Ανάλυση Συσχετίσεων
Βασικές έννοιες και αλγόριθμοι

Ανδρέας Λ. Συμεωνίδης
Αν. Καθηγητής
Τμήμα Ηλεκτρολόγων Μηχ/κών
&
Μηχ/κών Υπολογιστών, Α.Π.Θ.
Email: asymeon@eng.auth.gr

Διάρθρωση διάλεξης

- Ανάλυση συσχετίσεων
  - Κανόνες συσχέτισης – Βασικές έννοιες
  - Εξόρυξη κανόνων συσχέτισης
  - Το frequent itemset - Υποσύνολα
  - Ο αλγόριθμος Apriori
  - Maximal Frequent Itemsets
  - Δημιουργία κανόνων
  - Μετρικές Αξιολόγησης κανόνων

Εξόρυξη Κανόνων Συσχέτισης – Association Rule Mining

- Δεδομένου ενός σετ συναλλαγών (transactions), βρες τους κανόνες που προβλέπουν την ύπαρξη ενός αντικειμένου, δεδομένης της ύπαρξης άλλων αντικειμένων στη συναλλαγή.

Συναλλαγές καλαθιού αγοράς
(Market-Basket transactions)

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

Παράδειγμα κανόνων συσχέτισης

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

Η επαγωγή σημαίνει συνύπαρξη (co-occurrence), όχι αιτιότητα (causality)!

Ορισμός: Συχνό σύνολο αντικειμένων (Frequent Itemset)

- **Itemset**
  - Συλλογή από ένα ή περισσότερα αντικείμενα
    - Παράδειγμα: {Milk, Bread, Diaper}
  - k-itemset
    - Ένα itemset που περιέχει k αντικείμενα
- **Support count (σ)**
  - Συχνότητα εμφανίσεων ενός itemset
  - Π.χ. σ({Milk, Bread,Diaper}) = 2
- **Support**
  - Κλάσμα συναλλαγών που περιέχουν το itemset
  - Π.χ. s({Milk, Bread, Diaper}) = 2/5

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

Frequent Itemset:
Ένα itemset του οποίου το support είναι μεγαλύτερο ή ίσο από ένα κατώφλι **minsup.**

Ορισμός: Κανόνας συσχέτισης (Association rule)

- Association Rule
  - Μία συνεπαγωγή τύπου $X \rightarrow Y$, όπου X και Y είναι itemsets
  - Παράδειγμα:
    {Milk, Diaper} $\rightarrow$ {Beer}

- Μετρικές αξιολόγησης κανόνα
  - Support (s)
    - Κλάσμα συναλλαγών που περιέχουν και το X και το Y
  - Confidence (c)
    - Μετρά πόσο συχνά εμφανίζονται αντικείμενα του Y σε μία συναλλαγή που περιέχει το X

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

Παράδειγμα:

$$\{\text{Milk}, \text{Diaper}\} \Rightarrow \text{Beer}$$

$$s = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{|\text{T}|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{\sigma(\text{Milk}, \text{Diaper})} = \frac{2}{3} = 0.67$$

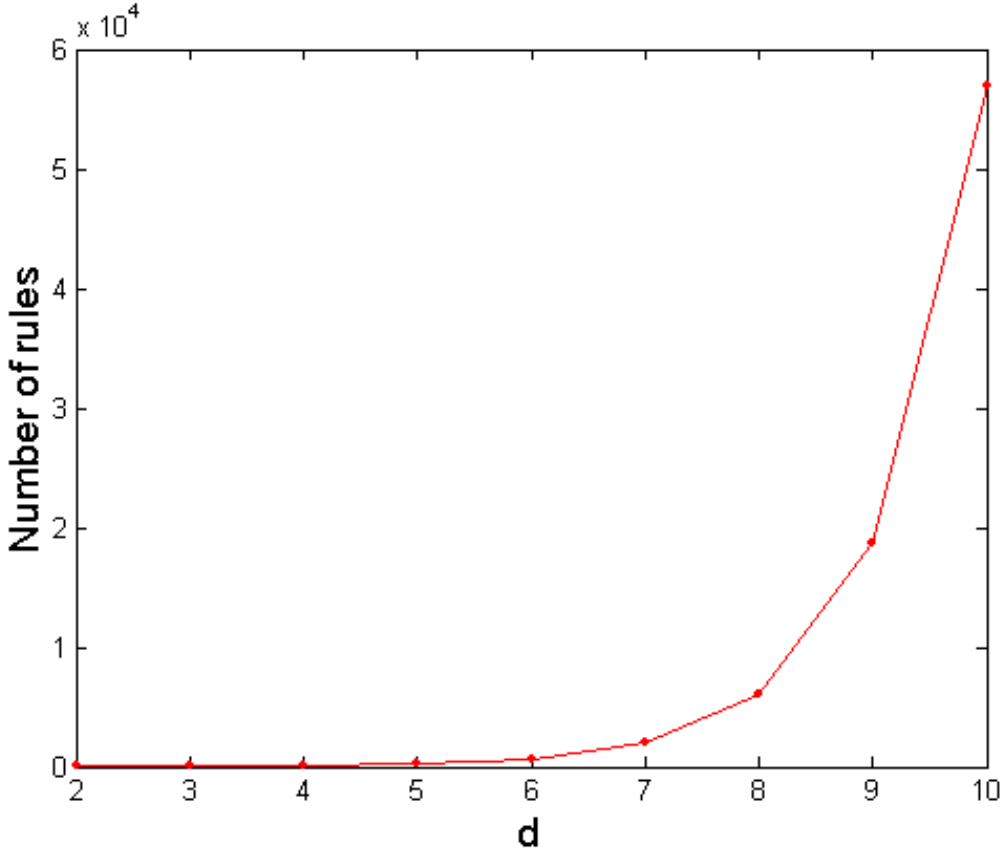
Εξόρυξης Κανόνων Συσχέτισης

- Δεδομένου ενός σετ συναλλαγών T, ο στόχος της εξόρυξης κανόνων συσχέτισης είναι να βρει όλους τους κανόνες που έχουν:
  - support ≥ ***minsup*** threshold
  - confidence ≥ ***minconf*** threshold
- Brute-force προσέγγιση:
  - Κατέγραψε όλους τους πιθανούς κανόνες συσχέτισης
  - Υπολόγισε τα support και confidence για κάθε κανόνα
  - Διέγραψε τους κανόνες που δεν επιτυγχάνουν να περάσουν τα κατώφλια ***minsup*** και ***minconf***

  ⇒ Υπολογιστικά απαγορευτικό!

Υπολογιστική πολυπλοκότητα

- Δεδομένων d μοναδικών αντικειμένων:
  - Συνολικός αριθμός itemsets = 2^d
  - Συνολικός αριθμός πιθανών association rules:



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

$$= 3^d - 2^{d+1} + 1$$

Αν d=6, R = 602 κανόνες

Εξόρυξη Κανόνων Συσχέτισης

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

Παραδείγματα κανόνων:

{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)

Παρατηρήσεις:

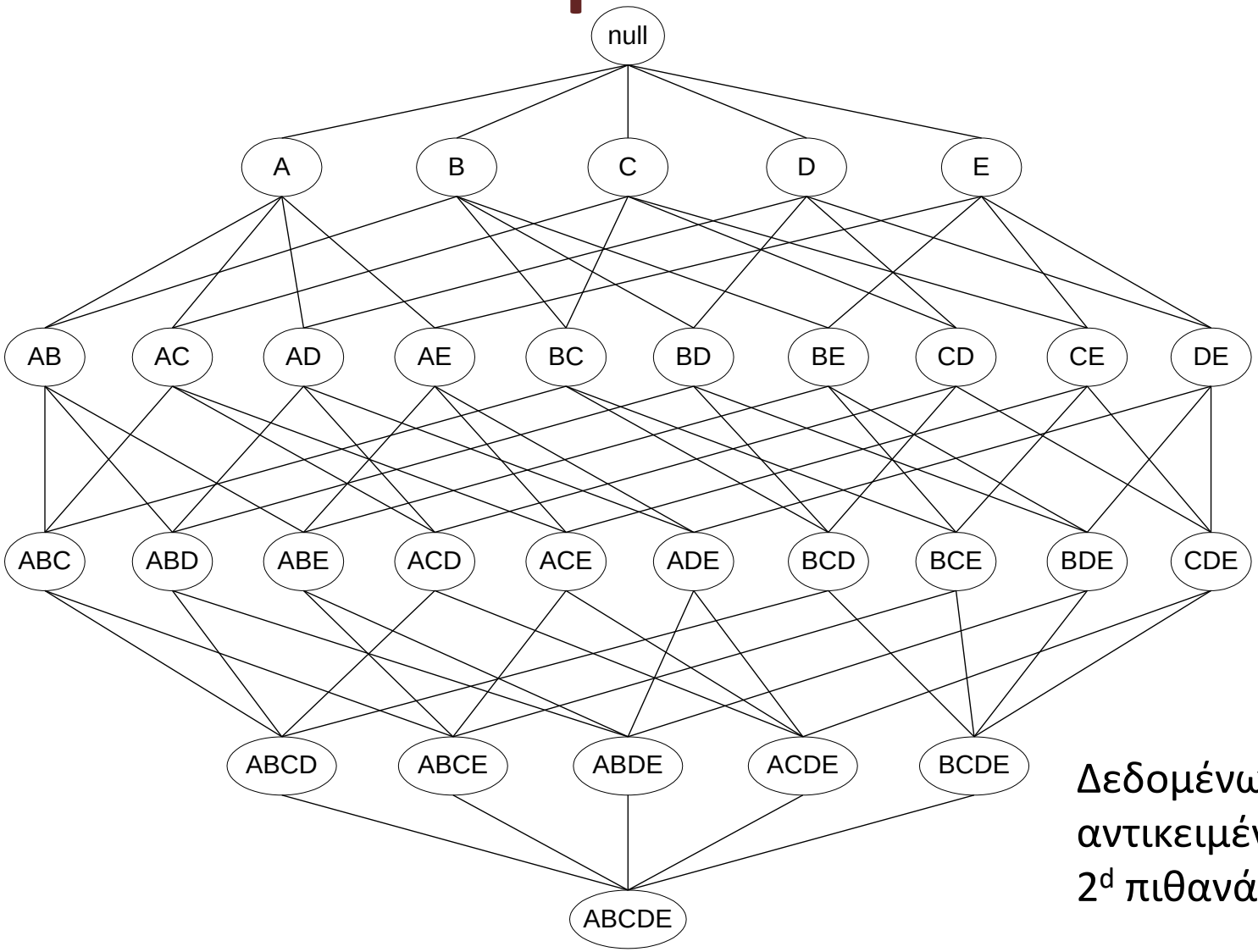
- Όλοι οι παραπάνω κανόνες είναι δυαδικές κατατμήσεις του ίδιου itemset: {Milk, Diaper, Beer}
- Οι κανόνες που εξάγονται από το ίδιο itemset έχουν ίδιο support αλλά μπορεί να έχουν διαφορετικό confidence
- Μπορούμε έτσι να βλέπουμε ξεχωριστά τις απαιτήσεις support και confidence

Εξόρυξη Κανόνων Συσχέτισης

- Προσέγγιση δύο βημάτων:
  1. Frequent Itemset Generation
     - Δημιούργησε όλα τα itemsets με support $\geq$ minsup
  2. Rule Generation
     - Δημιούργησε κανόνες υψηλού confidence από κάθε frequent itemset, όπου κάθε κανόνας είναι δυαδική κατάτμηση ενός frequent itemset
- Η δημιουργία frequent itemsets εξακολουθεί να είναι υπολογιστικά ακριβή

Τεχνικές εξόρυξης frequent itemsets

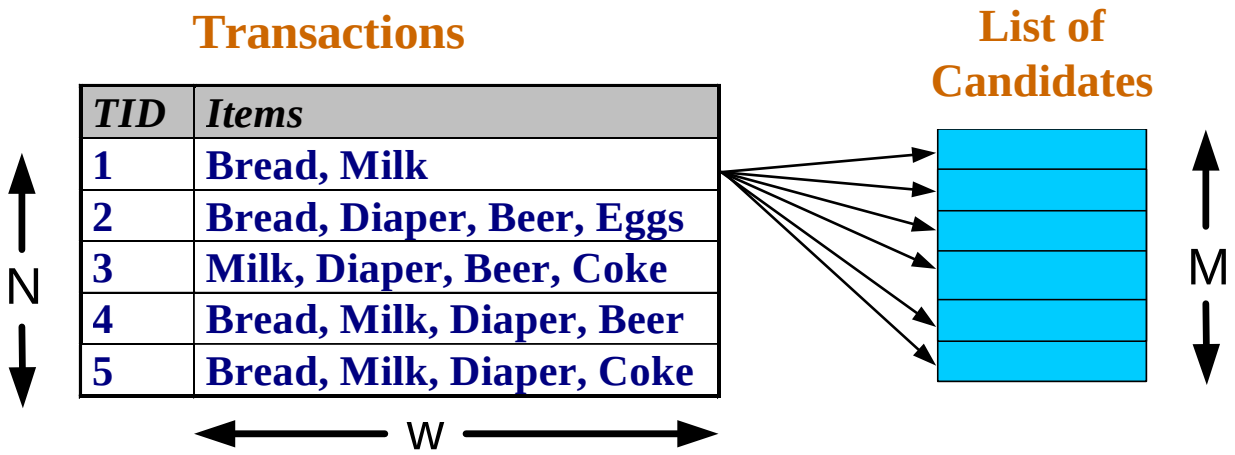
Frequent Itemsets



Δεδομένων d αντικειμένων, υπάρχουν 2^d πιθανά itemsets

Δημιουργία Frequent Itemset

- Προσέγγιση Brute-force:
  - Κάθε itemset στο lattice είναι υποψήφιο (candidate) frequent itemset
  - Υπολόγισε το support κάθε υποψηφίου σκανάροντας τη βάση δεδομένων



  - Συνέκρινε κάθε συναλλαγή και βρες τους υποψηφίους με τους οποίους ταιριάζει
  - Πολυπλοκότητα ~ O(NMw) => Ακριβό αφού M = 2^d !!!

Στρατηγικές Δημιουργίας Frequent Itemsets

- Μείωση του αριθμού των υποψηφίων (M)
  - Συνολική αναζήτηση: $M=2^d$
  - Χρήση τεχνικών κλαδέματος για να μειωθεί το M
- Μείωση του αριθμού των συγκρίσεων (NM)
  - Χρήση αποδοτικών δομών δεδομένων για την αποθήκευση των υποψηφίων ή των συναλλαγών
  - Μη απαίτηση σύγκρισης κάθε υποψηφίου με κάθε συναλλαγή

Μείωση του αριθμού των υποψηφίων

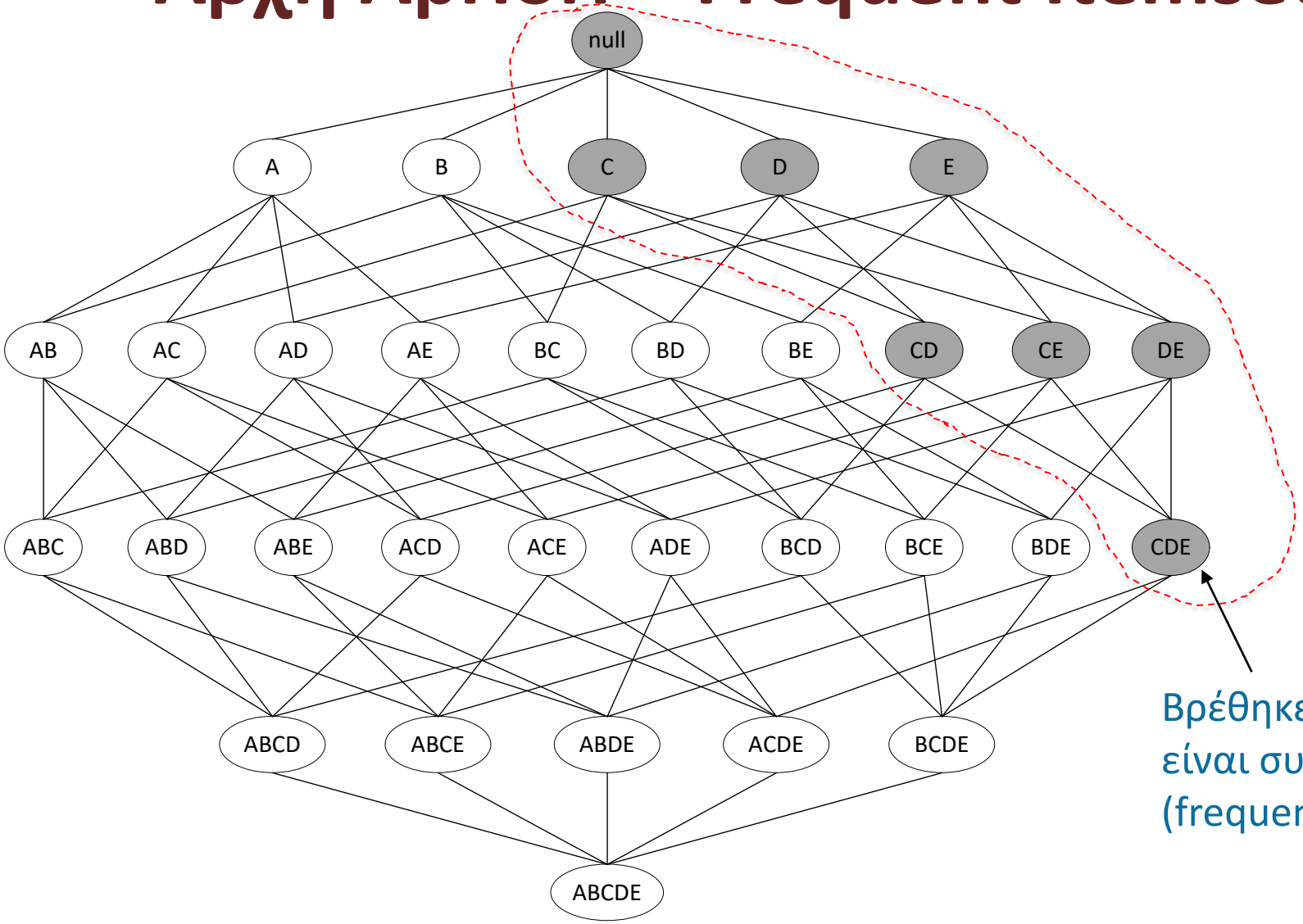
- Αρχή Apriori:
  - Αν ένα itemset είναι συχνό, τότε και όλα τα υποσύνολά του θα είναι συχνά

- Η αρχή Apriori ισχύει εξαιτίας της ακόλουθης ιδιότητας του support:

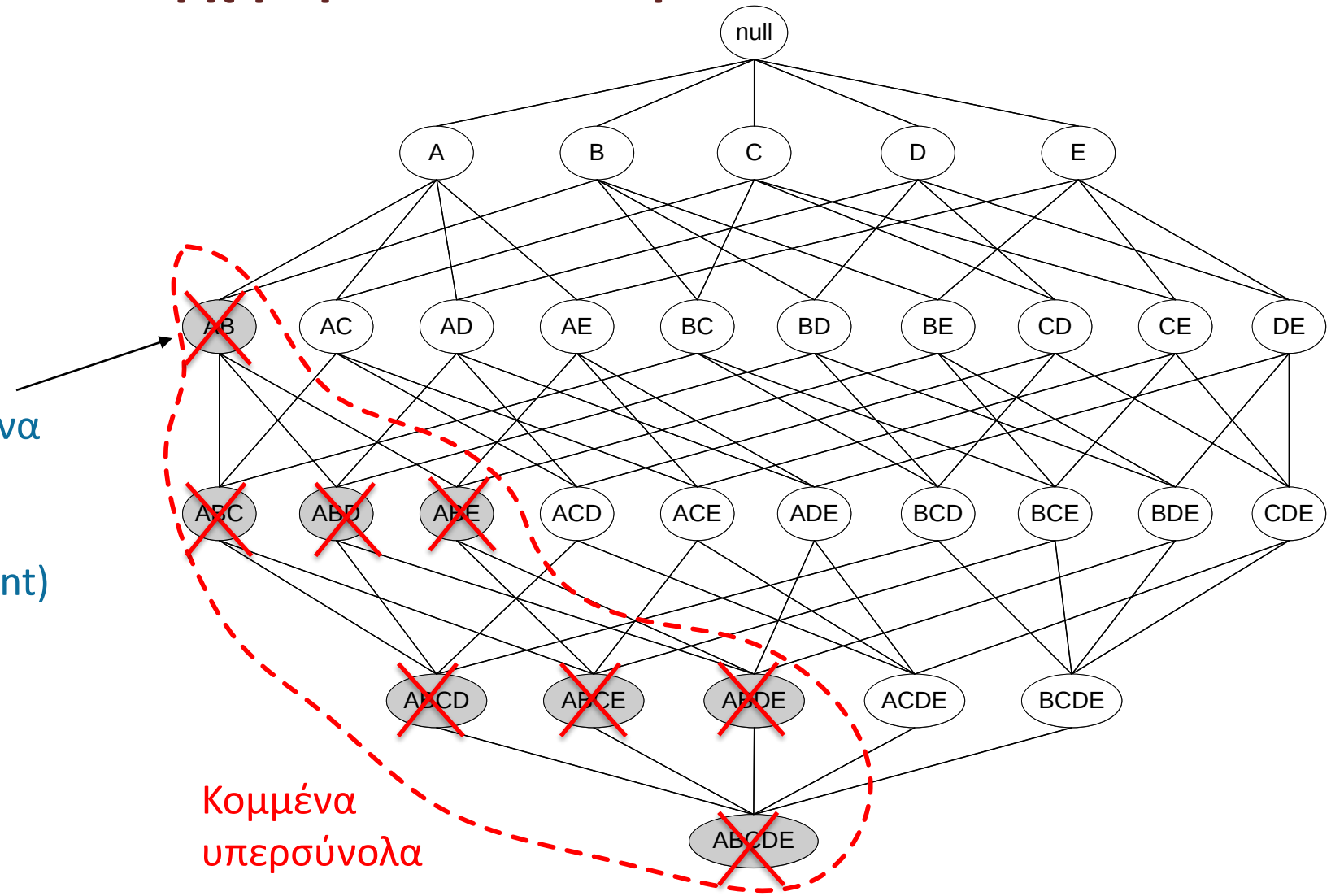
$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

  - Το support ενός itemset ποτέ δεν ξεπερνά το support των υποσυνόλων
  - Γνωστή και ως anti-monotone ιδιότητα του support

Αρχή Apriori – Frequent Itemset



Αρχή Apriori – Infrequent Itemset



Βρέθηκε να είναι όχι συχνό (infrequent)

Κομμένα υπερσύνολα

Αρχή Apriori - Παράδειγμα

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)

(Δε χρειάζεται η δημιουργία υποψηφίων που να περιέχουν Coke ή Eggs)

Minimum Support = 3

Triplets (3-itemsets)

Itemset	Count
{Bread,Milk,Diaper}	**3**

Brute force

$^6C_1 + {}^6C_2 + {}^6C_3 = 41$

Support-based pruning,

6 + 6 + 1 = 13

68% μείωση

Αλγόριθμος Apriori

- Μέθοδος:

  Let k=1

  – Δημιούργησε frequent itemsets μήκους 1

  – Επανέλαβε μέχρι να μη βρεθεί κανένα νέο frequent itemset

  1. Δημιούργησε υποψήφιους μήκους k+1 από frequent itemsets μήκους k
  2. Κόψε υποψήφια itemsets που περιέχουν υποσύνολα μήκους k που είναι infrequent
  3. Μέτρησε το support για κάθε υποψήφιο σκανάροντας τη βάση δεδομένων
  4. Απάλειψε τους υποψηφίους που είναι infrequent, αφήνοντας μόνο αυτούς που είναι frequent

Δημιουργία υποψηφίων – Brute Force

1

Item
Beer
Bread
Cola
Diapers
Milk
Eggs

Itemset
{Beer, Bread, Cola}
{Beer, Bread, Diapers}
{Beer, Bread, Milk}
{Beer, Bread, Eggs}
{Beer, Cola, Diapers}
…
{Diapers, Milk, Eggs}

Candidate Pruning

Itemset
{Beer, Diapers, Milk}

1 Δημιουργία υποψηφίων – F_{k-1} x F_1

Itemset
{Beer, Diapers}
{Bread, Diapers}
{Bread, Milk}
{Diapers, Milk}

Item
Beer
Bread
Diapers
Milk

Itemset
{Beer, Diapers, Bread}
{Beer, Diapers, Milk}
{Bread, Diapers, Milk}
{Bread, Milk, Beer}

Candidate Pruning

Itemset
{Beer, Diapers, Milk}

1 Δημιουργία υποψηφίων – F_{k-1} x F_{k-1}

Itemset
{Beer, Diapers}
{Bread, Diapers}
{Bread, Milk}
{Diapers, Milk}

Itemset
{Beer, Diapers}
{Bread, Diapers}
{Bread, Milk}
{Diapers, Milk}

Itemset
{Beer, Diapers, Milk}

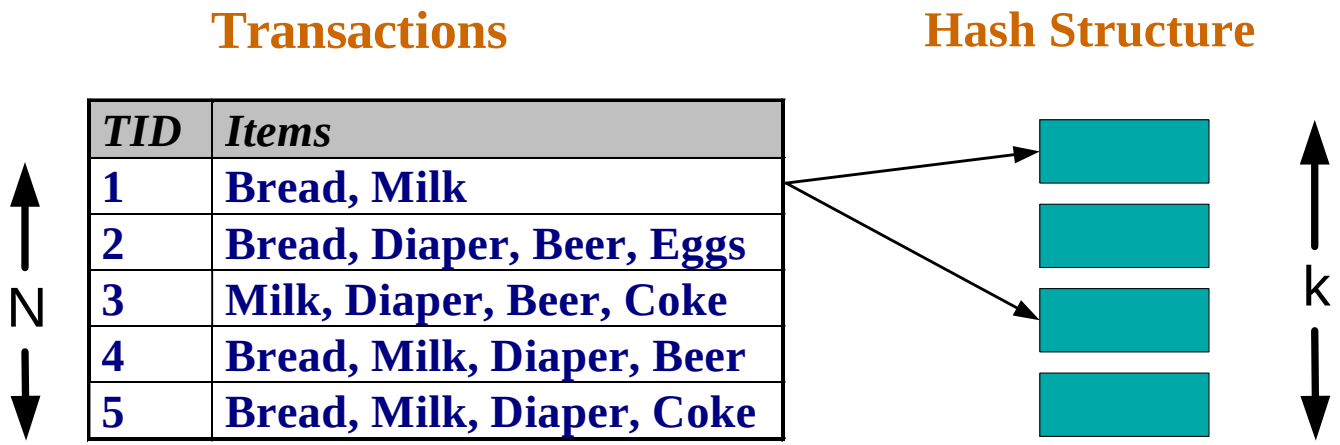
Candidate Pruning

Itemset
{Beer, Diapers, Milk}

Συγχωνεύει μόνο εάν τα πρώτα k-2 αντικείμενα είναι όμοια

Μείωση αριθμού συγκρίσεων

- Μέτρημα υποψηφίων:
  - Σκάναρε τη βάση συναλλαγών για τον υπολογισμό του support κάθε υποψηφίου itemset
  - Για να μειώσεις τον αριθμό των συγκρίσεων, αποθήκευσε τους υποψηφίους σε μία δομή κατακερματισμού (hash structure)
    - Αντί να ταιριάζεις κάθε συναλλαγή με κάθε υποψήφιο, ταίριαξε την με τους υποψηφίους στους κάδους κατακερματισμού



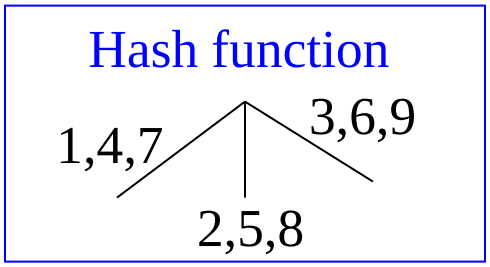
Δημιουργία δένδρου κατακερματισμού

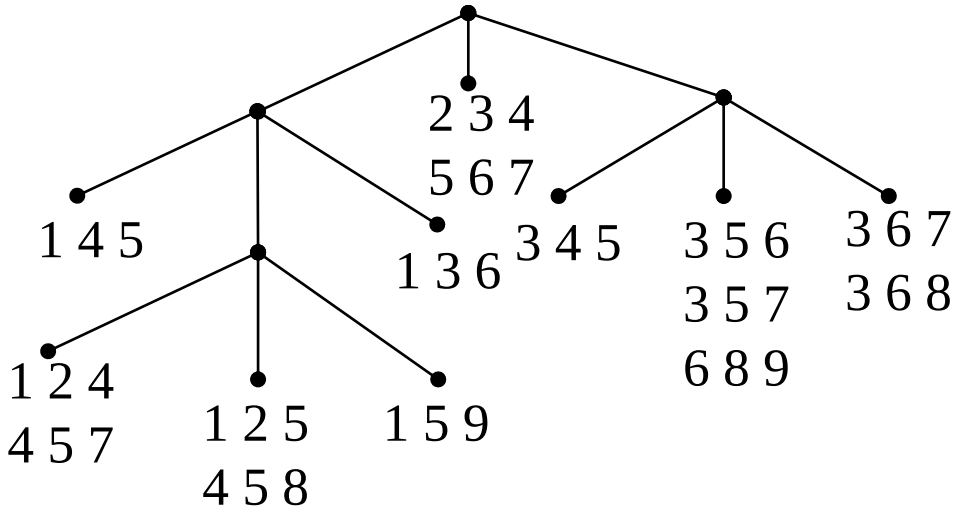
Έστω ότι έχουμε 15 υποψήφια itemsets μήκους 3:

{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}

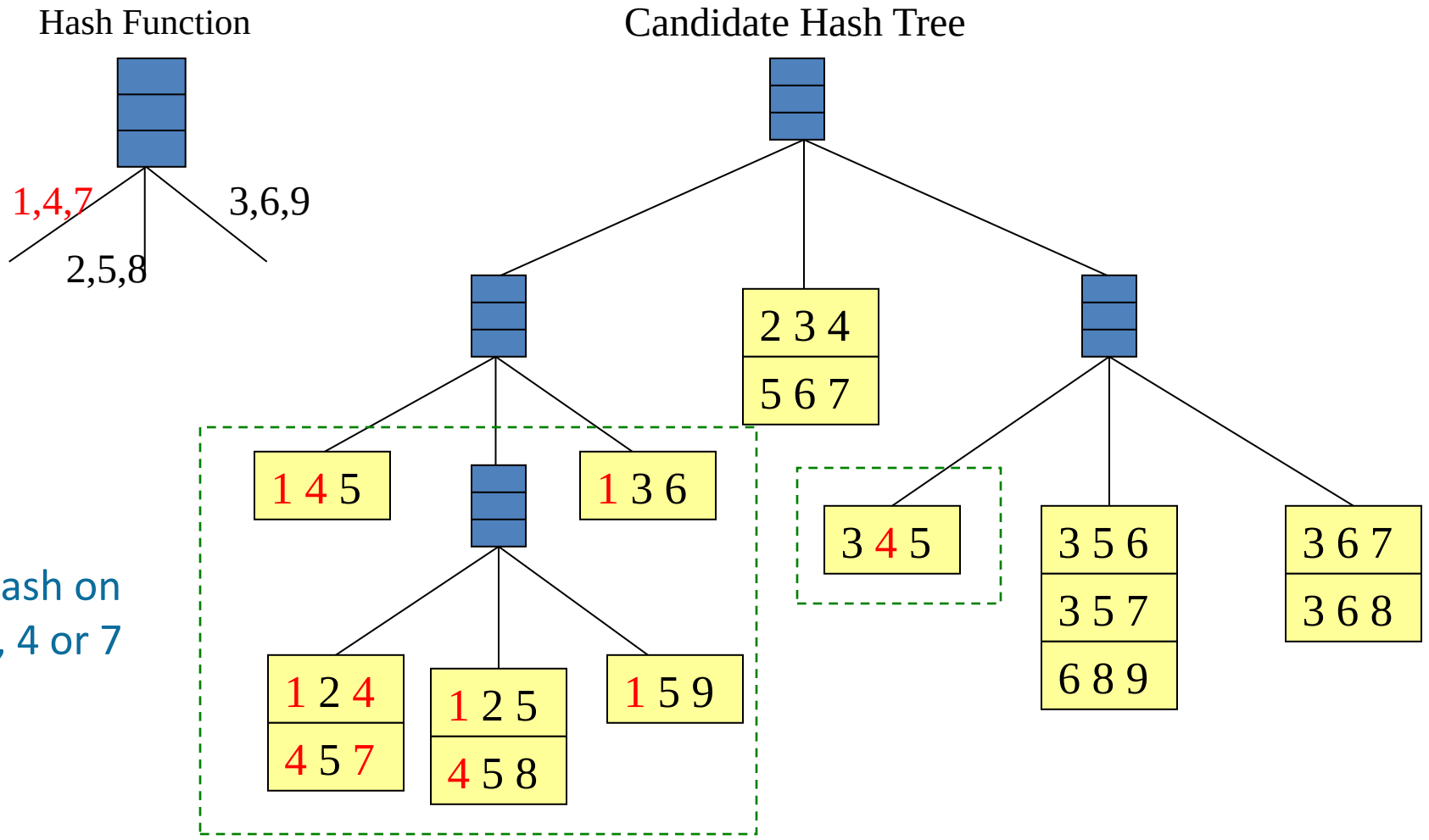
Χρειάζεται:

- Hash function
- Max leaf size: μέγιστος αριθμός itemsets που αποθηκεύονται σε ένα φύλλο (εάν ο αριθμός των υποψηφίων itemsets ξεπεράσει το max leaf size, διαχώρισε τον κόμβο)

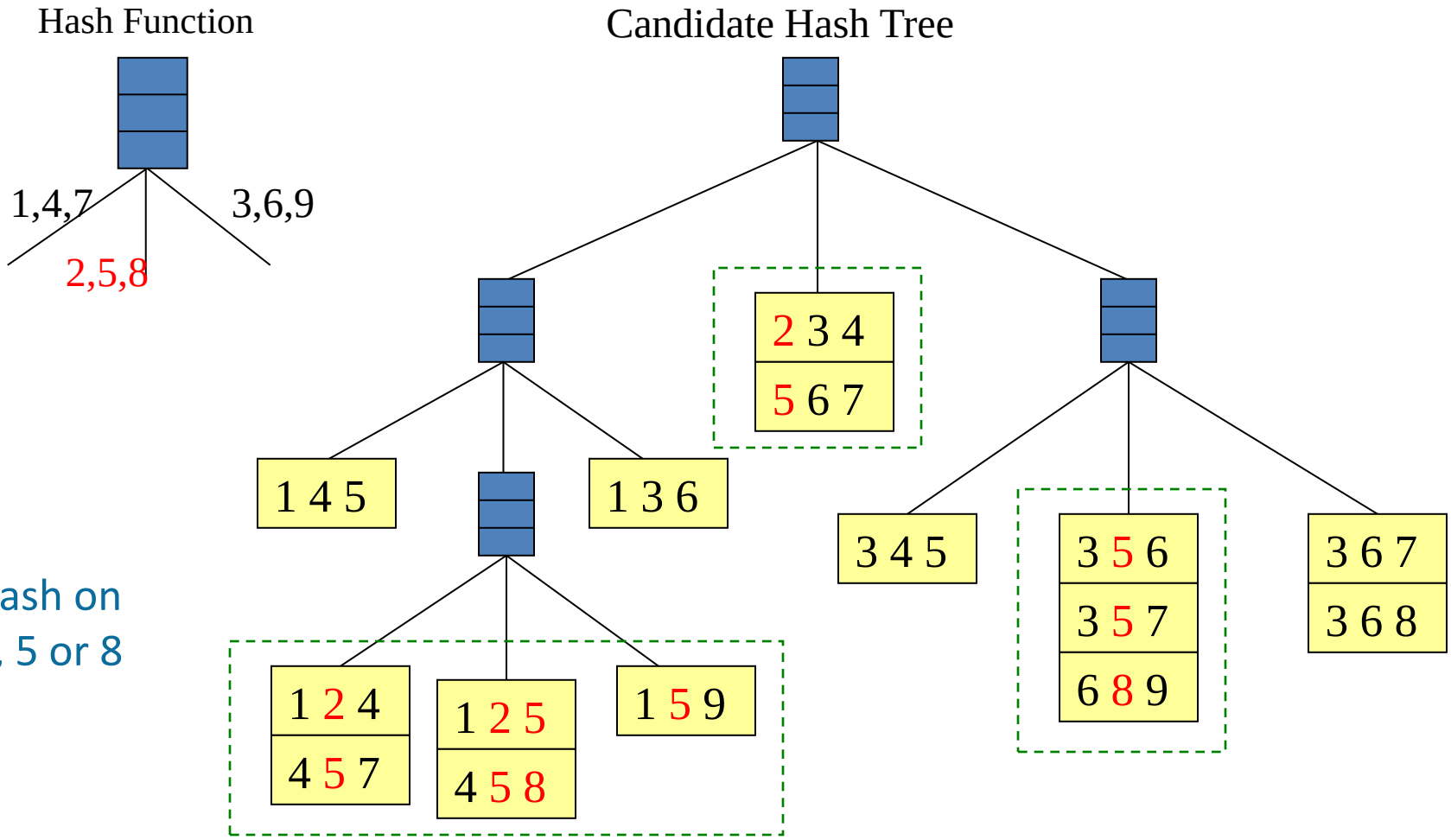




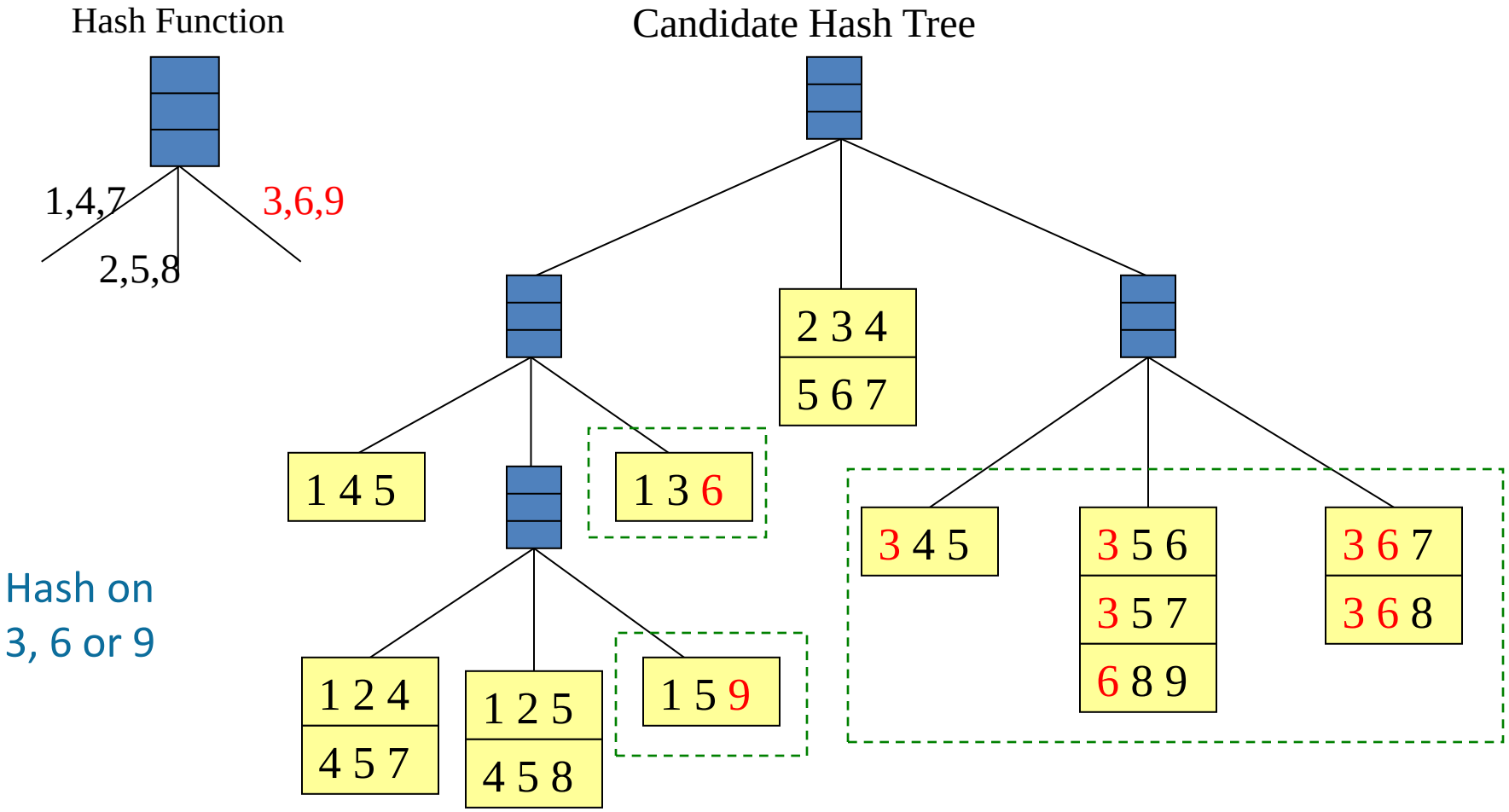
Δένδρο κατακερματισμού



Δένδρο κατακερματισμού

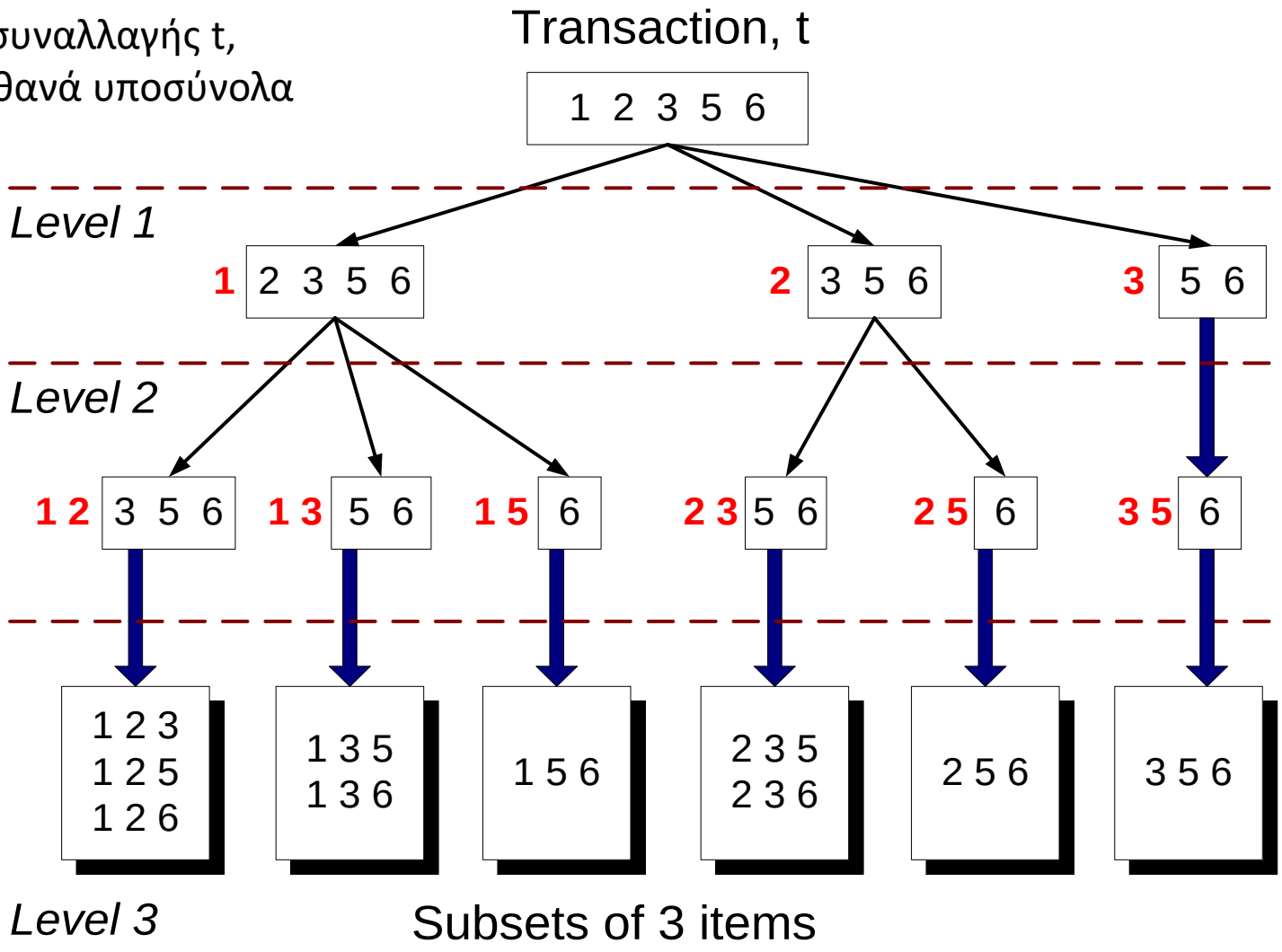


Δένδρο κατακερματισμού

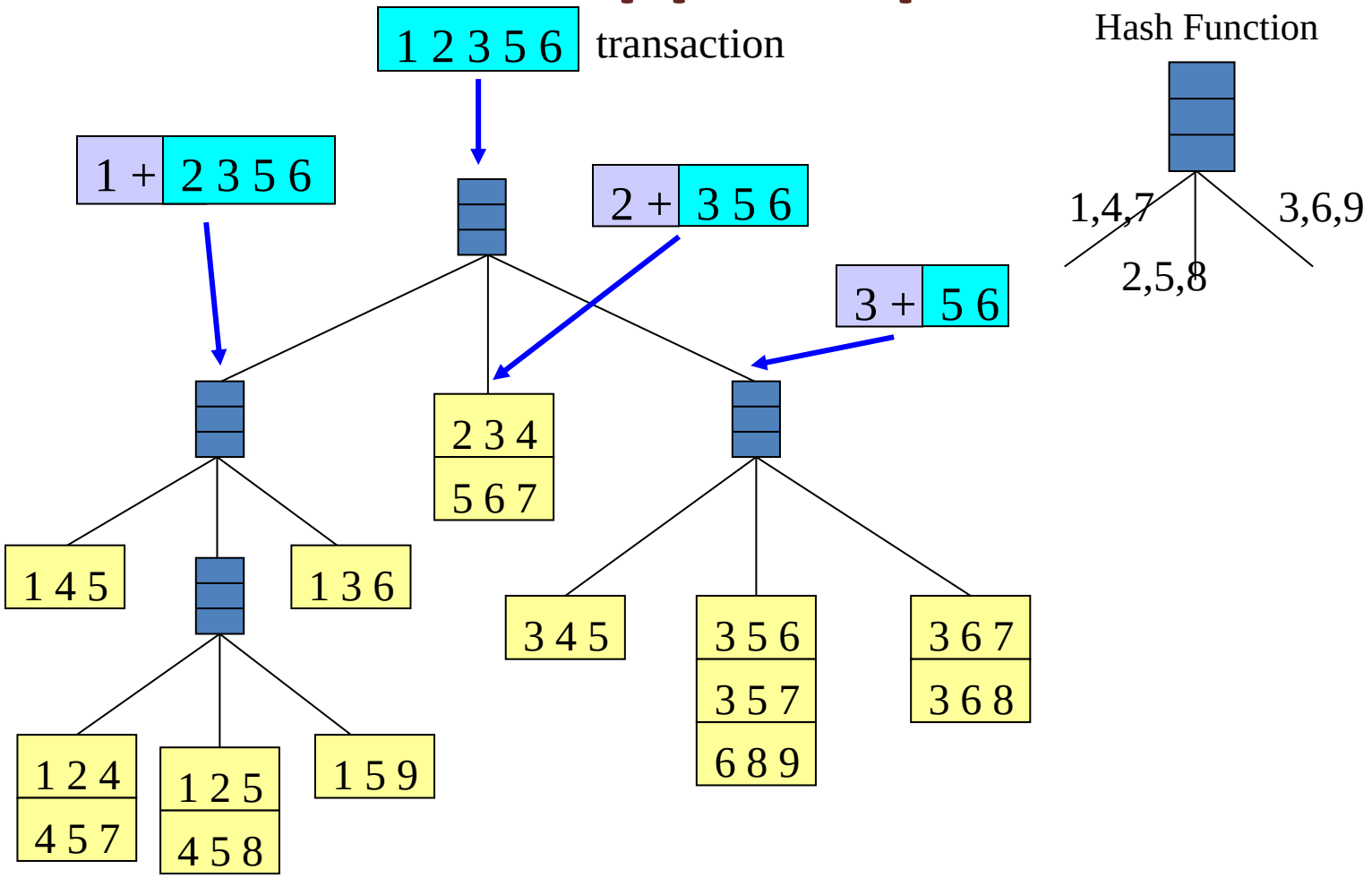


Υποσύνολα

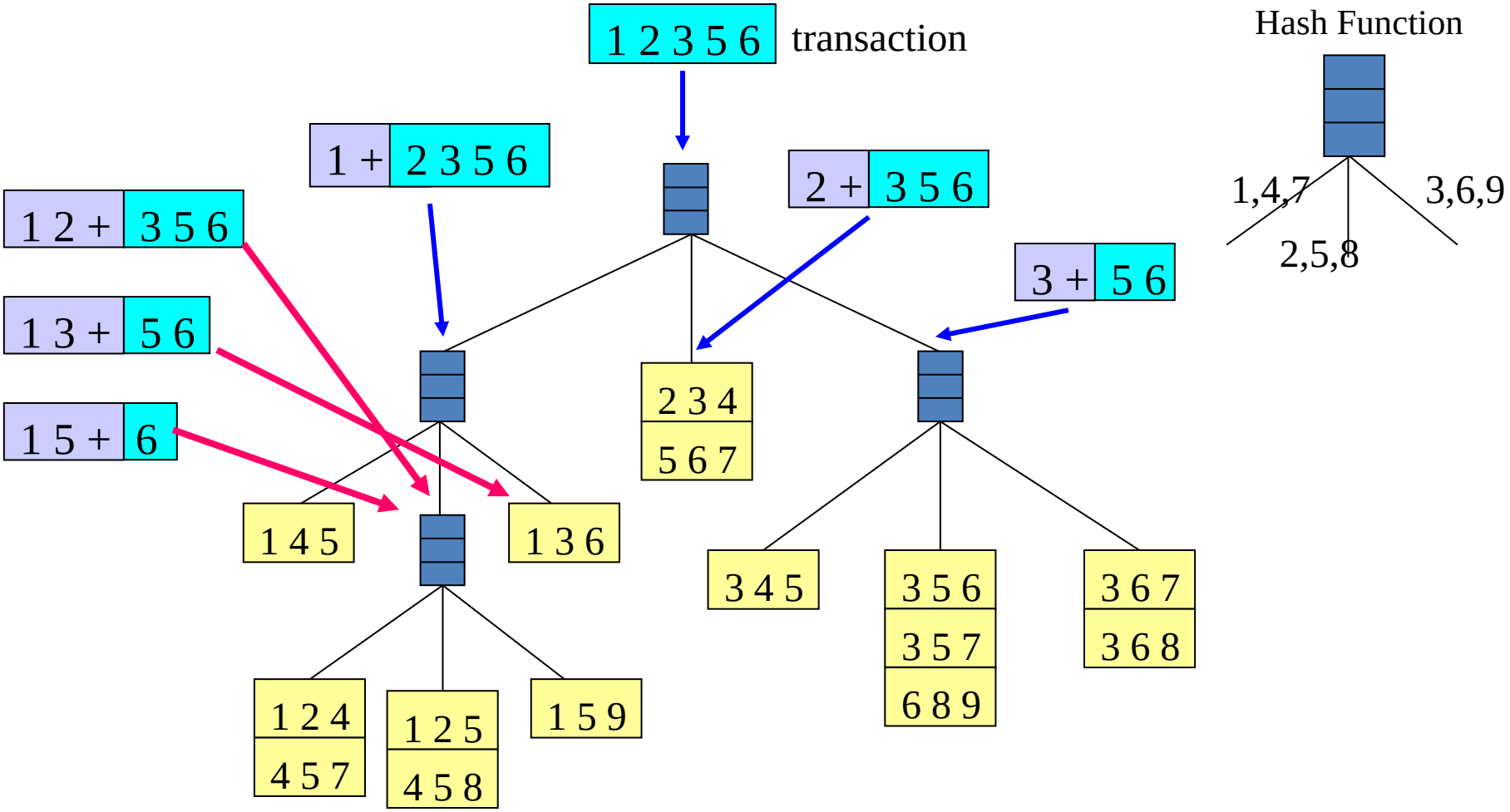
Δεδομένης μία συναλλαγής t, ποια είναι τα πιθανά υποσύνολα μήκους 3;



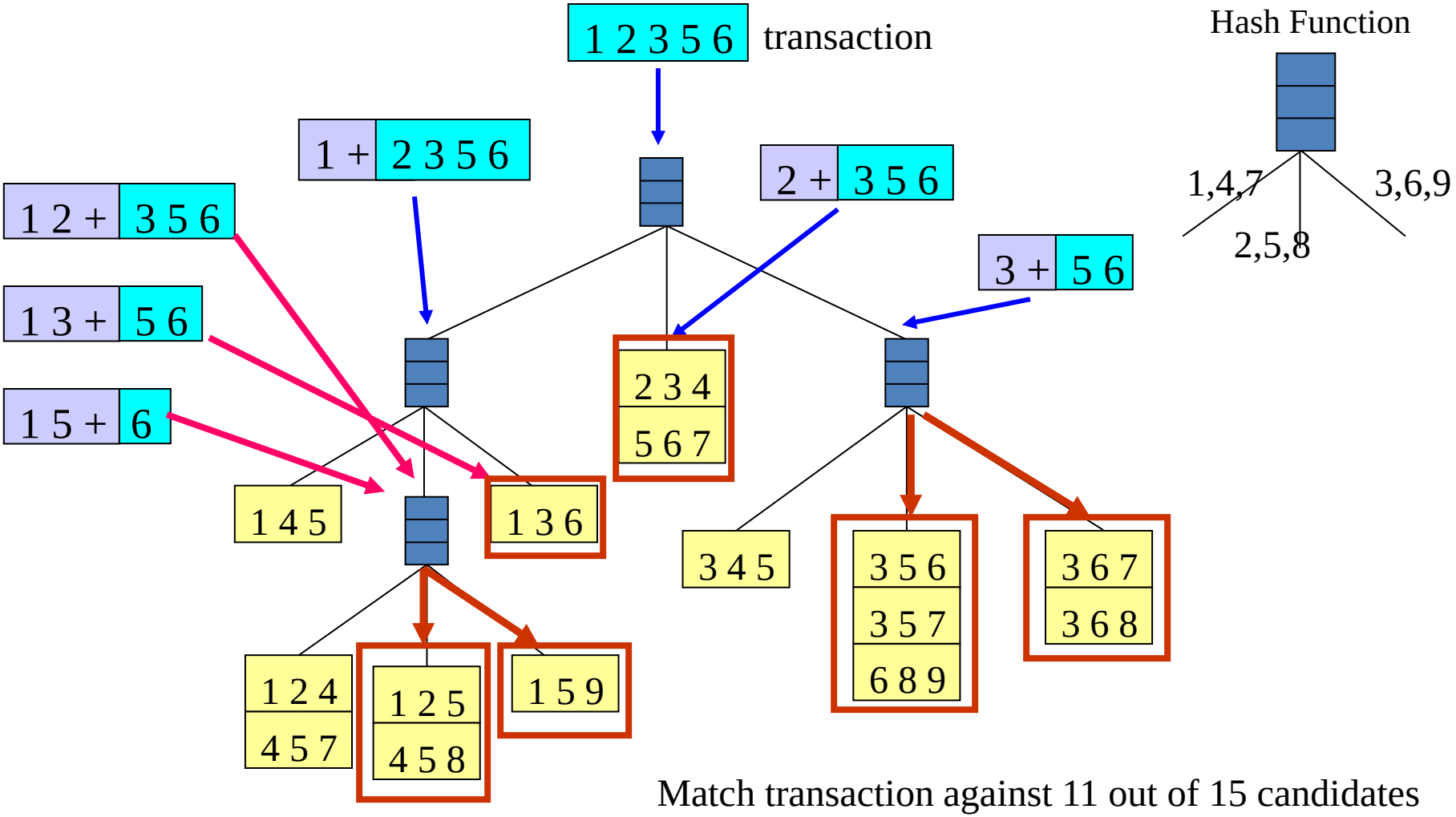
Υποσύνολα με τη χρήση δένδρου κατακερματισμού



Υποσύνολα με τη χρήση δένδρου κατακερματισμού



Υποσύνολα με τη χρήση δένδρου κατακερματισμού



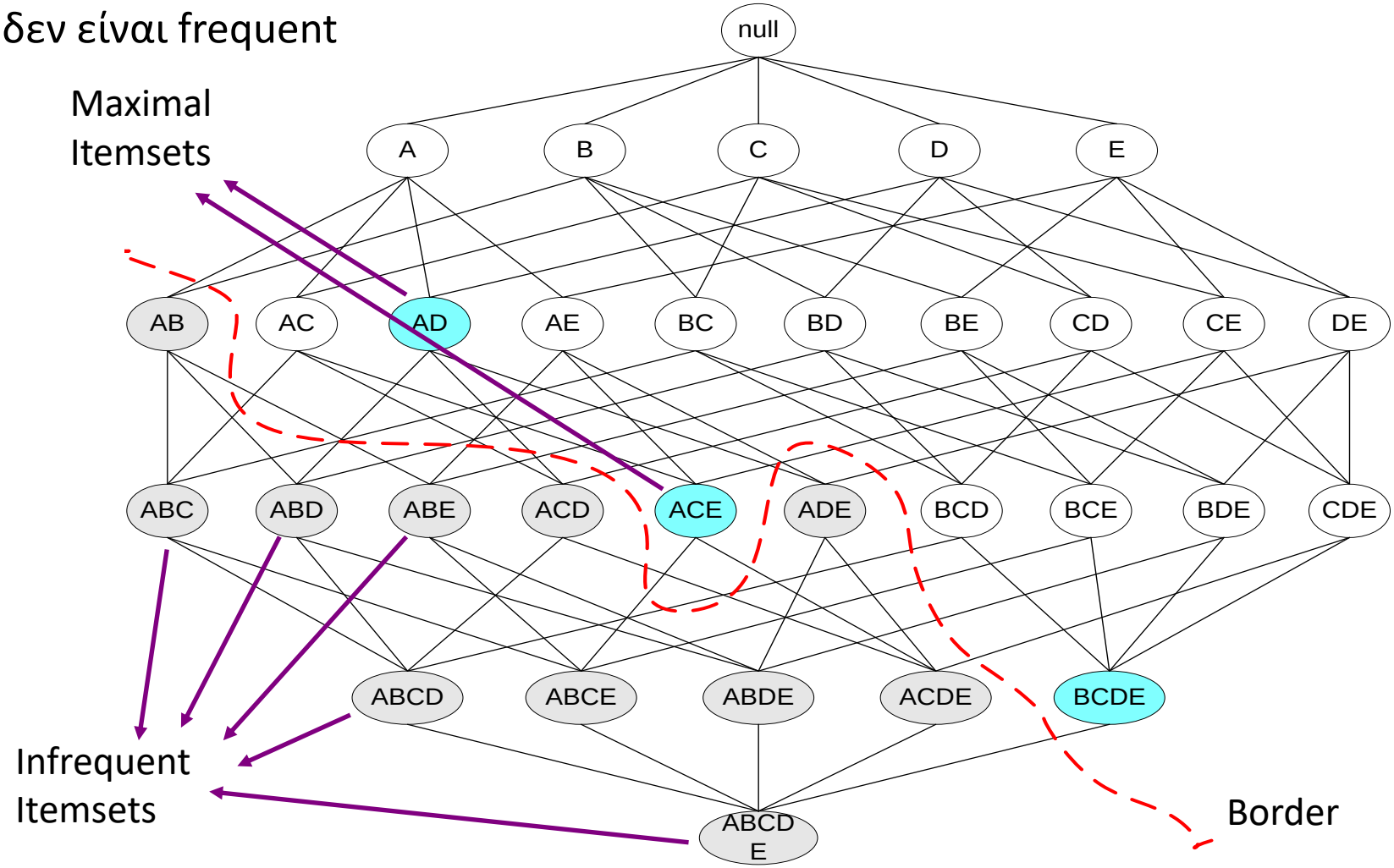
Match transaction against 11 out of 15 candidates

Παράγοντες που επηρεάζουν την πολυπλοκότητα

- **Επιλογή minimum support threshold**
  - χαμηλές τιμές → περισσότερα frequent itemsets
  - μπορεί να αυξήσει τον αριθμό των υποψηφίων και το μήκος των frequent itemsets
- **Διαστατικότητα (dimensionality - number of items) των δεδομένων**
  - χρειάζεται περισσότερος χώρος για την αποθήκευση του support count κάθε αντικειμένου
  - Αν αυξηθεί και ο αριθμός των frequent items τόσο θα αυξηθεί και το υπολογιστικό κόστος και το I/O
- **Μέγεθος βάσης**
  - Ο χρόνος εκτέλεσης αυξάνεται με τον αριθμό των συναλλαγών καθώς ο Apriori κάνει πολλαπλές διελεύσεις σε όλες τις συναλλαγές
- **Μέσο πλάτος συναλλαγής**
  - το πλάτος της συναλλαγής αυξάνει με πυκνά σετ δεδομένων
  - μπορεί να αυξήσει το μέγιστο μήκος των frequent itemsets και τον αριθμό διασχίσεων του δένδρου κατακερματισμού

Maximal Frequent Itemset

Ένα itemset είναι maximal frequent εάν κανένα από τα άμεσα υπερσύνολα δεν είναι frequent



Χρήση Maximal Frequent Itemset

- Το μικρότερο σετ itemsets από το οποίο μπορούν να προκύψουν frequent itemsets.
- Από maximal itemsets {AD, ACE, BCDE} μπορούν να προκύψουν:
  - Frequent itemsets που ξεκινάνε με a, c, d ή e
  - Frequent itemsets που ξεκινάνε με b, c, d ή e
- Δε μεταδίδουν πληροφορία support

Closed Itemset

- Ένα itemset X είναι is closed εάν κανένα από τα άμεσα υπερσύνολα δεν έχει το ίδιο support με το X

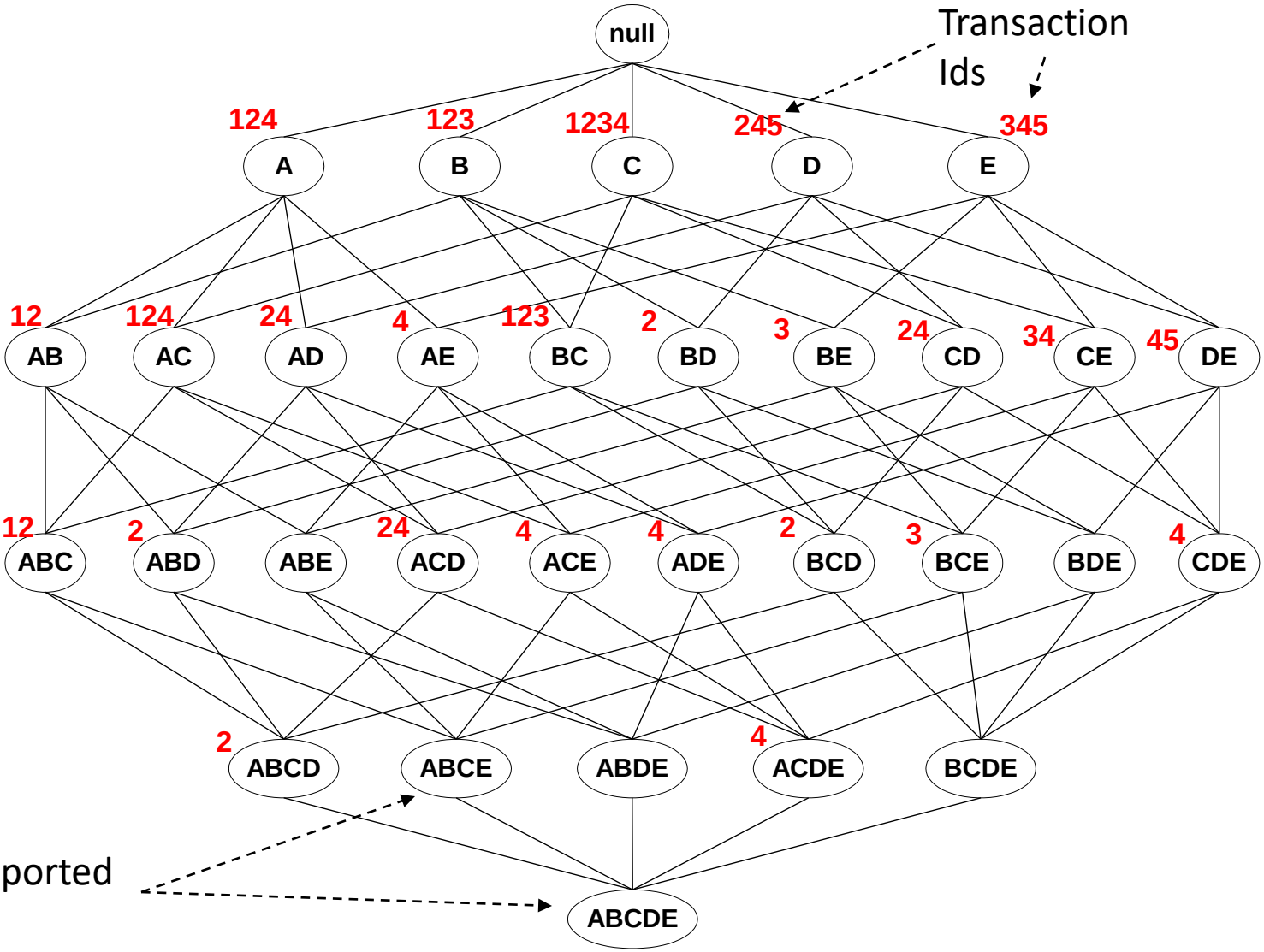
TID	Items
1	{A,B}
2	{B,C,D}
3	{A,B,C,D}
4	{A,B,D}
5	{A,B,C,D}

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

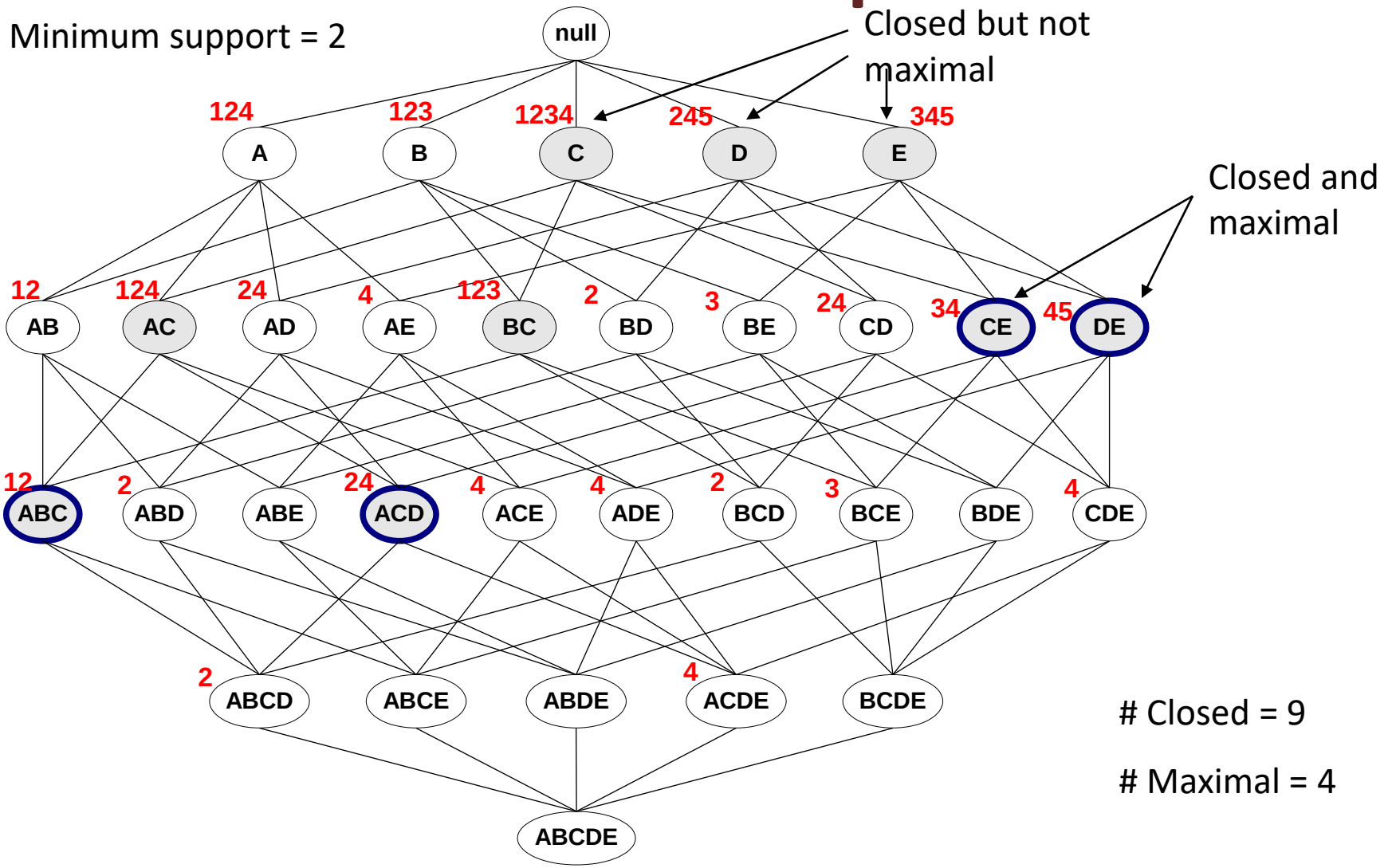
Itemset	Support
{A,B,C}	2
{A,B,D}	3
{A,C,D}	2
{B,C,D}	3
{A,B,C,D}	2

Maximal vs Closed Itemsets

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



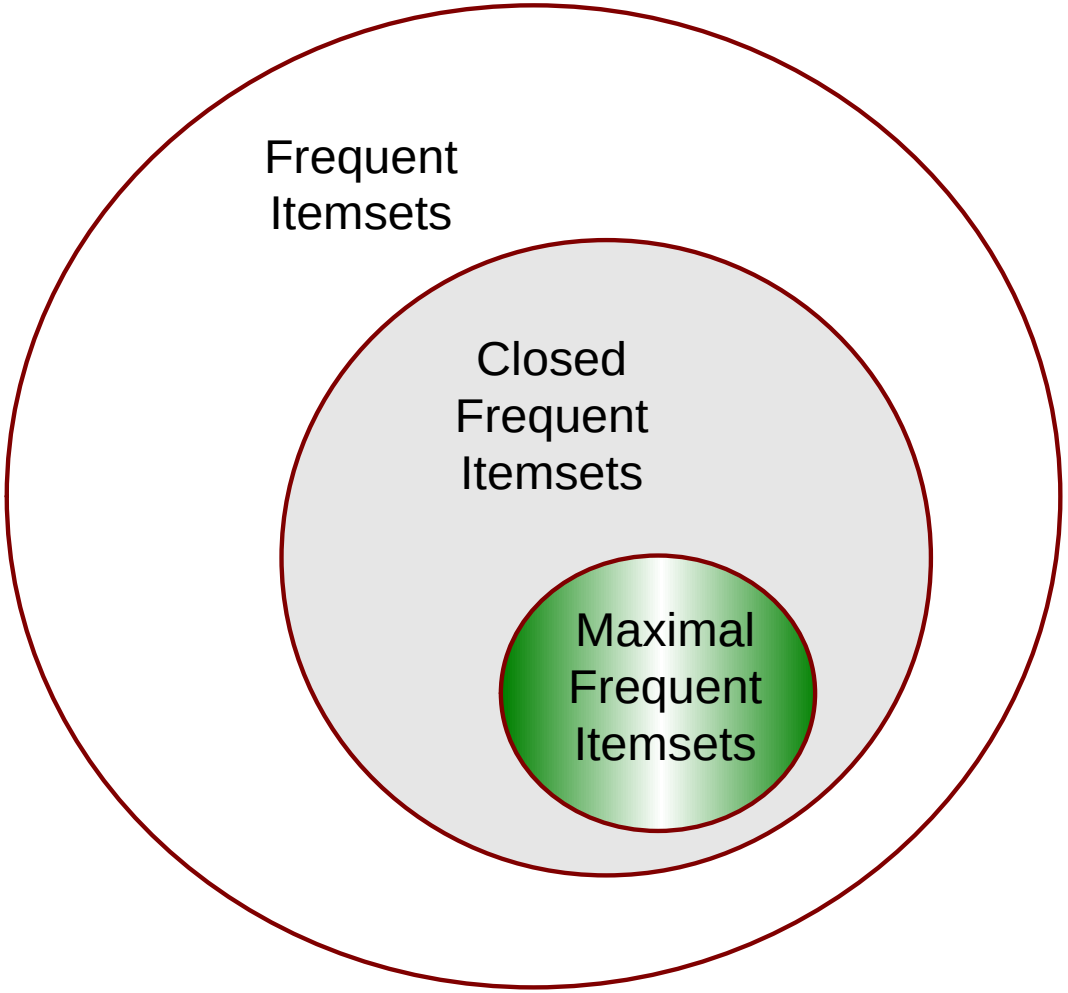
Maximal vs Closed Frequent Itemsets



Closed = 9

Maximal = 4

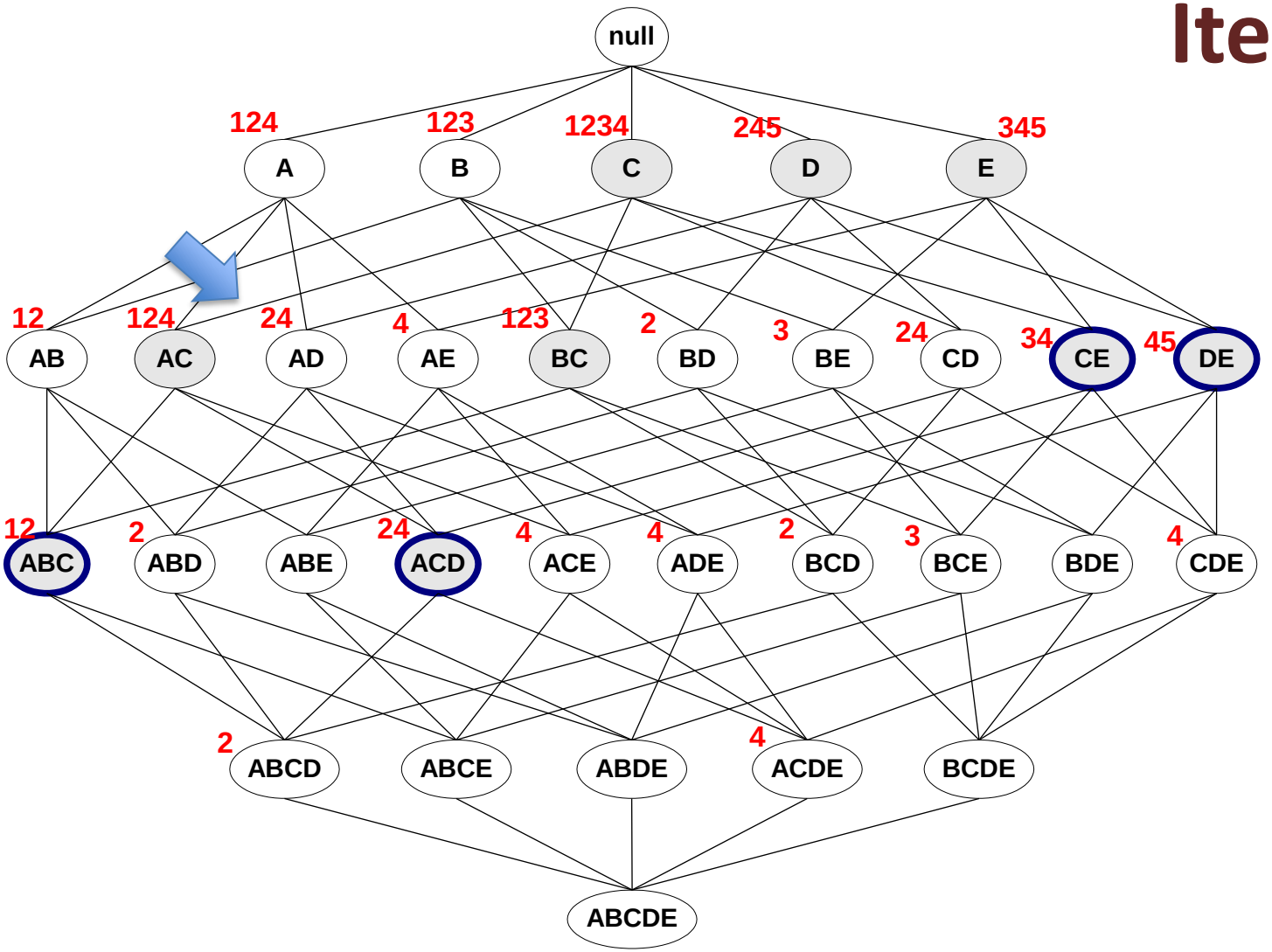
Maximal vs Closed Itemsets



Χρήση

- Υπάρχουν αλγόριθμοι που βρίσκουν τα closed frequent itemsets μέσα από ένα σετ συναλλαγών
- Από τα closed frequent itemsets μπορούμε να βρούμε το support των non-closed

Παράδειγμα χρήσης Closed Frequent Itemsets



- AD non-closed
- s(AD) = s(ABD) or s(ACD) or s(ADE)
- s(AD) = max(s(ABD), s(ACD), s(ADE))

Συμπαγής αναπαράσταση Frequent Itemsets

- Μερικά είναι περίσσεια καθώς έχουν ίδιο support όπως και τα υπερσύνολά τους

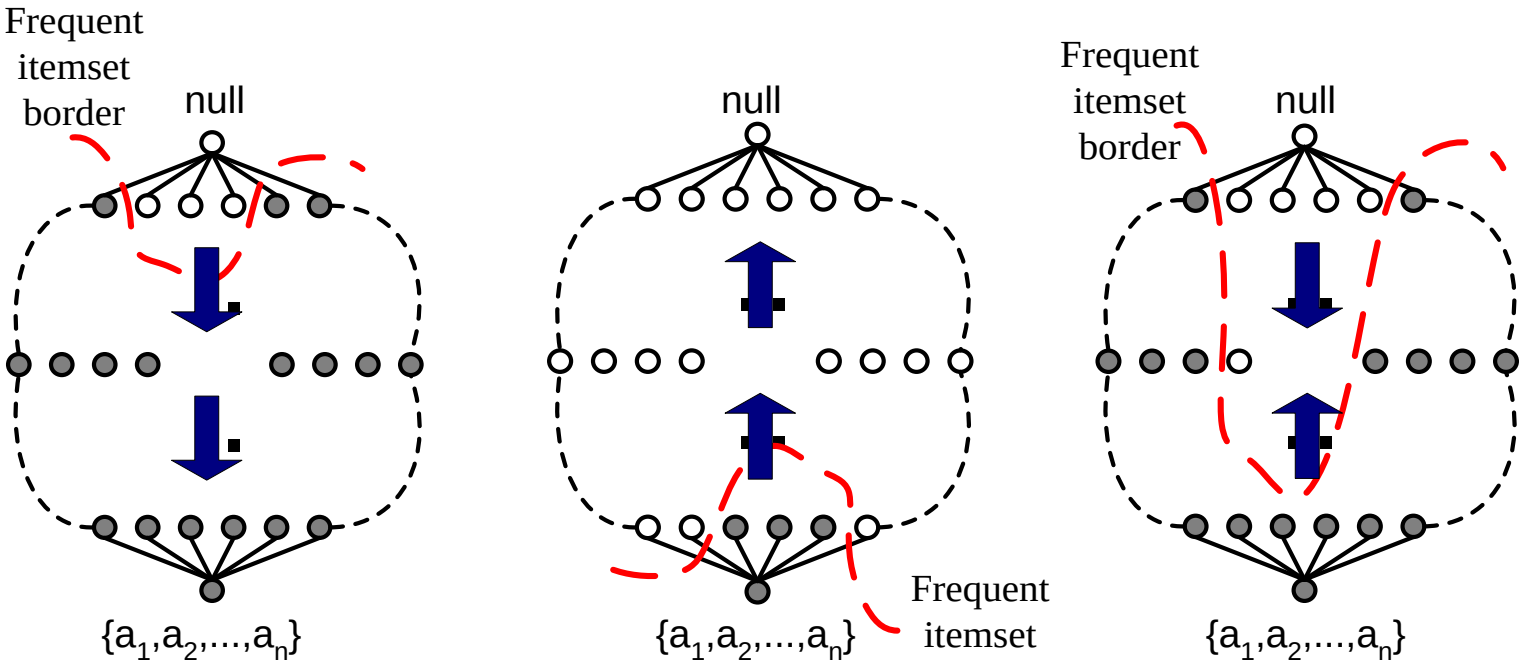
TID	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1

- Number of frequent itemsets = 3 x (2^10 - 1) = 3069

- Συμπαγής αναπαράσταση
  - {A1, A2, A3, A4, A5, A6, A7, A8, A9, A10}
  - {B1, B2, B3, B4, B5, B6, B7, B8, B9, B10}
  - {C1, C2, C3, C4, C5, C6, C7, C8, C9, C10}

Εναλλακτικοί μέθοδοι δημιουργίας Frequent Itemsets

- Apriori: Καλός, αλλά πολύ I/O, πολλές διασχίσεις των δεδομένων, αργός σε πυκνές συναλλαγές
- Διάσχιση του Itemset Lattice
  - General-to-specific vs Specific-to-general



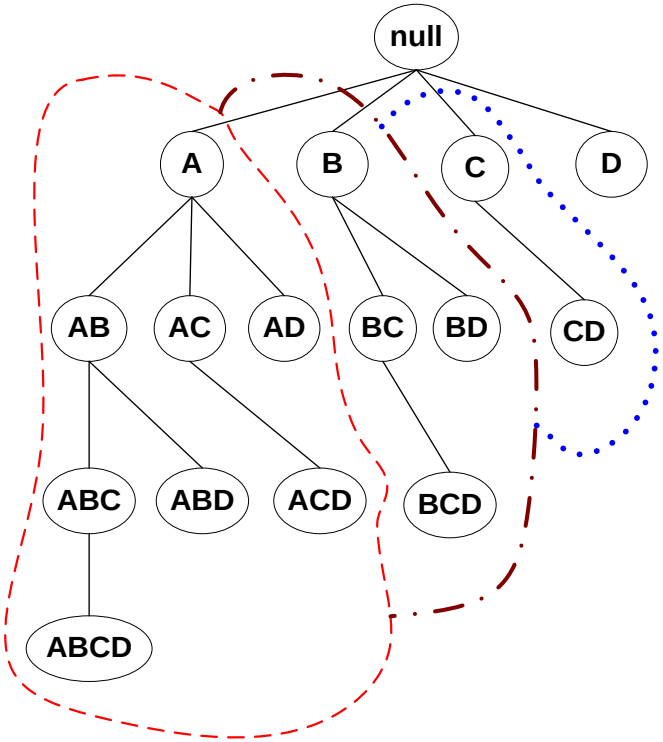
(a) General-to-specific (b) Specific-to-general (c) Bidirectional

Χρήση maximal frequent itemsets

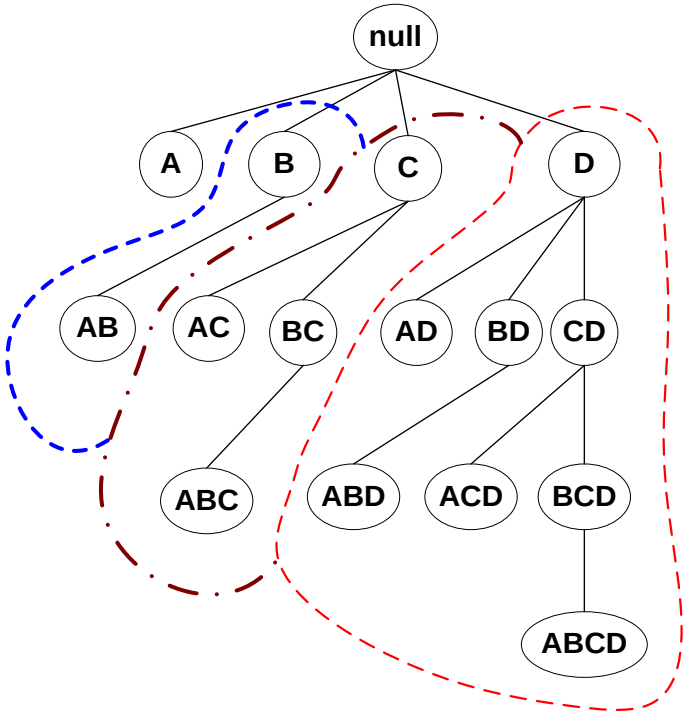
- General-to-specific (Apriori): καλή για μικρού μήκους frequent itemsets
- Specific-to-general: Καλή για πυκνά δεδομένα
  - Βρίσκω τα maximal frequent itemsets και δε χρειάζεται να ψάξω εσωτερικούς κόμβους για νέα maximal frequent itemsets
- Συνδυασμός:
  - περισσότερο χώρο
  - βρίσκω το όριο πιο γρήγορα

Εναλλακτικοί μέθοδοι δημιουργίας Frequent Itemsets

- Διάσχιση του Itemset Lattice
  - Equivalent Classes



(a) Prefix tree



(b) Suffix tree

Rule generation

Δημιουργία κανόνων

Δημιουργία κανόνων

- Δεδομένου ενός frequent itemset L, βρες όλα τα μη-κενά υποσύνολα $f \subset L$ τέτοια ώστε $f \rightarrow L - f$ να ικανοποιεί την απαίτηση minimum confidence
  - Εάν το {A,B,C,D} είναι ένα frequent itemset, οι υποψήφιοι κανόνες είναι:

    $ABC \rightarrow D$, $ABD \rightarrow C$, $ACD \rightarrow B$, $BCD \rightarrow A$,
    $A \rightarrow BCD$, $B \rightarrow ACD$, $C \rightarrow ABD$, $D \rightarrow ABC$
    $AB \rightarrow CD$, $AC \rightarrow BD$, $AD \rightarrow BC$, $BC \rightarrow AD$,
    $BD \rightarrow AC$, $CD \rightarrow AB$,

- Εάν $|L| = k$, τότε υπάρχουν $2^k - 2$ υποψήφιοι κανόνες συσχέτισης (αγνοώντας τα $L \rightarrow \varnothing$ και $\varnothing \rightarrow L$)

Δημιουργία κανόνων

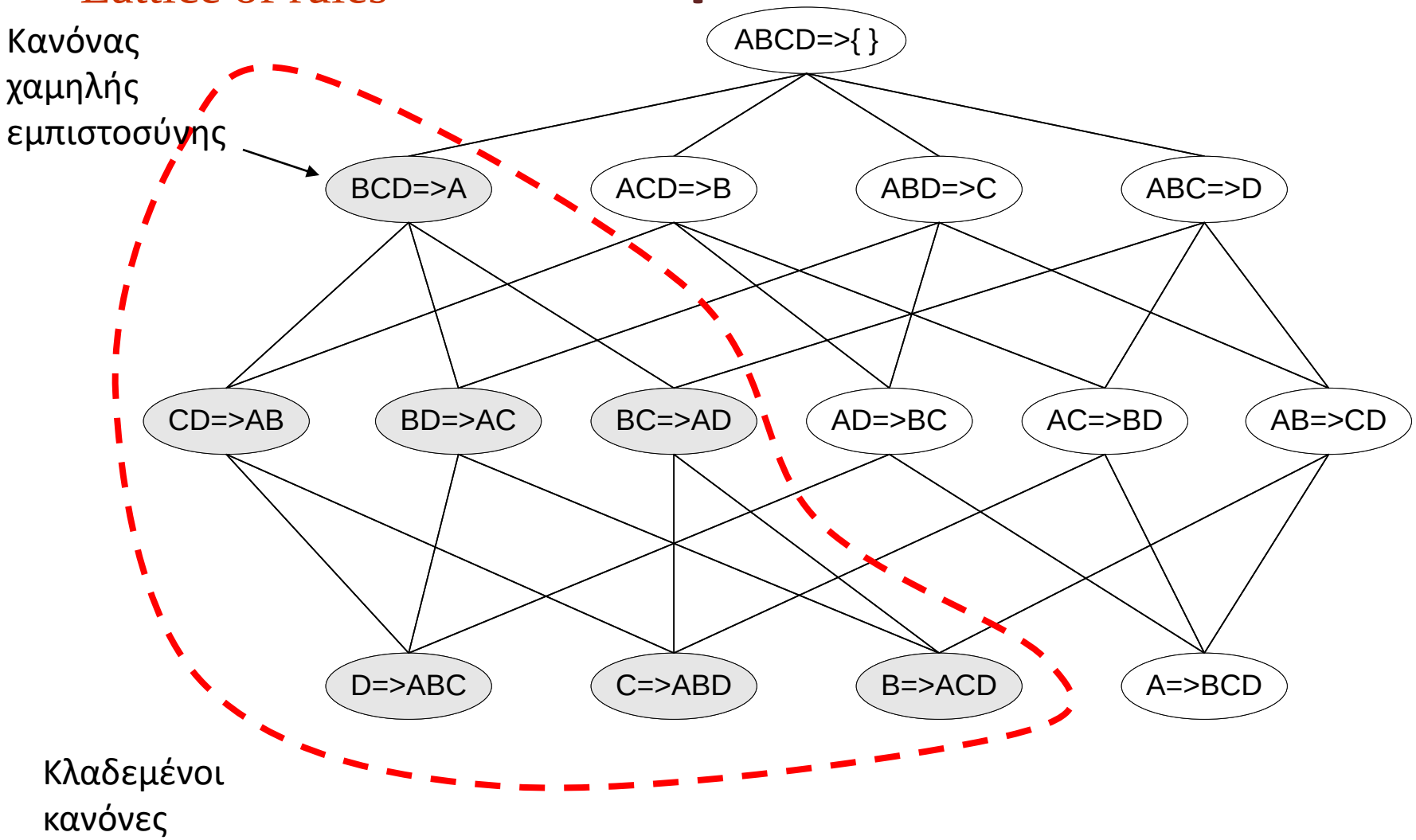
- Πώς δημιουργούνται αποδοτικά οι κανόνες από frequent itemsets;
  - Γενικά, η εμπιστοσύνη δεν έχει την ιδιότητα anti-monotone
    - c(ABC $\rightarrow$ D) μπορεί να είναι μεγαλύτερη ή μικρότερη από την c(AD $\rightarrow$ B) ή την c(AB $\rightarrow$ D)
  - Αλλά η εμπιστοσύνη από κανόνες που δημιουργήθηκαν από το ίδιο itemset έχουν την ιδιότητα αυτή
  - π.χ., L = {A,B,C,D}:

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

    - Η εμπιστοσύνη είναι αντι-μονότονη σε σχέση με τον αριθμό των αντικειμένων στο δεξί μέρος του κανόνα

Δημιουργία κανόνων για τον αλγόριθμο Apriori

Lattice of rules



Αξιολόγηση προτύπων

- Οι αλγόριθμοι κανόνων συσχέτισης τείνουν να παράγουν πάρα πολλούς κανόνες
  - πολλοί από αυτούς είναι περίσσιοι ή χωρίς ενδιαφέρον
  - Περίσσιος κανόνας είναι εάν ο $\{A,B,C\} \rightarrow \{D\}$ και ο $\{A,B\} \rightarrow \{D\}$ έχουν ίδιο support & confidence

- Μετρικές ενδιαφέροντος (Interestingness measures) μπορούν να χρησιμοποιηθούν για να κλαδέψουν/κατατάξουν κανόνες

Υπολογισμός μετρικής ενδιαφέροντος

- Δεδομένου ενός κανόνα X → Y, χρειαζόμαστε πληροφορία από τον παρακάτω πίνακα περιπτώσεων

Contingency table for X → Y

	Y	$\overline{Y}$	
X	f_{11}	f_{10}	f_{1+}
$\overline{X}$	f_{01}	f_{00}	f_{o+}
	f_{+1}	f_{+0}	|T|

f_{11}: support of X and Y
f_{10}: support of $\overline{X}$ and $\overline{Y}$
f_{01}: support of $\overline{X}$ and Y
f_{00}: support of X and $\overline{Y}$

Ο πίνακας χρησιμεύει σε διάφορες μετρικές

support, confidence, lift, Gini, J-measure, κτλ.

Μειονέκτημα της εμπιστοσύνης

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	75	5	80
	90	10	100

Κανόνας Συσχέτισης: Tea $\rightarrow$ Coffee

Confidence= P(Coffee|Tea) = 0.75

αλλά P(Coffee) = 0.9

$\Rightarrow$ Ενώ το confidence είναι μεγάλο, ο κανόνας είναι παραπλανητικός

$\Rightarrow$ P(Coffee|$\overline{\text{Tea}}$) = 0.9375

Στατιστική Ανεξαρτησία

- Πληθυσμός 1000 μαθητών
  - 600 μαθητές ξέρουν κολύμπι (S)
  - 700 μαθητές ξέρουν ποδήλατο (B)
  - 420 μαθητές ξέρουν και κολύμπι και ποδήλατο (S,B)

  - $P(S \wedge B) = 420/1000 = 0.42$
  - $P(S) \times P(B) = 0.6 \times 0.7 = 0.42$
  - $P(S \wedge B) = P(S) \times P(B)$ => Στατιστική ανεξαρτησία
  - $P(S \wedge B) > P(S) \times P(B)$ => Θετική συσχέτιση
  - $P(S \wedge B) < P(S) \times P(B)$ => Αρνητική συσχέτιση

Στατιστικές μετρικές

- Μετρικές που παίρνουν υπόψη τη στατιστική εξάρτηση

$$Lift = \frac{P(Y \mid X)}{P(Y)}$$

$$Interest = \frac{P(X,Y)}{P(X)P(Y)}$$

$$PS = P(X,Y) - P(X)P(Y)$$

$$\phi - coefficient = \frac{P(X,Y) - P(X)P(Y)}{\sqrt{P(X)[1-P(X)]P(Y)[1-P(Y)]}}$$

Παράδειγμα: Lift/Interest

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	75	5	80
	90	10	100

Association Rule: Tea $\rightarrow$ Coffee

Confidence= P(Coffee|Tea) = 0.75

but P(Coffee) = 0.9

$\Rightarrow$ Lift = 0.75/0.9= 0.8333 (< 1, άρα αρνητική συσχέτιση)

Μειονεκτήματα Lift & Interest

	Y	$\bar{Y}$	
X	10	0	10
$\bar{X}$	0	90	90
	10	90	100

$$Interest = \frac{0.1}{(0.1)(0.1)} = 10$$

	Y	$\bar{Y}$	
X	90	0	90
$\bar{X}$	0	10	10
	90	10	100

$$Interest = \frac{0.9}{(0.9)(0.9)} = 1.11$$

Statistical independence:

If P(X,Y)=P(X)P(Y) => Interest= 1

Αν και τα X, Y εμφανίζονται σπάνια μαζί στο πρώτο, δίνουν πολύ μεγάλο interest. Στην περίπτωση αυτή θα ήμασταν καλύτερα με confidence

Πολλές μετρικές στη βιβλιογραφία

Άλλα είναι καλά για κάποιες εφαρμογές, άλλα για κάποιες άλλες

Τι κριτήρια να χρησιμοποιήσουμε για το αν μία μετρική είναι καλή ή όχι;

#	Measure	Formula
1	ϕ-coefficient	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
2	Goodman-Kruskal's (λ)	$\frac{\sum_j \max_k P(A_j,B_k)+\sum_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-\max_k P(B_k)}$
3	Odds ratio (α)	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(A,\overline{B})P(\overline{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(\overline{AB})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{AB})+P(A,\overline{B})P(\overline{A},B)} = \frac{\alpha-1}{\alpha+1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(\overline{AB})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{AB})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$
6	Kappa (κ)	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
7	Mutual Information (M)	$\frac{\sum_i \sum_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i)\log P(A_i), -\sum_j P(B_j)\log P(B_j))}$
8	J-Measure (J)	$\max\left(P(A,B)\log(\frac{P(B\|A)}{P(B)})+P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}),\right.$ $\left.P(A,B)\log(\frac{P(A\|B)}{P(A)})+P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})\right)$
9	Gini index (G)	$\max\Big(P(A)[P(B\|A)^2+P(\overline{B}\|A)^2]+P(\overline{A})[P(B\|\overline{A})^2+P(\overline{B}\|\overline{A})^2]$ $-P(B)^2-P(\overline{B})^2,$ $P(B)[P(A\|B)^2+P(\overline{A}\|B)^2]+P(\overline{B})[P(A\|\overline{B})^2+P(\overline{A}\|\overline{B})^2]$ $-P(A)^2-P(\overline{A})^2\Big)$
10	Support (s)	$P(A,B)$
11	Confidence (c)	$\max(P(B\|A),P(A\|B))$
12	Laplace (L)	$\max\left(\frac{NP(A,B)+1}{NP(A)+2},\frac{NP(A,B)+1}{NP(B)+2}\right)$
13	Conviction (V)	$\max\left(\frac{P(A)P(\overline{B})}{P(A\overline{B})},\frac{P(B)P(\overline{A})}{P(B\overline{A})}\right)$
14	Interest (I)	$\frac{P(A,B)}{P(A)P(B)}$
15	cosine (IS)	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
16	Piatetsky-Shapiro's (PS)	$P(A,B)-P(A)P(B)$
17	Certainty factor (F)	$\max\left(\frac{P(B\|A)-P(B)}{1-P(B)},\frac{P(A\|B)-P(A)}{1-P(A)}\right)$
18	Added Value (AV)	$\max(P(B\|A)-P(B),P(A\|B)-P(A))$
19	Collective strength (S)	$\frac{P(A,B)+P(\overline{A}\overline{B})}{P(A)P(B)+P(\overline{A})P(\overline{B})}\times\frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{A}\overline{B})}$
20	Jaccard (ζ)	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
21	Klosgen (K)	$\sqrt{P(A,B)}\max(P(B\|A)-P(B),P(A\|B)-P(A))$

Αναφορά

- "Introduction to Data Mining", P. Tan, M. Steinbach and V. Kumar, Addison Wesley., 2005