Data Mining Techniques

CS 6220 - Section 3 - Fall 2016

Lecture 15: Association Rules

Jan-Willem van de Meent
(credit: Yijun Zhao, Tan et al.,
Leskovec et al.)

Association Rule Discovery

Market-basket model:

- Goal: Identify items that are bought together by sufficiently many customers
- Approach: Process the sales data to find dependencies among items
- A classic rule:
  - If someone buys diaper and milk, then he/she is likely to buy beer
  - Don't be surprised if you find six-packs next to diapers!

The Market-Basket Model

- A large set of items
  - e.g., things sold in a supermarket
- A large set of baskets
- Each basket is a small subset of items
  - e.g., the things one customer buys on one day
- Want to discover association rules
  - People who bought {x,y,z} tend to buy {v,w}

Input:

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

Output:

Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

Applications – (1)

- Items = products; Baskets = sets of products someone bought in one trip to the store
- Real market baskets: Chain stores keep TBs of data about what customers buy together
  - Tells how typical customers navigate stores, lets them position tempting items
  - Suggests tie-in "tricks", e.g., run sale on diapers + raise the price of beer
  - Need the rule to occur frequently, or no $$'s
- Amazon's people who bought *X* also bought *Y*

Applications – (2)

- Baskets = sentences; Items = documents containing those sentences
  - Items that appear together too often could represent plagiarism
  - Notice items do not have to be “in” baskets
- Baskets = patients; Items = drugs & side-effects
  - Has been used to detect combinations of drugs that result in particular side-effects
  - But requires extension: Absence of an item needs to be observed as well as presence

More generally

- A general many-to-many mapping (association) between two kinds of things
  - But we ask about connections among “items”, not “baskets”

- For example:
  - Finding communities in graphs (e.g., Twitter)

Frequent Itemsets

- Simplest question: Find sets of items that appear together "frequently" in baskets
- *Support* for itemset I: Number of baskets containing all items in I
  - (Often expressed as a fraction of the total number of baskets)
- Given a *support threshold* s, then sets of items that appear in at least s baskets are called *frequent itemsets*

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

Support of {Beer, Bread} = 2

Example: Frequent Itemsets

- Items = {milk, coke, pepsi, beer, juice}
- Support threshold = 3 baskets

$B_1 = \{m, c, b\}$ $B_2 = \{m, p, j\}$

$B_3 = \{m, b\}$ $B_4 = \{c, j\}$

$B_5 = \{m, c, b\}$ $B_6 = \{m, c, b, j\}$

$B_7 = \{c, b, j\}$ $B_8 = \{b, c\}$

- Frequent itemsets:
  {m}:5, {c}:6, {b}:6, {j}:4, {m,b}:4,
  {m,c}: 3, {c,b}:5, {c,j}:3, {m,c,b}:3

Association Rules

- If-then rules about the contents of baskets
- $\{i_1, i_2, \ldots, i_k\} \rightarrow j$ means: "if a basket contains all of $i_1, \ldots, i_k$ then it is *likely* to contain j"
- In practice there are many rules, want to find significant/interesting ones!
- *Confidence* of this association rule is the probability of j given $I = \{i_1, \ldots, i_k\}$

$$\text{conf}(I \rightarrow j) = \frac{\text{support}(I \cup j)}{\text{support}(I)}$$

Interesting Association Rules

- Not all high-confidence rules are interesting
  - The rule $X \rightarrow milk$ may have high confidence because milk is just purchased very often (independent of X)
- Interest of an association rule $I \rightarrow j$: difference between its confidence and the fraction of baskets that contain j

$$\text{Interest}(I \rightarrow j) = \text{conf}(I \rightarrow j) - \Pr[j]$$

  - Interesting rules are those with high positive or negative interest values (usually above 0.5)

Example: Confidence and Interest

$B_1 = \{m, c, b\}$ $B_2 = \{m, p, j\}$

$B_3 = \{m, b\}$ $B_4 = \{c, j\}$

$B_5 = \{m, c, b\}$ $B_6 = \{m, c, b, j\}$

$B_7 = \{c, b, j\}$ $B_8 = \{b, c\}$

- Association rule: $\{m\} \rightarrow b$
  - Confidence = 4/5
  - Interest = 4/5 – 6/8 = 1/20
    - Item *b* appears in 6/8 of the baskets
    - Rule is not very interesting!

Many measures of interest

Measure (Symbol)	Definition
Goodman-Kruskal (λ)	$\left(\sum_j \max_k f_{jk} - max_k f_{+k}\right)/\left(N - \max_k f_{+k}\right)$
Mutual Information (M)	$\left(\sum_i \sum_j \frac{f_{ij}}{N} \log \frac{N f_{ij}}{f_{i+} f_{+j}}\right)/\left(-\sum_i \frac{f_{i+}}{N} \log \frac{f_{i+}}{N}\right)$
J-Measure (J)	$\frac{f_{11}}{N} \log \frac{N f_{11}}{f_{1+} f_{+1}} + \frac{f_{10}}{N} \log \frac{N f_{10}}{f_{1+} f_{+0}}$
Gini index (G)	$\frac{f_{1+}}{N} \times (\frac{f_{11}}{f_{1+}})^2 + (\frac{f_{10}}{f_{1+}})^2] - (\frac{f_{+1}}{N})^2$
	$+ \frac{f_{0+}}{N} \times [(\frac{f_{01}}{f_{0+}})^2 + (\frac{f_{00}}{f_{0+}})^2] - (\frac{f_{+0}}{N})^2$
Laplace (L)	$\left(f_{11} + 1\right)/\left(f_{1+} + 2\right)$
Conviction (V)	$\left(f_{1+} f_{+0}\right)/\left(N f_{10}\right)$
Certainty factor (F)	$\left(\frac{f_{11}}{f_{1+}} - \frac{f_{+1}}{N}\right)/\left(1 - \frac{f_{+1}}{N}\right)$
Added Value (AV)	$\frac{f_{11}}{f_{1+}} - \frac{f_{+1}}{N}$

source: Tan, Steinbach & Kumar, "Introduction to Data Mining",
http://www-users.cs.umn.edu/~kumar/dmbook/ch6.pdf

Finding Association Rules

- Problem: Find all association rules with support $\geq s$ and confidence $\geq c$
  - Note: Support of an association rule is the support of the set of items on the left side
- Hard part: Finding the frequent itemsets!
  - If $\{i_1, i_2, \ldots, i_k\} \rightarrow j$ has high support and confidence, then both $\{i_1, i_2, \ldots, i_k\}$ and $\{i_1, i_2, \ldots, i_k, j\}$ will be "frequent"

$$\text{conf}(I \rightarrow j) = \frac{\text{support}(I \cup j)}{\text{support}(I)}$$

Mining Association Rules

- Step 1: Find all frequent itemsets $\boldsymbol{I}$
  - (we will explain this next)
- Step 2: Rule generation
  - For every subset $\boldsymbol{A}$ of $\boldsymbol{I}$, generate a rule $\boldsymbol{A \to I \setminus A}$
    - Since $\boldsymbol{I}$ is frequent, $\boldsymbol{A}$ is also frequent
    - Variant 1: Single pass to compute the rule confidence
      - confidence(***A,B→C,D***) = support(**A,B,C,D**) / support(**A,B**)
    - Variant 2:
      - Observation: If A,B,C→D is below confidence, so is A,B→C,D
      - Can generate "bigger" rules from smaller ones!
  - Output the rules above the confidence threshold

Example: Mining Association Rules

$B_1 = \{m, c, b\}$ $B_2 = \{m, p, j\}$

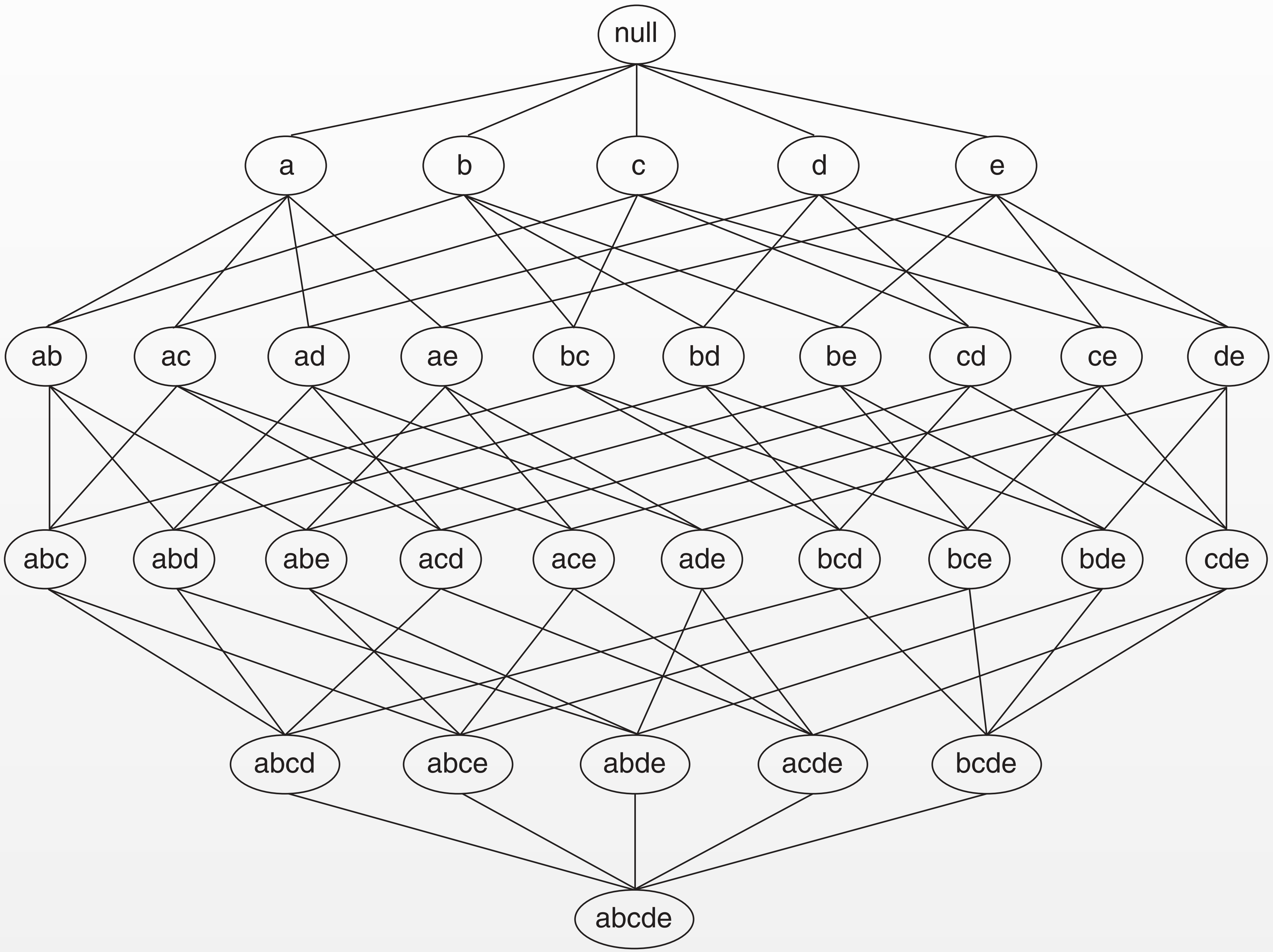
$B_3 = \{m, b\}$ $B_4 = \{c, j\}$

$B_5 = \{m, c, b\}$ $B_6 = \{m, c, b, j\}$

$B_7 = \{c, b, j\}$ $B_8 = \{b, c\}$

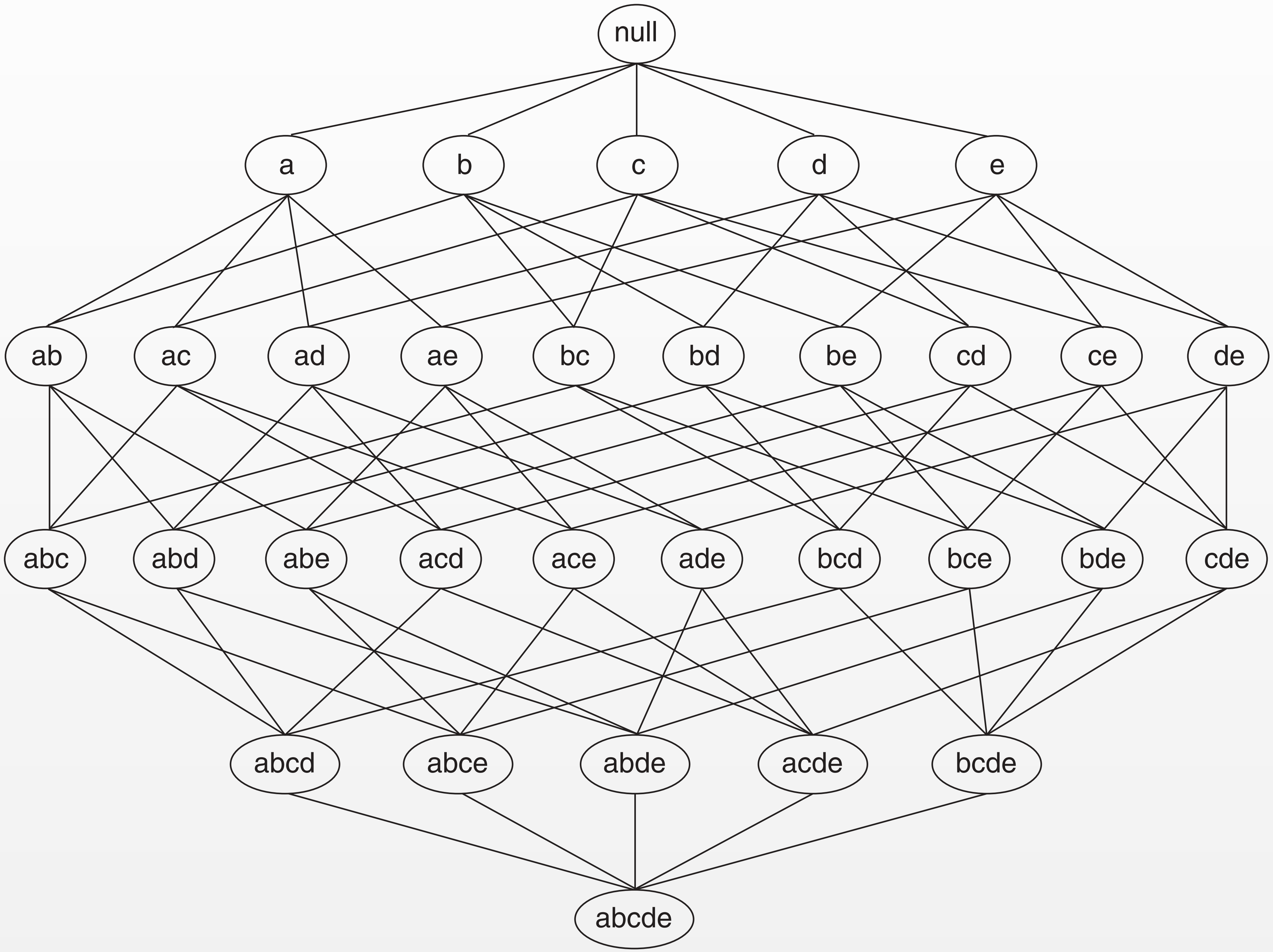
- Support threshold $s = 3$, confidence $c = 0.75$
- 1) Frequent itemsets:
  - {b,m}:4 {c,m}:3 {b,c}:5 {c,j}:3 {m,c,b}: 3
- 2) Generate rules:
  - $\cancel{b \rightarrow m: c=4/6}$ $b \rightarrow c$: $c=5/6$ $\cancel{b,c \rightarrow m: c=3/5}$
  - $m \rightarrow b$: $c=4/5$ … $b,m \rightarrow c$: $c=3/4$
  - $\cancel{b \rightarrow c,m: c=3/6}$

Finding Frequent Item Sets



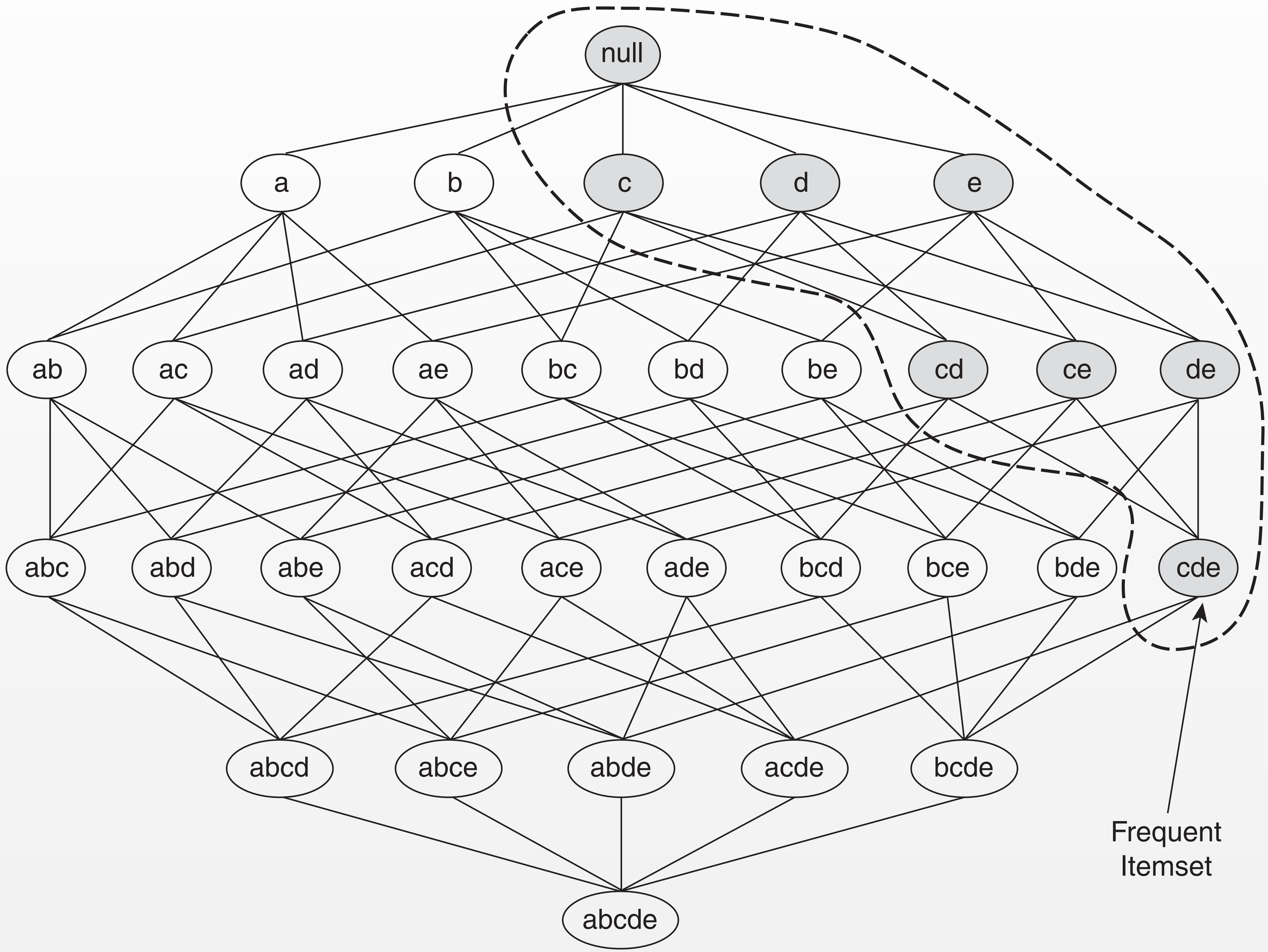
Given *k* products, how many possible item sets are there?

Finding Frequent Item Sets



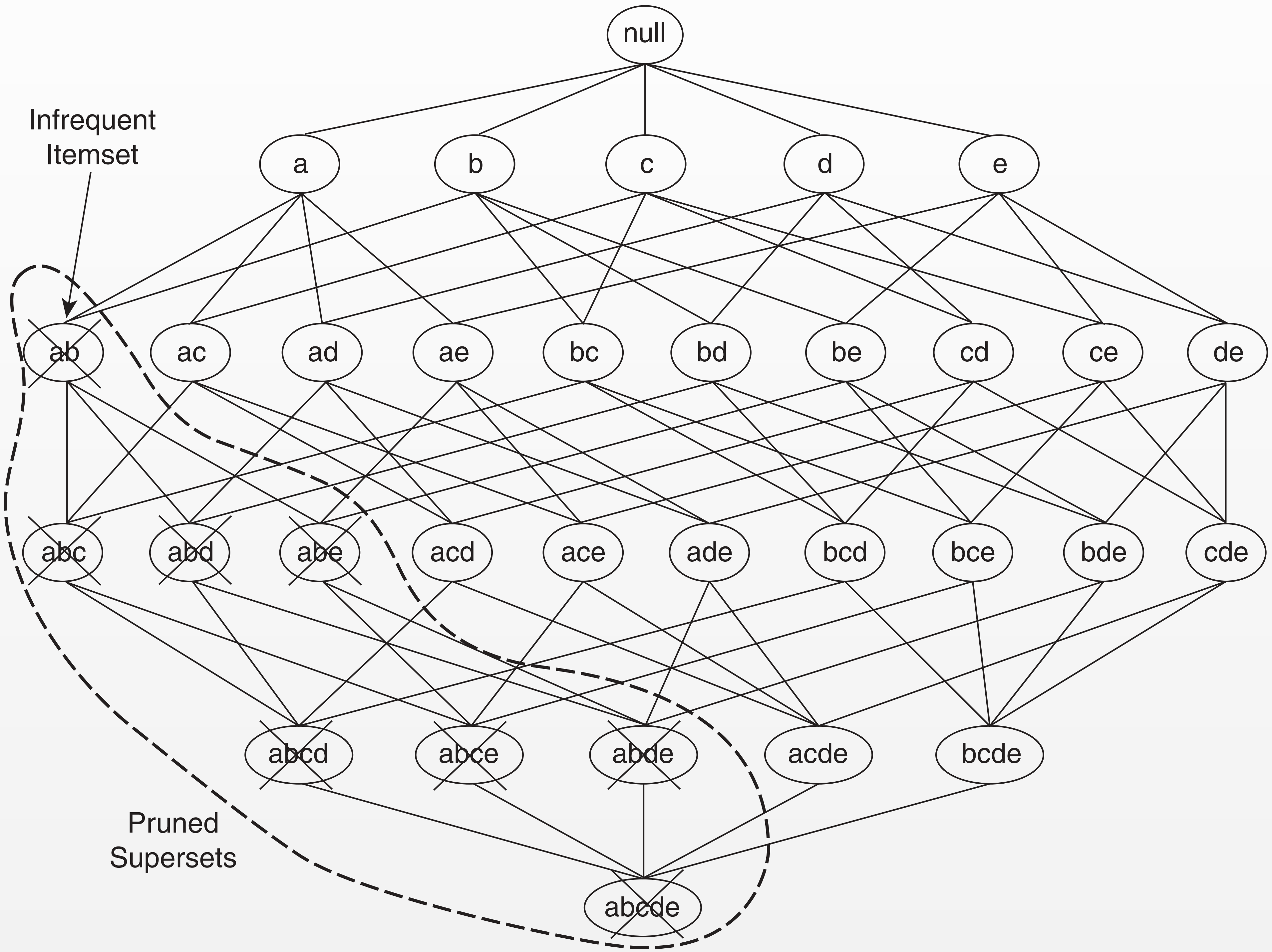
Answer: $2^k - 1$ -> Cannot enumerate all possible sets

Observation: A-priori Principle



Subsets of a frequent item set are also frequent

Corollary: Pruning of Candidates



If we know that a subset is not frequent, then we can ignore all its supersets

A-priori Algorithm

Algorithm 6.1 Frequent itemset generation of the *Apriori* algorithm.

$k = 1.$
$F_k = \{\ i \mid i \in I \wedge \sigma(\{i\}) \geq N \times minsup\}.$ {Find all frequent 1-itemsets}
repeat
    $k = k + 1.$
    $C_k =$ apriori-gen(F_{k-1}). {Generate candidate itemsets}
    for each transaction $t \in T$ **do**
        $C_t =$ subset(C_k, t). {Identify all candidates that belong to t}
        for each candidate itemset $c \in C_t$ **do**
            $\sigma(c) = \sigma(c) + 1.$ {Increment support count}
        end for
    end for
    $F_k = \{\ c \mid c \in C_k \wedge \sigma(c) \geq N \times minsup\}.$ {Extract the frequent k-itemsets}
until $F_k = \emptyset$
Result $= \bigcup F_k.$

Generating Candidates C_k

1. *Self-joining:* Find pairs of sets in L_{k-1} that differ by **one** element
2. *Pruning:* Remove all candidates with infrequent subsets

Example: Generating Candidates C_k

$B_1 = \{m, c, b\}$ $B_2 = \{m, p, j\}$

$B_3 = \{m, b\}$ $B_4 = \{c, j\}$

$B_5 = \{m, c, b\}$ $B_6 = \{m, c, b, j\}$

$B_7 = \{c, b, j\}$ $B_8 = \{b, c\}$

- *Frequent itemsets of size 2:*
  {m,b}:4, {m,c}:3, {c,b}:5, {c,j}:3
- *Self-joining:*
  {m,b,c}, {b,c,j}
- *Pruning:*
  ~~{b,c,j}~~ since ~~{b,j}~~ not frequent

Compacting the Output

- To reduce the number of rules we can post-process them and only output:
  - Maximal frequent itemsets:
    No immediate superset is frequent
    - Gives more pruning
  - Closed itemsets:
    No immediate superset has same count (> 0)
    - Stores not only frequent information, but exact counts

Example: Maximal vs Closed

B_1 = {m, c, b} B_2 = {m, p, j}

B_3 = {m, b} B_4= {c, j}

B_5 = {m, c, b} B_6 = {m, c, b, j}

B_7 = {c, b, j} B_8 = {b, c}

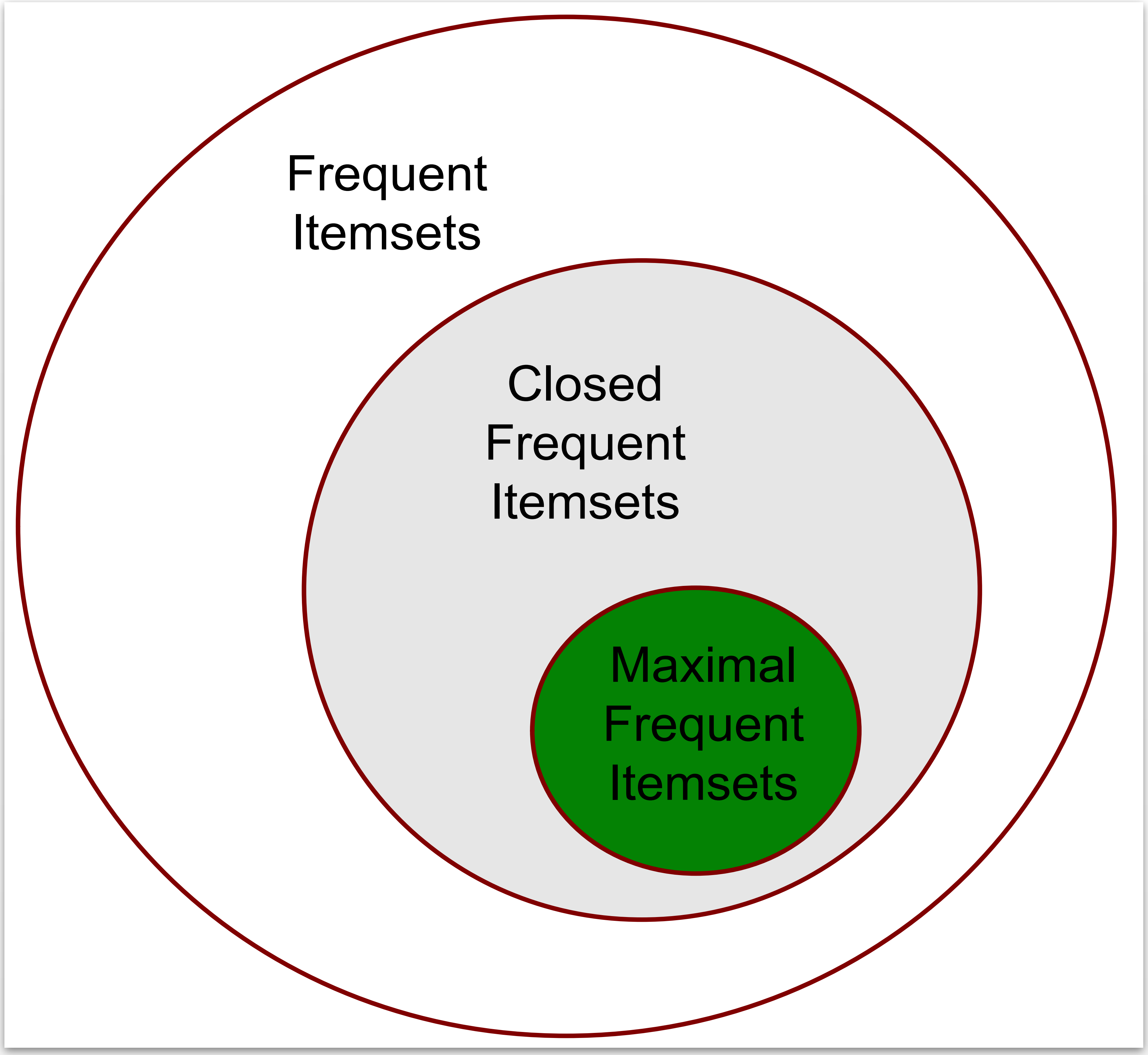
Frequent itemsets:

{m}:5, **{c}:6**, **{b}:6**, **{j}:4**,
{m,c}:3, **{m,b}:4**, **{c,b}:5**, **{c,j}:3**,
{m,c,b}:3

Closed

Maximal

Example: Maximal vs Closed



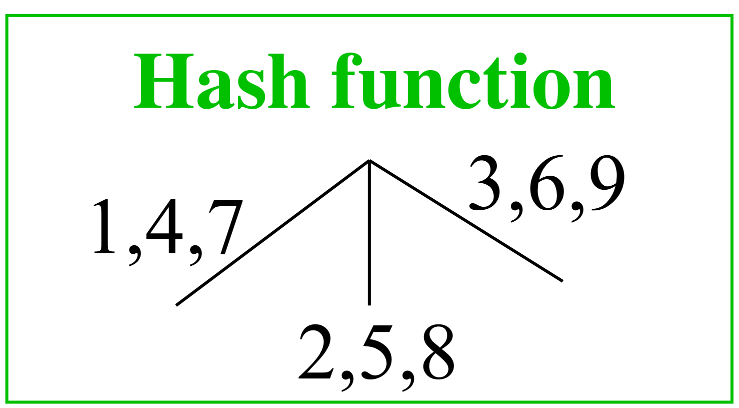
Hash Tree for Itemsets

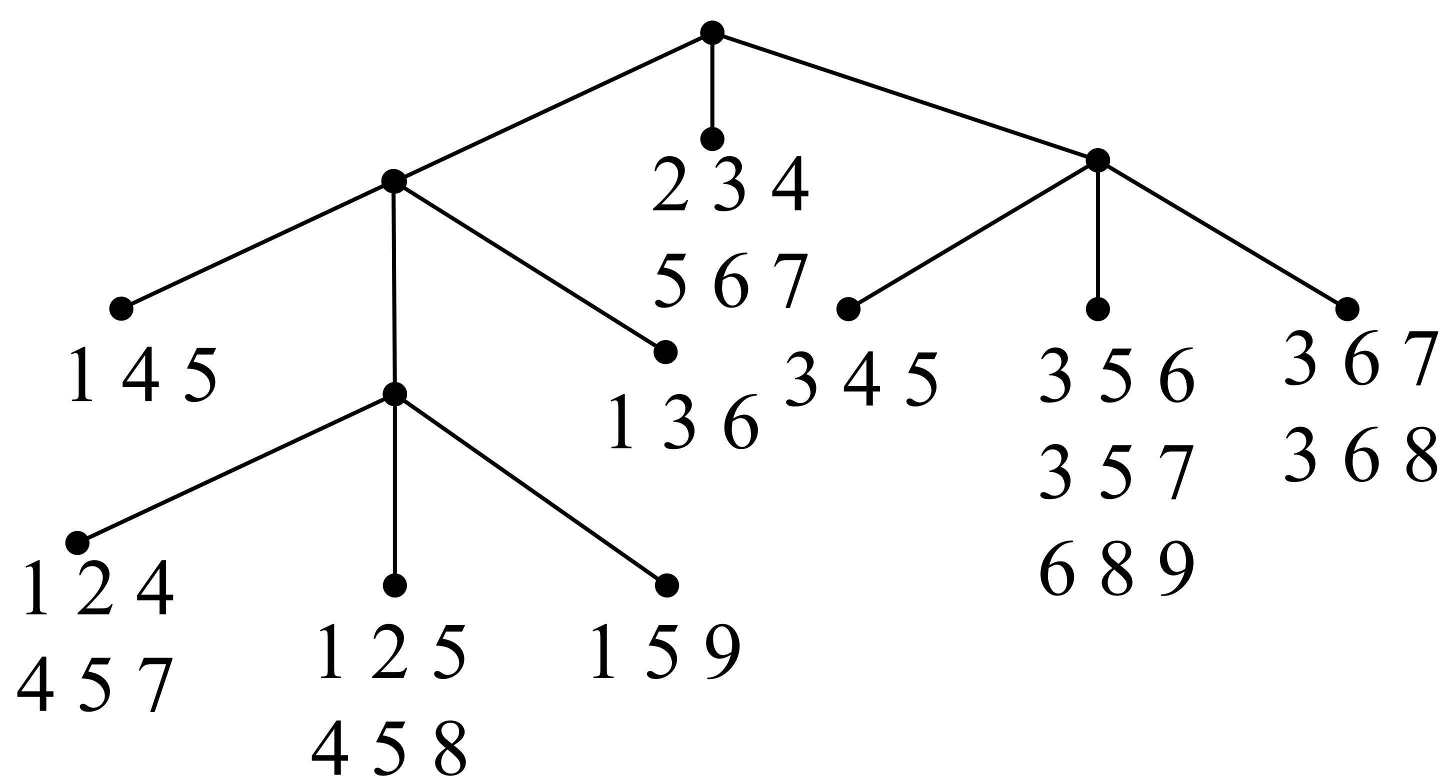
Suppose you have 15 candidate itemsets of length 3:

{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}

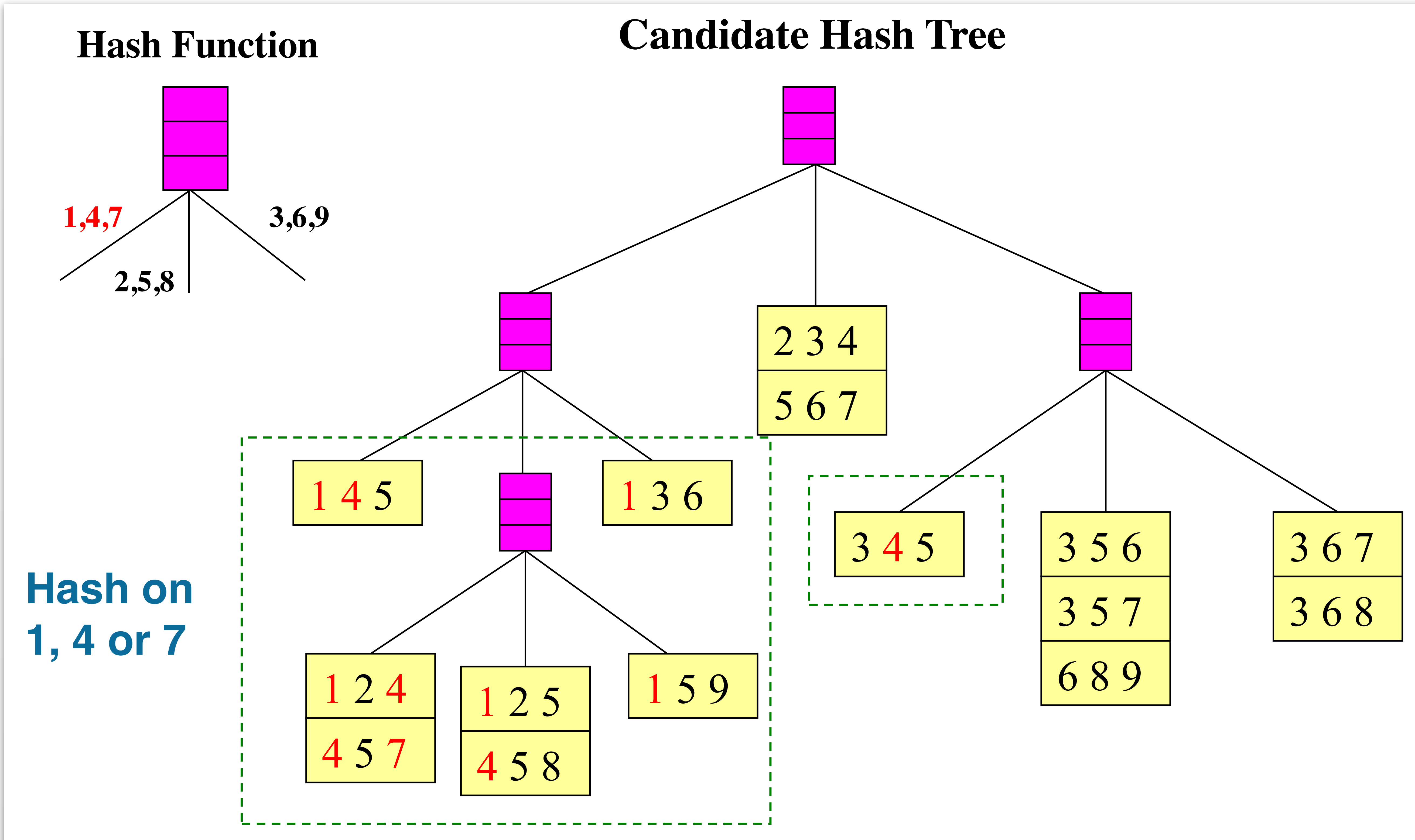
You need:

- **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)**

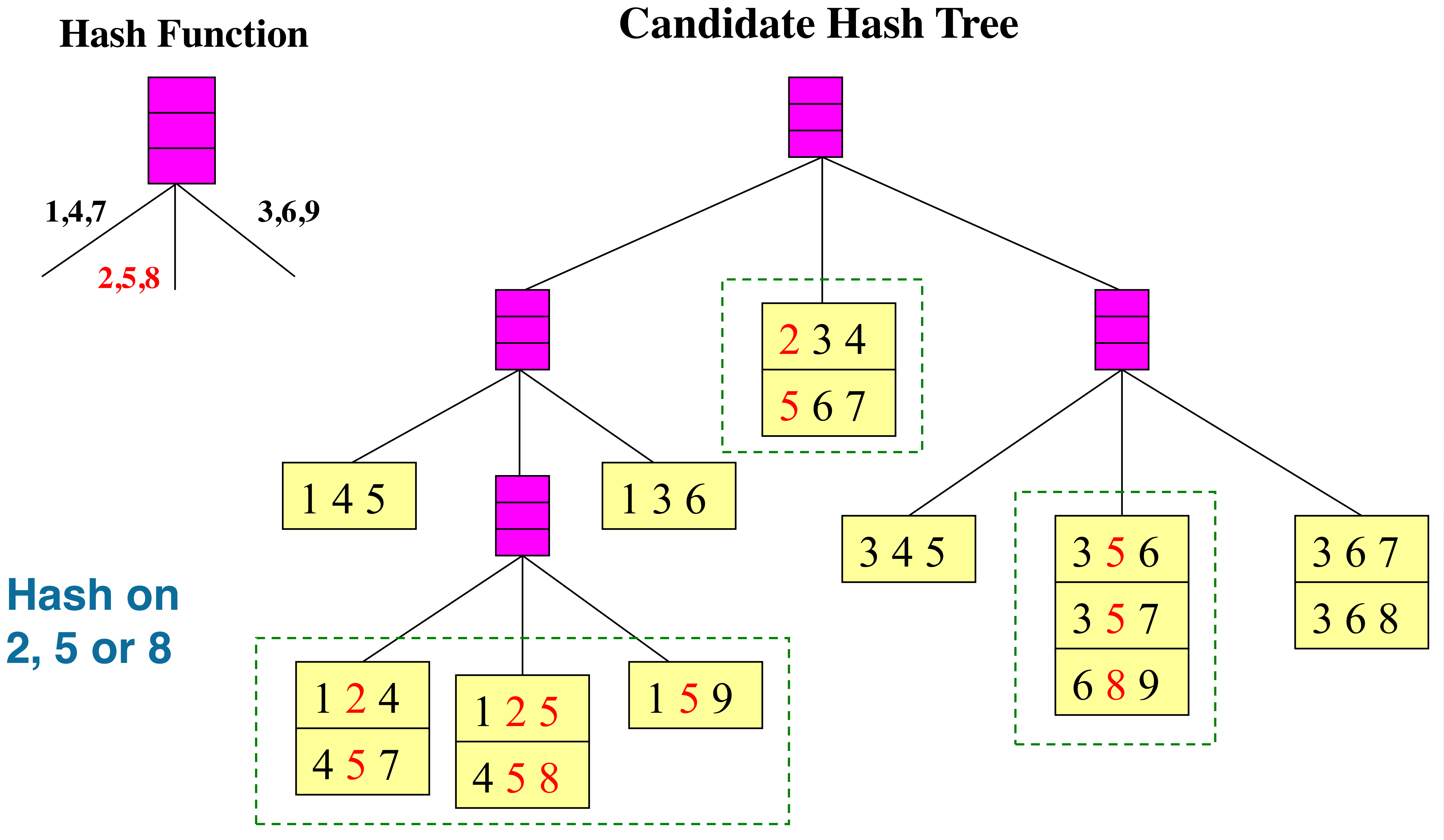




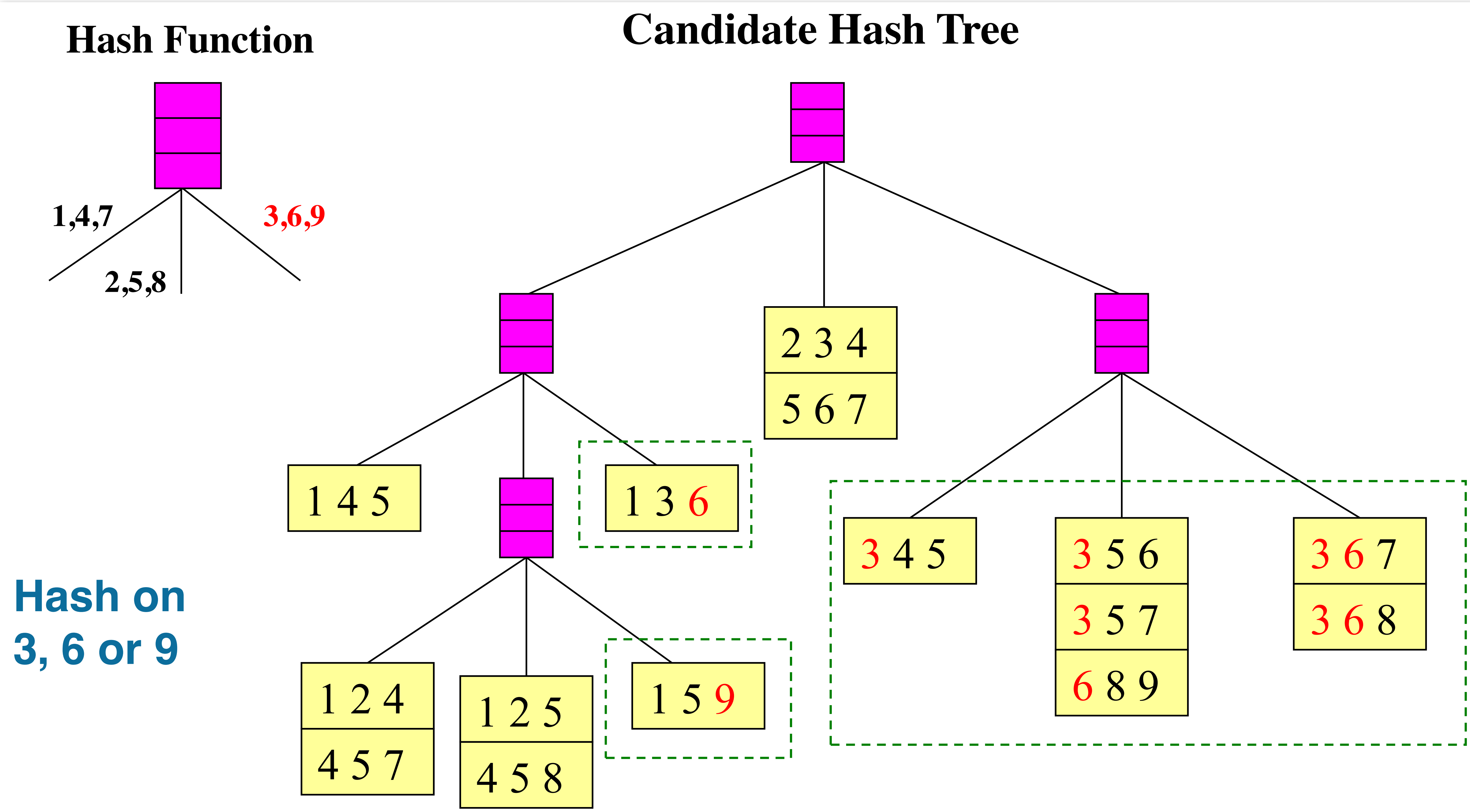
Hash Tree for Itemsets



Hash Tree for Itemsets

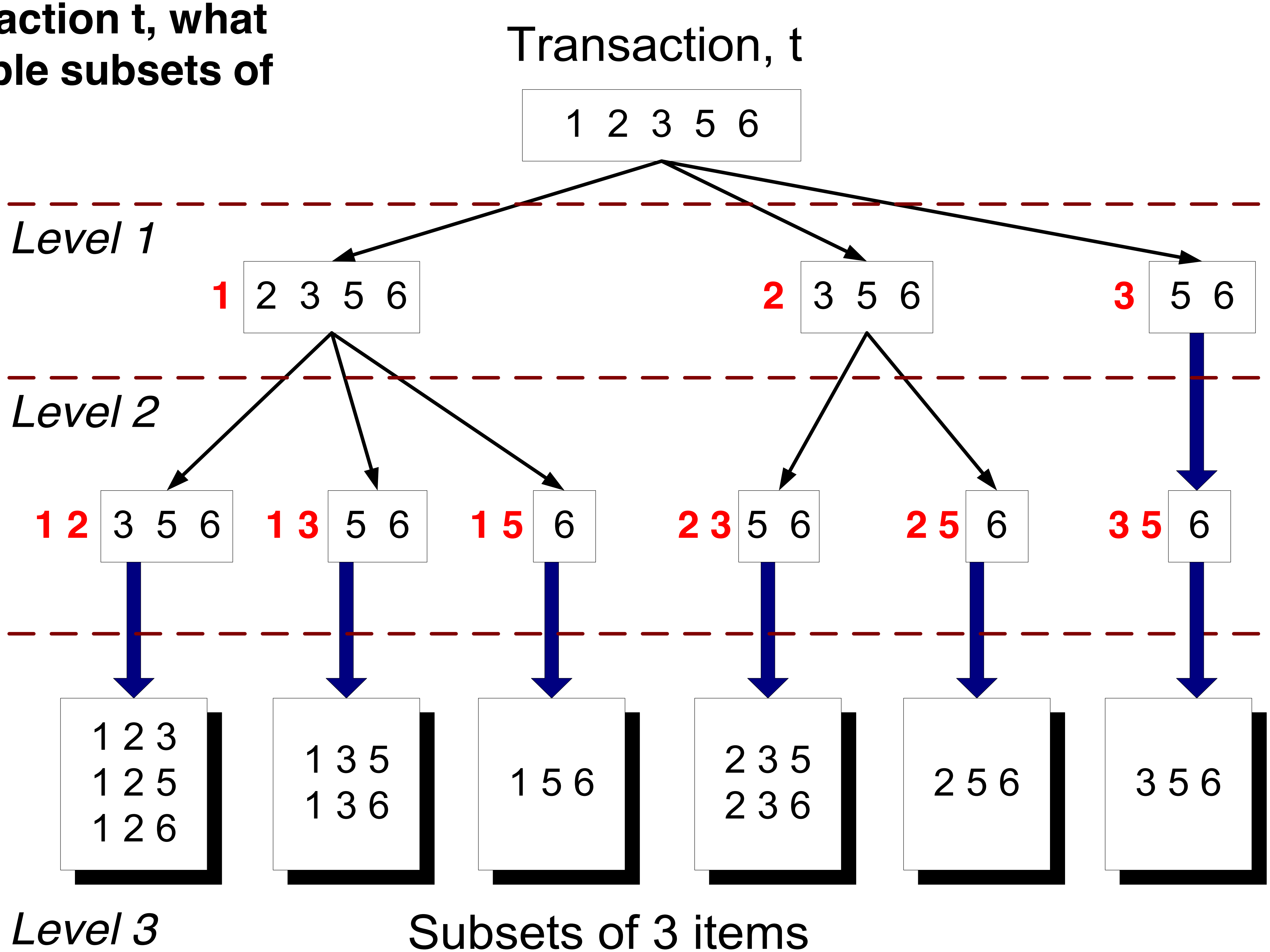


Hash Tree for Itemsets

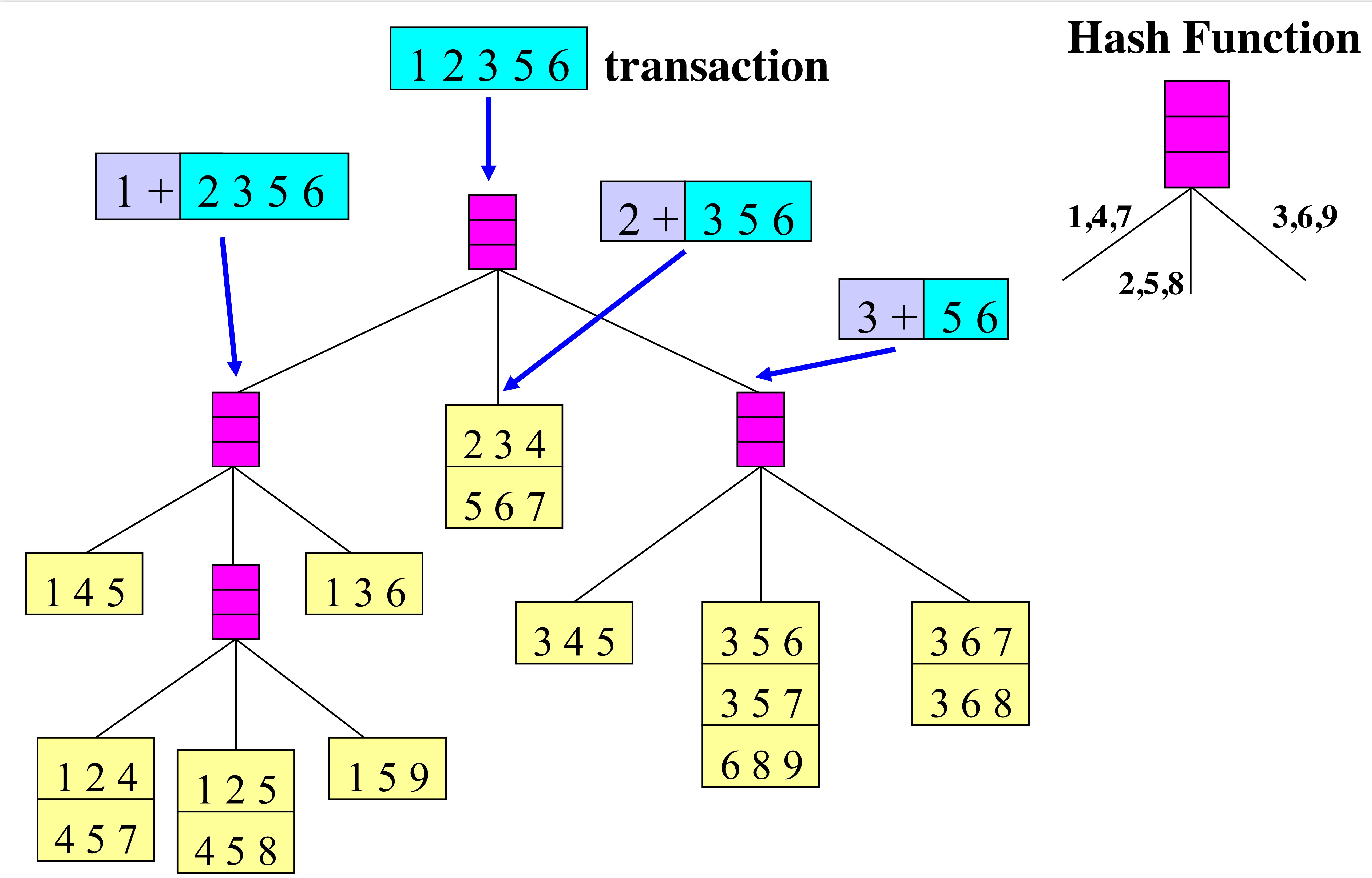


Subset Matching

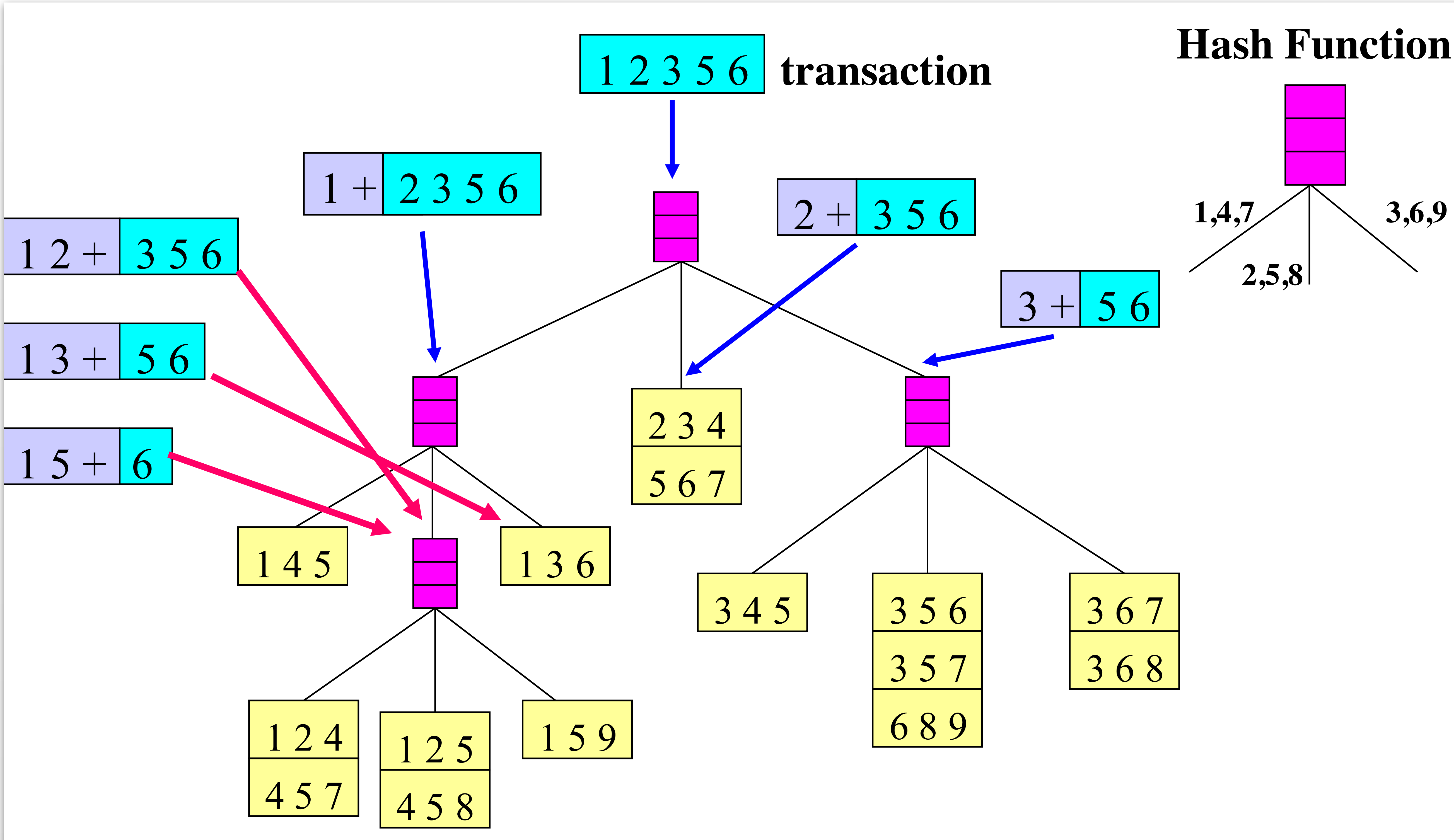
Given a transaction t, what are the possible subsets of size 3?



Subset Operation



Subset Operation



Subset Operation