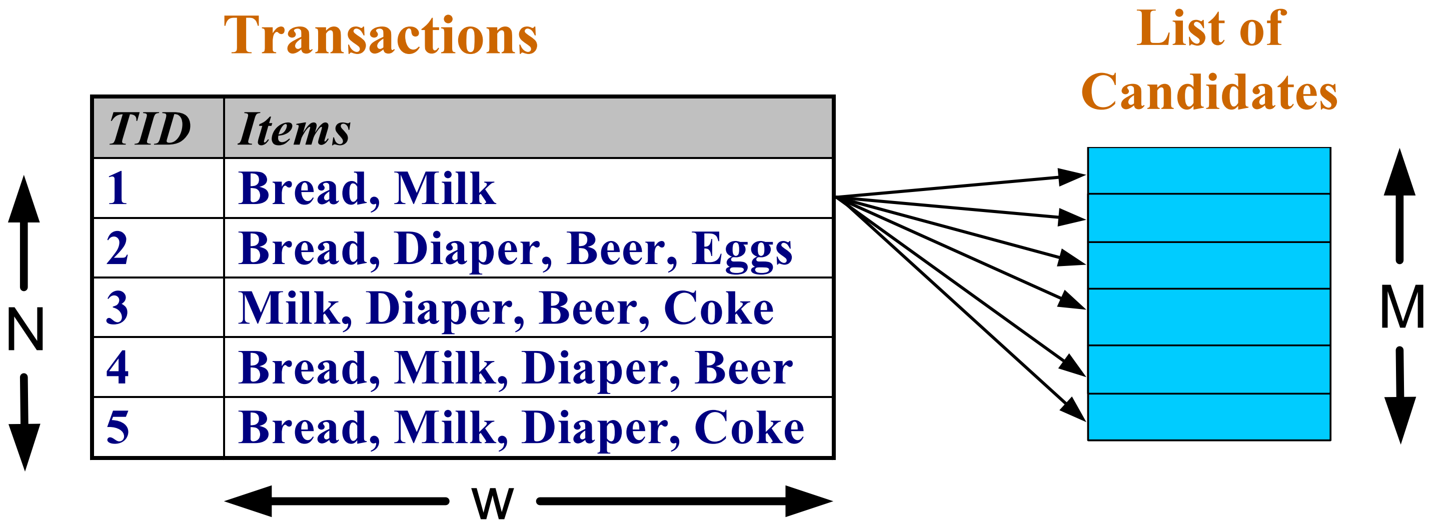
ASSOCIATION RULE MINING

BEIYU LIN

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



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

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

FREQUENT ITEMSET GENERATION STRATEGIES

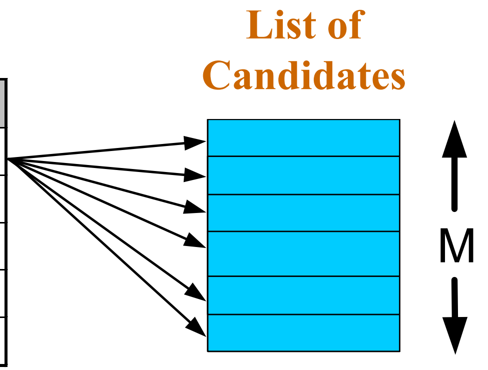
- Reduce the number of candidates (M)
  - Complete search: $M=2^d$
  - Use pruning techniques to reduce M
- Reduce the number of transactions (N)
  - Reduce size of N as the size of itemset increases
  - Used by DHP and vertical-based mining algorithms
- Reduce the number of comparisons (NM)
  - Use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

Transactions

N

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

w



Given a transaction {B, M, D, C}, find all possible subset with size 3 from this transaction.

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

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

Support

s({Milk, Bread, Diaper}) = 2/5 = # itemsets / total # of transaction

Support count: # of the itemsets that show in the transaction

Confidence

$\{A\}$ $\{AB\}$

X Y

$X \subseteq Y$

$s(X) \geq s(Y)$

$s(X) = \frac{\sigma(X)}{\#} = \frac{\sigma(\{A\})}{\#}$

$s(Y) = \frac{\sigma(Y)}{\#} = \frac{\sigma(\{AB\})}{\#}$

$\underbrace{\{A\}}_{X} \rightarrow \underbrace{\{BC\}}_{Y}$

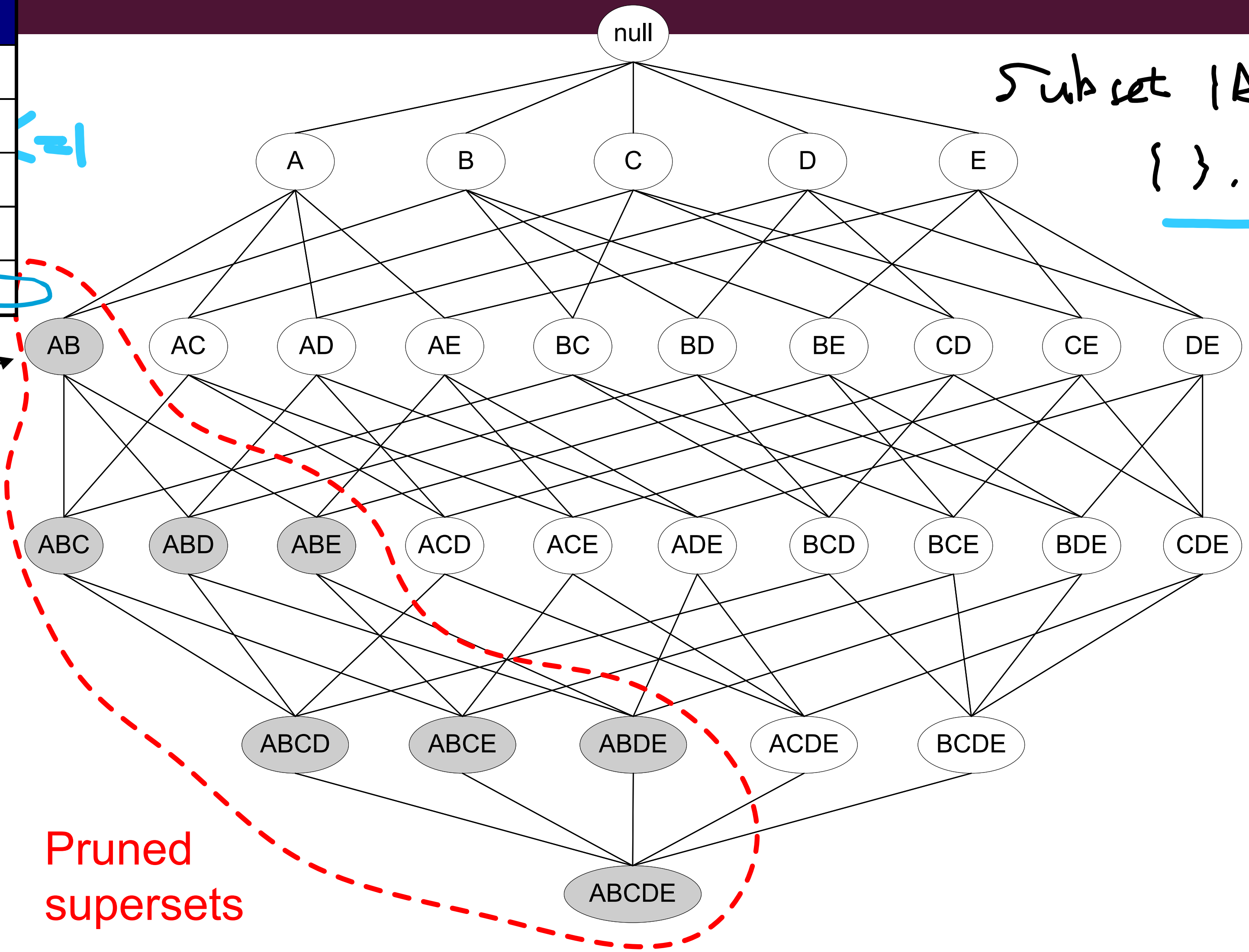
$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$

$P(ABC|A) = \frac{P(ABC)}{P(A)}$

ILLUSTRATING APRIORI PRINCIPLE

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**

Found to be Infrequent



Subset {ABC} 2^3

{ }, {A}, - - - -

supsets

{ABC}

{ABCD}

{ABCE}

SUPPORT COUNTING OF CANDIDATE ITEMSETS

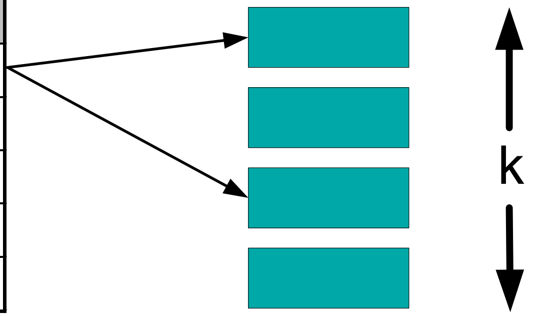
- To reduce number of comparisons, store the candidate itemsets in a hash structure / hash function
  - Instead of matching each transaction against every candidate, match it against candidates contained in the hashed buckets

Transactions

N

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

Hash Structure



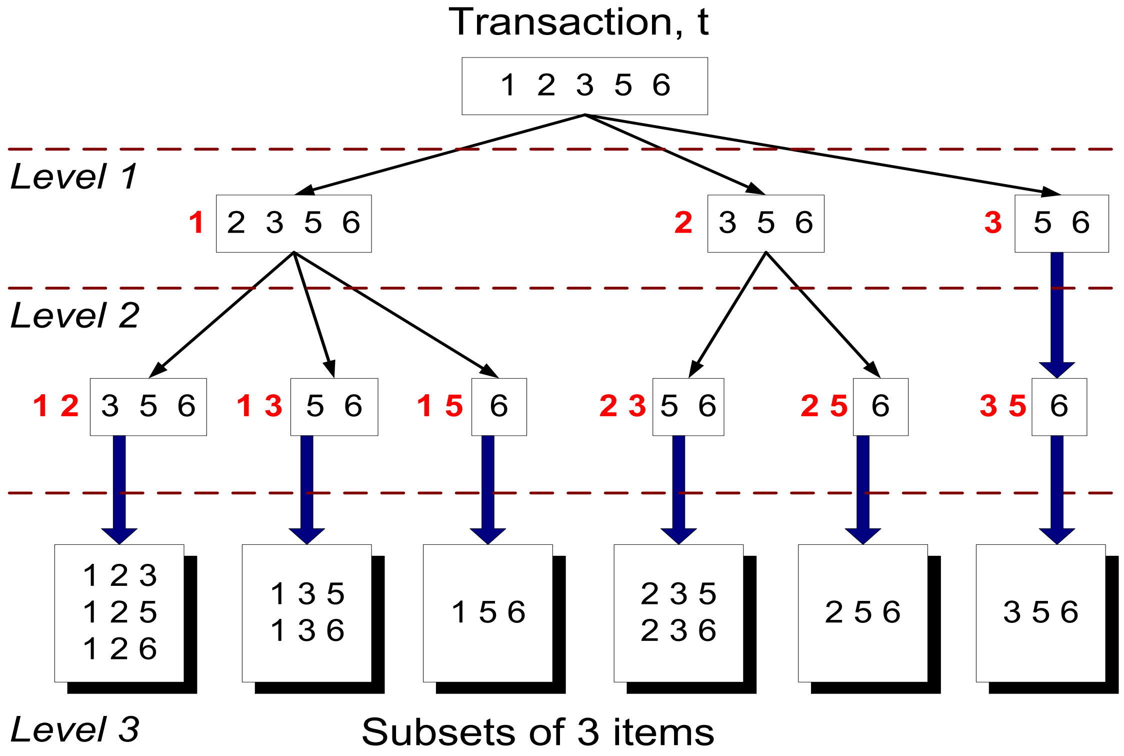
Buckets

SUPPORT COUNTING: AN EXAMPLE

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}

How many of these itemsets are supported by transaction (1,2,3,5,6)?



SUPPORT COUNTING: AN EXAMPLE

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}

How to find all the subsets with k items from an itemset?

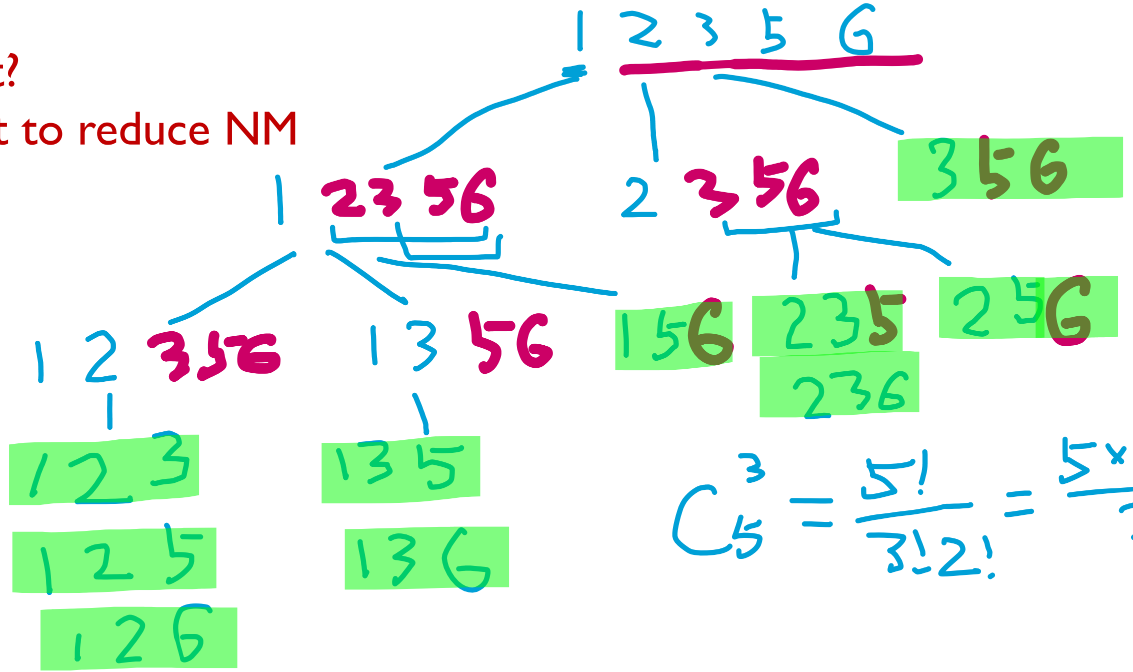
Given an itemset {1,2,3,5,6} ⇔ one transaction, we want to reduce NM

How many of these itemsets are supported by transaction (1,2,3,5,6)?

① Given Transaction (1, 2, 3, 5, 6)

We need list all the possible itemsets with k = 3 from this transaction.

②



$$C_5^3 = \frac{5!}{3!2!} = \frac{5\times 4}{2} = 10$$