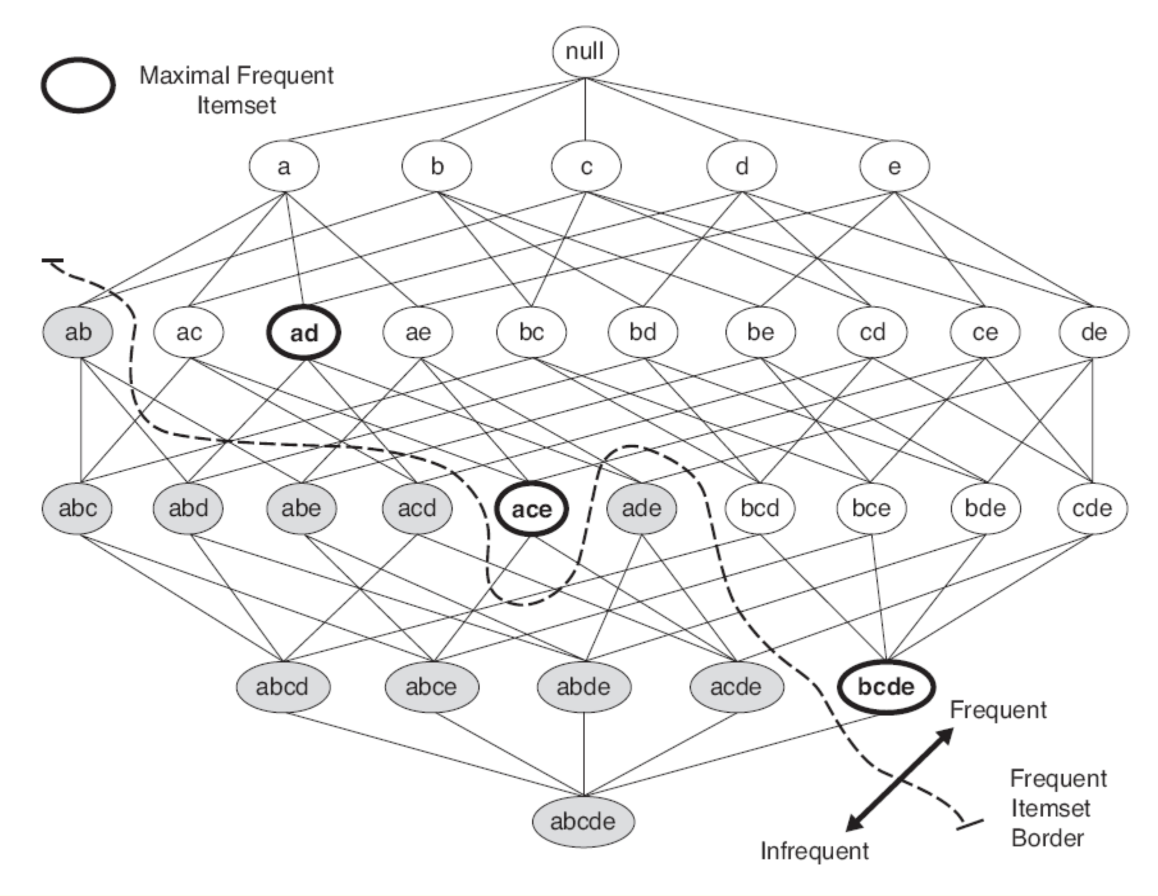
Compact representation of frequent itemsets

- The number of frequent itemsets produced from a transaction data set can be very large.
- It is useful to identify a small representative set of frequent itemsets from which all other frequent itemsets can be derived.
- Two representations are
  - Maximal frequent itemsets
  - Closed frequent itemsets

Maximal frequent itemsets

- A maximal frequent itemset is defined as a frequent itemset for which none of its immediate supersets are frequent.
- We consider the itemset lattice shown in the following figure.
- The itemsets in the lattice are divided into two groups
  - Those that are frequent
  - Those that are infrequent

Maximal frequent itemsets



Maximal frequent itemsets

- A frequent itemset border is also illustrated in the figure.
- Every itemset located above the border is frequent.
- On the other hand, those located below the border are infrequent.

Maximal frequent itemsets

- {a,d}, {a,c,e} and {b,c,d,e} are considered to be maximal frequent itemsets.
- This is because their immediate supersets are infrequent.
- In contrast, {a,c} is non-maximal because one of its immediate supersets, {a,c,e}, is frequent.

Maximal frequent itemsets

- Maximal frequent itemsets effectively provide a compact representation of frequent itemsets.
- They form the smallest set of itemsets from which all frequent itemsets can be derived.

Maximal frequent itemsets

- We can divide the frequent itemsets in the previous figure into two groups.
- The first group consists of frequent itemsets that
  - Begin with item a and
  - Followed by items c, d or e.
  - This group includes itemsets such as {a}, {a,c}, {a,d}, {a,e} and {a,c,e}.
- The second group consists of frequent itemsets that
  - Begin with b, c, d, or e.
  - This group includes itemsets such as {b}, {b,c}, {c,d}, {b,c,d,e}, etc.

Maximal frequent itemsets

- Frequent itemsets that belong to the first group are subsets of either {a,c,e} or {a,d}.
- Those that belong to the second group are subsets of {b,c,d,e}.
- Hence, the maximal frequent itemsets {a,c,e}, {a,d} and {b,c,d,e} provide a compact representation of the frequent itemsets.

Maximal frequent itemsets

- Maximal frequent itemsets do not contain the support information of their subsets.
- An additional pass over the data set is required to determine the support counts of the non-maximal frequent itemsets.

Closed frequent itemsets

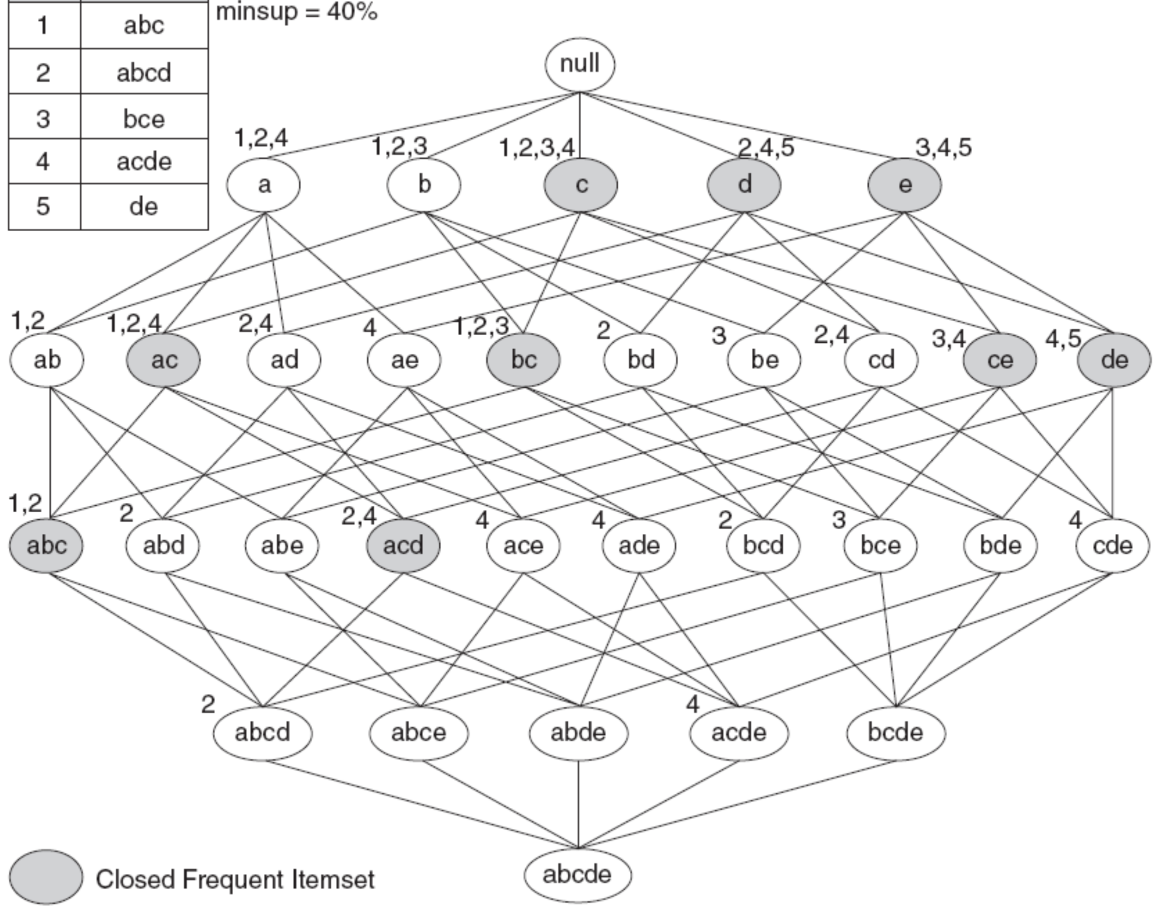
- An itemset X is closed if none of its immediate supersets has exactly the same support count as X.
- Put another way, X is not closed if at least one of its immediate supersets has the same support count as X.

Closed frequent itemsets

- Examples of closed itemsets are shown in the following figure.
- Each node (itemset) in the lattice is associated with a list of its corresponding TIDs.

Closed frequent itemsets

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



Closed frequent itemsets

- We notice that every transaction that contains b also contains c.
- Consequently, the support for {b} is identical to {b,c}.
- {b} should not be considered a closed itemset.

Closed frequent itemsets

- Similarly, the itemset {a,d} is not closed, since c occurs in every transaction that contains both a and d.
- On the other hand, {b,c} is a closed itemset.
- This is because it does not have the same support count as any of its supersets.

Closed frequent itemsets

- An itemset is a closed frequent itemset if
  - It is closed and
  - Its support is greater than or equal to minsup.
- In the previous example, assuming that the support threshold is 40%.
- {b,c} is a closed frequent itemset because its support is 60%.
- The rest of the closed frequent itemsets are indicated by the shaded nodes.

Closed frequent itemsets

- We can use the closed frequent itemsets to determine the support counts for the non-closed frequent itemsets.
- For example, we consider the frequent itemset {a,d} shown in the figure on slide 12.
- Because the itemset is not closed, its support count must be identical to one of its immediate supersets.
- The key is to determine which superset (among {a,b,d}, {a,c,d} or {a,d,e}) has exactly the same support count as {a,d}.

Closed frequent itemsets

- Any transaction that contains the superset of {a,d} must also contain {a,d}.
- However, any transaction that contains {a,d} does not have to contain the supersets of {a,d}.
- For this reason, the support for {a,d} must be equal to the largest support among its supersets.

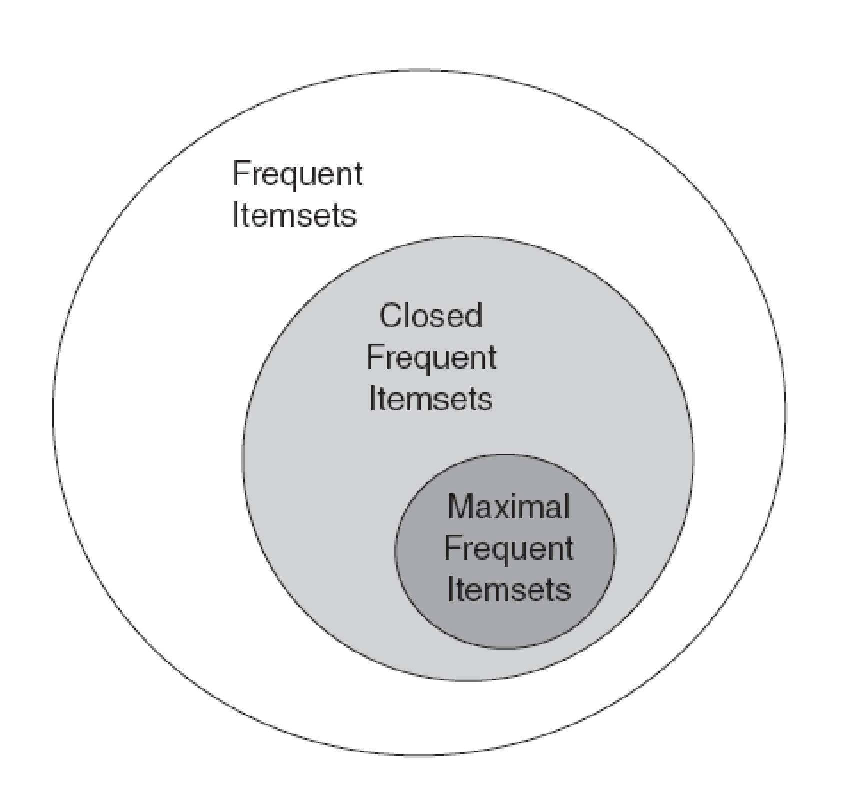
Closed frequent itemsets

- {a,c,d} has a larger support than both {a,b,d} and {a,d,e}.
- As a result, the support for {a,d} must be identical to the support for {a,c,d}.
- To find the support for a non-closed frequent itemset, the support for all of its supersets must be known.

Closed frequent itemsets

- All maximal frequent itemsets are closed.
- This is because none of the maximal frequent itemsets can have the same support count as their immediate supersets.
- The relationship among frequent, maximal frequent, and closed frequent itemsets are shown in the following figure.

Closed frequent itemsets



Evaluation of association patterns

- It is important to establish a set of well-accepted criteria for evaluating the quality of association patterns.
- An objective measure is a data-driven approach for evaluating the quality of association patterns.
- This kind of measure is usually computed based on the frequency counts tabulated in a contingency table.

Evaluation of association patterns

- We consider a contingency table for a pair of binary variables A and B.
- We use the notation $\overline{A}$($\overline{B}$) to indicate that A(B) is absent from a transaction.
- Each entry f_{ij} in this table denotes a frequency count.
- For example,
  - f_{11} is the number of times A and B appear together in the same transaction.
  - f_{01} is the number of transactions that contain B but not A.

Evaluation of association patterns

	B	$\overline{B}$	
A	f_{11}	f_{10}	f_{1+}
$\overline{A}$	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	N

Evaluation of association patterns

- The row sum f_{1+} represents the support count of A.
- The column sum f_{+1} represents the support count of B.
- N is the total number of transactions.

Evaluation of association patterns

- Existing association rule mining formulation relies on the support and confidence measures to eliminate uninteresting patterns.
- The drawback of using confidence for pattern evaluation is illustrated using the following example.

Evaluation of association patterns

- Suppose we are interested in analyzing the relationship between people who drink tea and coffee.
- We may summarize their preferences using the following contingency table.

Evaluation of association patterns

	Coffee	$\overline{\text{Coffee}}$	
Tea	150	50	200
$\overline{\text{Tea}}$	650	150	800
	800	200	1000

Evaluation of association patterns

- We can use the information in the table to evaluate the association rule {Tea}→{Coffee}.
- At first glance, it may appear that people who drink tea also tend to drink coffee.
- This is because the rule's confidence (75%) is reasonably high.

Evaluation of association patterns

- However, it is further observed that
  - The fraction of people who drink coffee, regardless of whether they drink tea, is 80%.
  - The fraction of tea drinkers who drink coffee is only 75%.
- Thus knowing that a person is a tea drinker actually decreases his/her probability of being a coffee drinker from 80% to 75%.
- The rule {Tea}$\rightarrow${Coffee} is therefore misleading despite its high confidence value.

Interest factor

- The tea-coffee example shows that high-confidence rules can sometimes be misleading.
- This is because the confidence measure ignores the support of the itemset appearing in the rule consequent.
- One way to address this problem is by applying a metric known as lift

$$Lift = \frac{c(A \rightarrow B)}{s(B)}$$

Interest factor

- This metric computes the ratio between
  - The rule's confidence and
  - The support of the itemset in the rule consequent.
- For binary variables, lift is equivalent to another objective measure called interest factor.
- The interest factor I(A,B) is defined as follows

$$I(A,B) = \frac{s(A,B)}{s(A)s(B)} = \frac{Nf_{11}}{f_{1+} f_{+1}}$$

Interest factor

- Interest factor compares the frequency of a pattern against a baseline frequency.
- This baseline frequency is computed under the statistical independence assumption.
- For a pair of mutually independent variables, we have the following relationship

$$\frac{f_{11}}{N} = \left(\frac{f_{1+}}{N}\right)\left(\frac{f_{+1}}{N}\right) \quad \text{or} \quad f_{11} = \frac{f_{1+} f_{+1}}{N}$$

Interest factor

- This equation follows from the standard approach of using simple fractions as estimates of probabilities.
- The fraction f_{11}/N is an estimate of the joint probability $P(A,B)$.
- The fractions f_{1+}/N and f_{+1}/N are the estimates of $P(A)$ and $P(B)$ respectively.
- If A and B are independent, then $P(A,B)=P(A)P(B)$.

Interest factor

- We can interpret the interest factor as follows:
  - If A and B are independent, then $I(A,B)=1$.
  - If A and B are positively correlated, then $I(A,B)>1$.
  - If A and B are negatively correlated, then $I(A,B)<1$.

Interest factor

- For the tea-coffee example,

$$I = \frac{(1000)(150)}{(200)(800)} = 0.9375$$

- This suggests a slight negative correlation between tea drinkers and coffee drinkers.