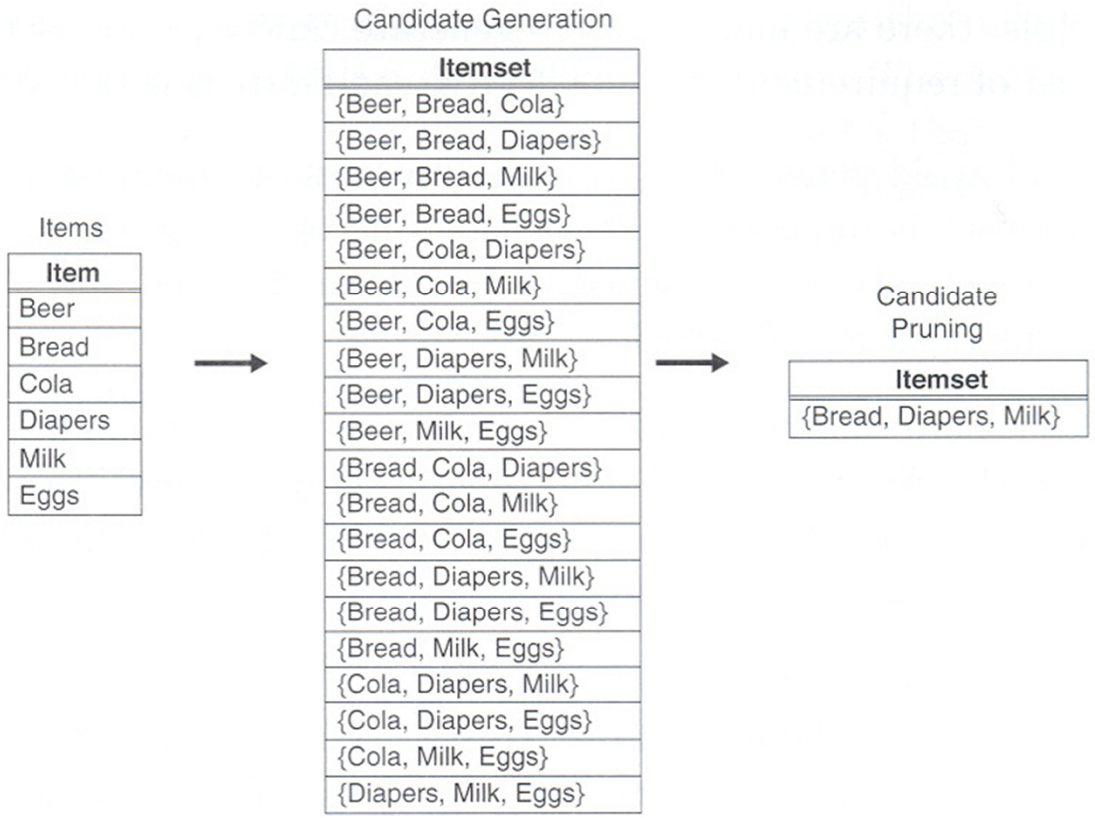


Figure 6.6. A brute-force method for generating candidate 3-itemsets.

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Candidate Generation (2)

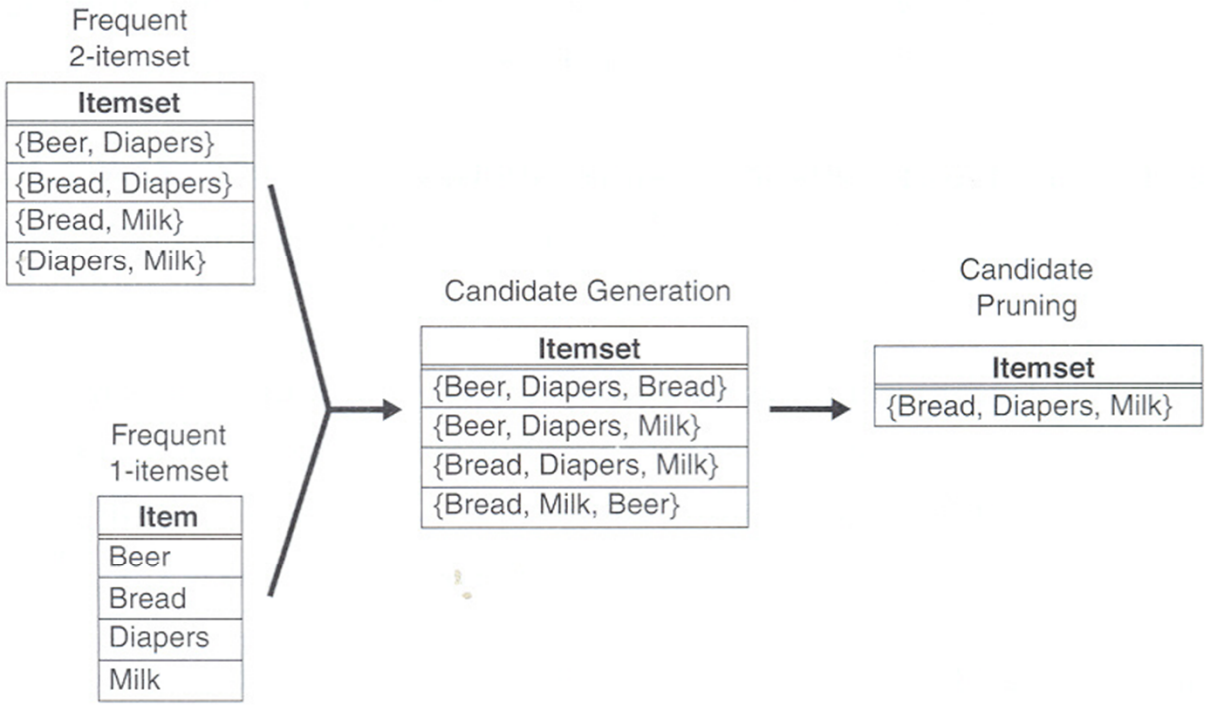


Figure 6.7. Generating and pruning candidate k-itemsets by merging a frequent $(k-1)$-itemset with a frequent item. Note that some of the candidates are unnecessary because their subsets are infrequent.

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Candidate Generation (3)

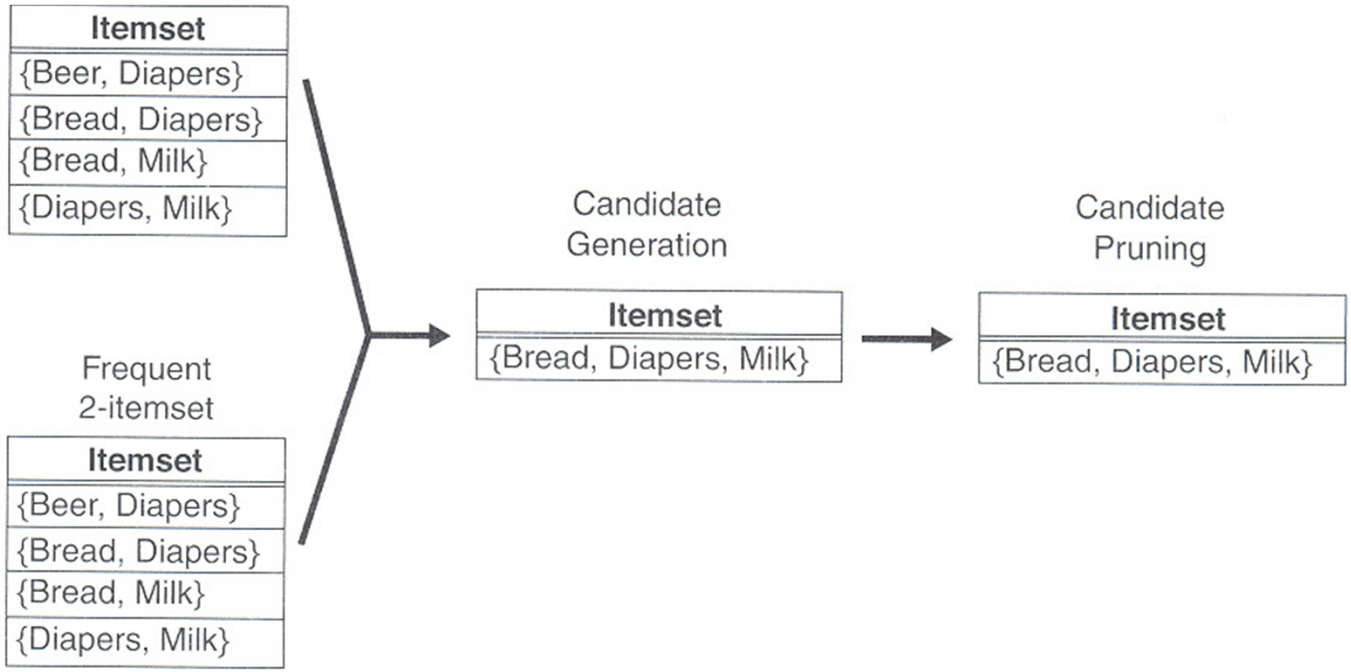
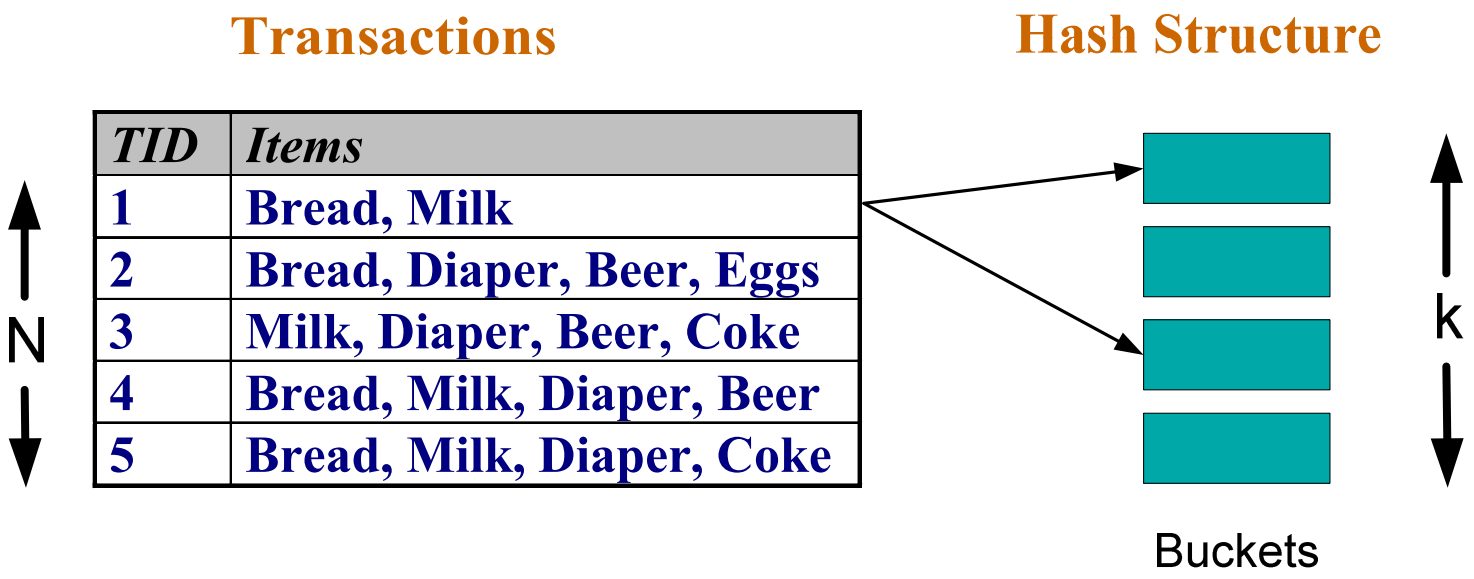


Figure 6.8. Generating and pruning candidate k-itemsets by merging pairs of frequent $(k-1)$-itemsets.

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Reducing Number of Comparisons

- Candidate counting:
  - Scan the database of transactions to determine the support of each candidate itemset
  - To reduce the number of comparisons, store the candidates in a hash structure
    - Instead of matching each transaction against every candidate, match it against candidates contained in the hashed buckets



Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

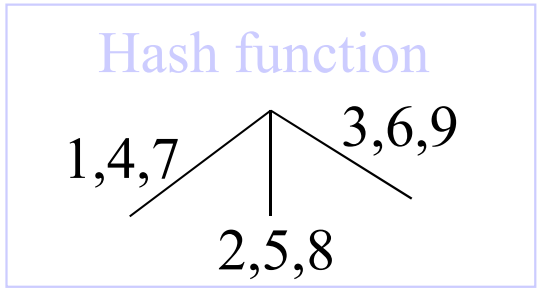
Generate Hash Tree

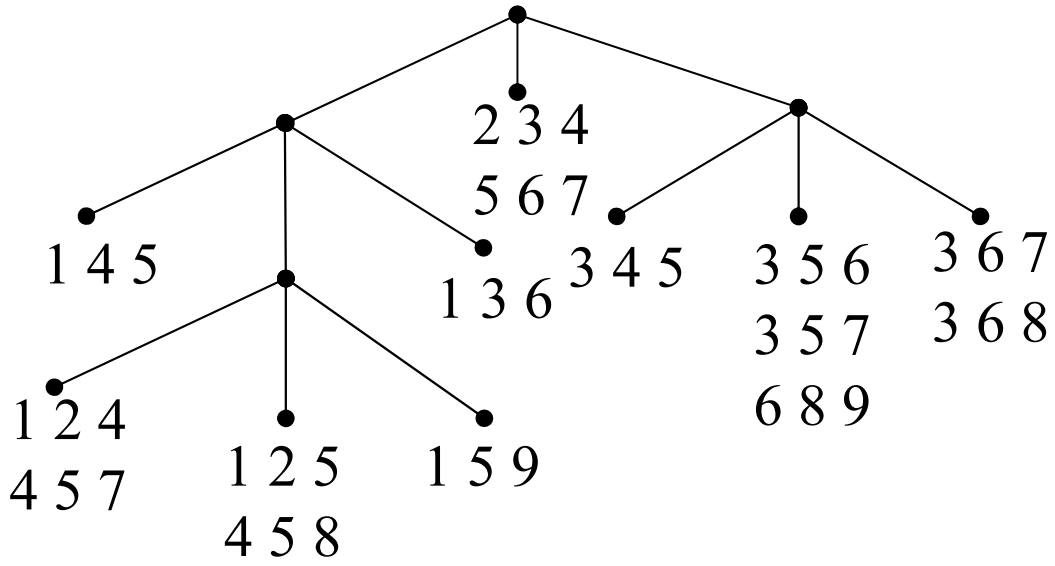
Suppose you have 15 candidate itemsets of length 3:

{1 4 5}, {1 2 4}, {4 5 7}, {1 2 5}, {4 5 8}, {1 5 9}, {1 3 6}, {2 3 4}, {5 6 7}, {3 4 5}, {3 5 6}, {3 5 7}, {6 8 9}, {3 6 7}, {3 6 8}

You need:

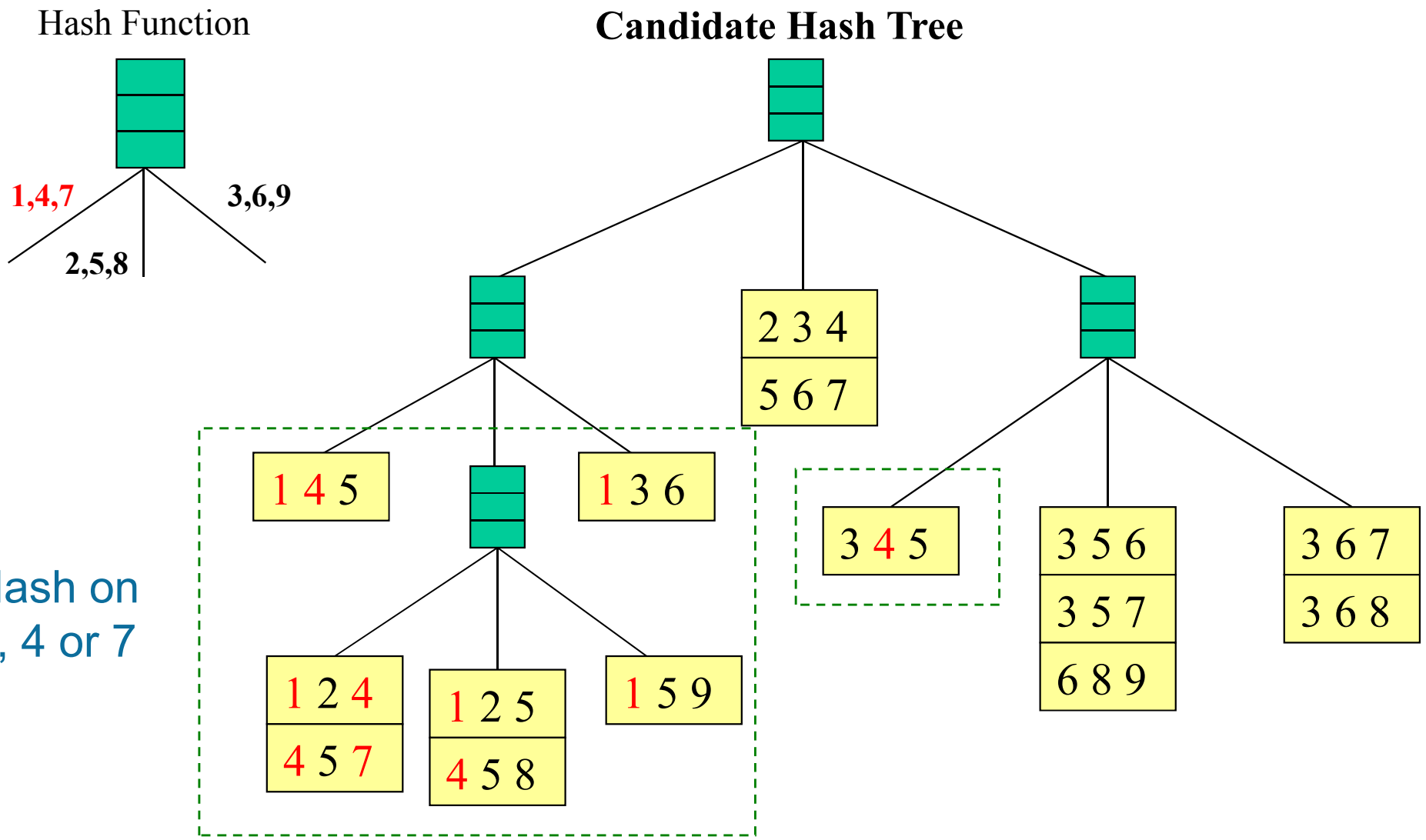
- **Hash function**
- **Max leaf size: max number of itemsets stored in a leaf node (if number of candidate itemsets exceeds max leaf size, split the node)**





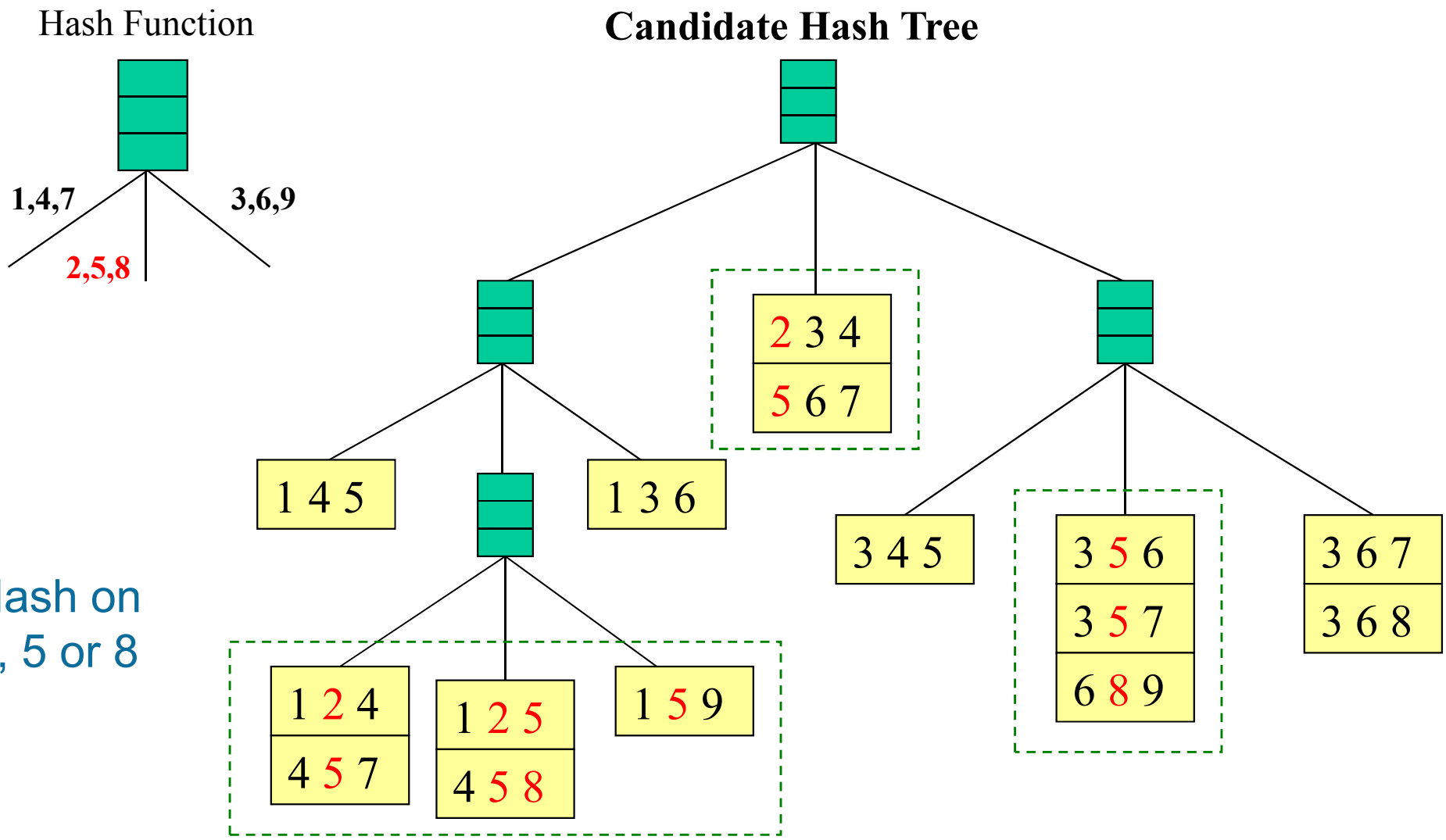
Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Association Rule Discovery: Hash tree



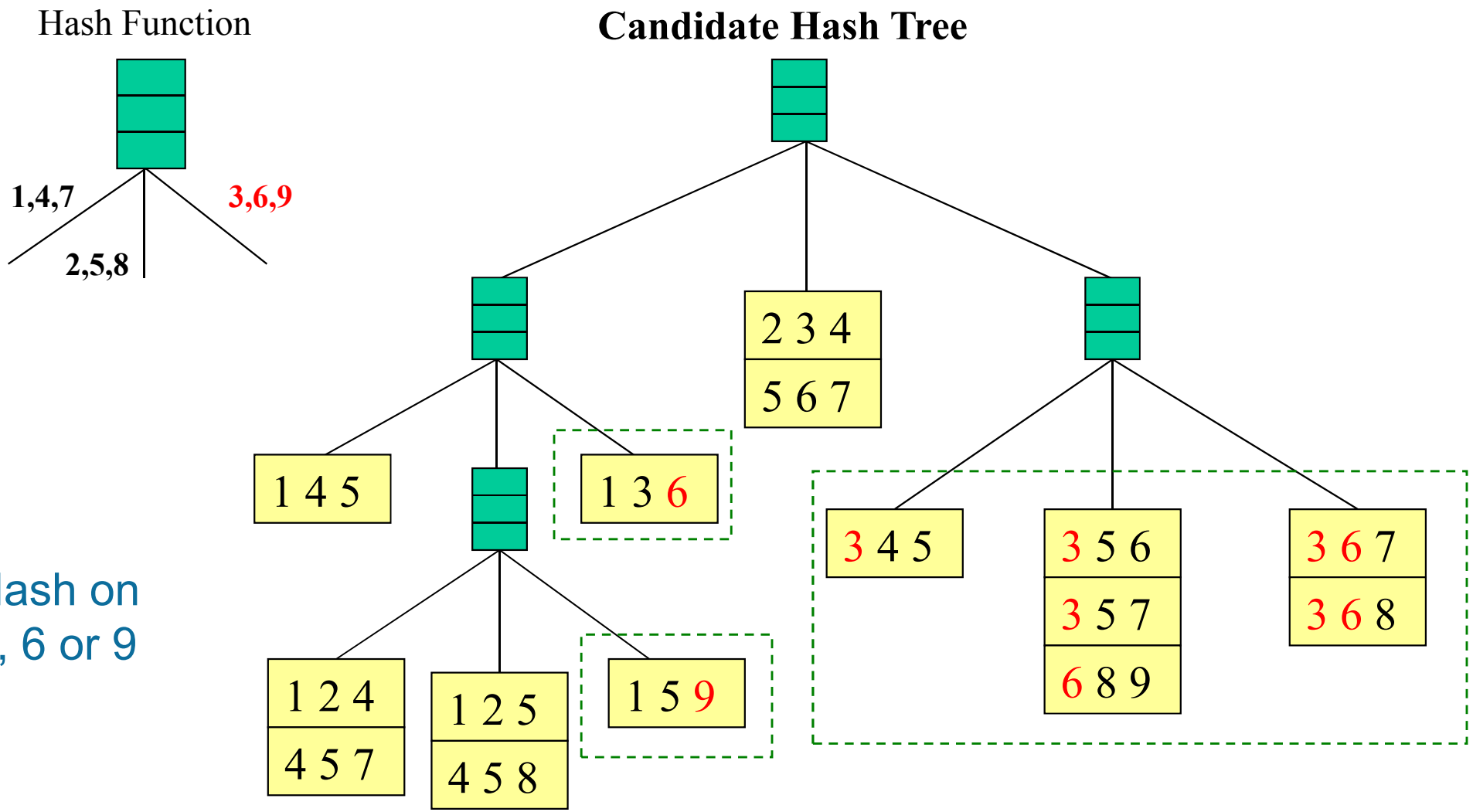
Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Association Rule Discovery: Hash tree



Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

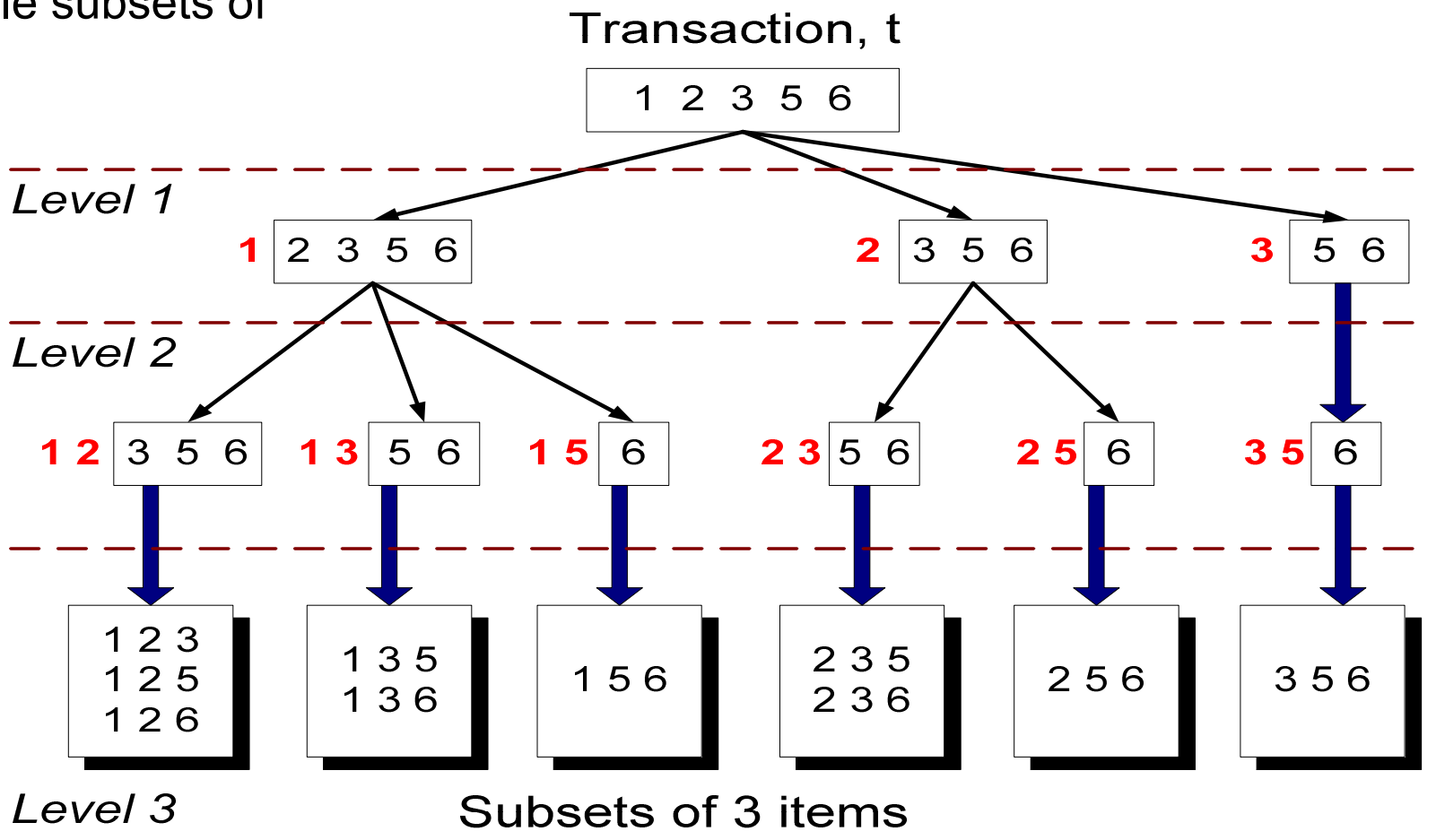
Association Rule Discovery: Hash tree



Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

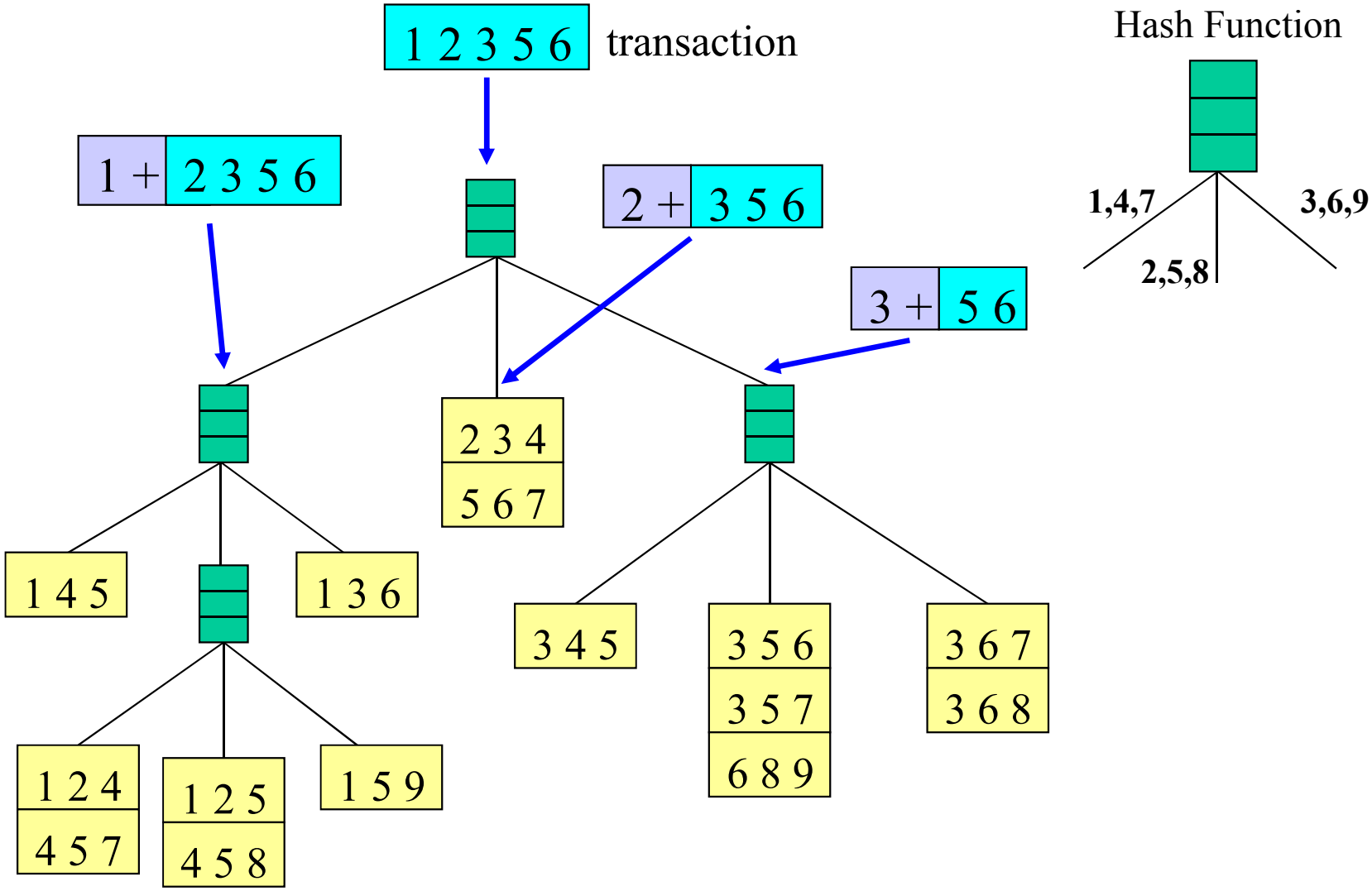
Subset Operation

Given a transaction t, what are the possible subsets of size 3?



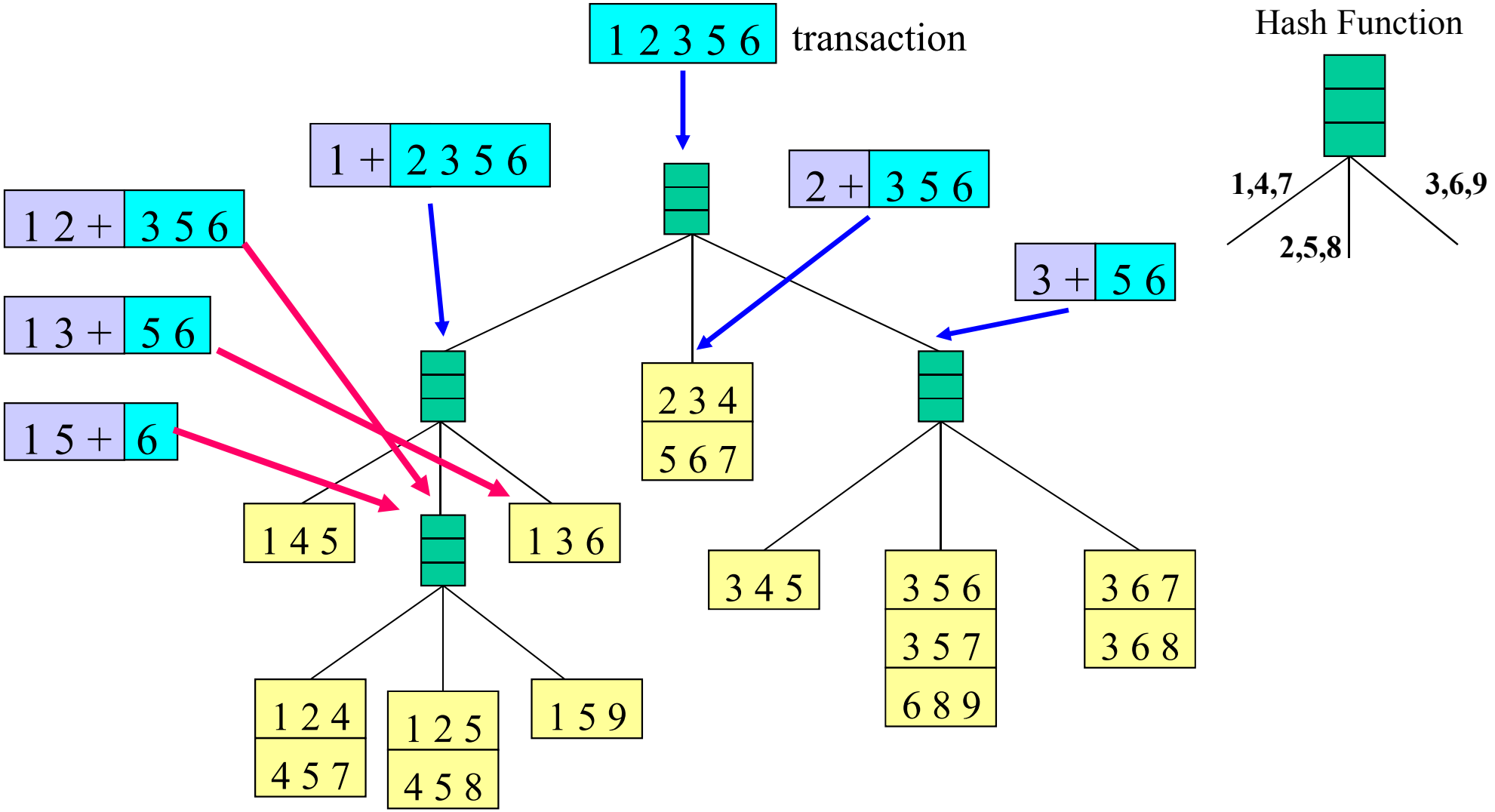
Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Subset Operation Using Hash Tree



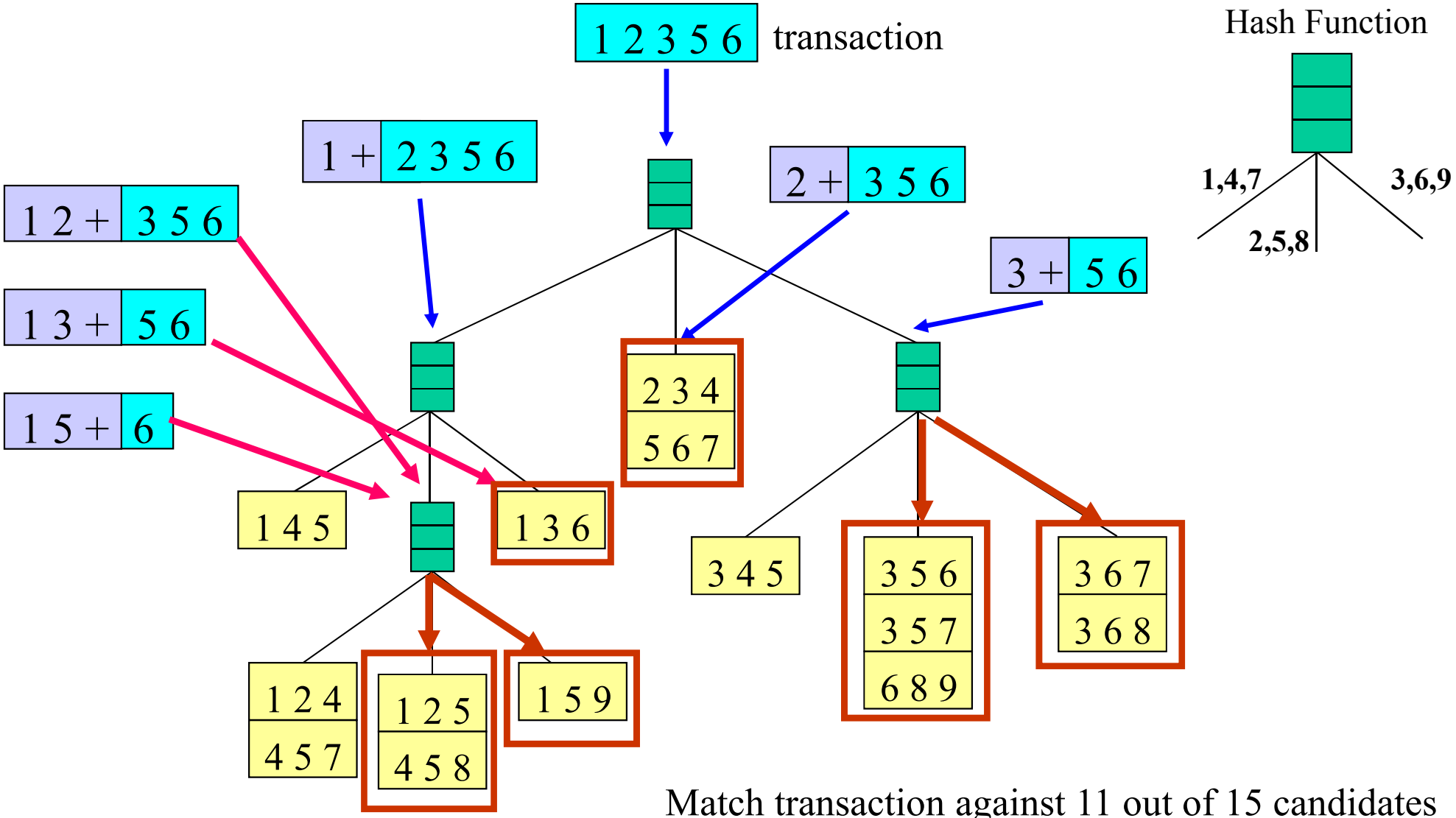
Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Subset Operation Using Hash Tree



Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

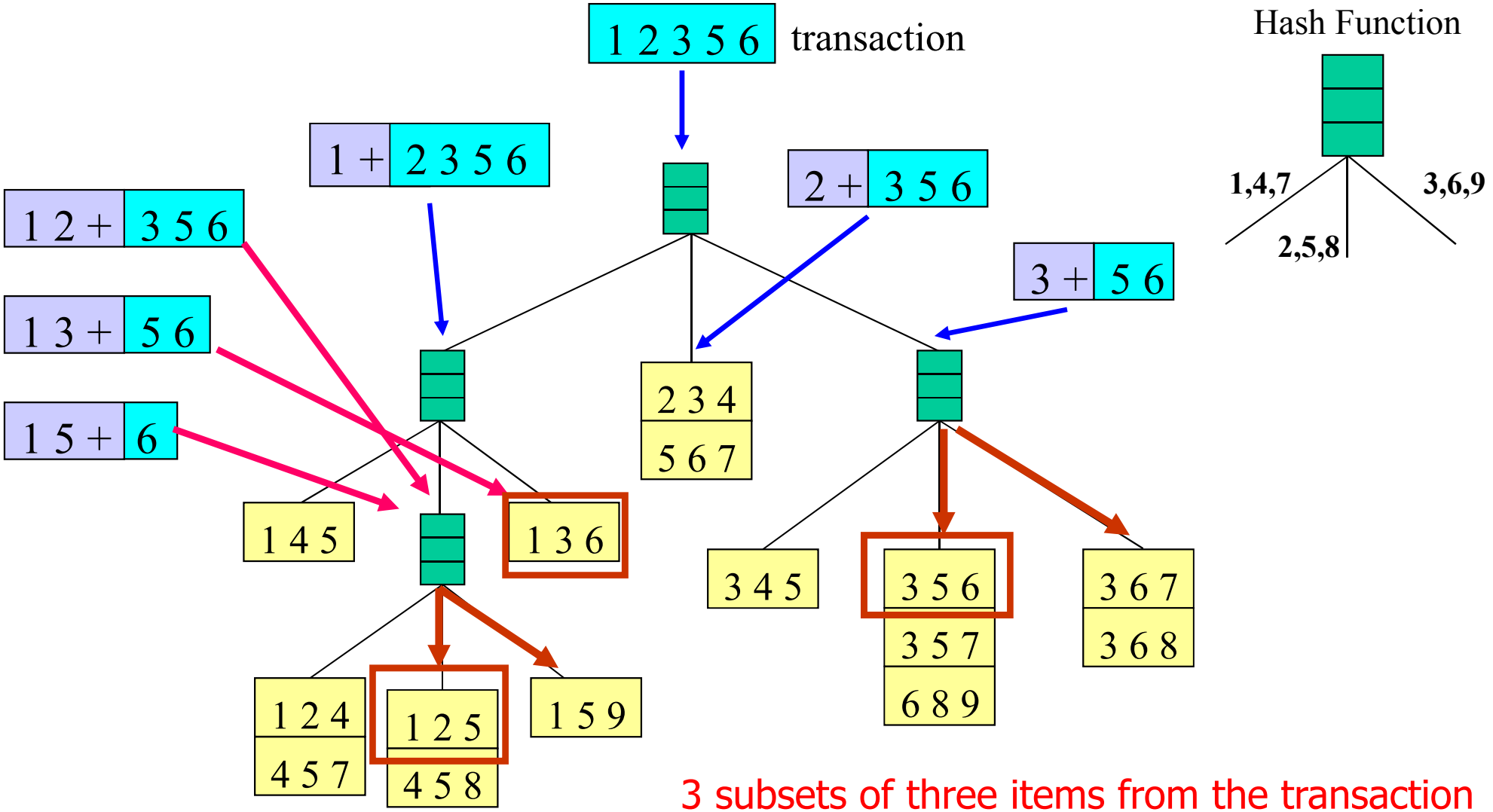
Subset Operation Using Hash Tree



Match transaction against 11 out of 15 candidates

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Subset Operation Using Hash Tree



3 subsets of three items from the transaction are among the candidate itemsets; used to increase support counting

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Factors Affecting Complexity

- 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 each item
  - if number of frequent items also increases, both computation and I/O costs may also increase
- Size of database
  - since Apriori makes multiple passes, run time of algorithm may increase with number of transactions
- Average transaction width
  - transaction width increases with denser data sets
  - This may increase max length of frequent itemsets and traversals of hash tree (number of subsets in a transaction increases with its width)

Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Rule Generation

- Given a frequent itemset L, find all non-empty subsets $f \subset L$ such that $f \rightarrow L - f$ satisfies the minimum confidence requirement
  - If {A,B,C,D} is a frequent itemset, candidate rules:

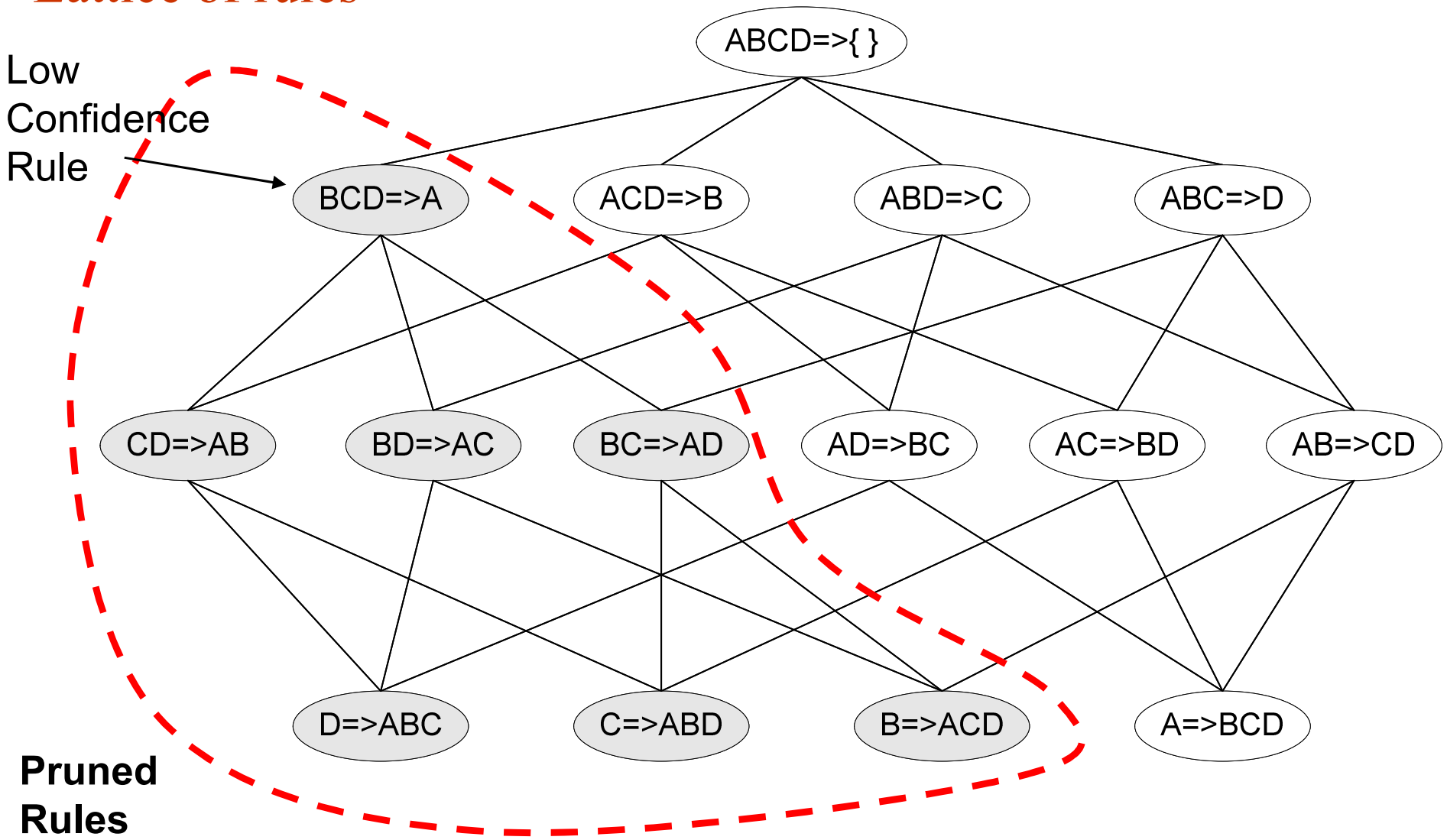
ABC $\rightarrow$D,	ABD $\rightarrow$C,	ACD $\rightarrow$B,	BCD $\rightarrow$A,
A $\rightarrow$BCD,	B $\rightarrow$ACD,	C $\rightarrow$ABD,	D $\rightarrow$ABC
AB $\rightarrow$CD,	AC $\rightarrow$ BD,	AD $\rightarrow$ BC,	BC $\rightarrow$AD,
BD $\rightarrow$AC,	CD $\rightarrow$AB,		

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

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Rule Generation for Apriori Algorithm

Lattice of rules



Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

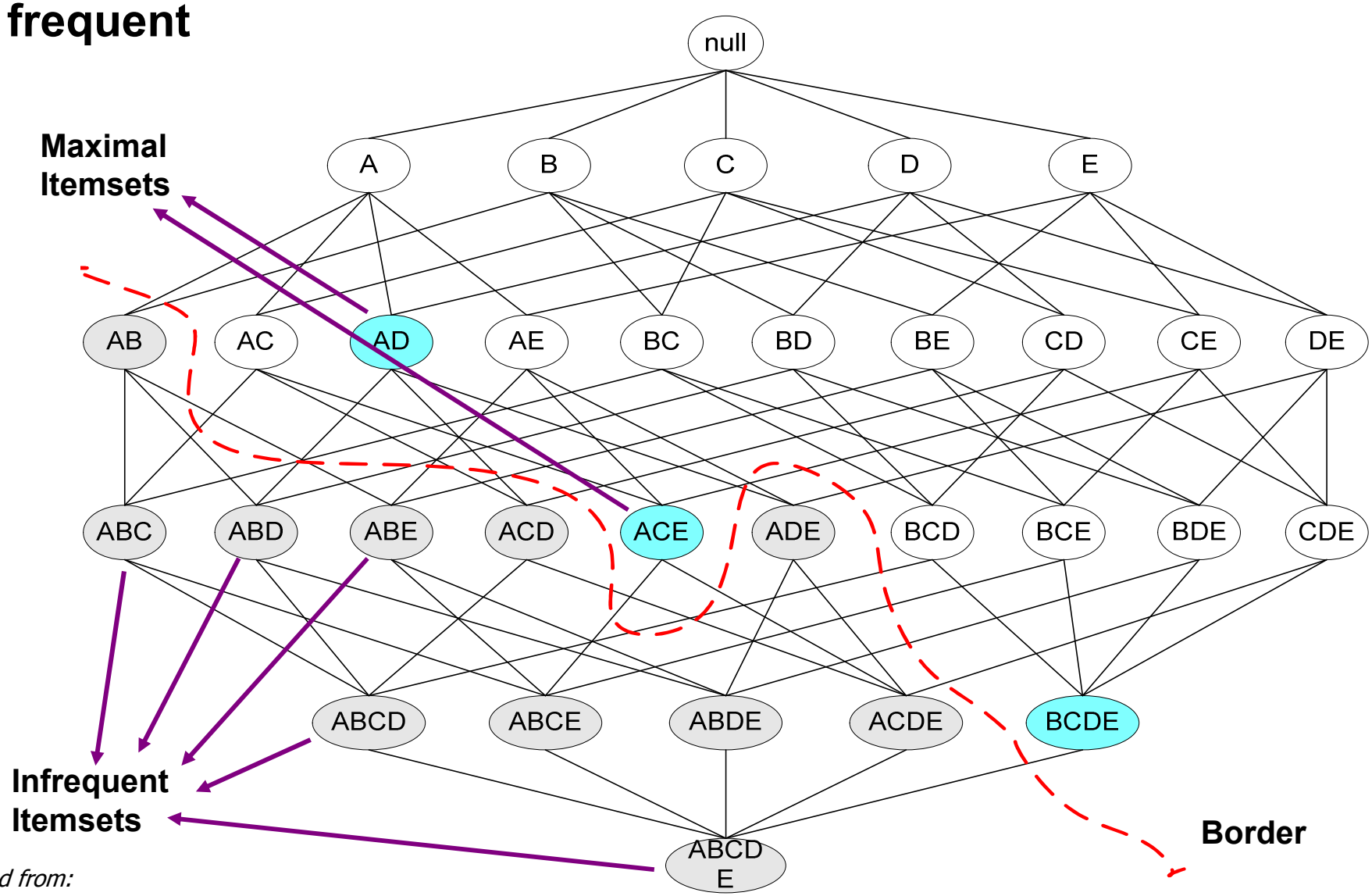
Compact Representation of Frequent Itemsets

- Some itemsets are redundant because they have identical support as their supersets

- Need a compact representation

Maximal Frequent Itemset

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



Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Closed Itemset

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

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

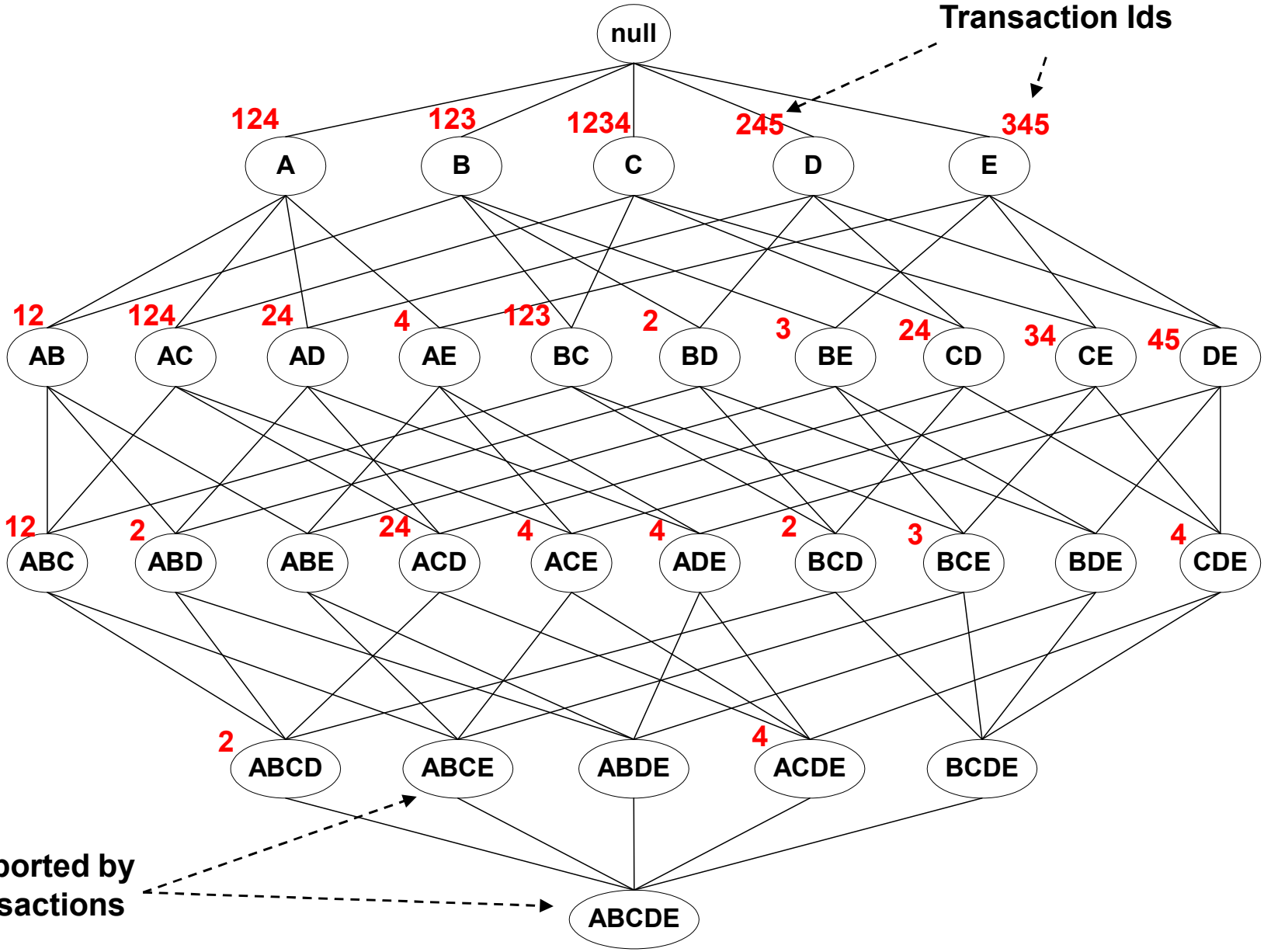
Itemset	Support
{A}	4
{B}	5
{C}	3
{D}	4
{A,B}	4
{A,C}	2
{A,D}	3
{B,C}	3
{B,D}	4
{C,D}	3

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

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Maximal vs Closed Itemsets

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

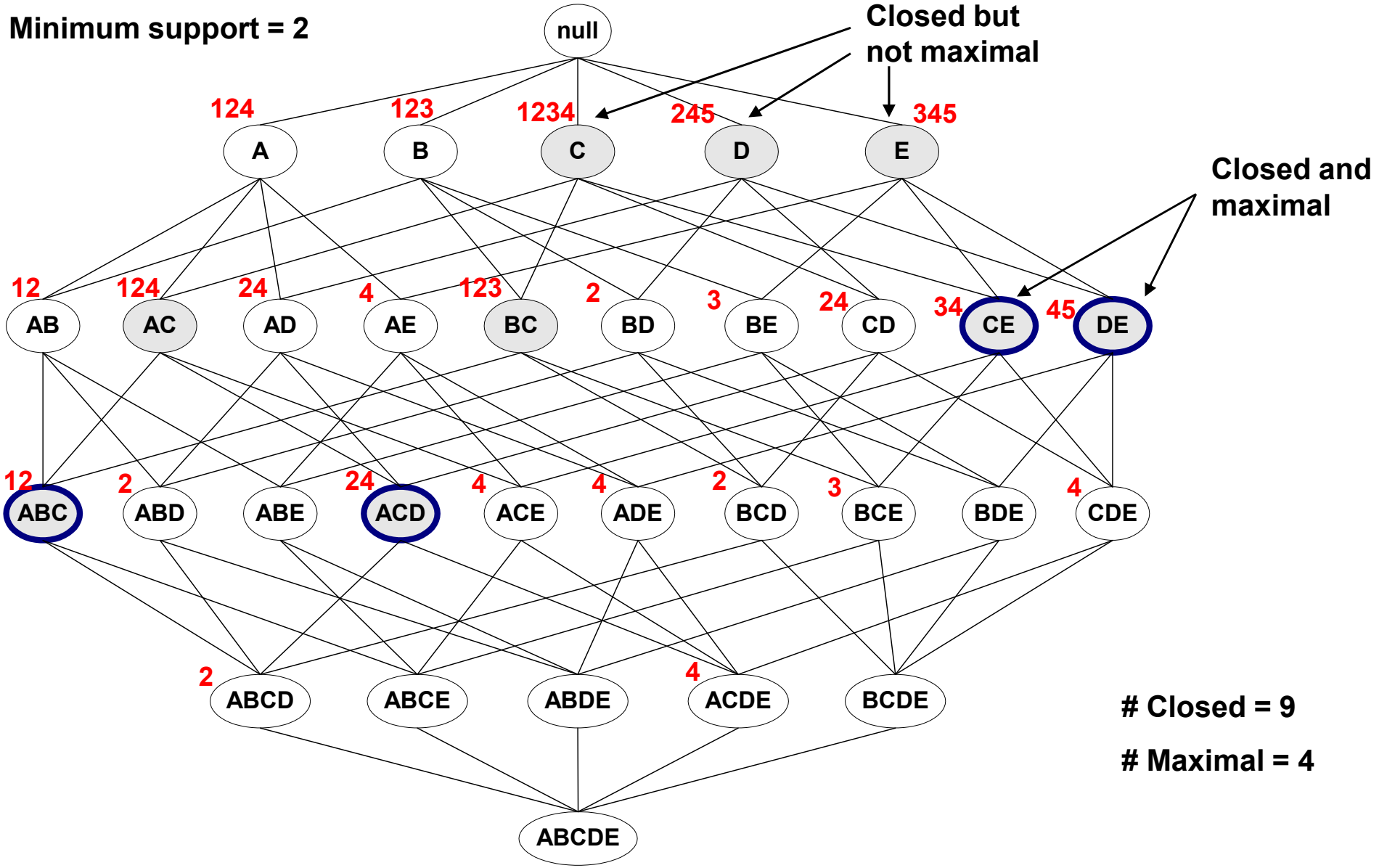


Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

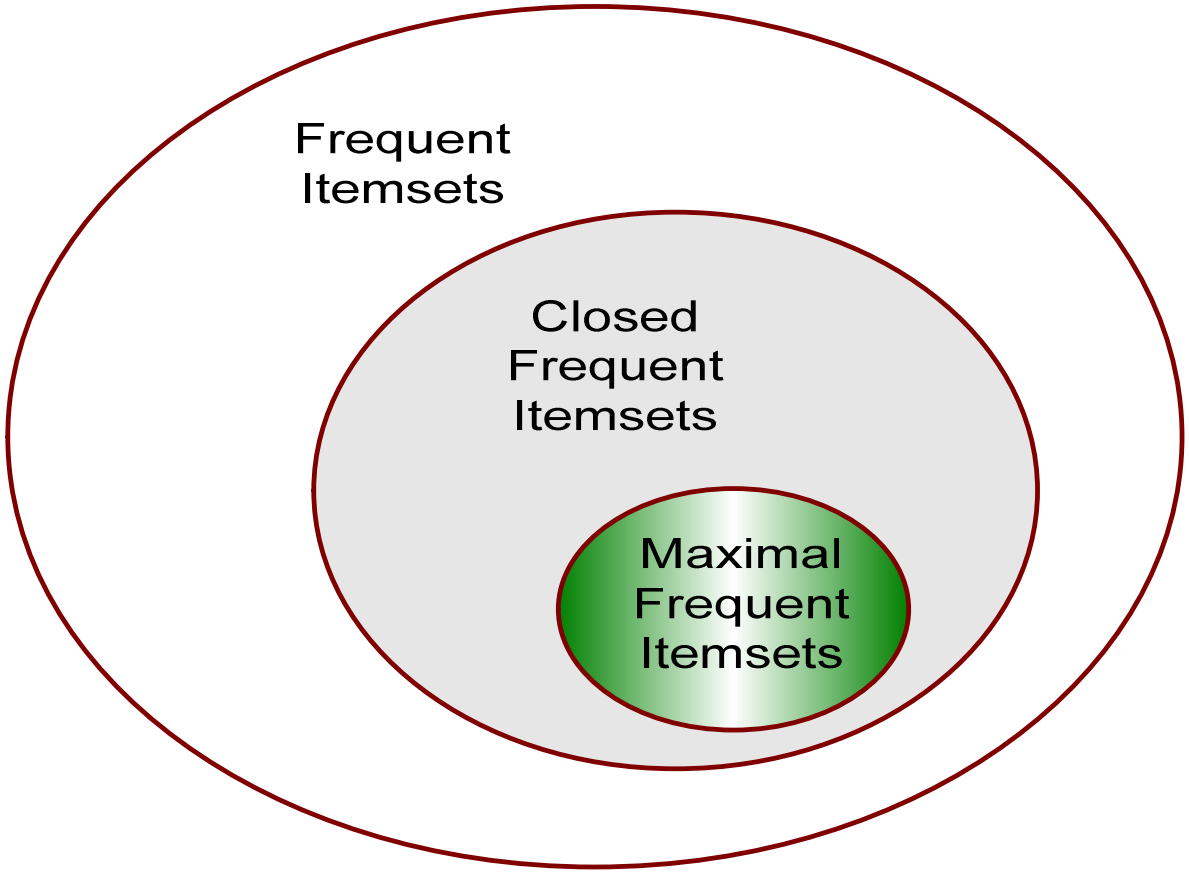
Han, Kamber - Data Mining: Concepts and Techniques

Maximal vs Closed Frequent Itemsets



Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Maximal vs Closed Itemsets



Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques