Frequent Pattern Mining

Based on: Introduction to Data Mining
by Tan, Steinbach, Kumar

Item sets
A New Type of Data

- Some notation:
  - All possible items: $I = \{i_1, \ldots, i_d\}$
  - Database: *T* is a bag of transactions $t_1, \ldots, t_N$
  - Transaction $t = (TID, X)$
    - transaction identifier: *TID*
    - item set $X \subseteq I$

ITEMSET DATABASE	
TID	*Items*
1	{ Bread, Milk }
2	{ Bread, Diapers, Beer, Eggs }
3	{ Milk, Diapers, Beer, Cola }
4	{ Bread, Milk, Diapers, Beer }
5	{ Bread, Milk, Diapers, Cola }

Definitions

ITEMSET DATABASE	
TID	*Items*
1	{ Bread, Milk }
2	{ Bread, Diapers, Beer, Eggs }
3	{ Milk, Diapers, Beer, Cola }
4	{ Bread, Milk, Diapers, Beer }
5	{ Bread, Milk, Diapers, Cola }

Definitions

- **Itemset**
  - A collection of one or more items
    - Example: {Milk, Bread, Diaper}
  - k-itemset
    - An itemset that contains k items

ITEMSET DATABASE	
TID	*Items*
1	{ Bread, Milk }
2	{ Bread, Diapers, Beer, Eggs }
3	{ Milk, Diapers, Beer, Cola }
4	{ Bread, Milk, Diapers, Beer }
5	{ Bread, Milk, Diapers, Cola }

Definitions

- **Itemset**
  - A collection of one or more items
    - Example: {Milk, Bread, Diaper}
  - k-itemset
    - An itemset that contains k items
- **Support count** $\sigma()$
  - Frequency of occurrence of an itemset
  - E.g. $\sigma(\{\text{Milk, Bread, Diaper}\}) = 2$

ITEMSET DATABASE	
TID	*Items*
1	{ Bread, Milk }
2	{ Bread, Diapers, Beer, Eggs }
3	{ Milk, Diapers, Beer, Cola }
4	{ Bread, Milk, Diapers, Beer }
5	{ Bread, Milk, Diapers, Cola }

Definitions

- **Itemset**
  - A collection of one or more items
    - Example: {Milk, Bread, Diaper}
  - k-itemset
    - An itemset that contains k items
- **Support count** $\sigma()$
  - Frequency of occurrence of an itemset
  - E.g. $\sigma(\{\text{Milk, Bread, Diaper}\}) = 2$
- **Support s()**
  - Fraction of transactions that contain an itemset
  - E.g. $s(\{\text{Milk, Bread, Diaper}\}) = 2/5$

ITEMSET DATABASE	
TID	*Items*
1	{ Bread, Milk }
2	{ Bread, Diapers, Beer, Eggs }
3	{ Milk, Diapers, Beer, Cola }
4	{ Bread, Milk, Diapers, Beer }
5	{ Bread, Milk, Diapers, Cola }

Definitions

ITEMSET DATABASE	
TID	*Items*
1	{ Bread, Milk }
2	{ Bread, Diapers, Beer, Eggs }
3	{ Milk, Diapers, Beer, Cola }
4	{ Bread, Milk, Diapers, Beer }
5	{ Bread, Milk, Diapers, Cola }

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

Association Rule Mining

- Given a set of transactions, find rules that predict the occurrence of an item based on occurrences of other items in the transaction

ITEMSET DATABASE	
TID	*Items*
1	{ Bread, Milk }
2	{ Bread, Diapers, Beer, Eggs }
3	{ Milk, Diapers, Beer, Cola }
4	{ Bread, Milk, Diapers, Beer }
5	{ Bread, Milk, Diapers, Cola }

Example of Association Rules

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

→ means co-occurrence

Definitions

ITEMSET DATABASE	
TID	*Items*
1	{ Bread, Milk }
2	{ Bread, Diapers, Beer, Eggs }
3	{ Milk, Diapers, Beer, Cola }
4	{ Bread, Milk, Diapers, Beer }
5	{ Bread, Milk, Diapers, Cola }

Definitions

- Association Rule
  - Expression of the form $X \rightarrow Y$

    (X and Y are itemsets)

ITEMSET DATABASE	
TID	*Items*
1	{ Bread, Milk }
2	{ Bread, Diapers, Beer, Eggs }
3	{ Milk, Diapers, Beer, Cola }
4	{ Bread, Milk, Diapers, Beer }
5	{ Bread, Milk, Diapers, Cola }

Definitions

- Association Rule
  - Expression of the form X → Y

    (X and Y are itemsets)
  - Example:

    {Milk, Diaper} → {Beer}

ITEMSET DATABASE	
TID	*Items*
1	{ Bread, Milk }
2	{ Bread, Diapers, Beer, Eggs }
3	{ Milk, Diapers, Beer, Cola }
4	{ Bread, Milk, Diapers, Beer }
5	{ Bread, Milk, Diapers, Cola }

- Rule Evaluation Measures
  - Support (s)
    - Fraction of transactions that contain both X and Y
  - Confidence (c)

Definitions

- Association Rule
  - Expression of the form X → Y

    (X and Y are itemsets)
  - Example:

    {Milk, Diaper} → {Beer}

ITEMSET DATABASE	
TID	Items
1	{ Bread, Milk }
2	{ Bread, Diapers, Beer, Eggs }
3	{ Milk, Diapers, Beer, Cola }
4	{ Bread, Milk, Diapers, Beer }
5	{ Bread, Milk, Diapers, Cola }

- Rule Evaluation Measures
  - 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

Definitions

- Association Rule
  - Expression of the form X → Y

    (X and Y are itemsets)
  - Example:

    {Milk, Diaper} → {Beer}

ITEMSET DATABASE	
TID	Items
1	{ Bread, Milk }
2	{ Bread, Diapers, Beer, Eggs }
3	{ Milk, Diapers, Beer, Cola }
4	{ Bread, Milk, Diapers, Beer }
5	{ Bread, Milk, Diapers, Cola }

- Rule Evaluation Measures
  - 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

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 Mining Task

ITEMSET DATABASE	
TID	*Items*
1	{ Bread, Milk }
2	{ Bread, Diapers, Beer, Eggs }
3	{ Milk, Diapers, Beer, Cola }
4	{ Bread, Milk, Diapers, Beer }
5	{ Bread, Milk, Diapers, Cola }

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)

Association Rule Mining Task

- Given set of transactions T, find all rules having
  - support ≥ *minsup* threshold
  - confidence ≥ *minconf* threshold

ITEMSET DATABASE	
TID	*Items*
1	{ Bread, Milk }
2	{ Bread, Diapers, Beer, Eggs }
3	{ Milk, Diapers, Beer, Cola }
4	{ Bread, Milk, Diapers, Beer }
5	{ Bread, Milk, Diapers, Cola }

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)

Mining Association Rules

- 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 (combinatorial explosion)
- Can we do better?

ITEMSET DATABASE	
TID	*Items*
1	{ Bread, Milk }
2	{ Bread, Diapers, Beer, Eggs }
3	{ Milk, Diapers, Beer, Cola }
4	{ Bread, Milk, Diapers, Beer }
5	{ Bread, Milk, Diapers, Cola }

{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)

Mining Association Rules

Observations:

- All 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

ITEMSET DATABASE	
TID	*Items*
1	{ Bread, Milk }
2	{ Bread, Diapers, Beer, Eggs }
3	{ Milk, Diapers, Beer, Cola }
4	{ Bread, Milk, Diapers, Beer }
5	{ Bread, Milk, Diapers, Cola }

{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)

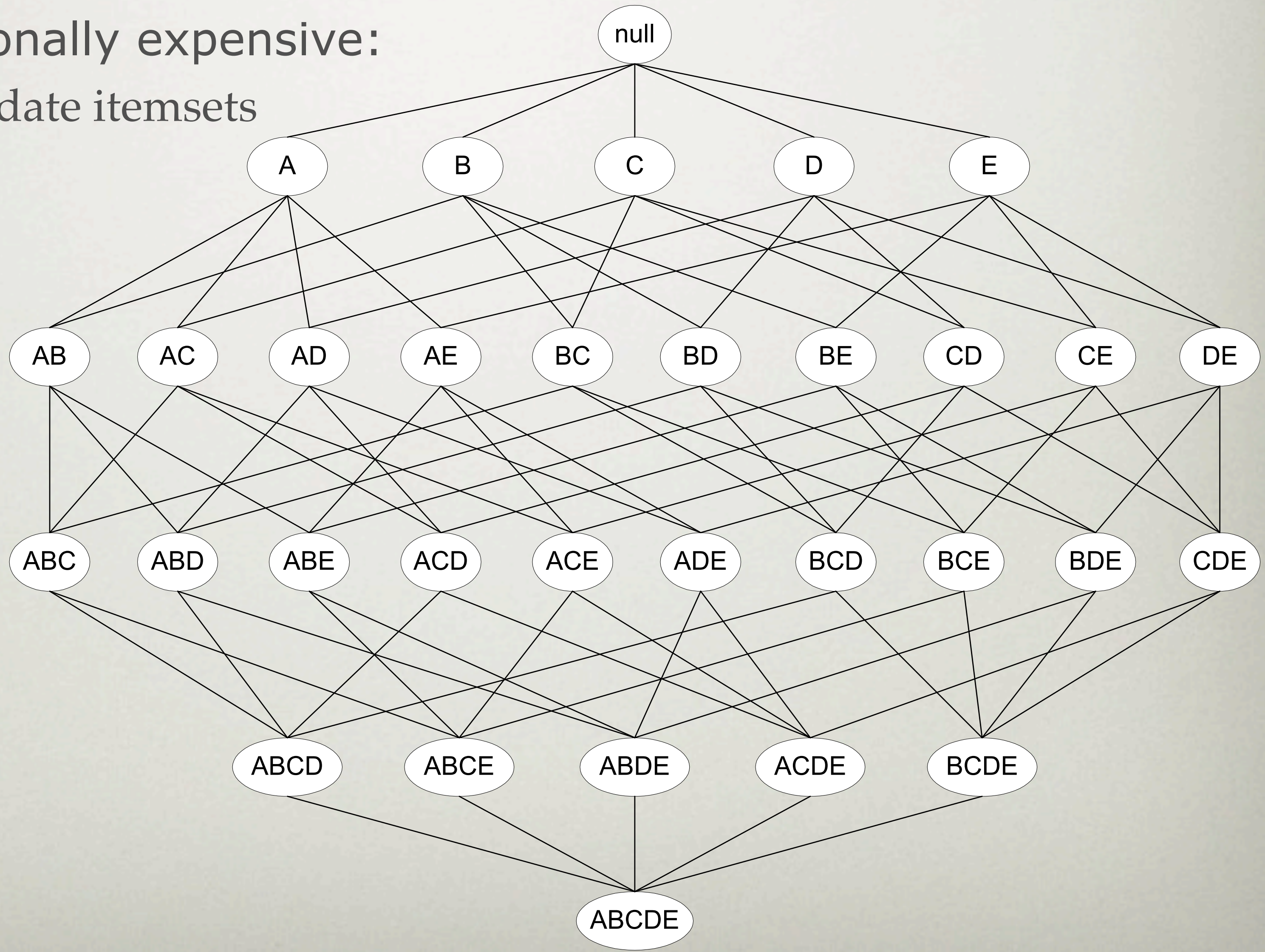
Mining Association Rules

- Two-step approach:
  - Frequent Itemset Generation (count support)
    - Generate *all* itemsets whose support ≥ *minsup*
  - Rule Generation (count confidence)
    - Generate high confidence rules from each frequent itemset, where each rule is a *binary partitioning* of a frequent itemset

Frequent Itemset Generation

Still computationally expensive:

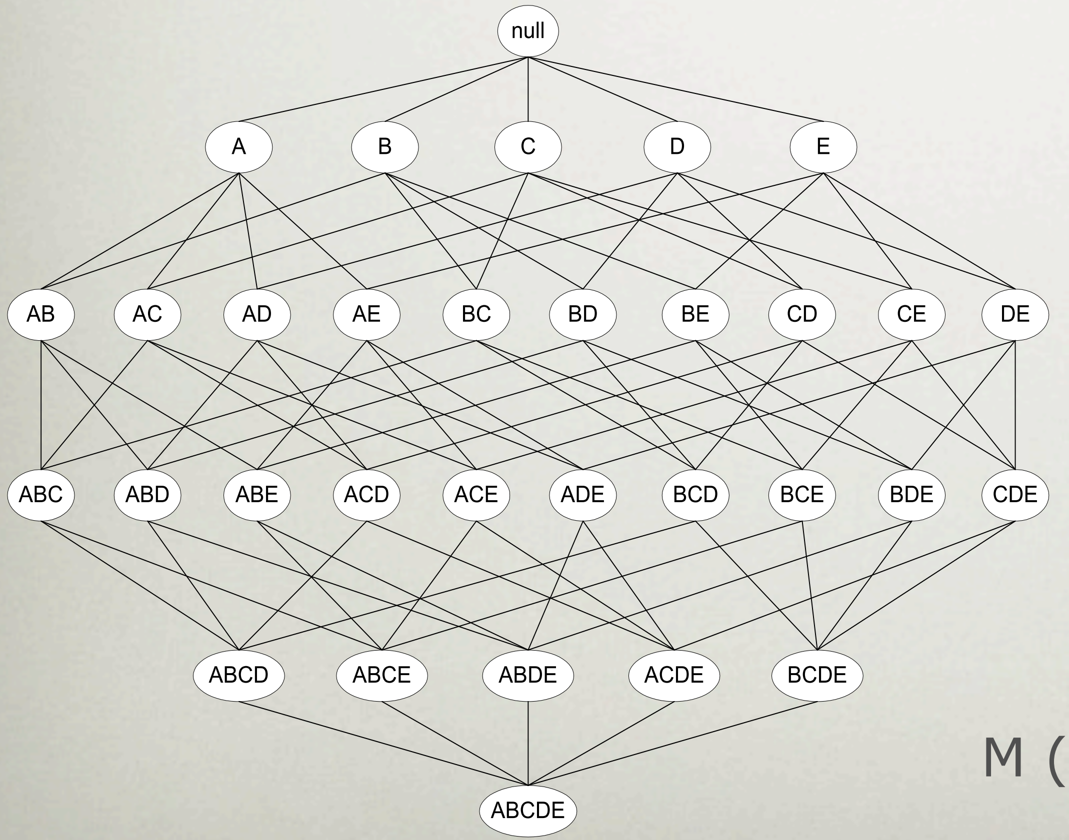
d items: 2^d candidate itemsets



Frequent Itemset Generation

Brute force:

Match every itemset against transaction database



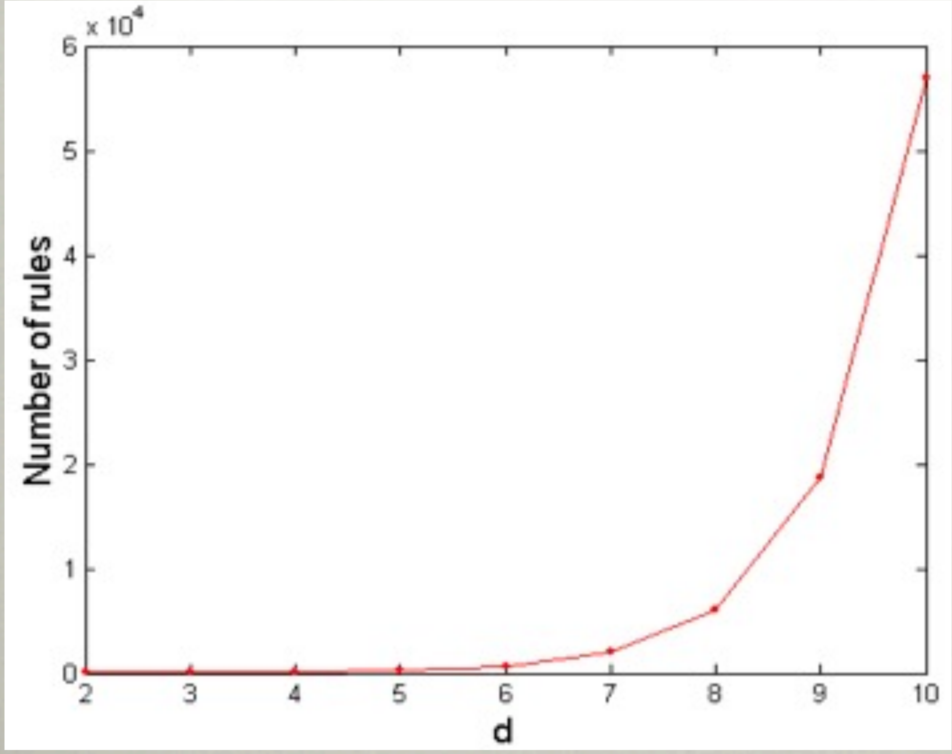
ITEMSET DATABASE	
TID	*Items*
1	{ Bread, Milk }
2	{ Bread, Diapers, Beer, Eggs }
3	{ Milk, Diapers, Beer, Cola }
4	{ Bread, Milk, Diapers, Beer }
5	{ Bread, Milk, Diapers, Cola }

M (2^d) itemsets, N transactions:

Complexity ~ O(NMw) !

Computational Complexity

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



$$R = \sum_{k=1}^{d-1}\left[\binom{d}{k} \times \sum_{j=1}^{d-k}\binom{d-k}{j}\right]$$
$$= 3^d - 2^{d+1} + 1$$

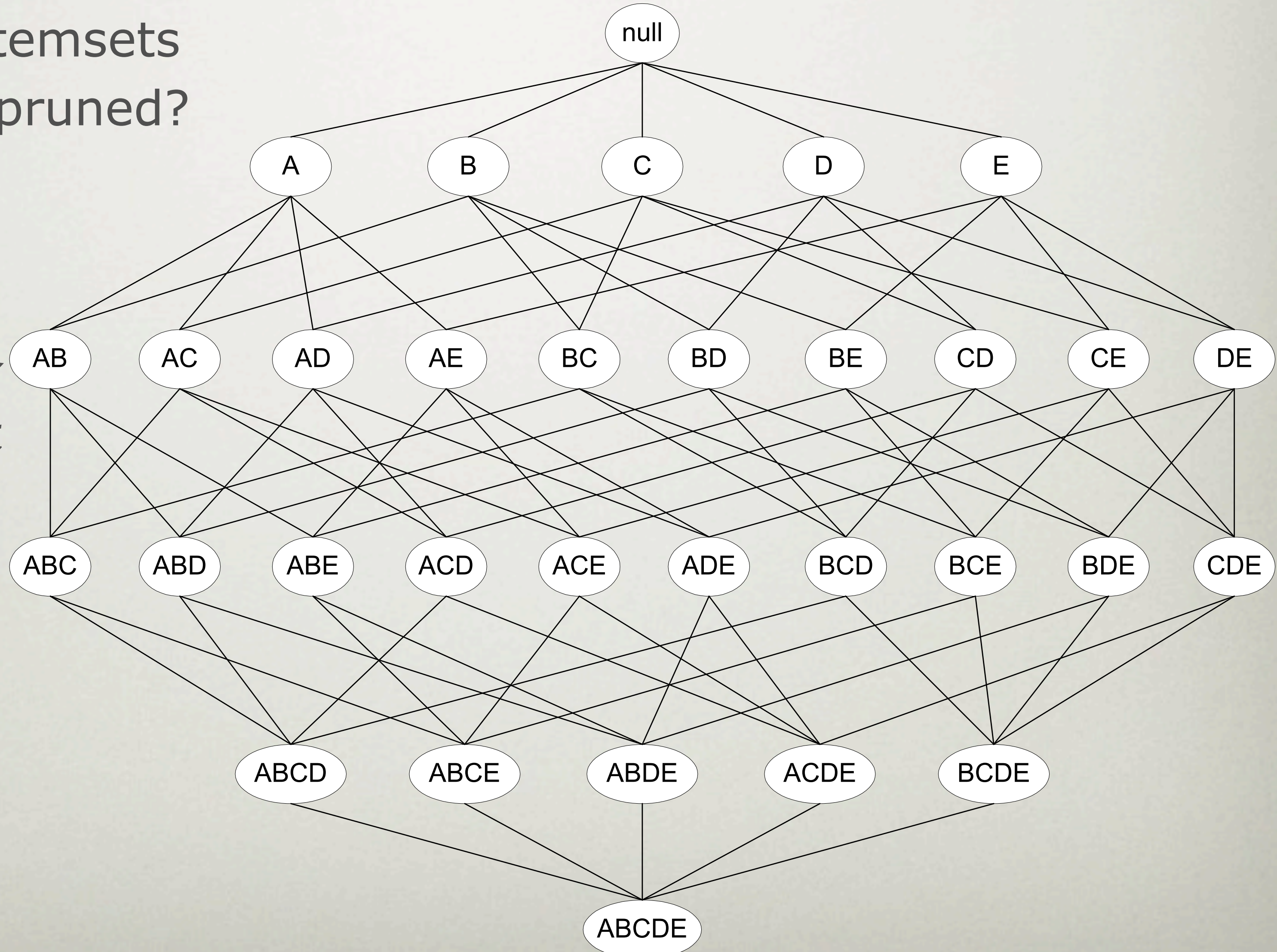
If d=6, R = 602 rules

Frequent Itemset Generation Strategies

- Need to:
  - Reduce the number of candidates (M)
    - *Pruning*
  - Reduce the number of comparisons (NM)
    - *Efficient support counting*

1) Reduce number of candidates

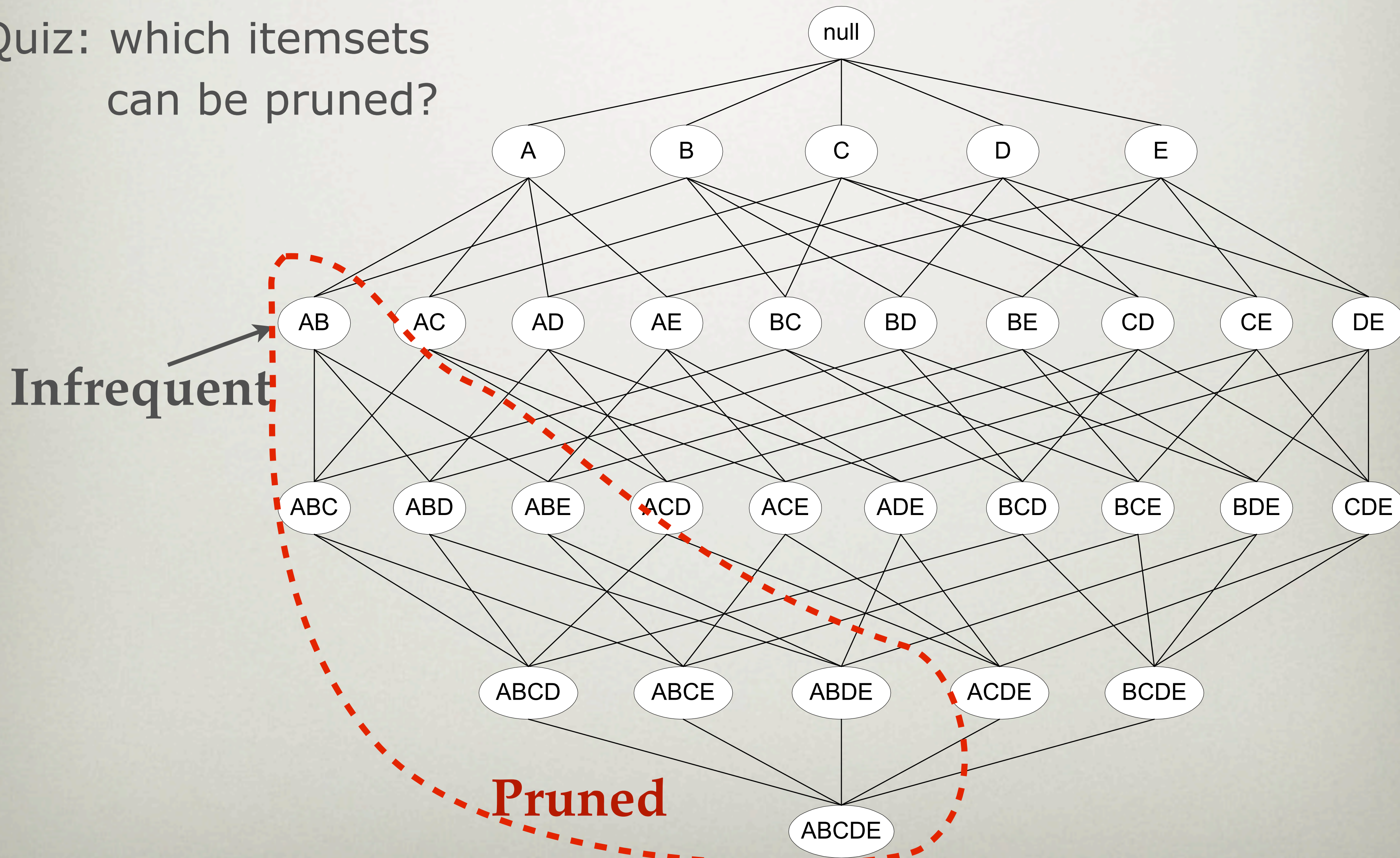
Quiz: which itemsets can be pruned?



Infrequent

1) Reduce number of candidates

Quiz: which itemsets can be pruned?



Apriori Principle

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

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

- Support of itemset never exceeds the support of its subsets
- Anti-monotonicity

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}	**2**

If every subset is considered,
$^6C_1 + {}^6C_2 + {}^6C_3 = 41$
With support-based pruning,
$6 + 6 + 1 = 13$

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 infrequent candidates**, leaving only those that are frequent

Apriori Algorithm: Level 1

candidate	support	Frequent?
A	6	Y
B	7	Y
C	6	Y
D	2	Y
E	2	Y

TID	Items
1	ABE
2	BD
3	BC
4	ABD
5	AC
6	BC
7	AC
8	ABCE
9	ABC

Minimum Support = 2

Level 2 candidates?

Apriori Algorithm: Level 2

candidate	support	Frequent?
AB	4	Y
AC	4	Y
~~AD~~	~~1~~	~~N~~
AE	2	Y
BC	4	Y
BD	2	Y
BE	2	Y
~~CD~~	~~0~~	~~N~~
~~CE~~	~~1~~	~~N~~
~~DE~~	~~0~~	~~N~~

TID	Items
1	ABE
2	BD
3	BC
4	ABD
5	AC
6	BC
7	AC
8	ABCE
9	ABC

Level 3?

Apriori Algorithm: Level 3

candidate	support	Frequent?
ABC	2	Y
ABE	2	Y

TID	Items
1	ABE
2	BD
3	BC
4	ABD
5	AC
6	BC
7	AC
8	ABCE
9	ABC

Level 4?

2) Reduce Number of Comparisons

candidate	support	Frequent
AB	4	Y
AC	4	Y
AD	1	N
AE	2	Y
BC	4	Y
BD	2	Y
BE	2	Y
CD	0	N
CE	1	N
DE	0	N

Look-up →

TID	Items
1	ABE
2	BD
3	BC
4	ABD
5	AC
6	BC
7	AC
8	ABCE
9	ABC

Quiz: efficient data structure to look up items in a list?

2) Reduce Number of Comparisons

candidate	support	Frequent?
AB	4	Y
AC	4	Y
AD	1	N
AE	2	Y
BC	4	Y
BD	2	Y
BE	2	Y
CD	0	N
CE	1	N
DE	0	N

Look-up

TID	Items
1	ABE
2	BD
3	BC
4	ABD
5	AC
6	BC
7	AC
8	ABC
9	ABC

Hash maps!

Key	Value
A,C,E	AB, AC, AD, AE, CD, CE
B,D	BC, BD, BE, DE

Factors Affecting Complexity

- Minimum support threshold
  - lower support threshold: more frequent itemsets
  - increases number and length of candidate itemsets
- Dimensionality (number of items)
  - more space needed to store support counts
  - if many frequent items: computation and I/O cost increase
- Size of database (number of transactions)
  - increases run time (Apriori makes multiple passes)
- Transaction width (dense datasets)
  - increases number of subsets
  - increases length of frequent itemsets and hash tree traversals

Back to Association Rules

- 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 →D, ABD →C, ACD →B, BCD →A,
      A →BCD, B →ACD, C →ABD, D →ABC
      AB →CD, AC → BD, AD → BC, BC →AD,
      BD →AC, CD →AB
- If $|L| = k$, then there are $2^k - 2$ candidate association rules (ignoring $L \rightarrow \emptyset$ and $\emptyset \rightarrow L$)

Rule Generation

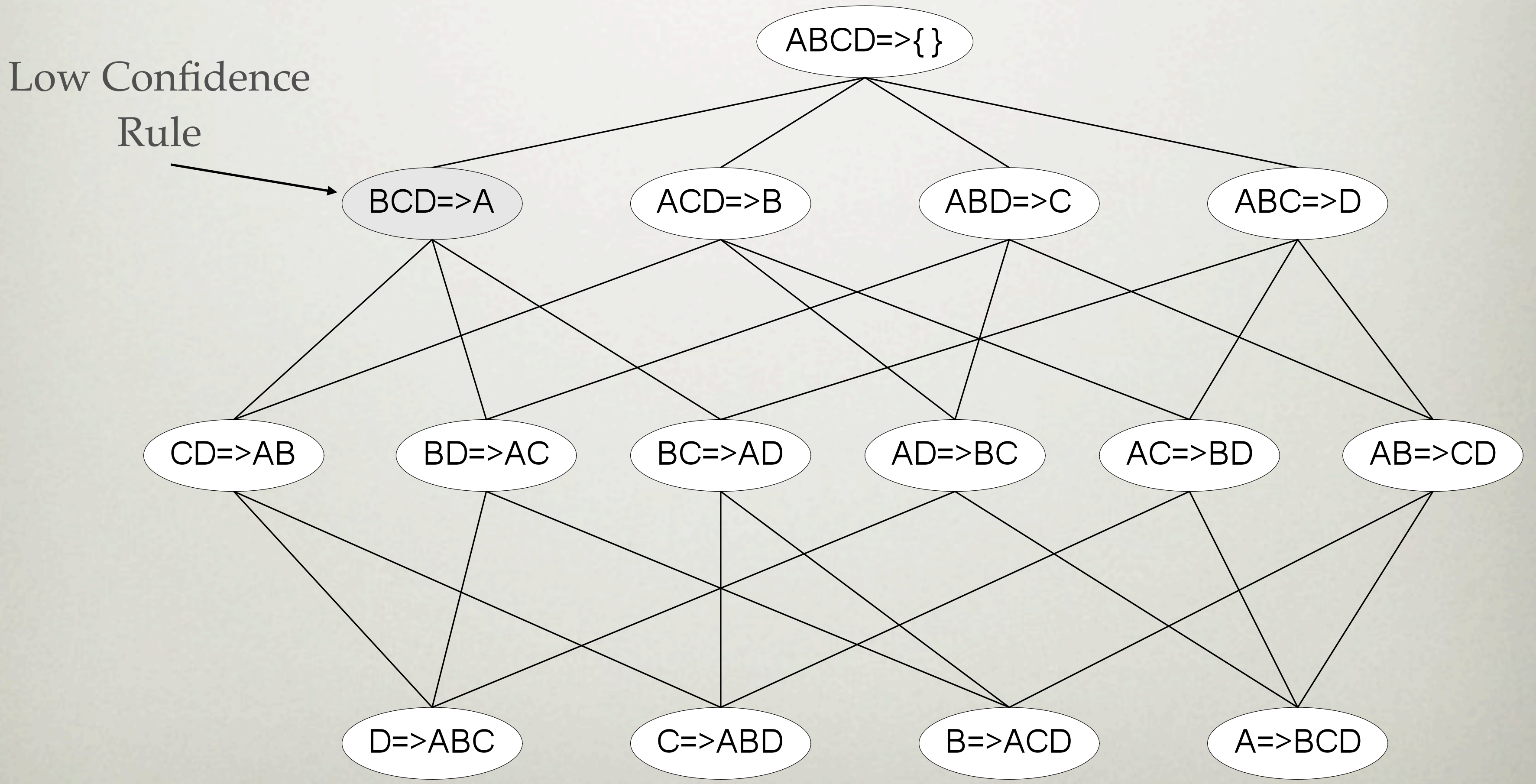
- How to efficiently generate rules?
  - **Confidence** is *not* anti-monotone
    - $c(ABC \rightarrow D)$ can be larger or smaller than $c(AB \rightarrow D)$
  - **Confidence of rules** generated from the **same** itemset *is*!
  - e.g., $L = \{A,B,C,D\}$:

    $$c(ABC \rightarrow D) \geq c(AB \rightarrow CD) \geq c(A \rightarrow BCD)$$

    - Anti-monotone w.r.t. number of items on the RHS of the rule

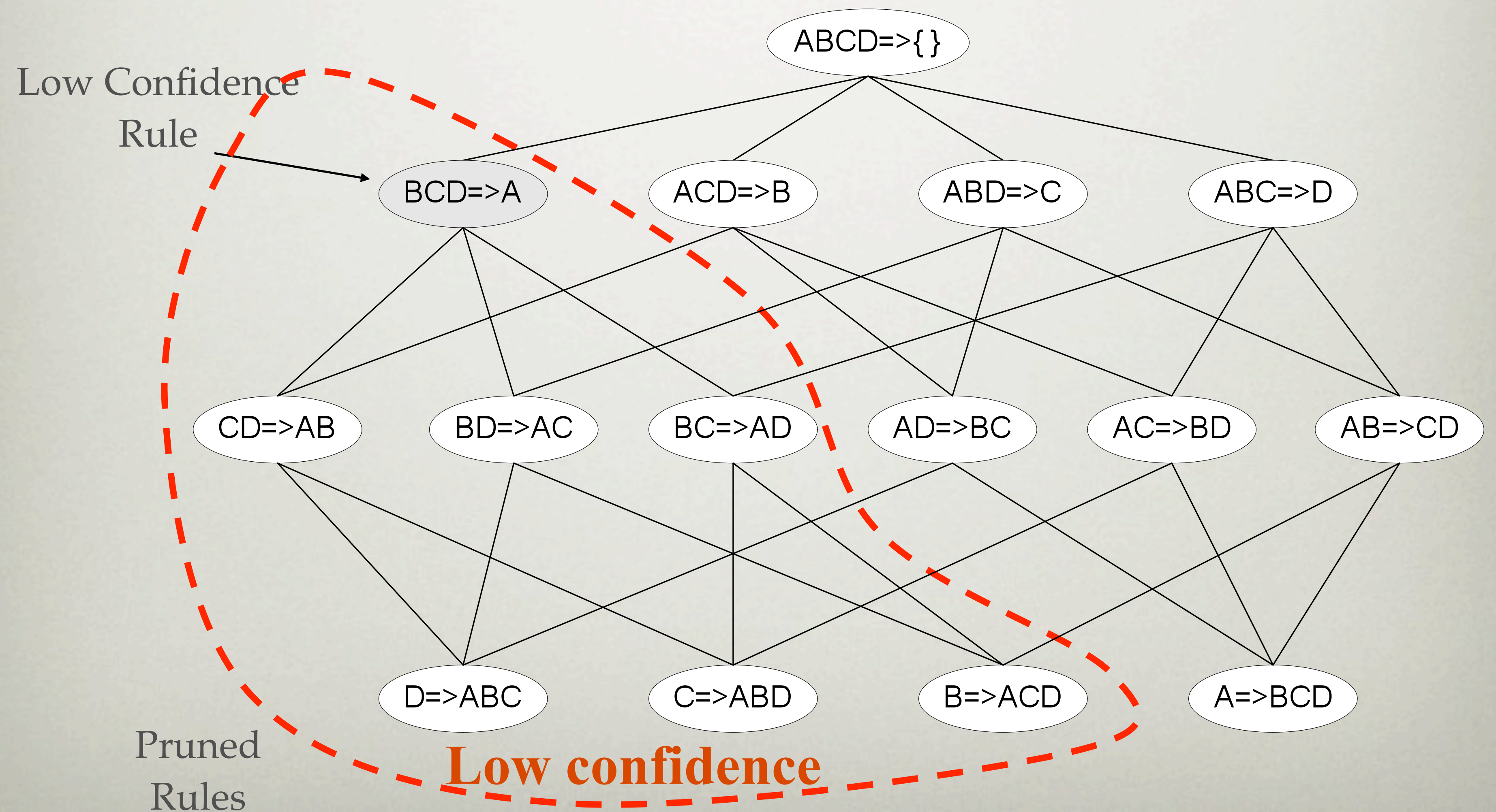
Rule Generation

Quiz: which rules have low confidence?



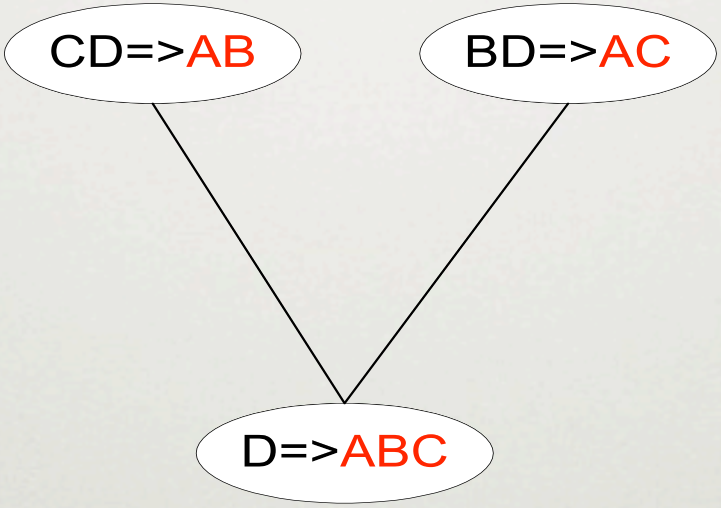
Rule Generation

Quiz: which rules have low confidence?



Rule Generation for Apriori Algorithm

- Candidate rule generated by merging two high-confidence rules with same prefix on right hand side

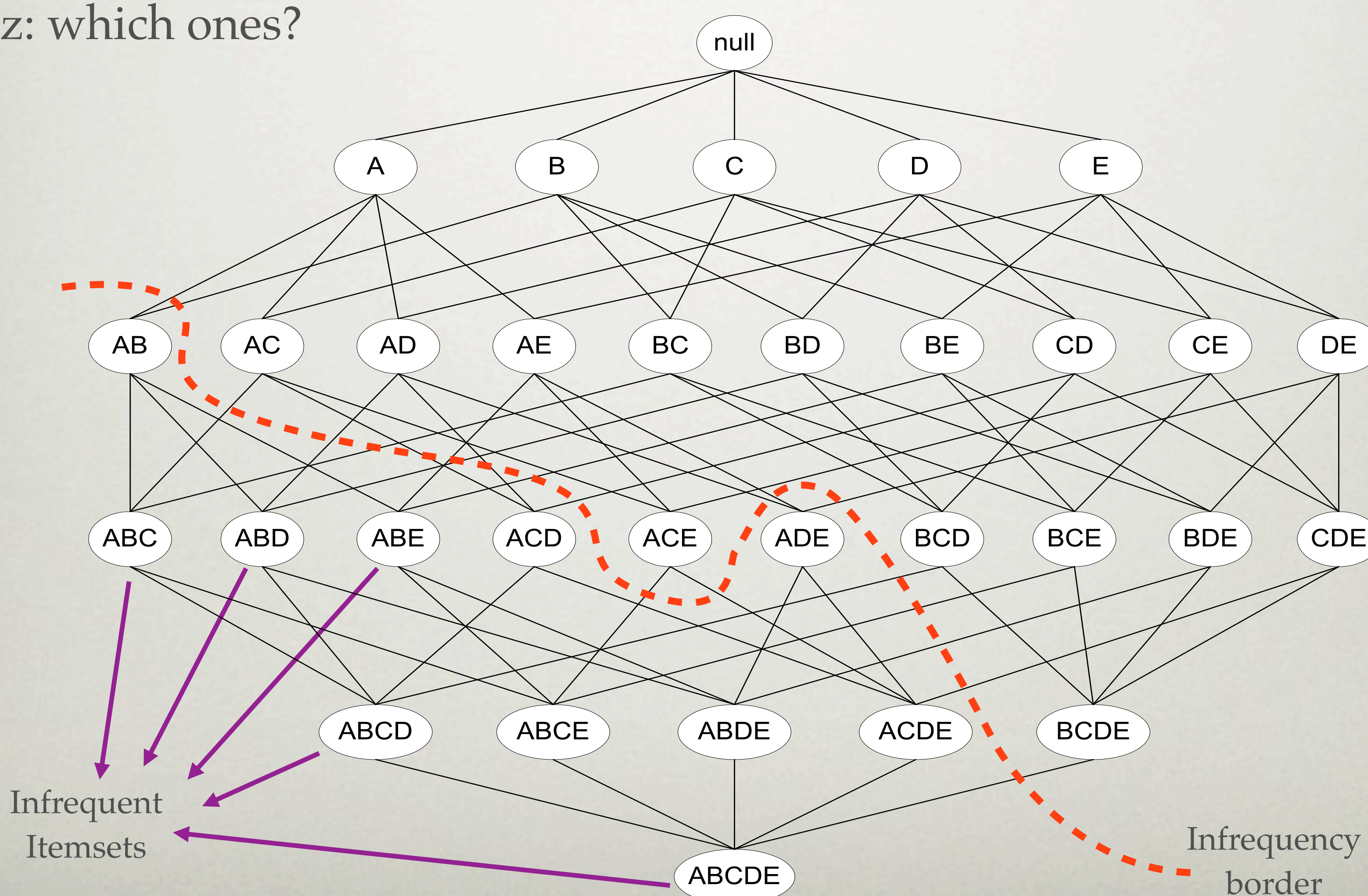


- Prune D=>ABC if subset AD=>BC has low confidence

Maximal Frequent Itemset

Itemset is *maximal frequent* if **none** of its immediate supersets is frequent

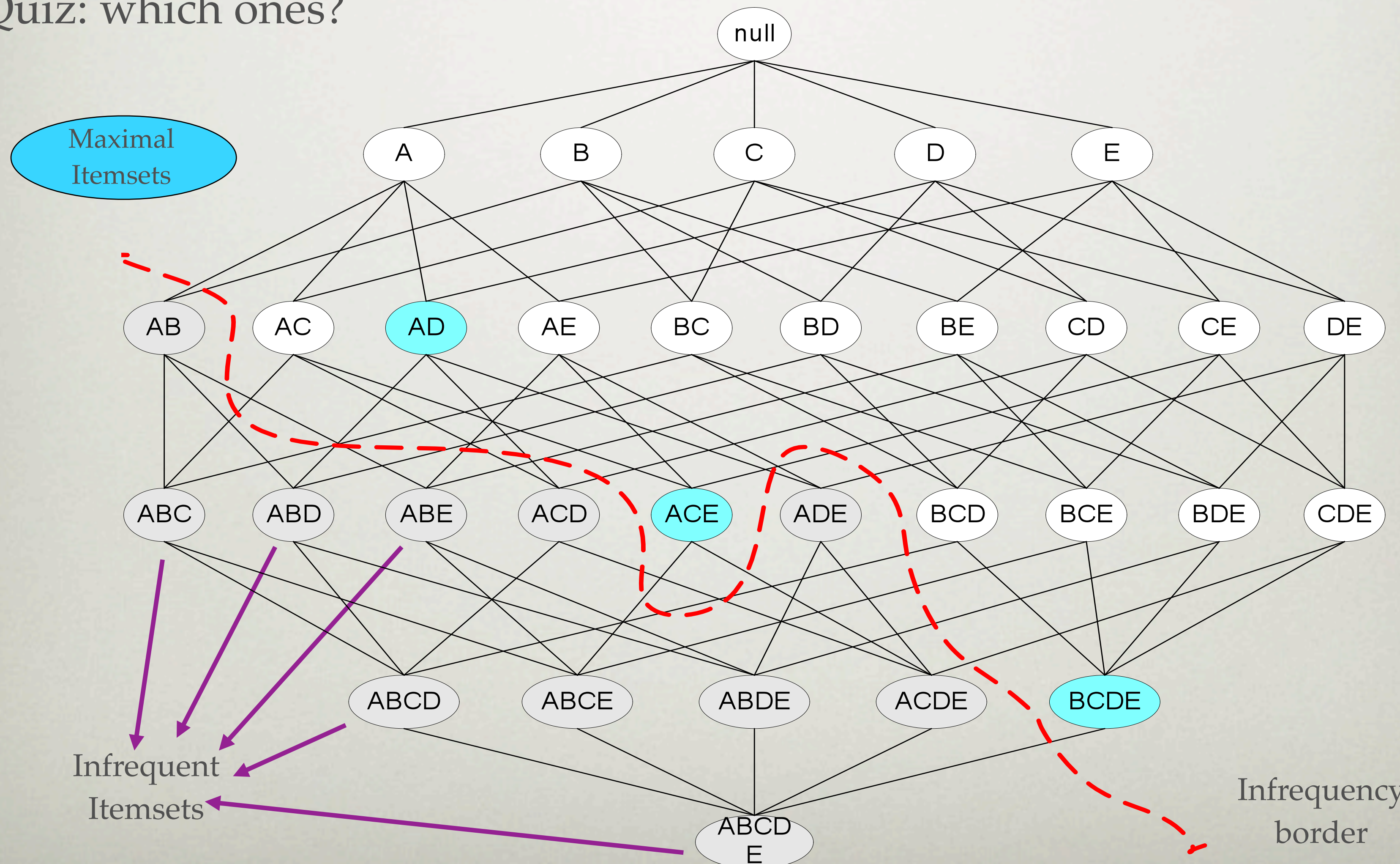
Quiz: which ones?



Maximal Frequent Itemset

Itemset is *maximal frequent* if **none** of its immediate supersets is frequent

Quiz: which ones?



Closed Itemset

- An itemset is *closed* if **none** of its (immediate) supersets has the same support
- Quiz: which ones are closed?

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

Closed Itemset

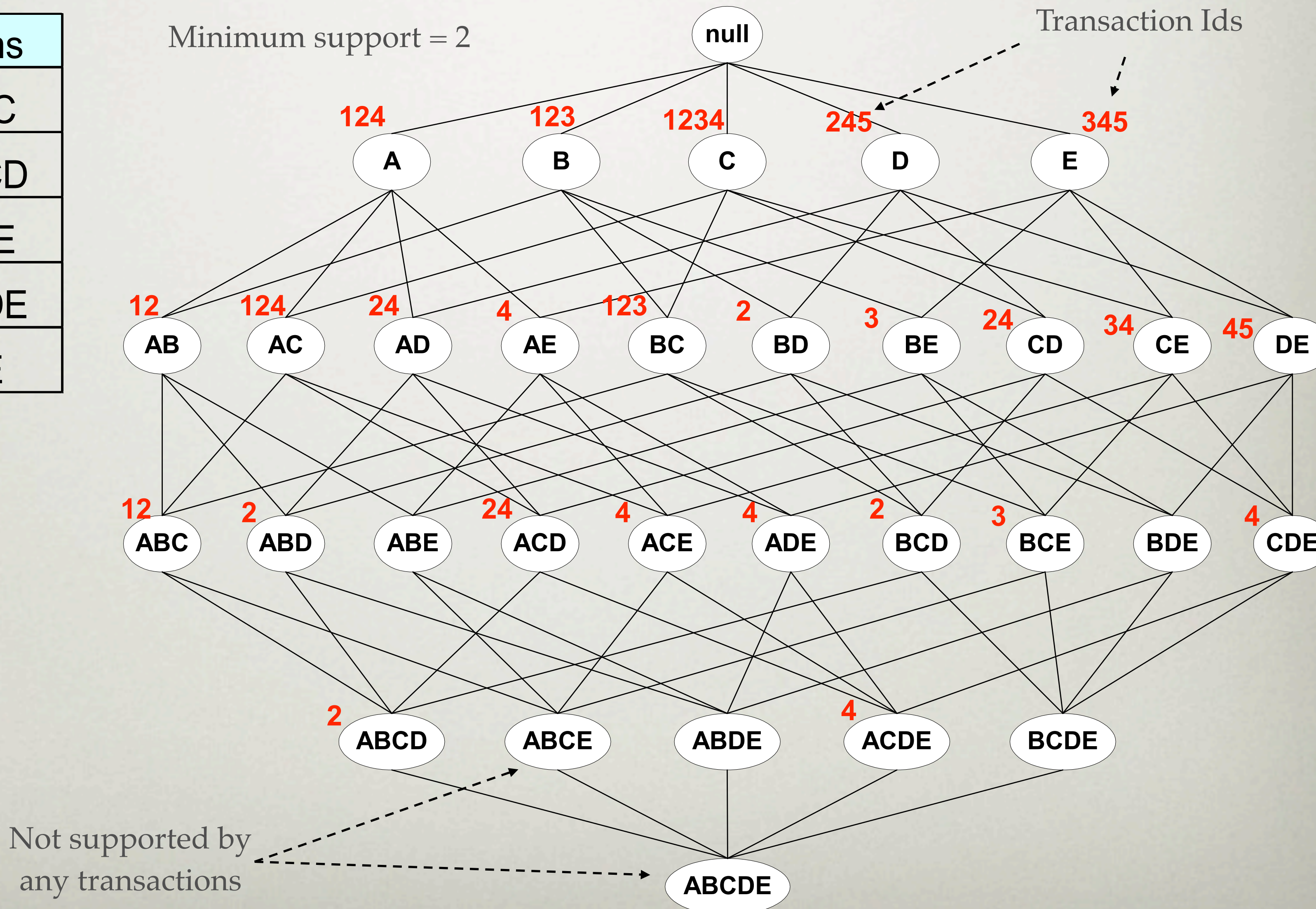
- An itemset is *closed* if **none** of its (immediate) supersets has the same support
- Quiz: which ones are closed?

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

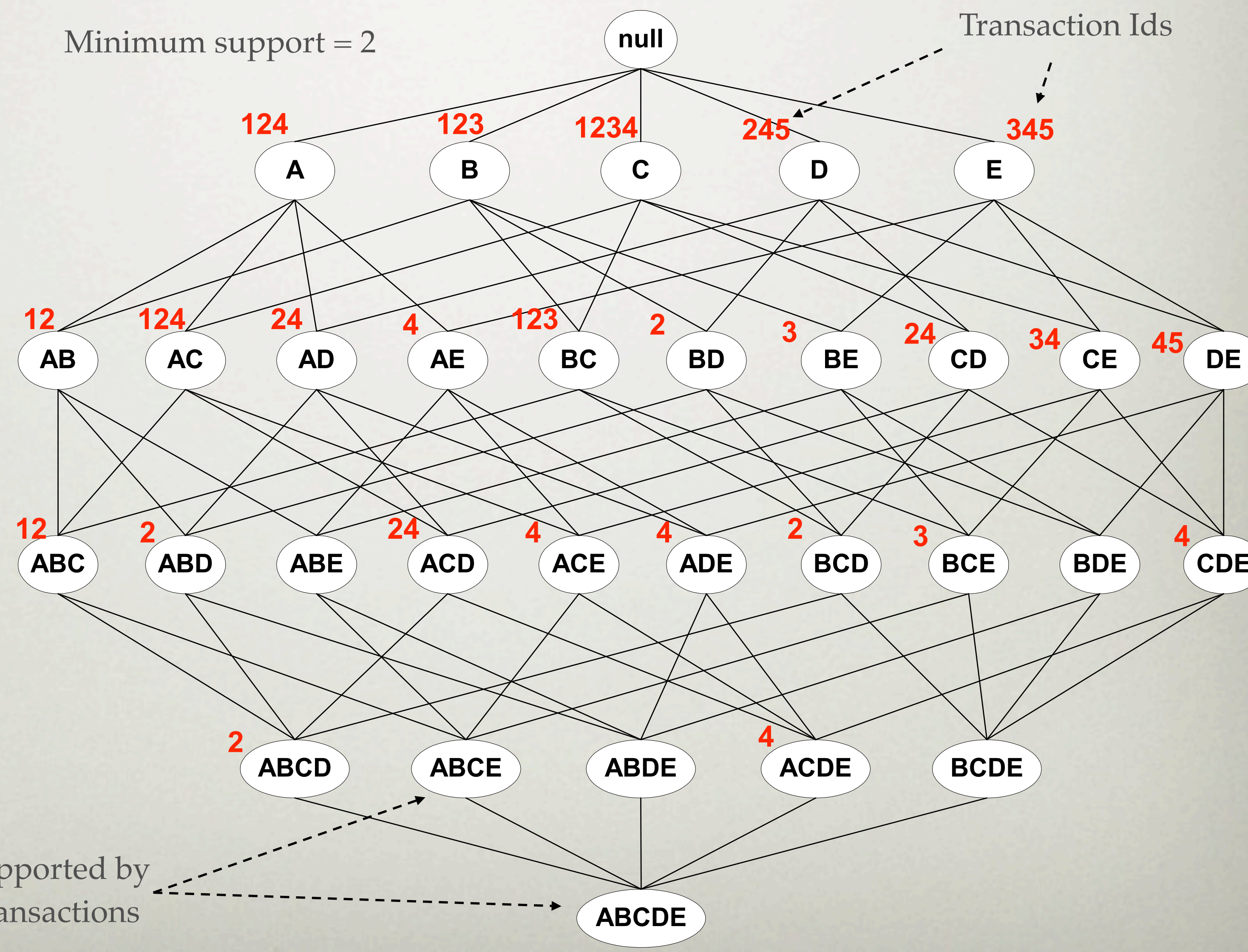
Quiz: which are frequent, closed, maximal?

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



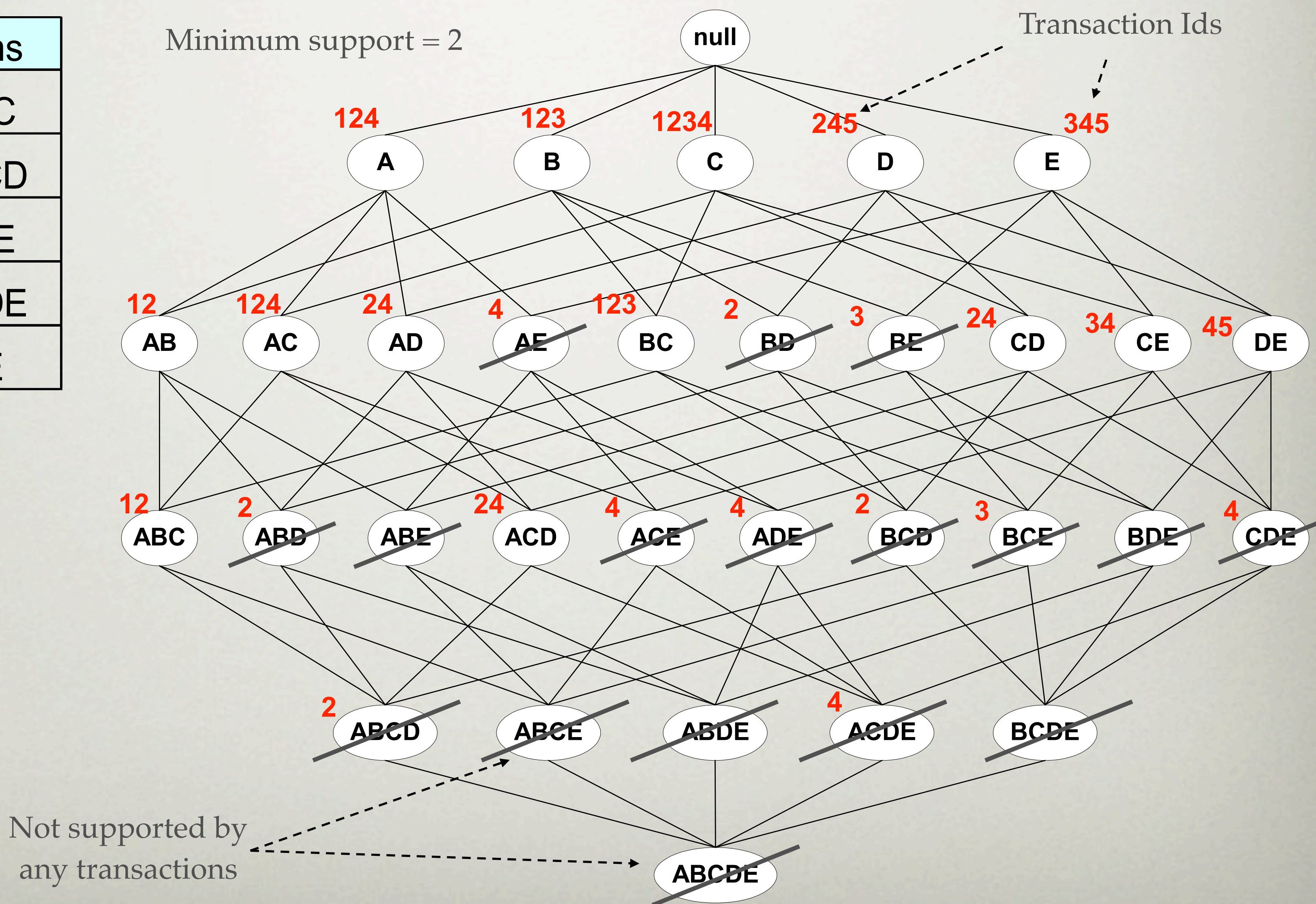
Quiz: which are FREQUENT, closed, maximal?

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



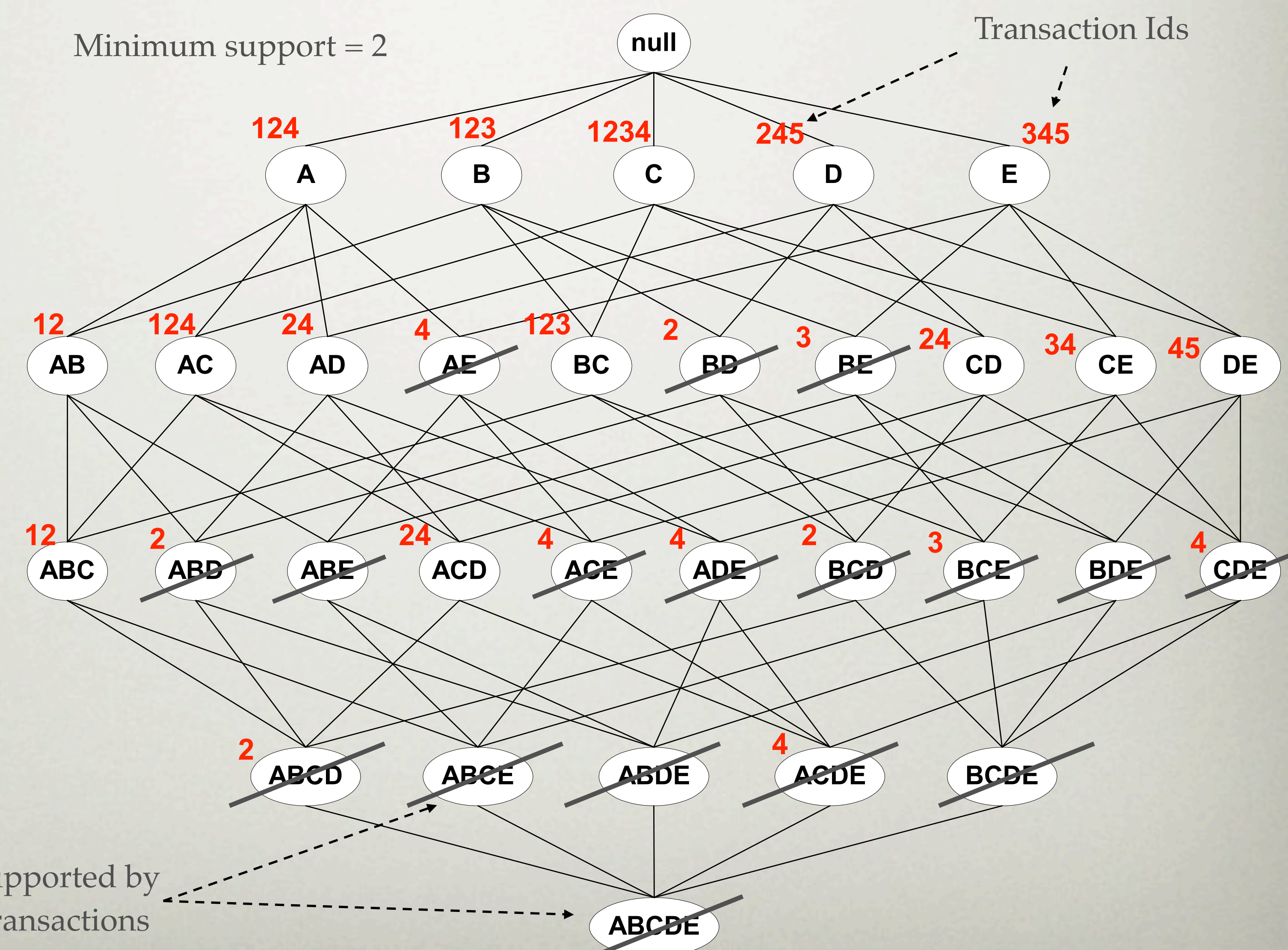
Quiz: which are frequent, closed, maximal?

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



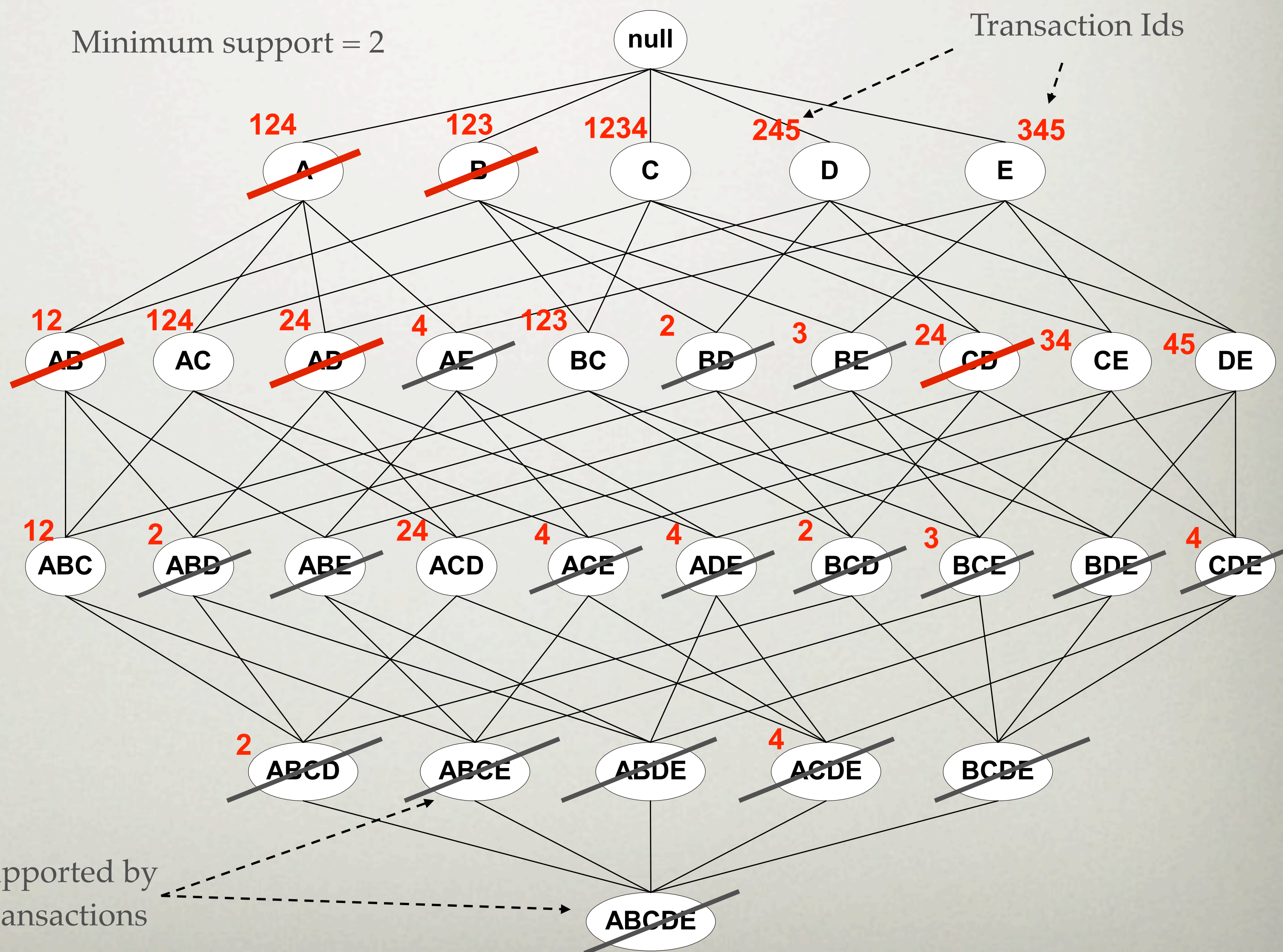
Quiz: which are frequent, closed, maximal?

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



Quiz: which are frequent, closed, maximal?

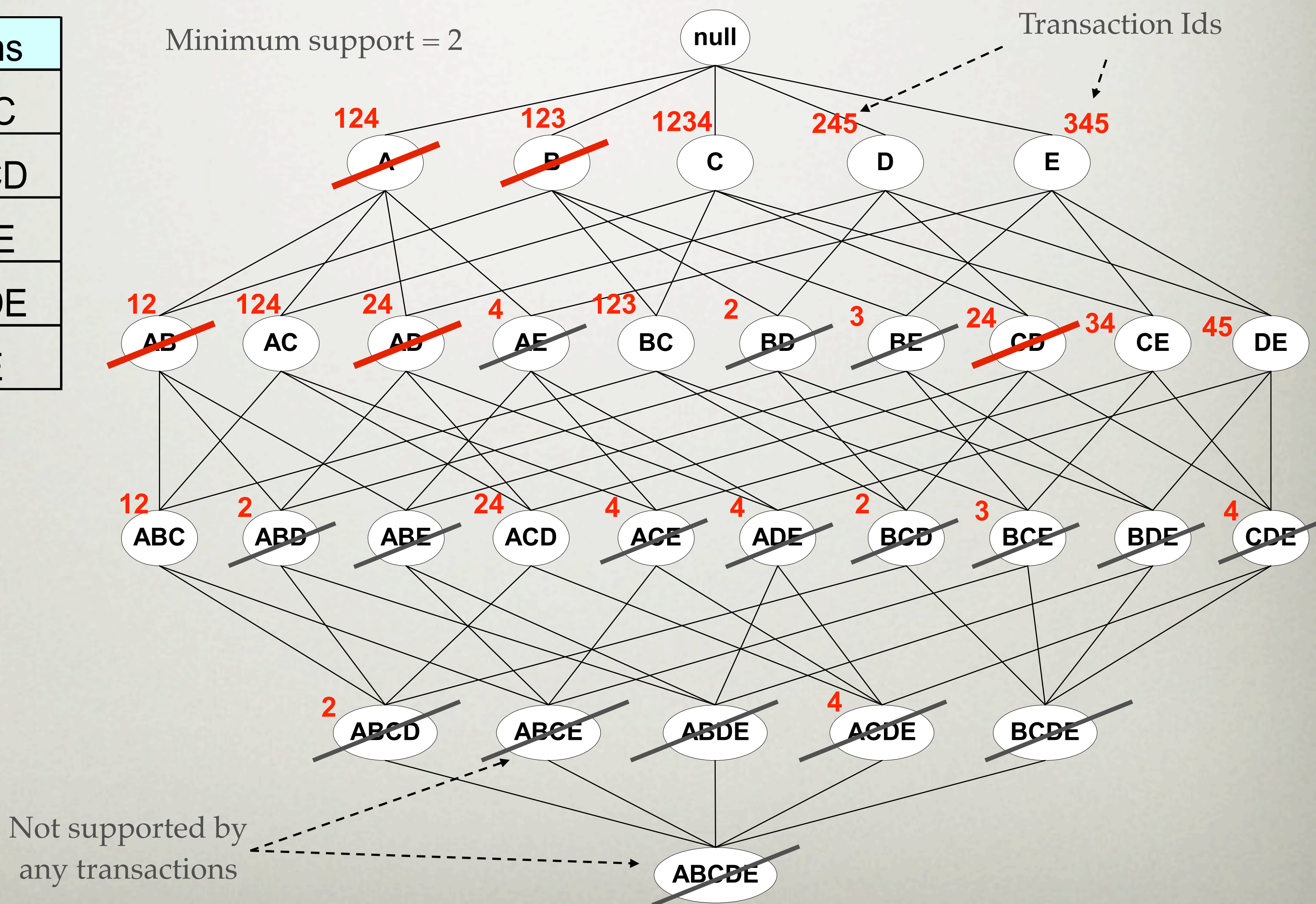
TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE



Not supported by
any transactions

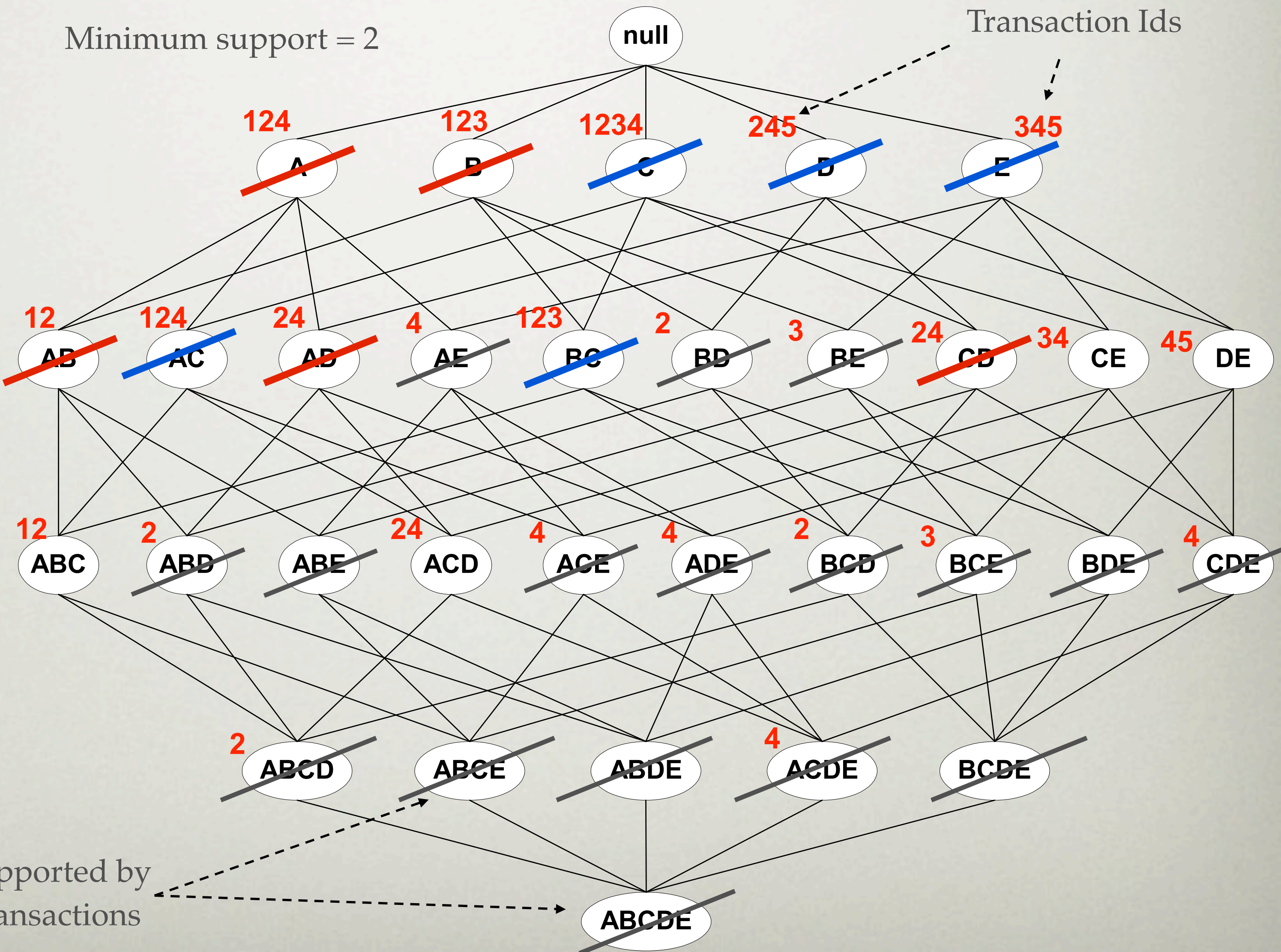
Quiz: which are frequent, closed, maximal?

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



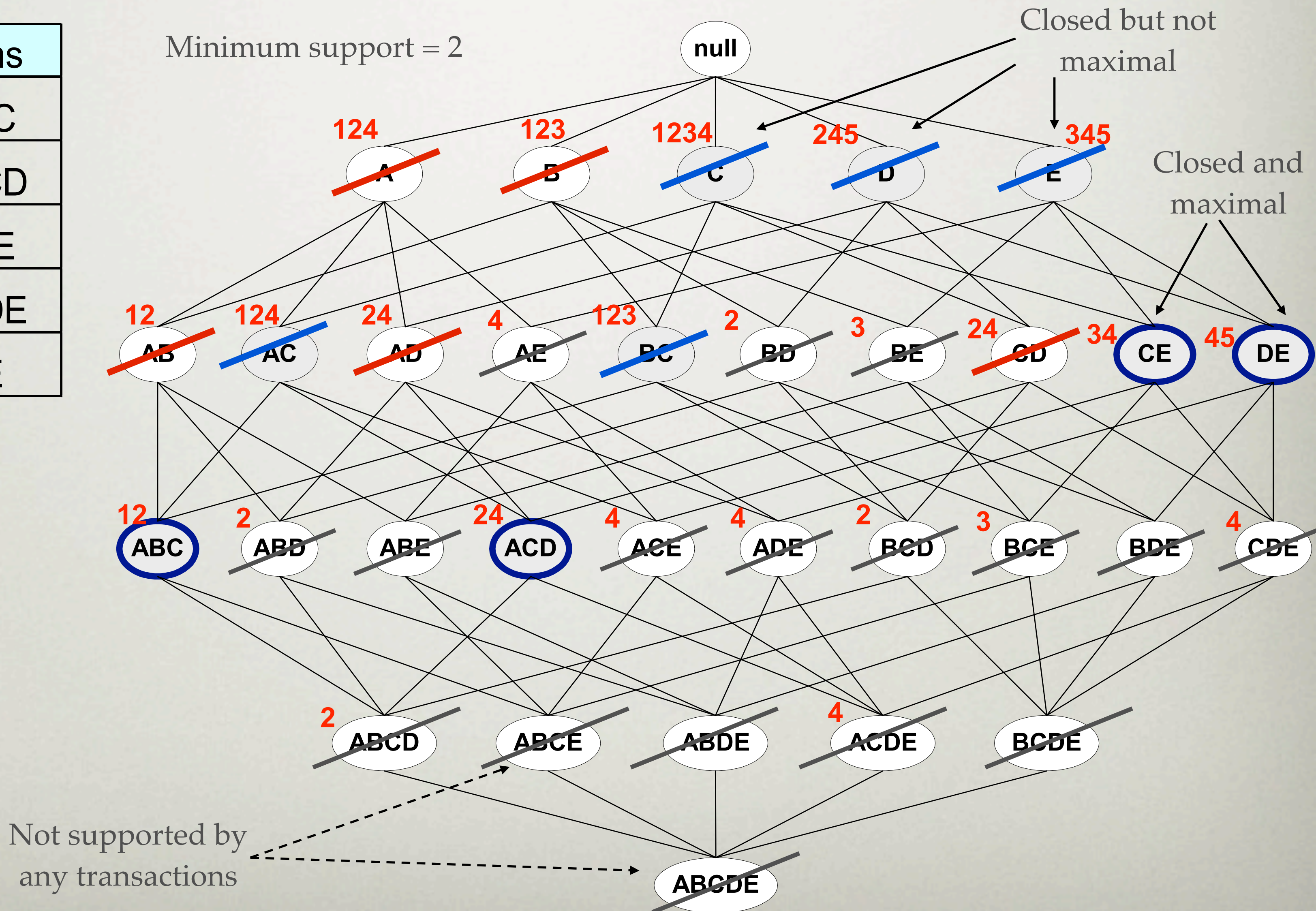
QUIZ: WHICH ARE FREQUENT, CLOSED, MAXIMAL?

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

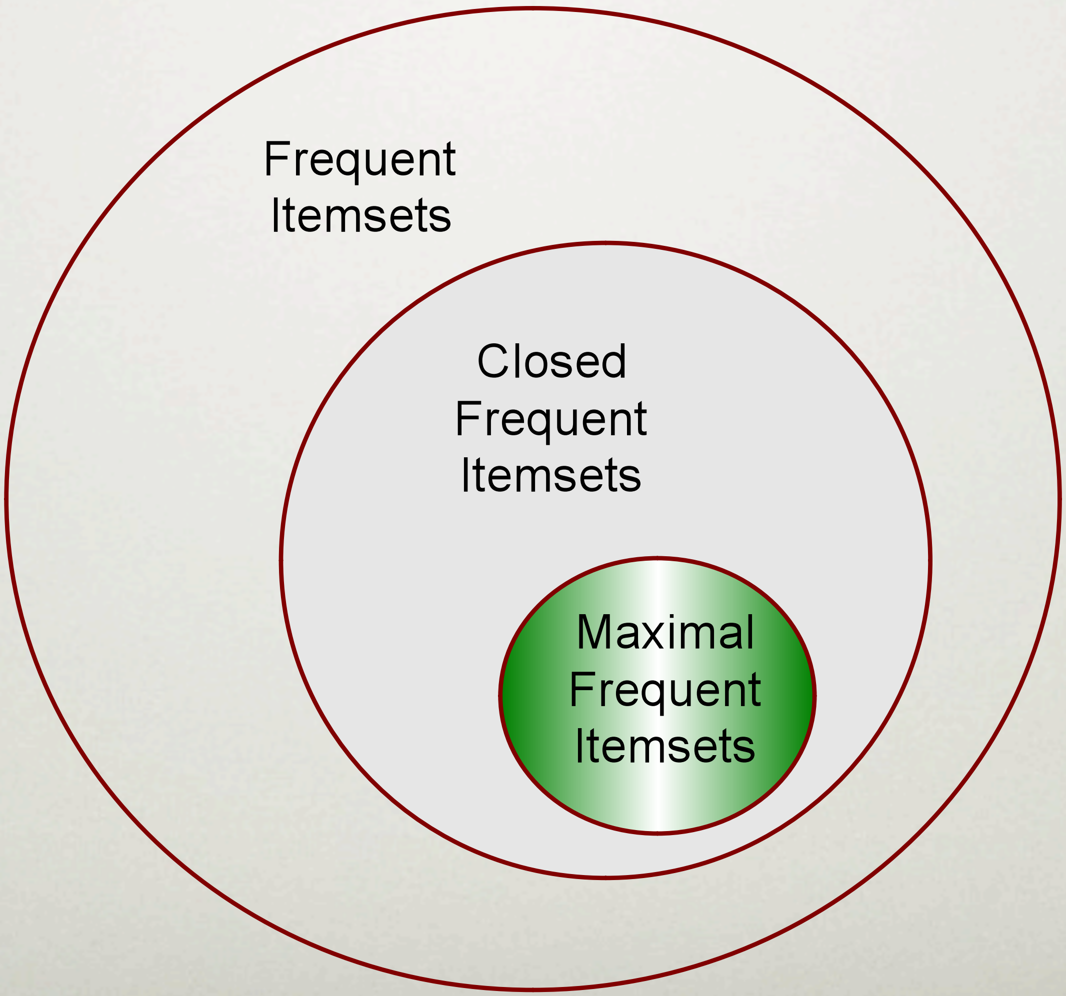


Quiz: which are frequent, closed, maximal?

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



Maximal vs Closed Itemsets



Exam question

For *minsup*=0.3, draw the itemset lattice and label each node with:

I = infrequent itemset
C = closed itemset
M = maximal itemset

tid	Items
1	{a, b, c}
2	{a, d, e}
3	{a, c}
4	{d, e}
5	{b, c}
6	{a, c, d, e}
7	{c, d, e}
8	{b, c}
9	{a, c, d, e}
10	{b}

Compute support and confidence for:

$$\{b\} \rightarrow \{c\}, \{a, d\} \rightarrow \{e\}, \text{ en } \{c\} \rightarrow \{d, e\}$$