Chapter VII.3: Association Rules

1. Generating the Association Rules
2. Measures of Interestingness
   - 2.1. Problems with confidence
   - 2.2. Some other measures
3. Properties of Measures
4. Simpson's Paradox

Zaki & Meira, Chapter 10; Tan, Steinbach & Kumar, Chapter 6

Generating association rules

- We can generate the association rules from the frequent itemsets
  - If Z is a frequent itemset and $X \subset Z$ is its proper subset, we have rule $X \rightarrow Y$, where $Y = Z \setminus X$
- These rules are frequent because

  $supp(X \rightarrow Y) = supp(X \cup Y) = supp(Z)$

  - We still need to compute the confidence as $supp(Z)/supp(X)$
- If rule $X \rightarrow Z \setminus X$ is not confident, no rule of type $W \rightarrow Z \setminus W$, with $W \subseteq X$, is confident
  - We can use this to prune the search space

Pseudo-code for generating association rules

Algorithm 8.6: Algorithm AssociationRules

```
AssociationRules (F, minconf):
1  foreach Z ∈ F, such that |Z| ≥ 2 do
2      A ← {X | X ⊂ Z, X ≠ ∅}
3      while A ≠ ∅ do
4          X ← maximal element in A
5          A ← A \ X // remove X from A
6          c ← sup(Z)/sup(X)
7          if c ≥ minconf then
8              print X ⟶ Y, sup(Z), c
9          else
10             A ← A \ {W | W ⊂ X} // remove all subsets of X from A
```

Algorithm 8.6 of Zaki & Meira

Measures of Interestingness

- Consider the following example:

	Coffee	Not Coffee	∑
Tea	150	50	200
Not Tea	650	150	800
∑	800	200	1000

- The rule {Tea} → {Coffee} has 15% support and 75% confidence
  - Reasonably good numbers
- Is this a good rule?
- The overall fraction of coffee drinkers is 80%
  ⇒ Drinking tea reduces the probability of drinking coffee!

Problems with Confidence

- Support–Confidence framework doesn't take into account the support of the consequent (tail)
  - Rules with relatively small support for the antecedent and high support for the consequent often have high confidence
- To fix this, many other measures have been proposed
- Most measures are easy to express using **contingency tables**

	B	¬B	∑
A	f_{11}	f_{10}	f_{1+}
¬A	f_{01}	f_{00}	f_{0+}
∑	f_{+1}	f_{+0}	N

Interest Factor

- The **interest factor** I of rule $A \rightarrow B$ is defined as

$$I(A, B) = \frac{N \times supp(AB)}{supp(A) \times supp(B)} = \frac{N f_{11}}{f_{1+} f_{+1}}$$

  - It is equivalent to **lift** $conf(A \rightarrow B)/supp(B)$

- Interest factor compares the frequencies against the assumption that A and B are independent
  - If A and B are independent, $f_{11} = \frac{f_{1+} f_{+1}}{N}$
- Interpreting interest factor:
  - $I(A, B) = 1$ if A and B are independent
  - $I(A, B) > 1$ if A and B are positively correlated
  - $I(A, B) < 1$ if A and B are negatively correlated

The *IS* measure

- The *IS* measure of rule $A \rightarrow B$ is defined as

$$IS(A,B) = \sqrt{I(A,B) \times supp(AB)/N} = \frac{f_{11}}{\sqrt{f_{1+}f_{+1}}}$$

- If we think A and B as binary vectors, *IS* is their cosine
- *IS* is also the geometric mean between confidences of $A \rightarrow B$ and $B \rightarrow A$

$$IS(A,B) = \sqrt{\frac{supp(AB)}{supp(A)} \times \frac{supp(AB)}{supp(B)}}$$

$$= \sqrt{conf(A \rightarrow B) \times conf(B \rightarrow A)}$$

Examples (1)

	Coffee	Not Coffee	∑
Tea	150	50	200
Not Tea	650	150	800
∑	800	200	1000

- The interest factor of {Tea} → {Coffee} is $(1000\times150)/(800\times200) = 0.9375$
  - Slight negative correlation
- The *IS* of the rule is 0.375

Examples (2)

	p	¬*p*	∑
q	880	50	930
¬*q*	50	20	70
∑	930	70	1000

	r	¬*r*	∑
s	20	50	70
¬*s*	50	880	930
∑	70	930	1000

- $I(p, q) = 1.02$ and $I(r, s) = 4.08$
  - p and q are close to independent
  - r and s have higher interest factor

But p and q appear together in 88% of cases

But r and s seldom appear together

- Now $conf(p \rightarrow q) = 0.946$ and $conf(r \rightarrow s) = 0.286$

Measures for pairs of itemsets

Measure (Symbol)	Definition
Correlation (ϕ)	$\frac{Nf_{11}-f_{1+}f_{+1}}{\sqrt{f_{1+}f_{+1}f_{0+}f_{+0}}}$
Odds ratio (α)	$\left(f_{11}f_{00}\right)/\left(f_{10}f_{01}\right)$
Kappa (κ)	$\frac{Nf_{11}+Nf_{00}-f_{1+}f_{+1}-f_{0+}f_{+0}}{N^2-f_{1+}f_{+1}-f_{0+}f_{+0}}$
Interest (I)	$\left(Nf_{11}\right)/\left(f_{1+}f_{+1}\right)$
Cosine (IS)	$\left(f_{11}\right)/\left(\sqrt{f_{1+}f_{+1}}\right)$
Piatetsky-Shapiro (PS)	$\frac{f_{11}}{N}-\frac{f_{1+}f_{+1}}{N^2}$
Collective strength (S)	$\frac{f_{11}+f_{00}}{f_{1+}f_{+1}+f_{0+}f_{+0}}\times\frac{N-f_{1+}f_{+1}-f_{0+}f_{+0}}{N-f_{11}-f_{00}}$
Jaccard (ζ)	$f_{11}/\left(f_{1+}+f_{+1}-f_{11}\right)$
All-confidence (h)	$\min\left[\frac{f_{11}}{f_{1+}},\frac{f_{11}}{f_{+1}}\right]$

Tan, Steinbach & Kumar Table 6.11

Measures for association rules

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)	$\frac{f_{11}}{f_{1+}} - \frac{f_{+1}}{N}$

Tan, Steinbach & Kumar Table 6.12

Properties of Measures

- The measures do not agree on how they rank itemset pairs or rules
- To understand how they behave, we need to study their properties
  - Measures that share some property behave similarly under that property's conditions

Three properties

- Measure has the **inversion property** if its value stays the same if we exchange f_{11} with f_{00} and f_{10} with f_{01}
  - The measure is invariant for flipping the bits
- Measure has the **null addition property** if it is not affected by increasing f_{00} if other values stay constant
  - The measure is invariant on adding new transactions that don't have the items in the itemsets
- Measure has the **scaling invariance property** if it is not affected by replacing the values f_{11}, f_{10}, f_{01}, and f_{00} with values $k_1k_3f_{11}$, $k_2k_3f_{10}$, $k_1k_4f_{01}$, and $k_2k_4f_{00}$
  - k's are positive constants

Which properties hold?

Symbol	Measure	Inversion	Null Addition	Scaling
ϕ	ϕ-coefficient	Yes	No	No
α	odds ratio	Yes	No	Yes
κ	Cohen's	Yes	No	No
I	Interest	No	No	No
IS	Cosine	No	Yes	No
PS	Piatetsky-Shapiro's	Yes	No	No
S	Collective strength	Yes	No	No
ζ	Jaccard	No	Yes	No
h	All-confidence	No	No	No
s	Support	No	No	No

Tan, Steinbach & Kumar Table 6.17

Simpson's Paradox

- Consider the following data on who bought HDTVs and exercise machines

	Exercise Machine	No Exercise Machine	Σ
HDTV	99	81	180
No HDTV	54	66	120
Σ	153	147	300

- {HDTV} → {Exercise mach.} has confidence 0.55
- {¬HDTV} → {Exercise mach.} has confidence 0.45

  ⇒ Customers who buy HDTVs are more likely to buy exercise machines than those who don't buy HDTVs

Deeper analysis

		Exerc. mach.		
Group	**HDTV**	**Yes**	**No**	**∑**
College	**Yes**	1	9	10
	No	4	30	34
Working	**Yes**	98	72	170
	No	50	36	86

- For college students
  - conf(HDTV → Exerc. mach.) = 0.10
  - conf(¬HDTV → Exerc. mach.) = 0.118
- For working adults
  - conf(HDTV → Exerc. mach.) = 0.577
  - conf(¬HDTV → Exerc. mach.) = 0.581

No HDTV is more likely to by exercise machine!

The paradox and why it happens

- In the combined data, HDTVs and exercise machines correlate positively
- In the stratified data, they correlate negatively
  - This is the Simpson's paradox
- The explanation:
  - Most customers were working adults
    - They also bought most HDTVs and exercise machines
  - In the combined data this increased the correlation between HDTVs and exercise machines
- Moral of the story: stratify your data properly!

Chapter VII.4: Summarizing Itemsets

1. The flood of itemsets
2. Maximal and closed frequent itemsets
   - 2.1. Definitions
   - 2.2. Algorithms
3. Non-derivable itemsets
   - 3.1. Inclusion-exclusion principle
   - 3.2. Non-derivability

Zaki & Meira, Chapter 11; Tan, Steinbach & Kumar, Chapter 6

The Flood of Itemsets

- Consider the following table:

tid	A	B	C	D	E	F	G	H
1	✔	✔	✔	✔	✔			
2		✔	✔	✔	✔	✔	✔	
3			✔	✔	✔	✔	✔	✔
4	✔	✔			✔	✔	✔	✔
5		✔	✔		✔	✔		✔
6	✔			✔	✔	✔		✔
7	✔	✔	✔	✔	✔	✔	✔	✔

- How many itemsets with minimum frequency of 1/7 it has?
- 255!
- Still 31 frequent itemsets with 50% minfreq

- ”Data mining is … to summarize the data”
  - Hardly a summarization!

Maximal and closed frequent itemsets

- Let $\mathcal{F}$ be the collection of all frequent itemsets of some data set
- Itemset $X \in \mathcal{F}$ is **maximal** it has no frequent supersets
  - I.e. for all $Y \supset X$, $freq(Y) <$ **minfreq**
- We can use the set of all maximal itemsets to decide whether an itemset is frequent
  - X is frequent if and only if there exists a maximal frequent itemset M such that $X \subseteq M$
  - This does not tell us what is the frequency of X

Example of maximal frequent itemsets

Not maximal because of {a, c, e}

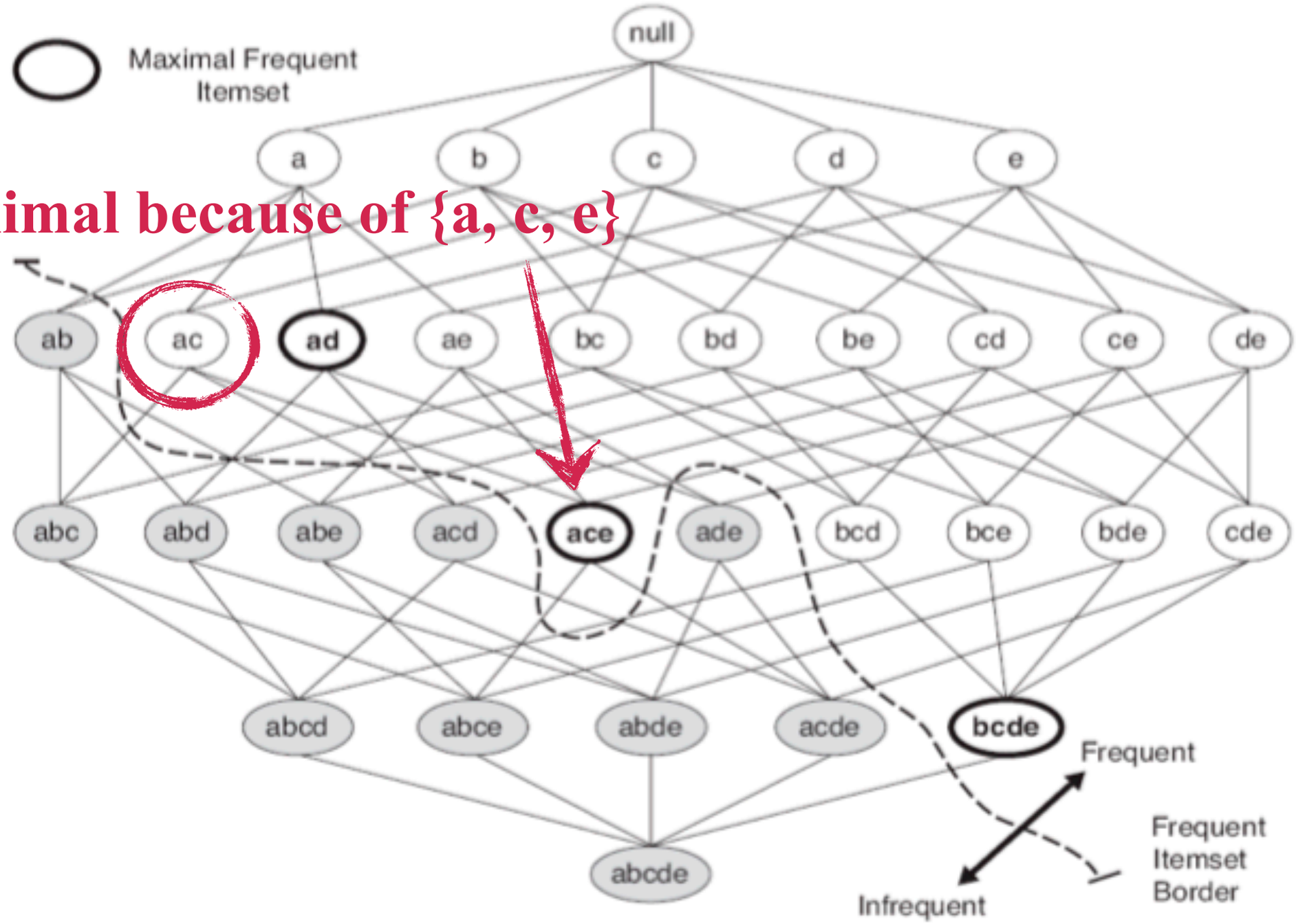


Figure 6.16. Maximal frequent itemset.

Closed frequent itemsets

- Let $\mathcal{F}$ be the collection of all frequent itemsets of some data set
- Itemset $X \in \mathcal{F}$ is **closed** if all its supersets are less frequent
  - I.e. for all $Y \supset X$, $freq(Y) < freq(X)$
  - All maximal itemsets are also closed itemsets
- Given the set of all frequent closed itemsets, we can decide if an itemset is frequent and its frequency
  - X is frequent if it is a subset of a frequent closed itemset
  - $supp(X) = \max\{supp(Z) : X \subseteq Z, Z \text{ is frequent and closed}\}$

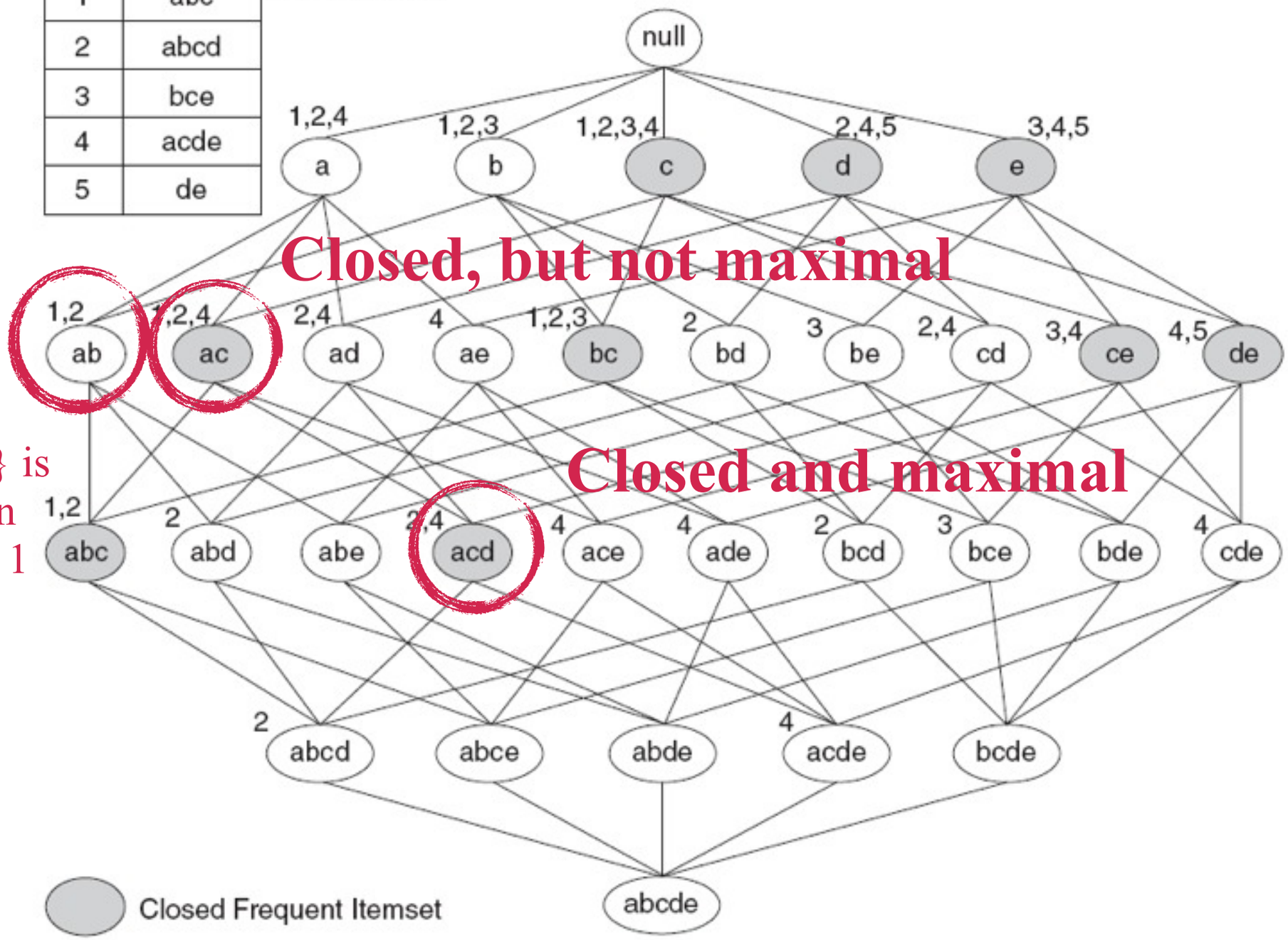
Why "closed"?

- Consider the following functions
  - $\mathbf{t}(X)$ returns all transactions that contain itemset X
  - $\mathbf{i}(T)$ returns all items that are contained in all transactions in T
- The **closure** function $\mathbf{c}(X)$ maps itemsets to itemsets by $\mathbf{c}(X) = \mathbf{i} \circ \mathbf{t}(X) = \mathbf{i}(\mathbf{t}(X))$
- Closure function satisfies the following properties
  - Extensive: $X \subseteq \mathbf{c}(X)$
  - Monotonic: if $X \subseteq Y$, then $\mathbf{c}(X) \subseteq \mathbf{c}(Y)$
  - Idempotent: $\mathbf{c}(\mathbf{c}(X)) = \mathbf{c}(X)$
- Itemset $\mathbf{X}$ is closed if and only if $X = \mathbf{c}(X)$

Example of closed frequent itemsets

TID	Items
1	abc
2	abcd
3	bce
4	acde
5	de

minsup = 40%



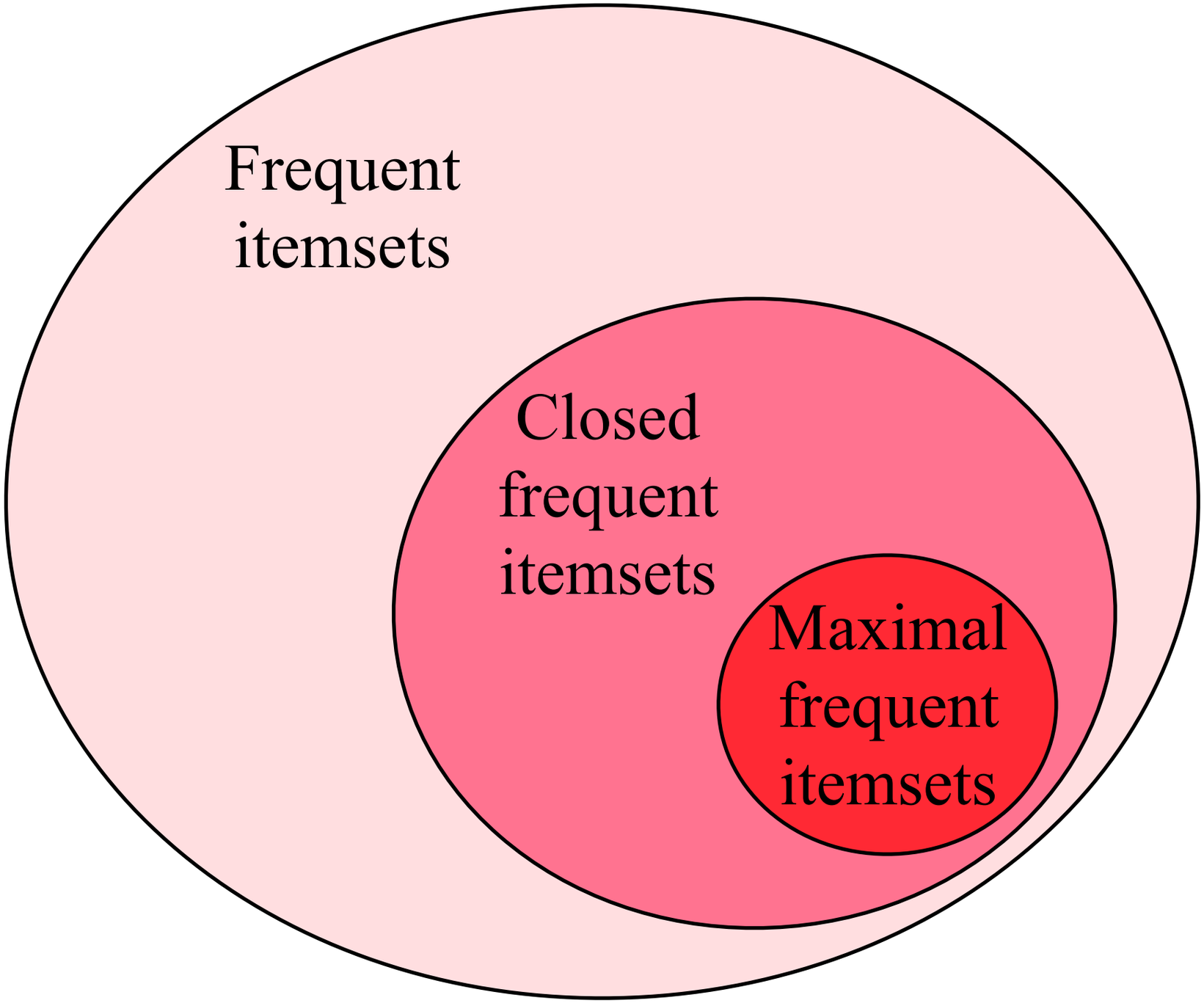
Closed, but not maximal

Closed and maximal

Itemset {a, b} is contained in transactions 1 and 2

Figure 6.17. An example of the closed frequent itemsets (with minimum support count equal to 40%).

Itemset taxonomy



Mining maximal and closed itemsets

- Frequent maximal and closed itemsets can be found by post-processing the set of frequent itemsets
- To find maximal itemsets:
  - Start with empty set of candidate maximal itemsets $\mathcal{M}$
  - **For each** frequent itemset F
    - **If** a superset of F is in $\mathcal{M}$, **continue**
    - **Else** insert F in $\mathcal{M}$ and remove all subsets of F from $\mathcal{M}$
  - **Return** set $\mathcal{M}$

Mining frequent closed itemsets

- Closed itemsets can be found from the frequent itemsets by computing their closure
  - This can be very time consuming
- The **Charm** algorithm avoids testing all frequent itemsets by using the following properties:
  - If $\mathbf{t}(X) = \mathbf{t}(Y)$, then $\mathbf{c}(X) = \mathbf{c}(Y) = \mathbf{c}(X \cup Y)$
    - We can replace X with $X \cup Y$ and prune Y
  - If $\mathbf{t}(X) \subset \mathbf{t}(Y)$, then $\mathbf{c}(X) \neq \mathbf{c}(Y)$, but $\mathbf{c}(X) = \mathbf{c}(X \cup Y)$
    - We can replace X with $X \cup Y$, but not prune Y
  - If $\mathbf{t}(X) \neq \mathbf{t}(Y)$, then $\mathbf{c}(X) \neq \mathbf{c}(Y) \neq \mathbf{c}(X \cup Y)$
    - We cannot prune anything

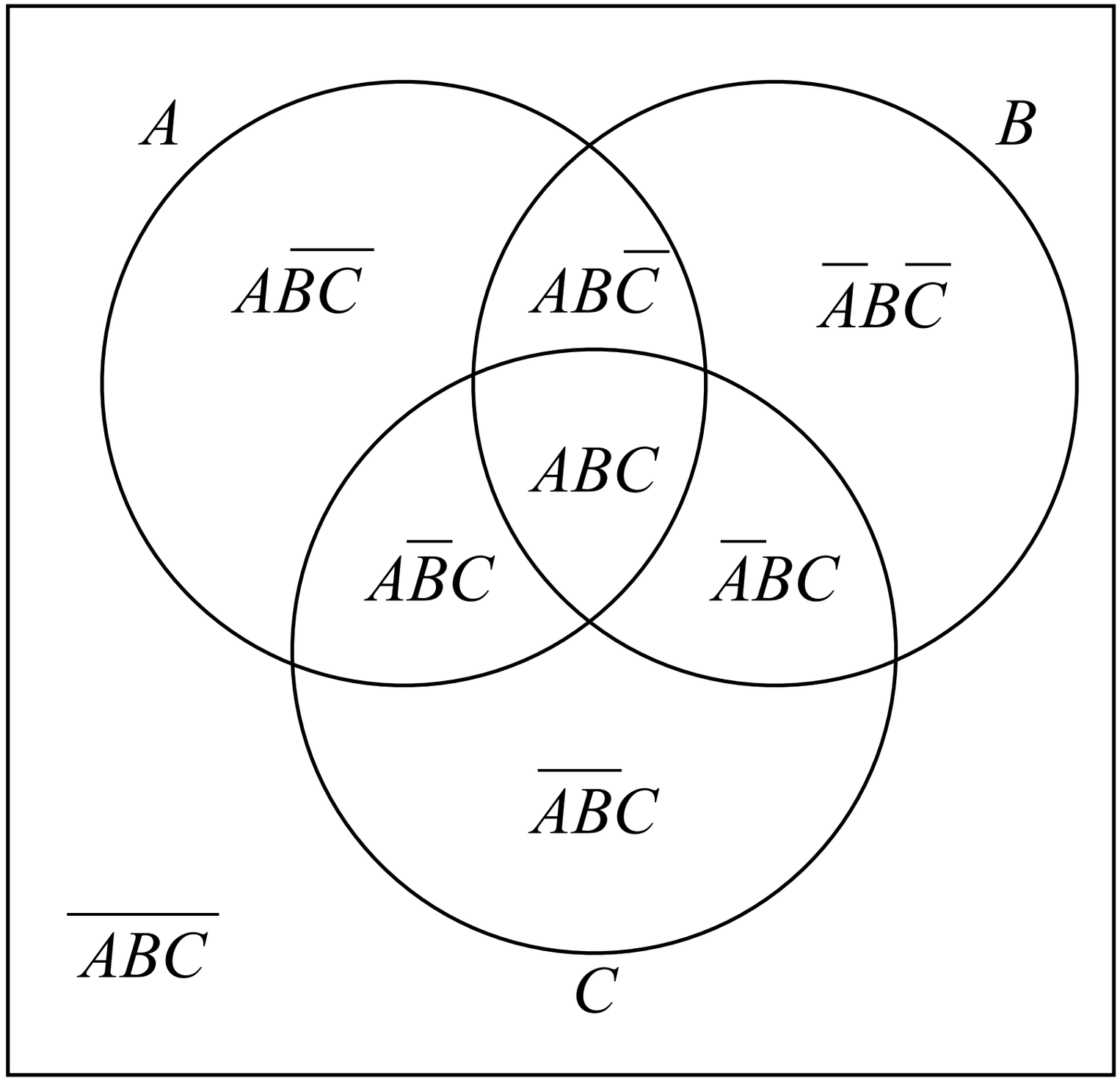
Non-Derivable Itemsets

- Let F be the set of all frequent itemsets. Itemset $X \in F$ is **non-derivable** if we cannot derive its support from its subsets.
  - We can derive the support of X from its subsets if, by knowing the supports of all of the subsets of X we can compute the support of X
- If X is derivable, it doesn't add any new information
  - Knowing just the non-derivable frequent itemsets, we can construct every frequent itemset
  - We only return itemsets that add new information on top of what we already knew

The Support of a Generalized Itemset

- A **generalized itemset** is an itemset of form $X\bar{Y}$
  - All items is X and no items in Y
- The *support* of a generalized itemset $X\bar{Y}$ is the number of transactions that contain all the items in X, but *no* items in Y
- To compute the support of a generalized itemset $A\overline{BC}$, we can
  - Take the support of A
  - Remove the supports of AB and AC
  - Add the support of ABC that was removed twice
  - $supp(A\overline{BC}) = supp(A) - supp(AB) - supp(AC) + supp(ABC)$

Generalized Itemsets



The Inclusion-Exclusion Principle

- Let $X\bar{Y}$ be a generalized itemset and let $I = X \cup Y$
- Now $supp(X\bar{Y})$ can be expressed as a combination of supports of supersets $J \supseteq X$ such that $J \subseteq I$ using the **inclusion-exclusion principle**

$$supp(X\bar{Y}) = \sum_{X \subseteq J \subseteq I} (-1)^{|J \setminus X|} supp(J)$$

  - Example:

$$\begin{aligned} supp(\overline{ABC}) = supp(\emptyset) \\ - supp(A) - supp(B) - supp(C) \\ + supp(AB) + supp(AC) + supp(BC) \\ - supp(ABC) \end{aligned}$$

Support Bounds

- The inclusion-exclusion formula gives us bounds for the supports of itemsets in $X \cup Y$ that are supersets of X
  - All supports are non-negative!
  - $supp(A\overline{BC}) = supp(A) - supp(AB) - supp(AC) + supp(ABC)$ ≥ 0 implies $supp(ABC) \geq -supp(A) + supp(AB) + supp(AC)$
    - This is a lower bound, but we can also get upper bounds
- In general the bounds for itemset I w.r.t. $X \subset I$:
  - If $|I \setminus X|$ is odd: $supp(I) \leq \sum_{X \subseteq J \subset I} (-1)^{|I \setminus J|+1} supp(J)$
  - If $|I \setminus X|$ is even: $supp(I) \geq \sum_{X \subseteq J \subset I} (-1)^{|I \setminus J|+1} supp(J)$

Deriving the Support

- Given the formula for the bounds, we can define
  - the *least upper bound lub(I)* and
  - the *greatest lower bound glb(I)* for itemset *I*
- We know that $supp(I) \in [glb(I), lub(I)]$
- If $glb(I) = lub(I)$, then we can compute $supp(I)$ by just knowing its subsets' supports
  - Hence, *I* is derivable
- Otherwise *I* is non-derivable

Example on deriving support (blackboard)

tid	*A*	*B*	*C*	*D*	*E*
1	1	1	0	1	1
2	0	1	1	0	1
3	1	1	0	1	1
4	1	1	1	0	1
5	1	1	1	1	1
6	0	1	1	1	0

Question: Is itemset *ACD* derivable?

Conclusions

- Association rules tell us which items we will probably see given that we've seen some other items
  - Many business applications
- Frequent itemsets tell which items appear together
  - Also, mining them is the first step on mining anything else
    - ⇒ Many algorithms for efficient freq. itemset mining
- The number of freq. itemsets is usually too large to study by itself
  - Maximal, closed, and non-derivable itemsets provide a summarisation of the frequent itemsets