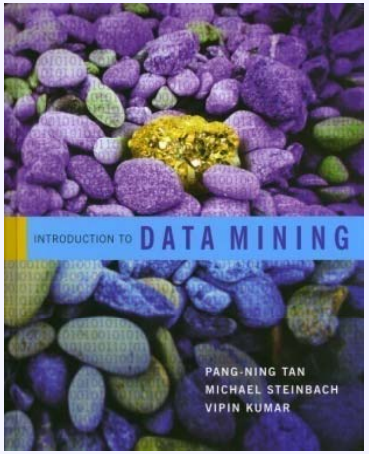
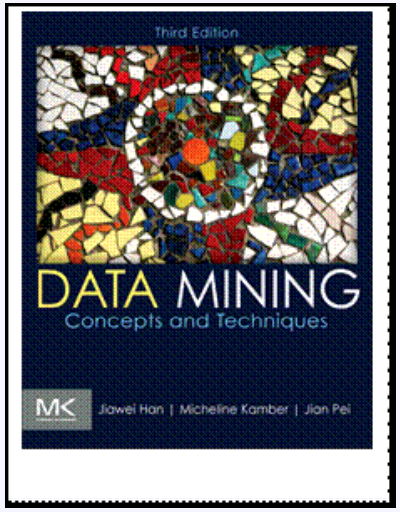
Αποθήκες Δεδομένων και Εξόρυξη Δεδομένων:

Κανόνες Συσχέτισης: FP-Growth

Ευχαριστίες

- Χρησιμοποιήθηκε επιπλέον υλικό από τα βιβλία
  - «Εισαγωγή στην Εξόρυξη και τις Αποθήκες Δεδομένων»
  - «Introduction to Data Mining» των Tan, Steinbach, Kumar , και
  - «Data Mining: Concepts and Techniques» των Jiawei Han, Micheline Kamber.





Είναι γρήγορος ο Apriori? — Bottlenecks στην απόδοση

- Ο βασικός αλγόριθμος Apriori:
  - Χρησιμοποιεί συχνά (k – 1)-στοιχειοσύνολα για την παραγωγή υποψηφίων συχνών k-στοιχειοσυνόλων.
  - Χρήση τεχνικών σαρώματος ΒΔ και ταύτισης προτύπων για τη μέτρηση της υποστήριξης των υποψηφίων συνόλων.
- Το bottleneck του Apriori: δημιουργία υποψηφίων
  - Πολύ μεγάλα υποψήφια σύνολα.
  - Πολλαπλές σαρώσεις της ΒΔ.

Ο Αλγόριθμος FP-Growth

- Χρησιμοποιεί μια συμπιεσμένη αναπαράσταση της βάσης
- με τη μορφή ενός FP-δένδρου (*FP: frequent pattern)*
- Το δένδρο μοιάζει με προθεματικό δένδρο - prefix tree (trie).
- Ο αλγόριθμος κατασκευής διαβάζει μια συναλλαγή τη φορά, και απεικονίζει τη συναλλαγή σε ένα μονοπάτι του FP-δένδρου.
- Μερικά μονοπάτια μπορεί να επικαλύπτονται: όσο περισσότερα μονοπάτια επικαλύπτονται, τόσο καλύτερη συμπίεση.
- Τα συχνά στοιχειοσύνολα βρίσκονται με μια αναδρομική διαίρει-και-βασίλευε προσέγγιση.

Κατασκευή FP-δένδρου (1)

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{B,C}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

- Το FP-δένδρο είναι ένα προθεματικό δένδρο
- Άρα τα στοιχεία σε κάθε σύνολο πρέπει να ακολουθούν κάποια **διάταξη**, έστω τη *λεξικογραφική*
- Θα δούμε αργότερα ότι κάτι άλλο συμφέρει περισσότερο

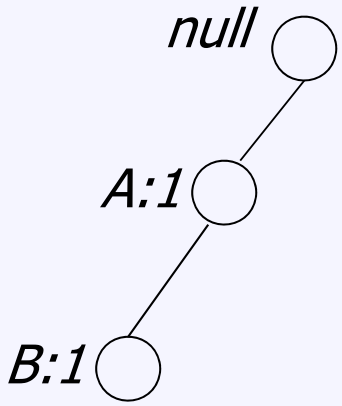
Αρχικά, το δένδρο είναι κενό

null

Κατασκευή FP-δένδρου (2)

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{B,C}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

Διάβασμα TID=1:

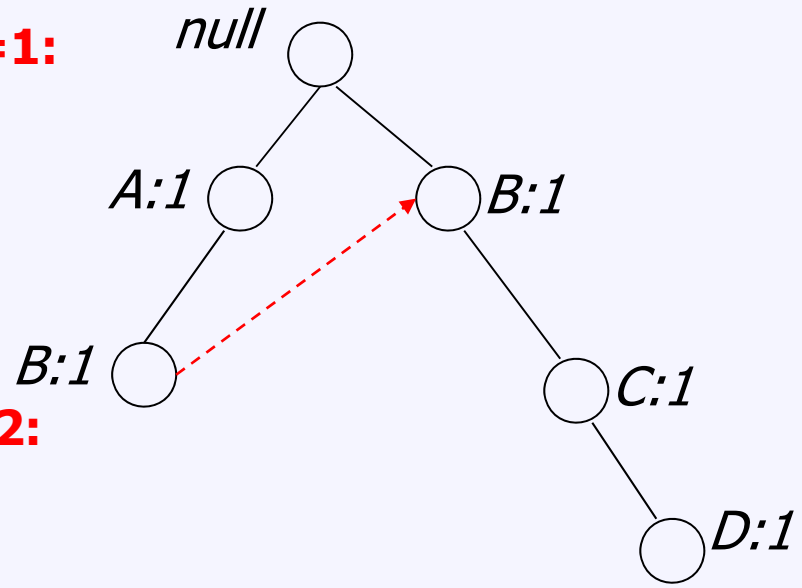


Κάθε κόμβος έχει μια ετικέτα που δείχνει πόσες συναλλαγές φτάνουν σε αυτόν, δηλαδή πόσα μονοπάτια καταλήγουν σε αυτόν τον κόμβο.

Κατασκευή FP-δένδρου (3)

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{B,C}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

Διάβασμα TID=1:



Διάβασμα TID=2:

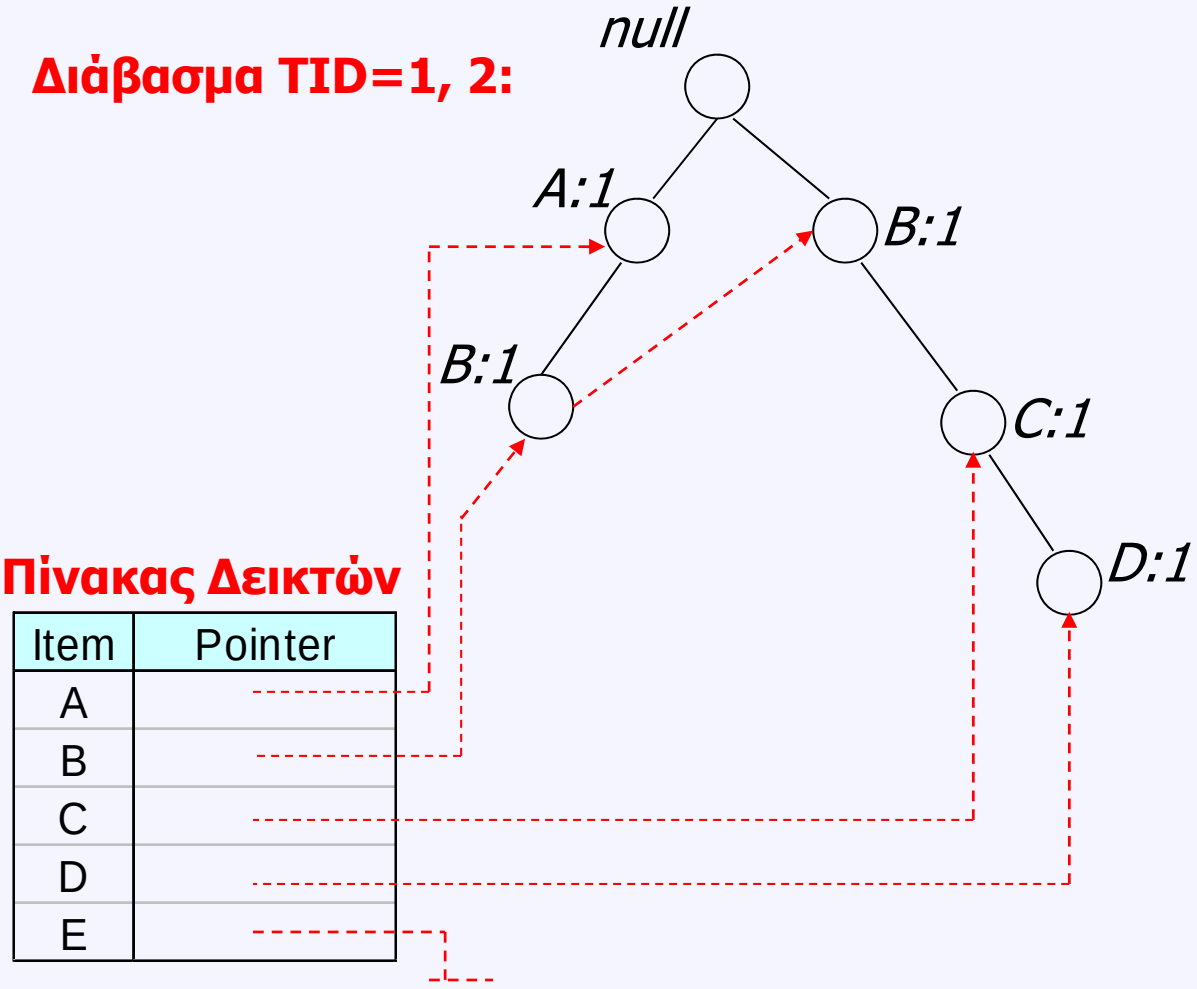
Κάθε κόμβος έχει μια ετικέτα που δείχνει πόσες συναλλαγές φτάνουν σε αυτόν.

Επίσης, υπάρχουν δείκτες μεταξύ των κόμβων που αναφέρονται στο ίδιο στοιχείο

Κατασκευή FP-δένδρου (4)

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{B,C}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

Διάβασμα TID=1, 2:



Πίνακας Δεικτών

Item	Pointer
A	
B	
C	
D	
E	

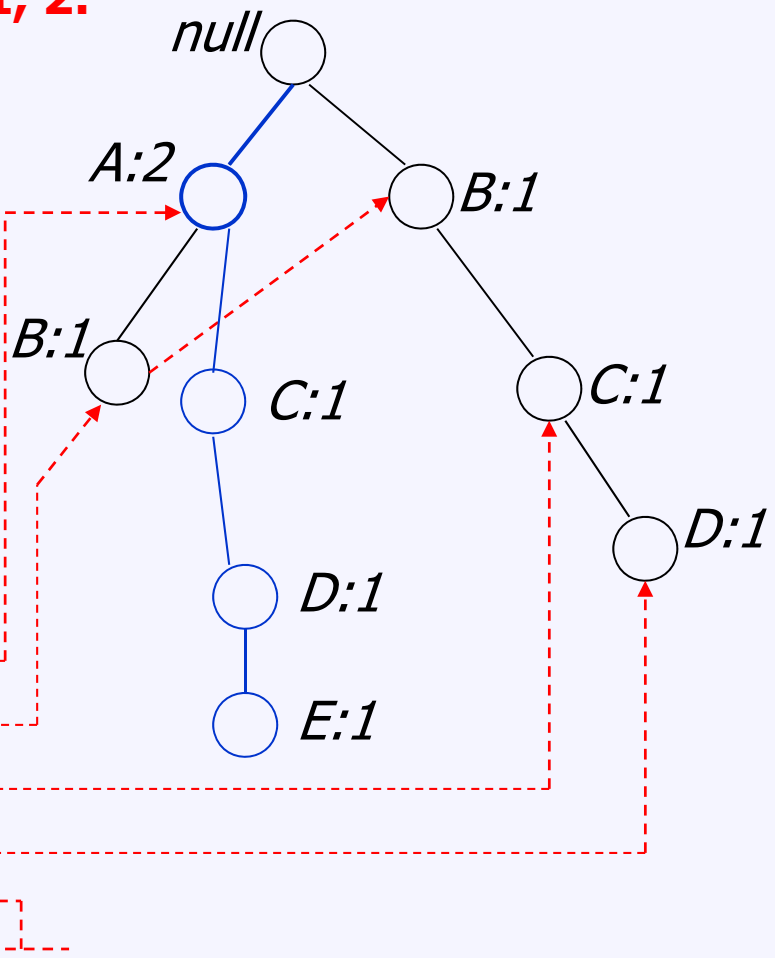
Επίσης, κρατάμε πίνακα δεικτών για να βοηθήσουν στον υπολογισμό των συχνών στοιχειοσυνόλων.

Κατασκευή FP-δένδρου (5)

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{B,C}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

Διάβασμα TID=1, 2:

Διάβασμα TID=3



Πίνακας Δεικτών

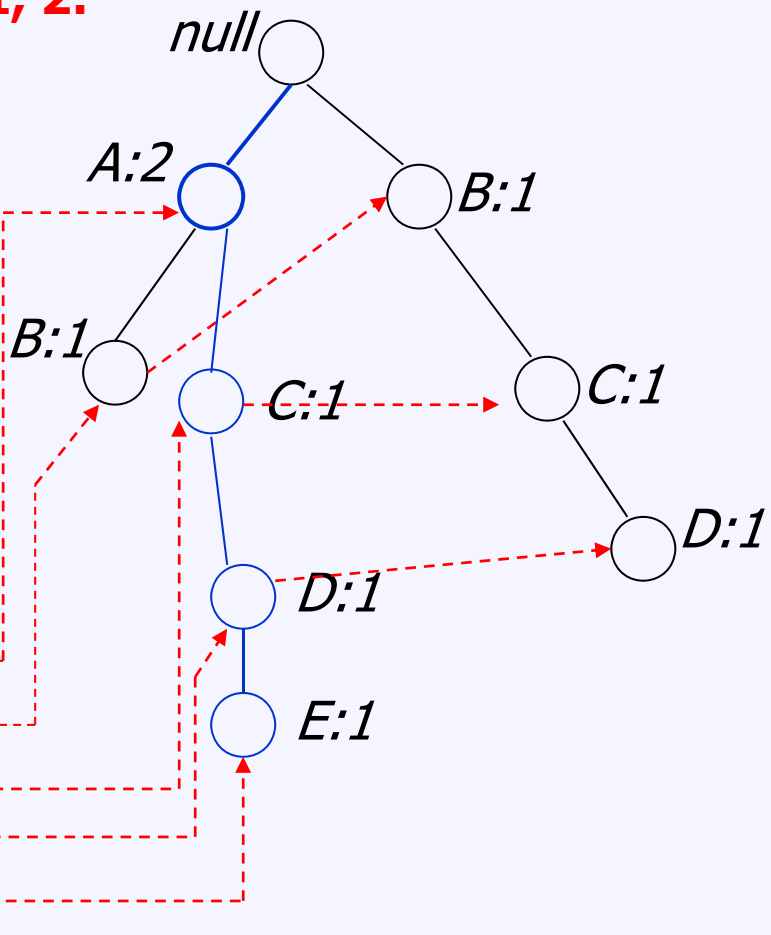
Item	Pointer
A	
B	
C	
D	
E	

Κατασκευή FP-δένδρου (6)

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{B,C}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

Διάβασμα TID=1, 2:

Διάβασμα TID=3



Πίνακας Δεικτών

Item	Pointer
A	
B	
C	
D	
E	

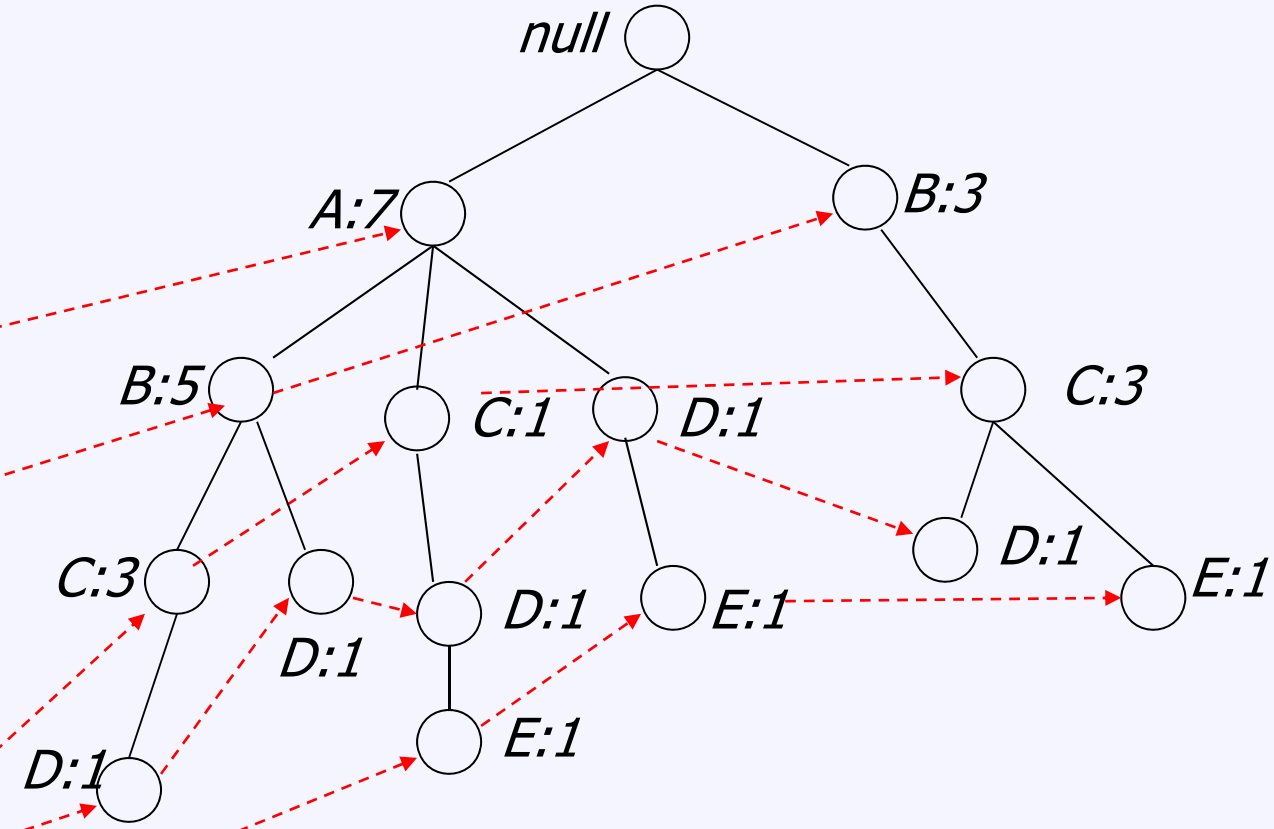
Κατασκευή FP-δένδρου (7)

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{B,C}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

Πίνακας Δεικτών

Item	Pointer
A	
B	
C	
D	
E	

Αφού έχουν διαβαστεί όλες οι συναλλαγές...



Μέγεθος FP-δένδρου

- Κάθε *συναλλαγή* αντιστοιχεί σε *ένα μονοπάτι* από τη ρίζα
- Το μέγεθος του δένδρου είναι συνήθως μικρότερο των δεδομένων, αν υπάρχουν κοινά προθέματα.
  - Αν όλες οι συναλλαγές περιέχουν τα ίδια δεδομένα, τότε υπάρχει μόνο ένα κλαδί.
  - Αν όλες είναι διαφορετικές, ο χώρος είναι μεγαλύτερος...
  - ...γιατί αποθηκεύεται περισσότερη πληροφορία, όπως δείκτες μεταξύ των κόμβων αλλά και συχνότητες εμφάνισης.

data engineering

Επιλογή προθέματος

- Το τελικό δένδρο, εξαρτάται από τη **διάταξη**:
  - άλλη διάταξη → άλλα προθέματα.
- (Συνήθως) μικρότερο δένδρο, αν δεν διατάσουμε τα αντικείμενα λεξικογραφικά, αλλά σύμφωνα με τη συχνότητα εμφάνισης.
- Αρχικά, διαβάζουμε όλα τα δεδομένα μια φορά ώστε να υπολογιστεί ο μετρητής υποστήριξης κάθε στοιχείου, και διατάσουμε τα στοιχεία με βάση αυτό (αγνοούμε όσα στοιχεία είναι μη συχνά)

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{B,C}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

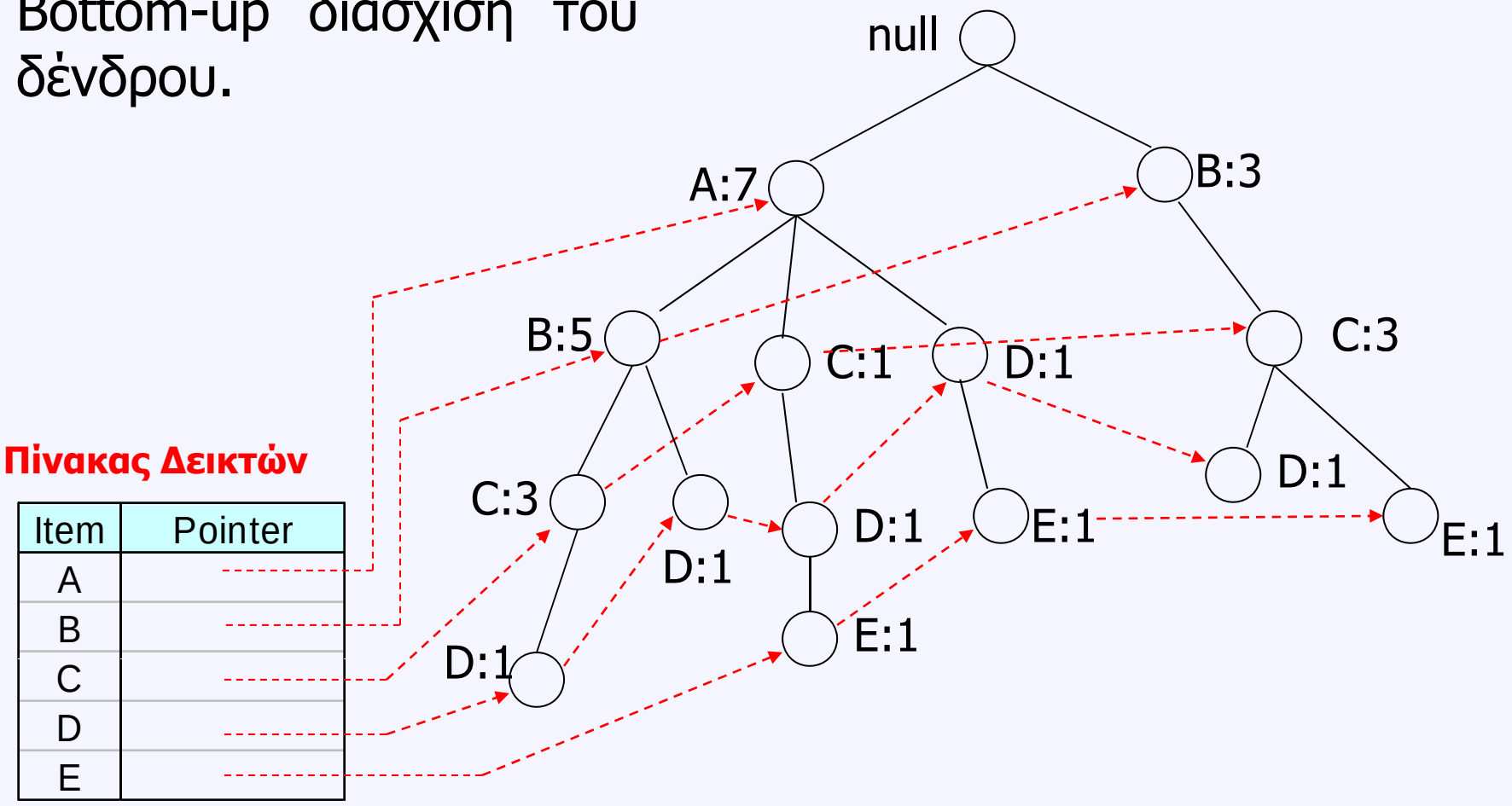
TID	Items
1	{B,A}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{B,A,C}
6	{B,A,C,D}
7	{B,C}
8	{B,A,C}
9	{B,A,D}
10	{B,C,E}

Εύρεση συχνών στοιχειοσυνόλων

- Είσοδος: FP-δένδρο
- Έξοδος: Συχνά στοιχειοσύνολα και η υποστήριξη τους
- Μέθοδος Διαίρει-και-Βασίλευε:
  - Χωρίζουμε τα στοιχειοσύνολα σε αυτά που τελειώνουν σε E, D, C, B, A
  - Μετά αυτά που τελειώνουν σε E σε αυτά σε DE, CE, BE, AE κ.ο.κ.
  - Αν η διάταξη είναι βάσει της συχνότητας εμφάνισης, τότε χωρίζουμε τα στοιχειοσύνολα σε αυτά που τελειώνουν στο πιο σπάνιο στοιχείο, μετά στο δεύτερο πιο σπάνιο κ.ο.κ.

Εύρεση συχνών στοιχειοσυνόλων με χρήση του FP-δένδρου

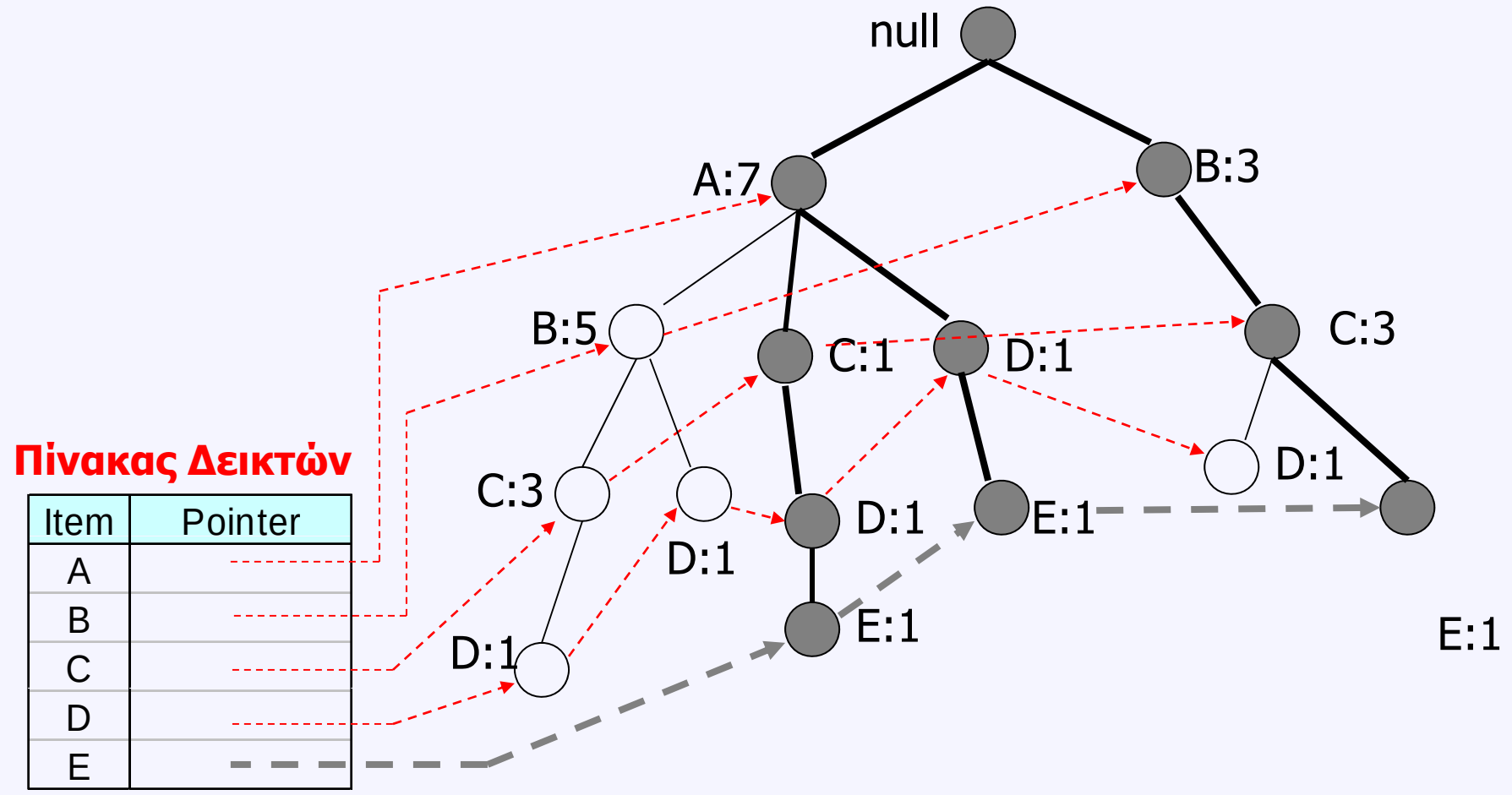
Bottom-up διάσχιση του δένδρου.



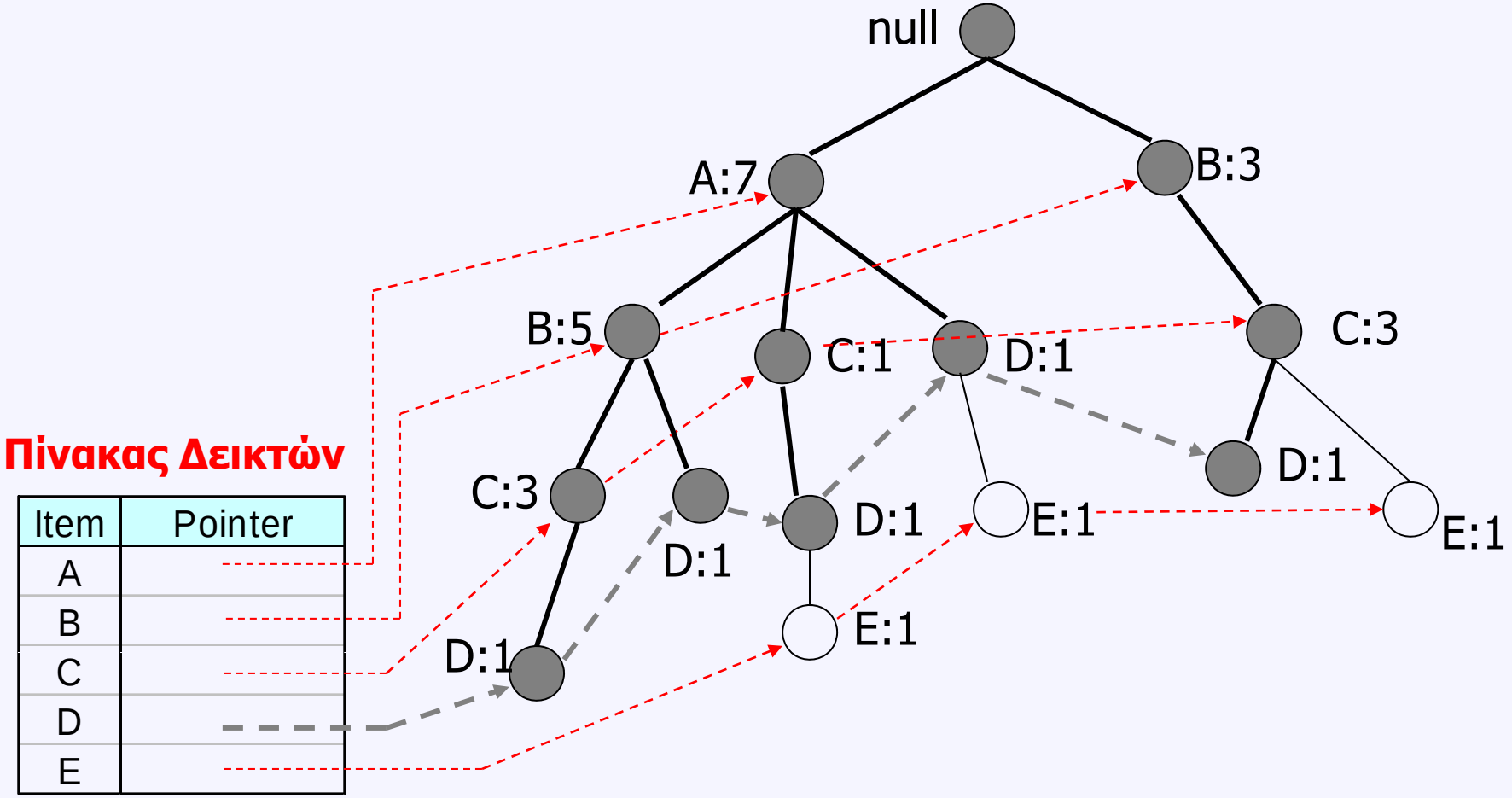
Πίνακας Δεικτών

Item	Pointer
A	
B	
C	
D	
E	

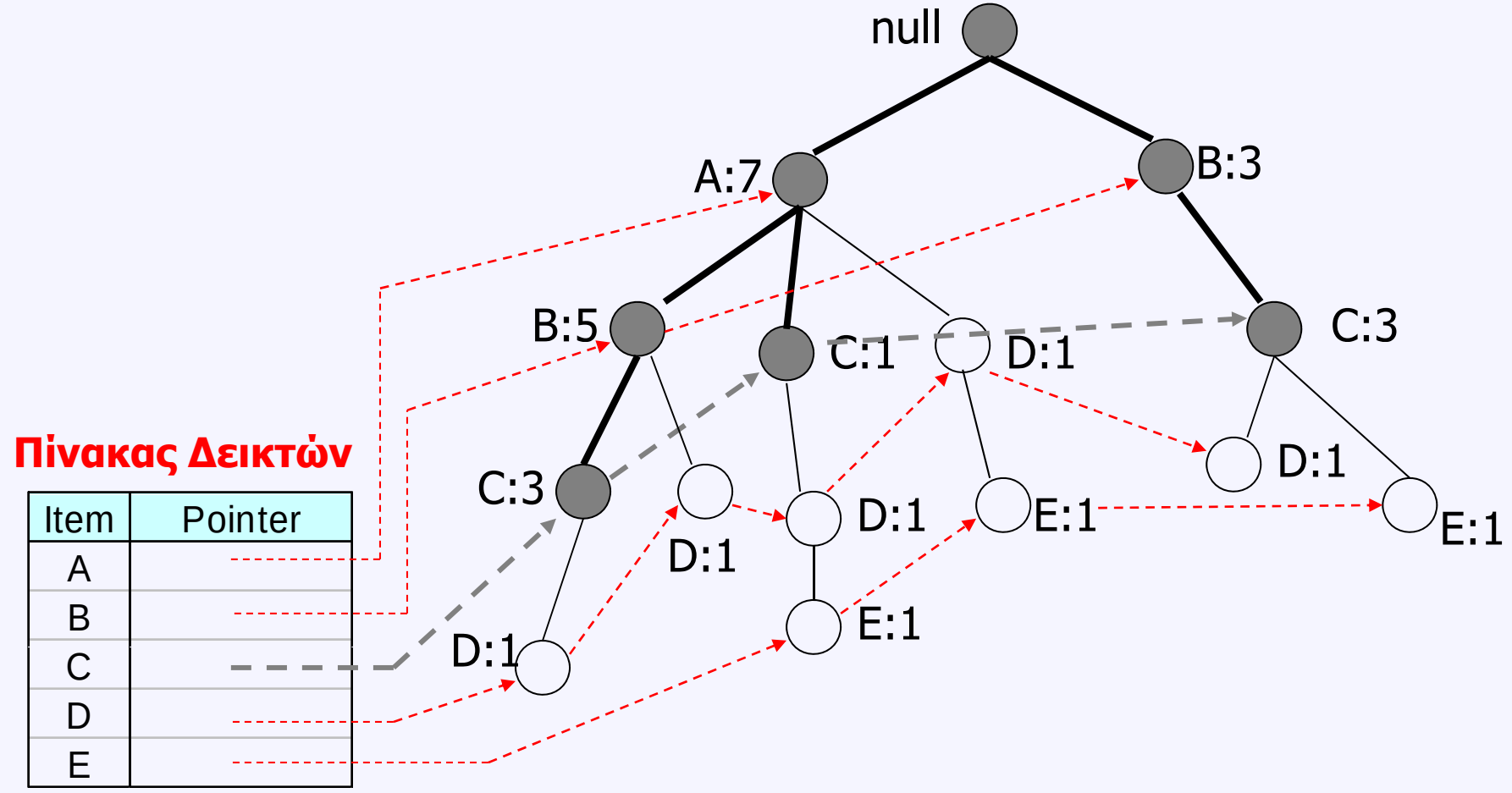
Συχνά στοιχειοσύνολα που τελειώνουν σε E



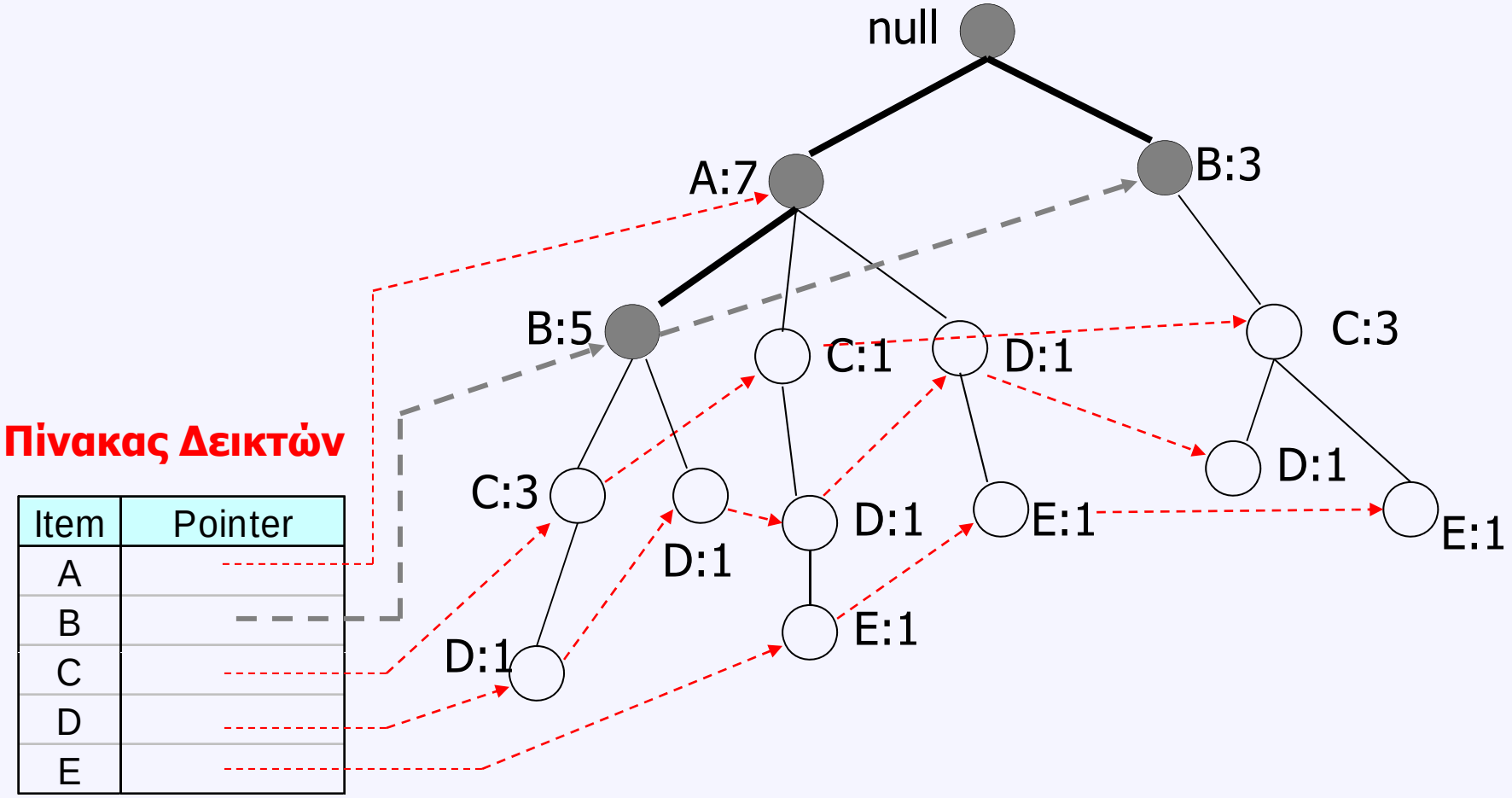
Για το D


null
A:7
B:3
B:5
C:1
D:1
C:3
C:3
D:1
D:1
D:1
E:1
E:1
D:1
E:1
Πίνακας Δεικτών
Item
Pointer
A
B
C
D
E

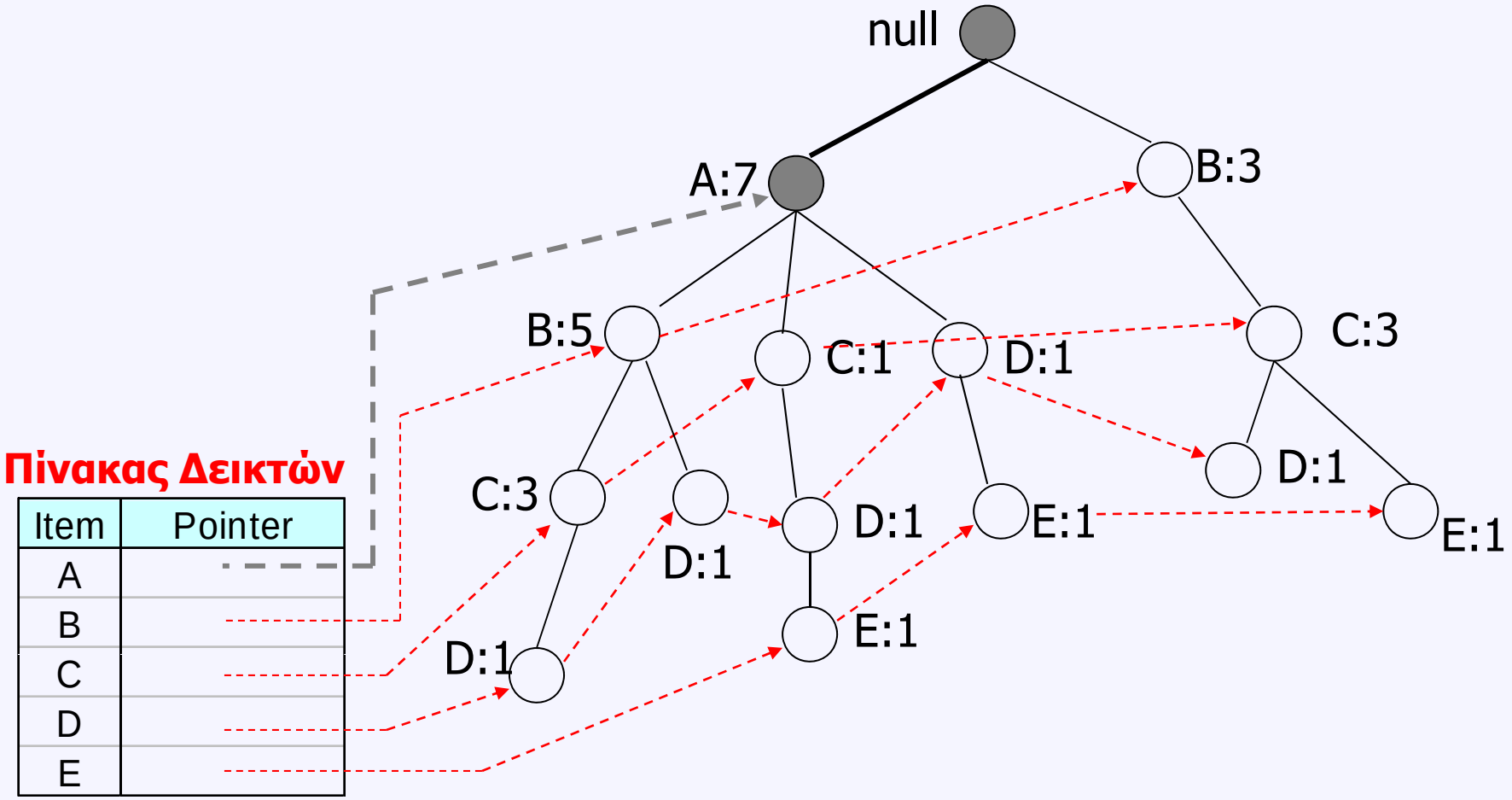
Για το C


null
A:7
B:3
B:5
C:1
D:1
C:3
C:3
D:1
D:1
D:1
E:1
E:1
D:1
E:1
Πίνακας Δεικτών
Item
Pointer
A
B
C
D
E

Για το B


null
A:7
B:3
B:5
C:1
D:1
C:3
C:3
D:1
D:1
E:1
E:1
D:1
E:1
Πίνακας Δεικτών
Item
Pointer
A
B
C
D
E

Για το A


null
A:7
B:3
B:5
C:1
D:1
C:3
C:3
D:1
D:1
D:1
E:1
E:1
D:1
E:1
Πίνακας Δεικτών
Item
Pointer
A
B
C
D
E

Συνοπτικά ο αλγόριθμος

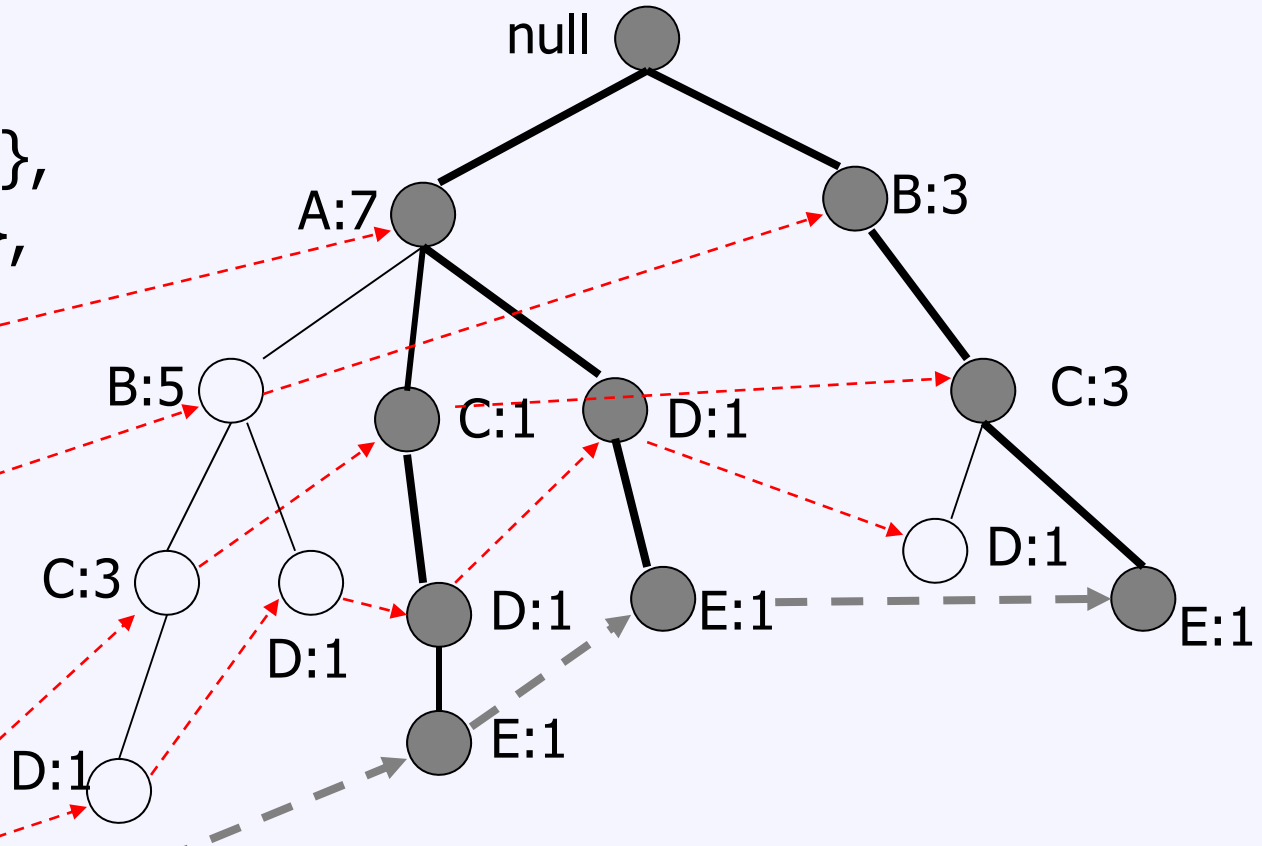
Σε κάθε βήμα, για το επίθεμα (suffix) X

- Φάση 1
  - κατασκευάζουμε το προθεματικό δένδρο για το X και υπολογίζουμε την υποστήριξη χρησιμοποιώντας τον πίνακα
- Φάση 2
  - Αν είναι συχνό, κατασκευάζουμε το υπο-συνθήκη δένδρο για το X, σε βήματα
    - επανα-υπολογισμός υποστήριξης
    - περικοπή κόμβων με μικρή υποστήριξη
    - περικοπή φύλλων

Φάση 1

Προθεματικά μονοπάτια του E:

{E}, {D,E}, {C,D,E}, {A,D,E}, {A,C,D,E}, {C,E}, {B,C,E}



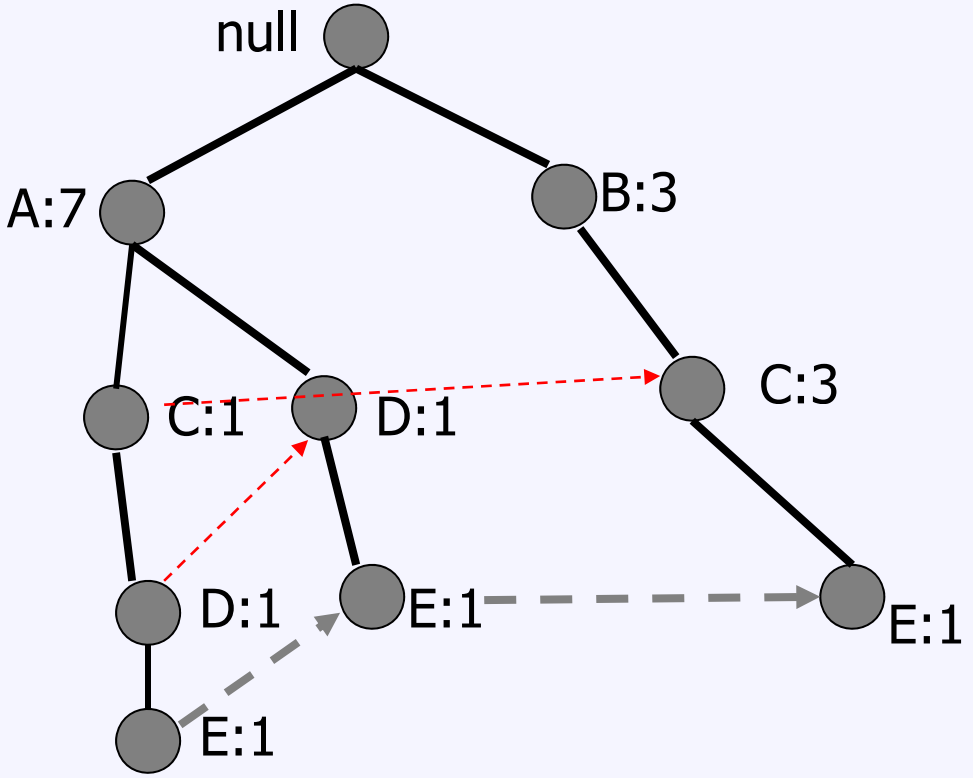
Πίνακας Δεικτών

Item	Pointer
A	
B	
C	
D	
E	

Φάση 1

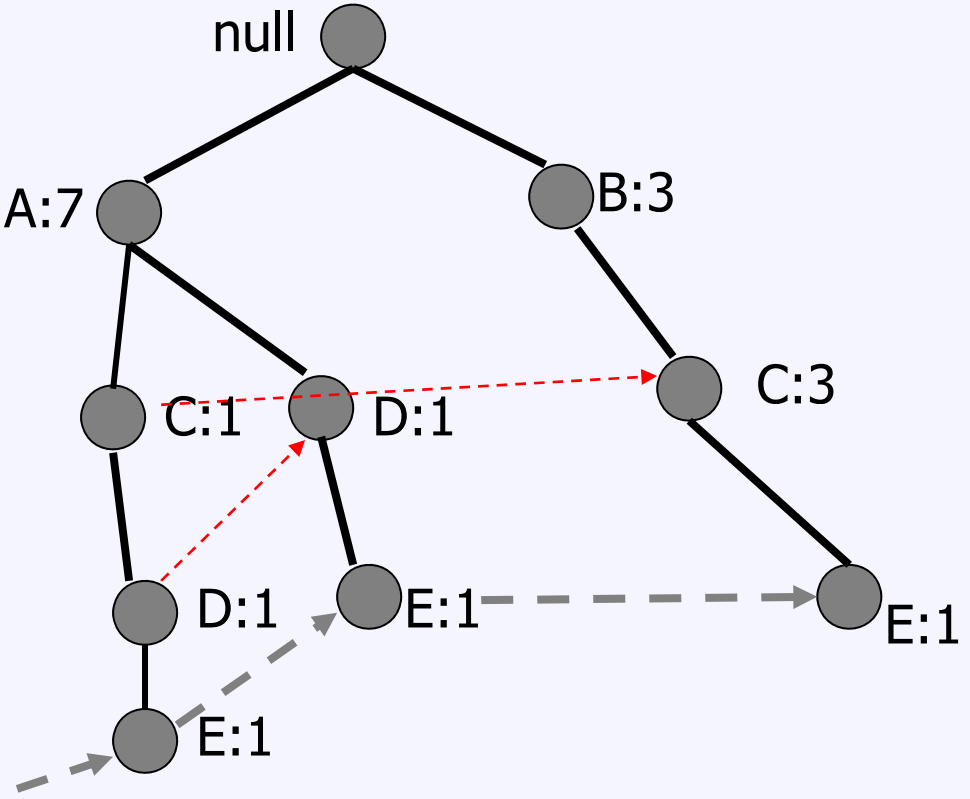
Προθεματικά μονοπάτια του E:

{E}, {D,E}, {C,D,E}, {A,D,E}, {A,C,D,E}, {C,E}, {B,C,E}



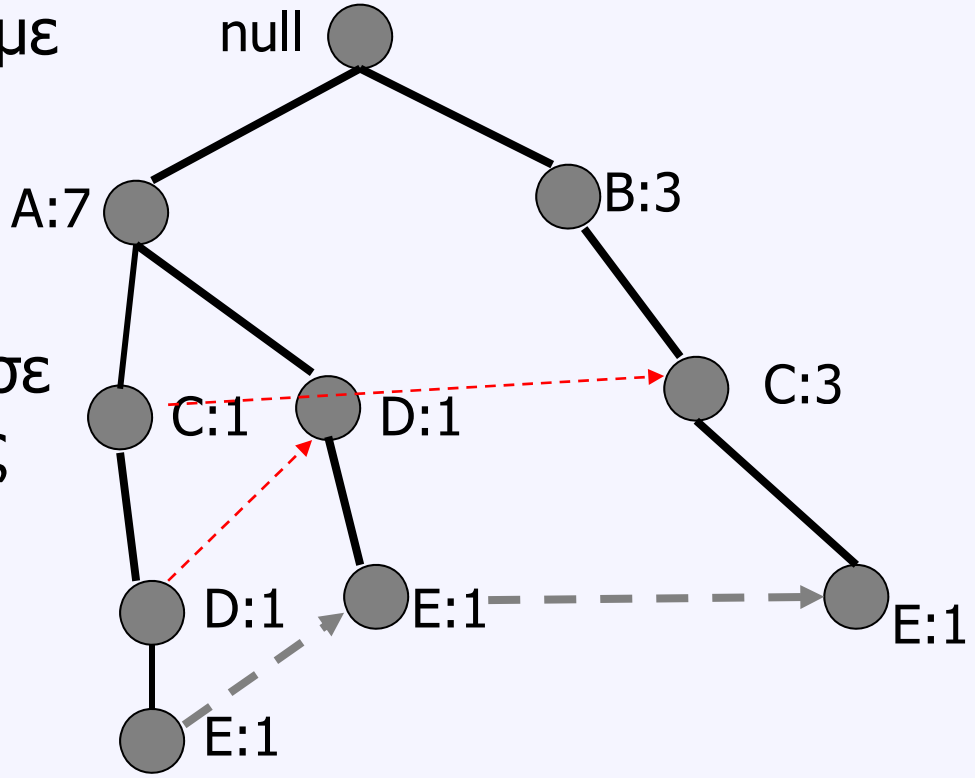
Μέτρηση Υποστήριξης

- Έστω minsup = 2
- Ακολουθούμε τους συνδέσμους αθροίζοντας 1+1+1=3>2
- Οπότε {E} συχνό
- … άρα προχωράμε για DE, CE, BE, AE



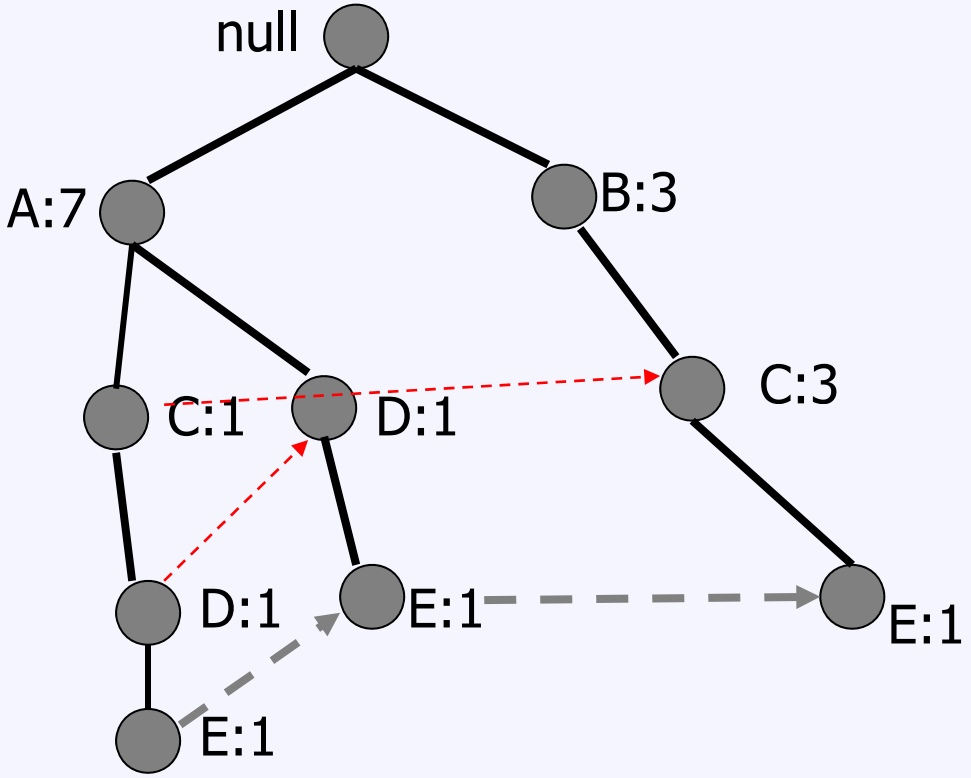
Φάση 2

- {E} συχνό άρα προχωράμε για DE, CE, BE, AE
- Μετατροπή των προθεματικών δένδρων σε FP-δένδρο υπό συνθήκες (conditional FP-tree)
- Δύο αλλαγές
  (1) Αλλαγή των μετρητών
  (2) Περικοπή

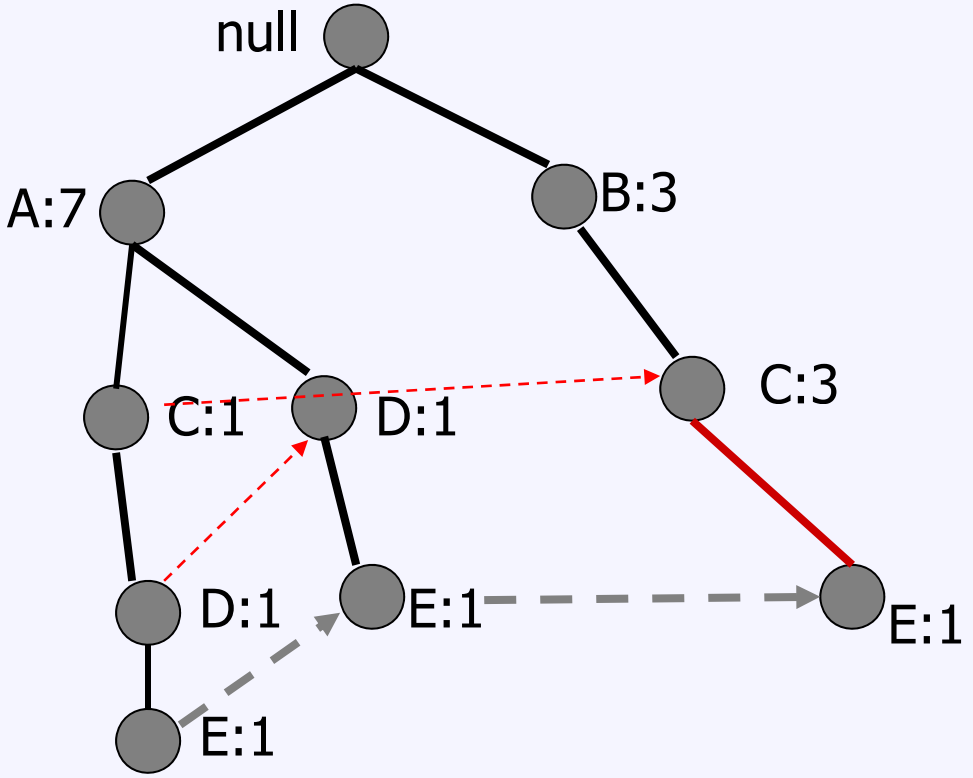


Αλλαγή μετρητών

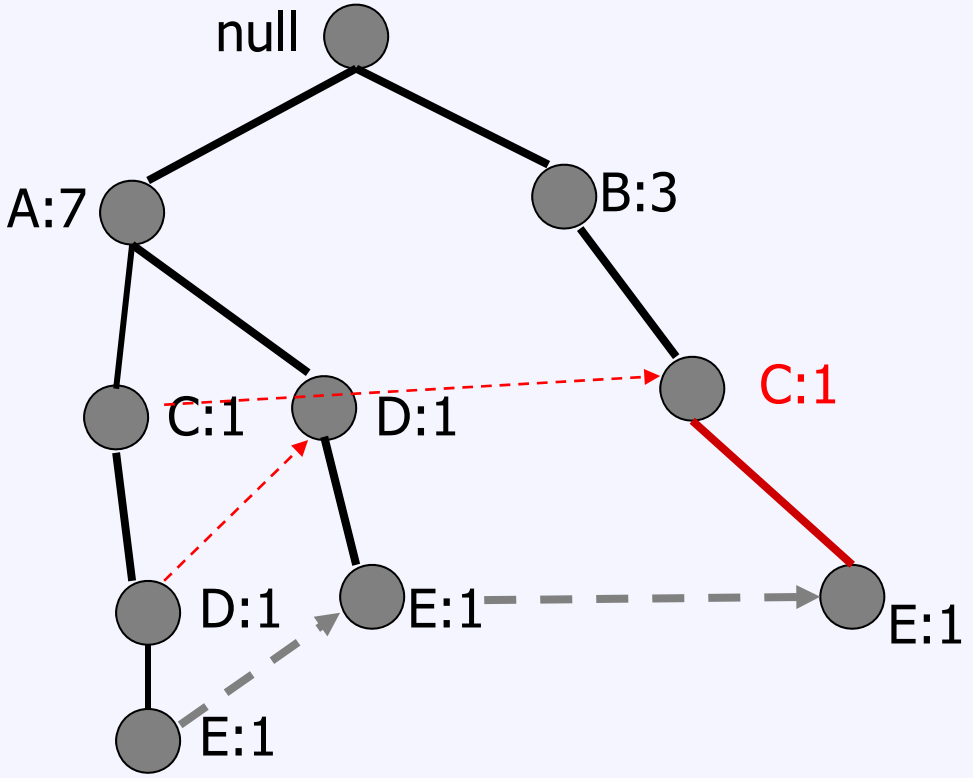
- Οι μετρητές σε κάποιους κόμβους περιλαμβάνουν δοσοληψίες που δεν έχουν το E
- Πχ στο null→B→C→E μετράμε και την συναλλαγή {B, C}



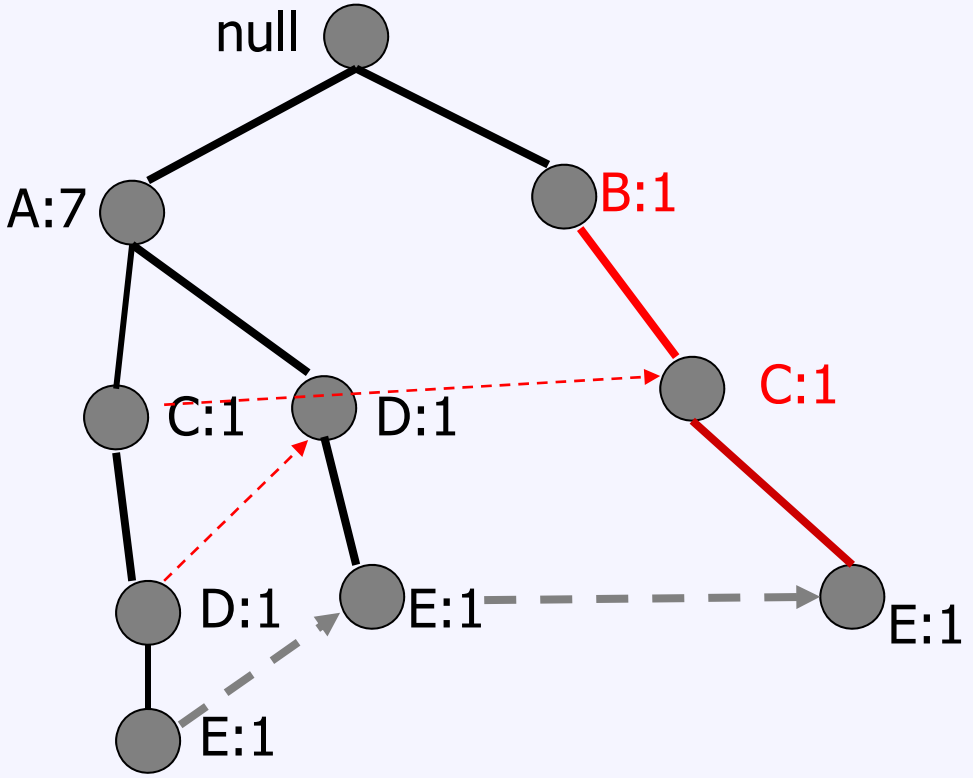
Αλλαγή μετρητών



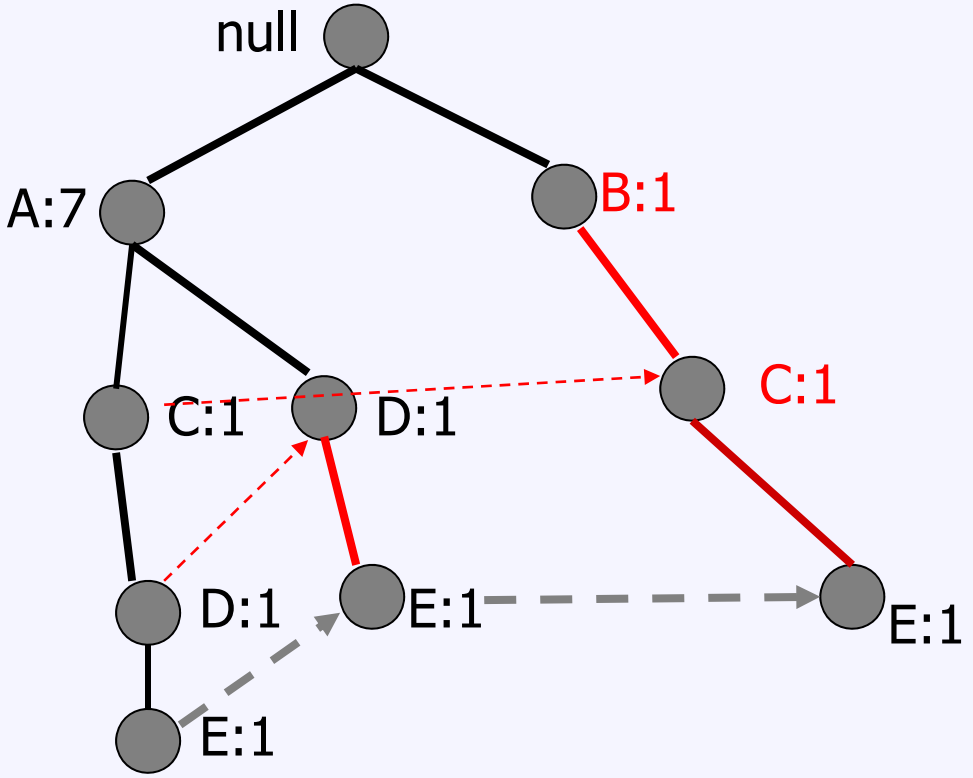
Αλλαγή μετρητών



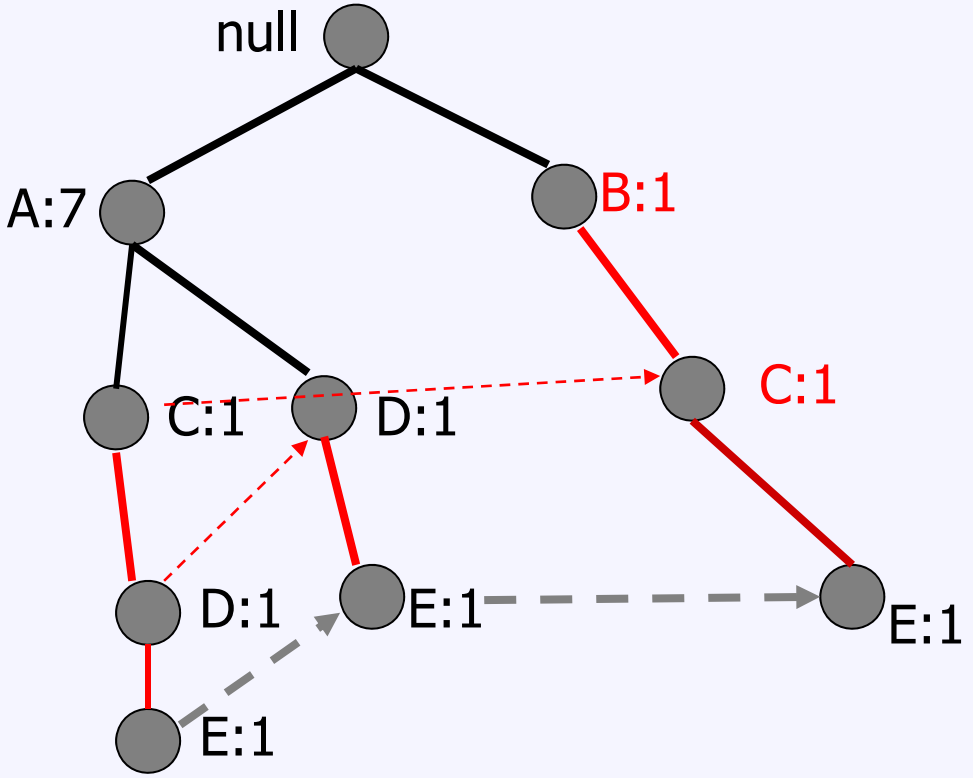
Αλλαγή μετρητών



Αλλαγή μετρητών

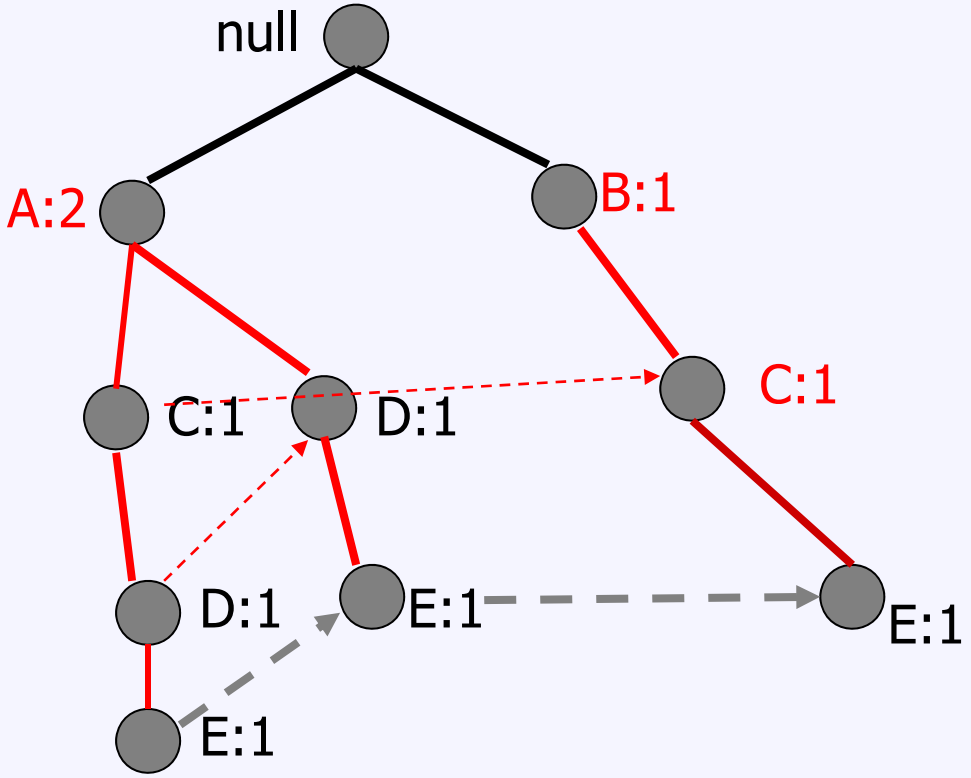


Αλλαγή μετρητών



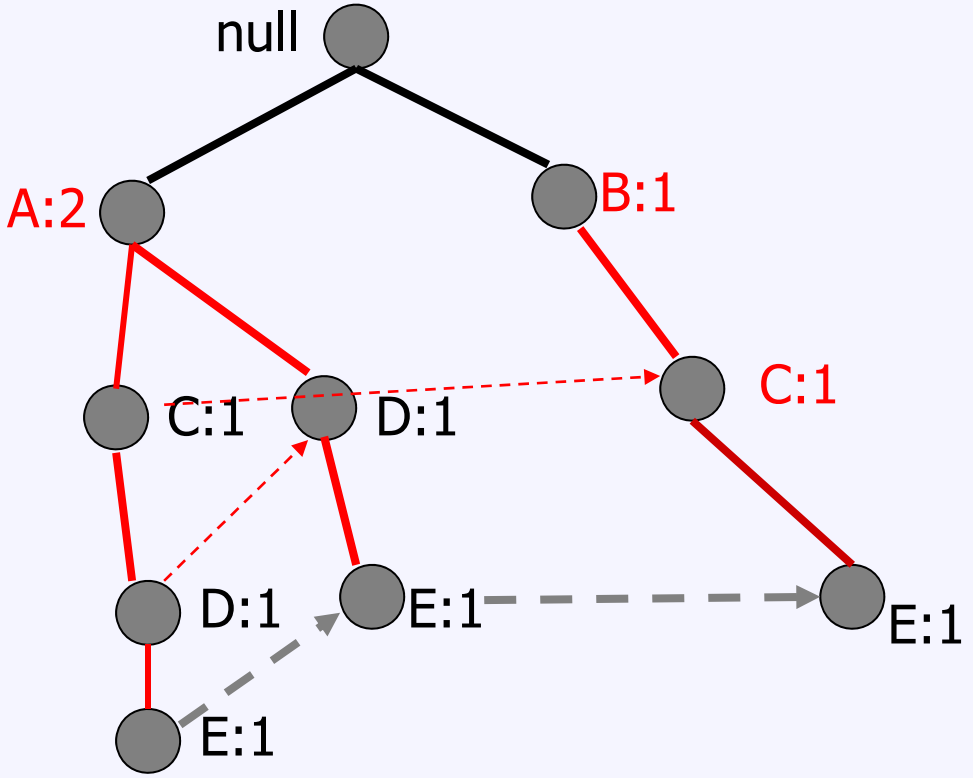
Αλλαγή μετρητών

Περικοπή (truncate):
Σβήσε τους κόμβους του E



Αλλαγή μετρητών

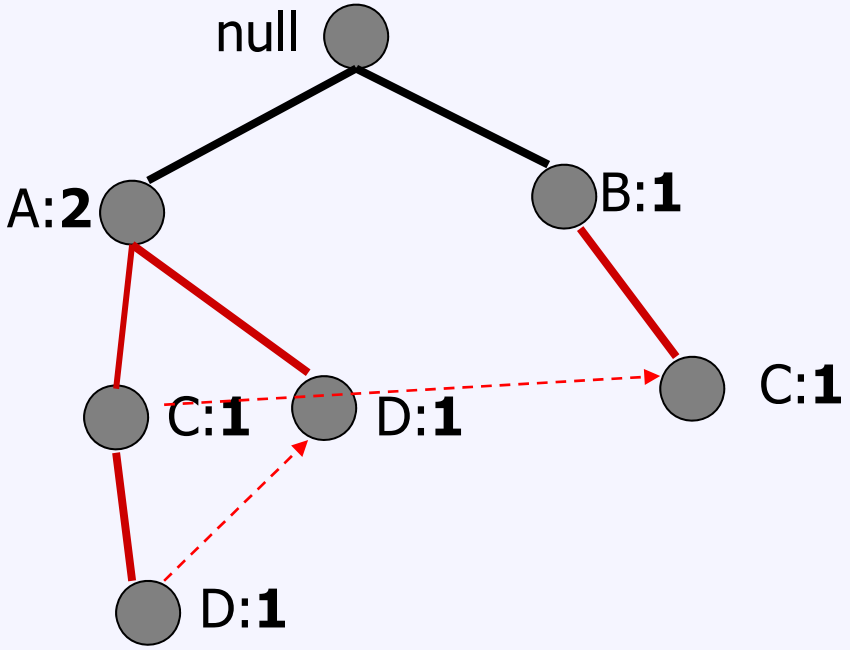
Περικοπή (truncate):
Σβήσε τους κόμβους του E



Περικοπή

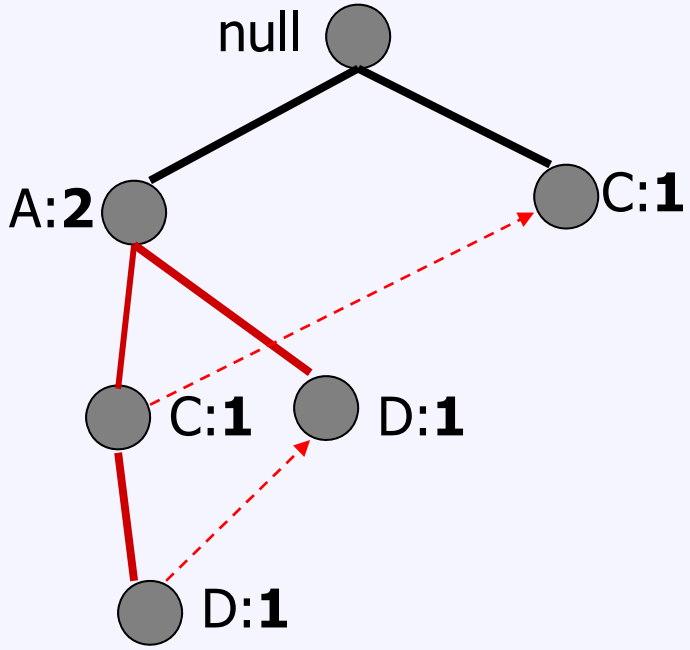
Κάποια στοιχεία μπορεί να έχουν υποστήριξη μικρότερη της ελάχιστης (π.χ., B).

Αυτό σημαίνει ότι το B εμφανίζεται μαζί με το E λιγότερο από minsup φορές



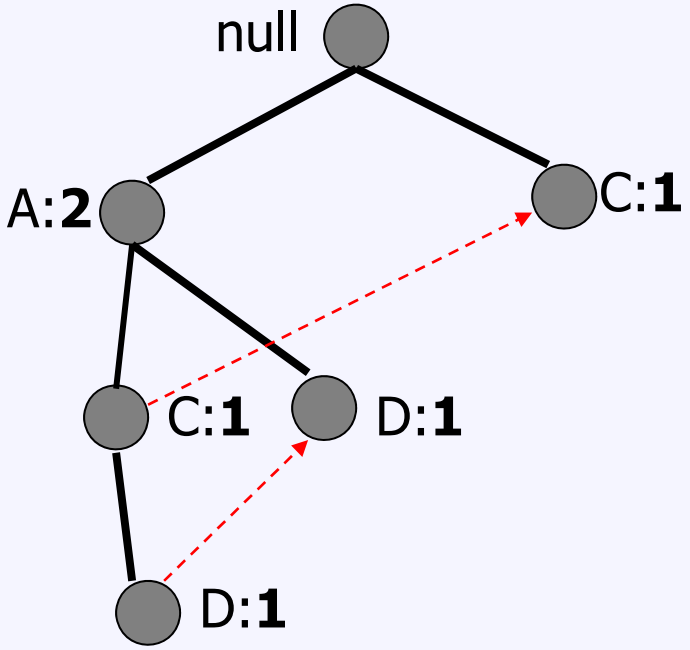
Άρα B → περικοπή

Περικοπή



Αναδρομή

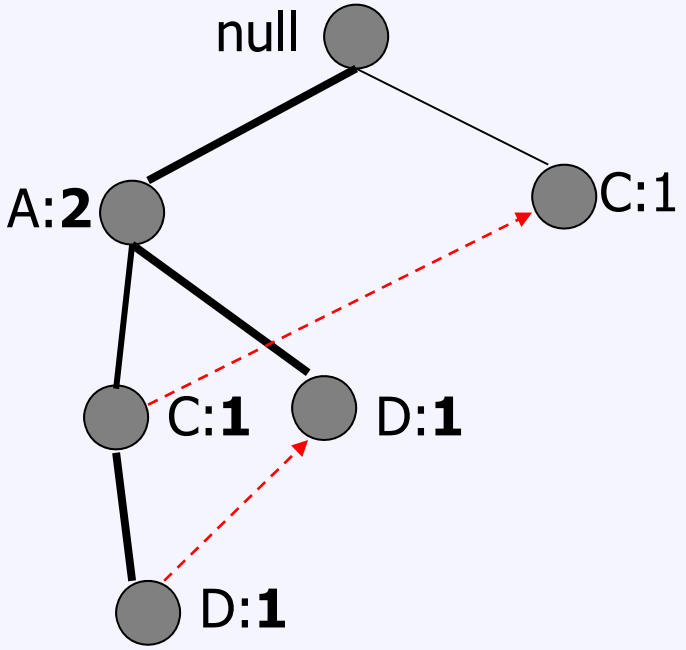
Υπο-συνθήκη FP-δένδρο για το E.



Ο αλγόριθμος επαναλαμβάνεται για το {D, E}, {C, E}, {A, E}

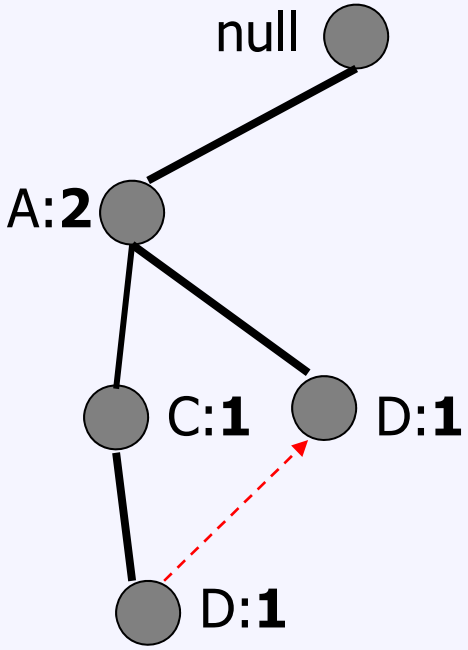
Φάση 1

Βρίσκουμε όλα τα μονοπάτια που περιέχουν το D (DE)



Φάση 1

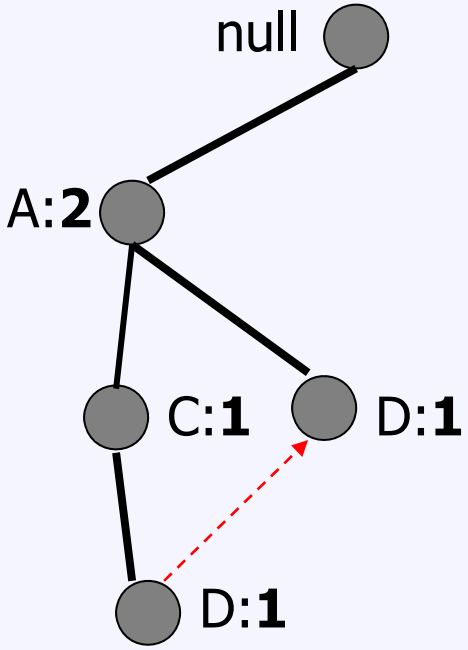
Βρίσκουμε όλα τα μονοπάτια που περιέχουν το D (DE)



Υποστήριξη DE

Ακολουθούμε τους συνδέσμους αθροίζοντας: 1+1=2≥2

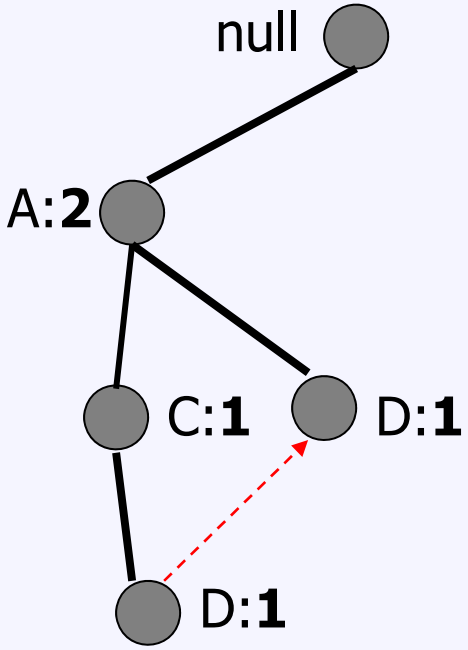
Οπότε {D, E} συχνό.



Φάση 2: Υπο-συνθήκη δένδρο

Κατασκεύασε το υπο-συνθήκη FP-δένδρο για το {D, E}

1. Αλλαγή υποστήριξης
2. Περικοπές κόμβων



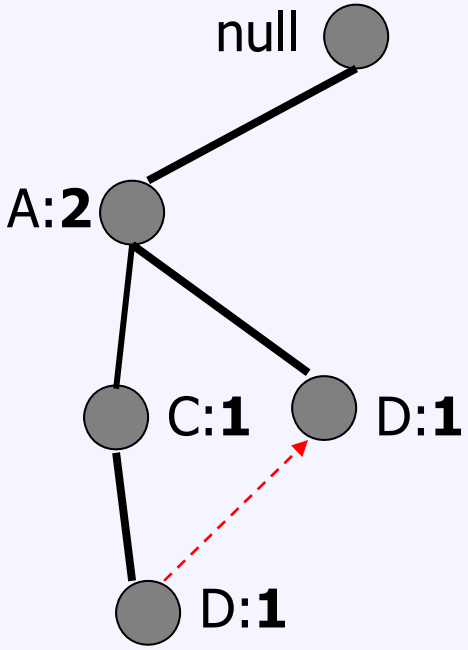
Φάση 2: Υπο-συνθήκη δένδρο

Κατασκεύασε το υπο-συνθήκη FP-δένδρο για το {D, E}

1. Αλλαγή υποστήριξης:

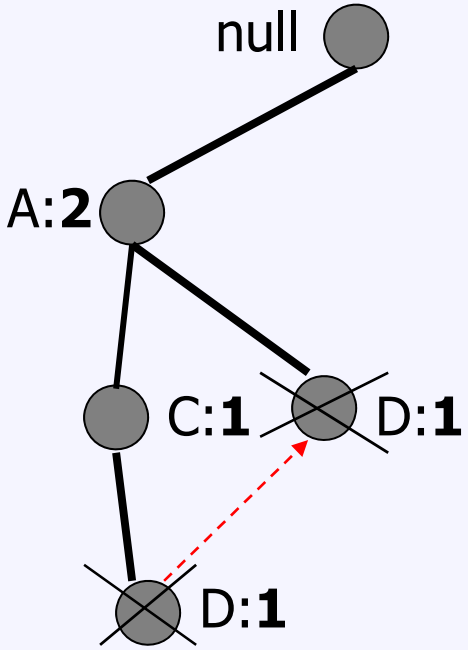
   Δεν υπάρχει καμία

2. Περικοπές κόμβων



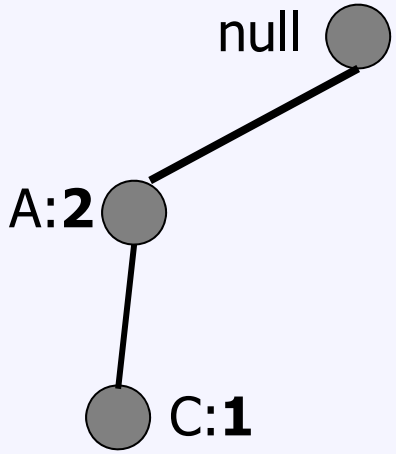
Φάση 2: Υπο-συνθήκη δένδρο

2. Περικοπές κόμβων



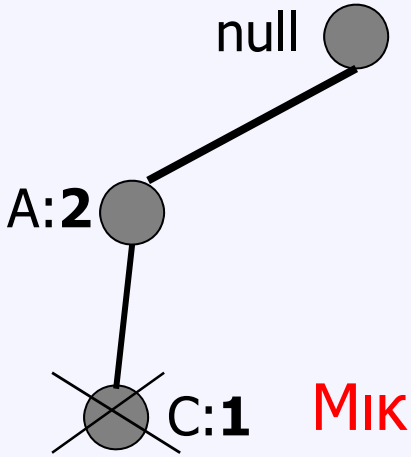
Φάση 2: Υπο-συνθήκη δένδρο

2. Περικοπές κόμβων



Φάση 2: Υπο-συνθήκη δένδρο

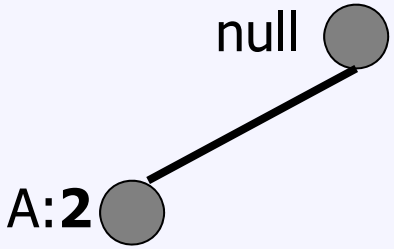
2. Περικοπές κόμβων



Μικρή υποστήριξη

Φάση 2: Υπο-συνθήκη δένδρο

Τελικό υπο-συνθήκη FP-δένδρο για το {D, E}



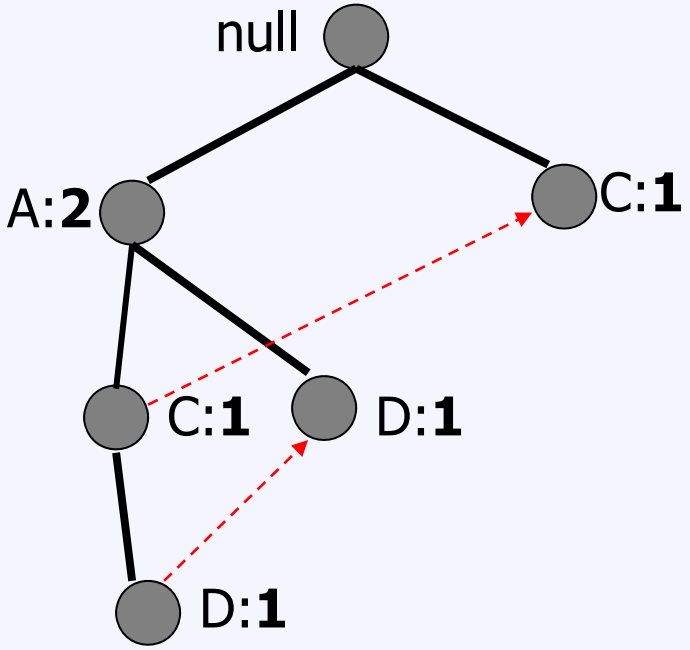
Υποστήριξη του A είναι ≥ minsup -> {A, D, E} συχνό

Αφού μόνο ένας κόμβος απέμεινε, επιστροφή στο επόμενο υποπρόβλημα.

Αναδρομή

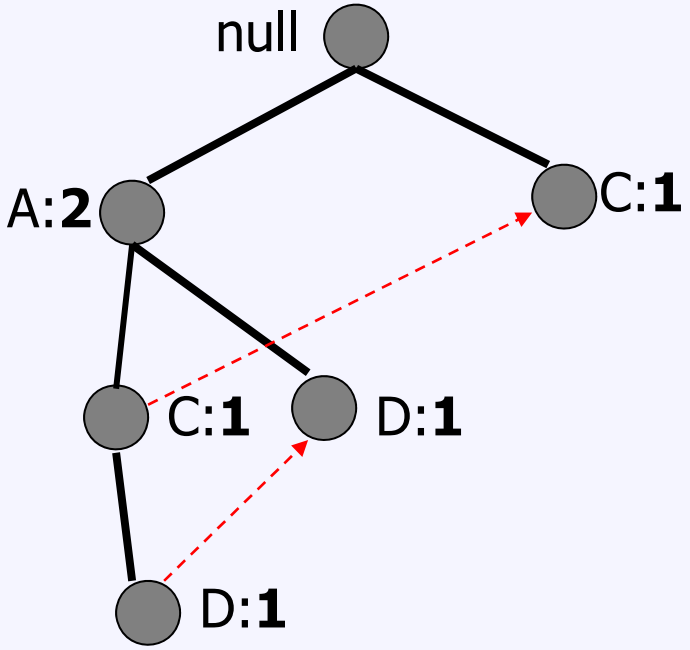
Υπο-συνθήκη FP-δένδρο για το E.

Ο αλγόριθμος επαναλαμβάνεται για το {D, E}, **{C, E},** {A, E}



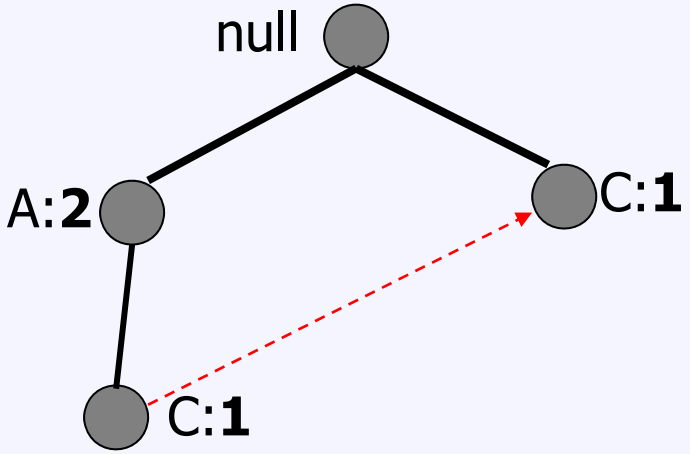
Φάση 1

Όλα τα μονοπάτια που περιέχουν το C (CE)



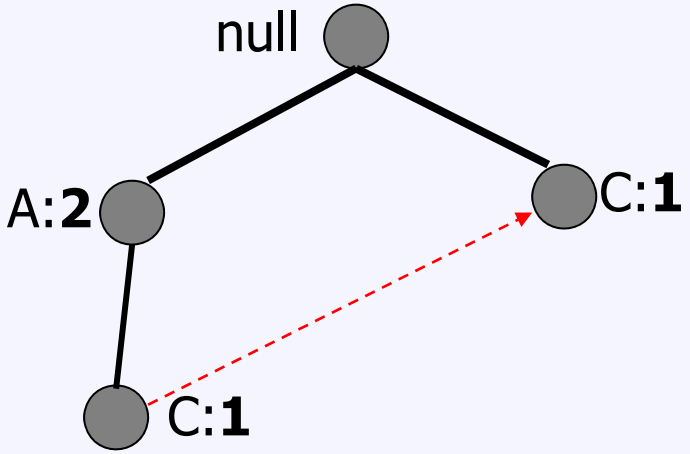
Φάση 1

Όλα τα μονοπάτια που περιέχουν το C (CE)



Υποστήριξη CE

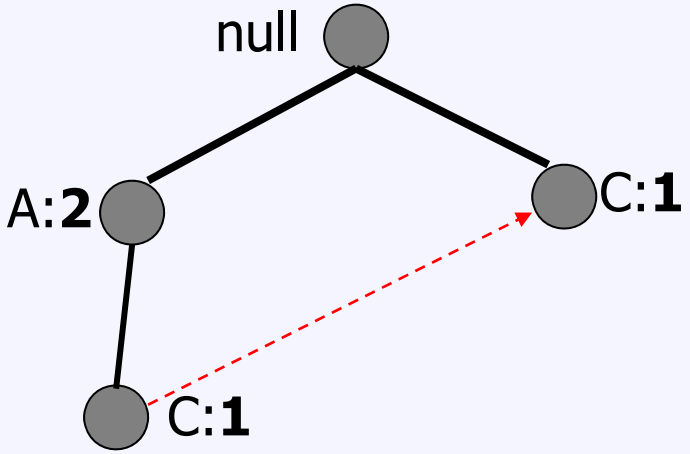
{C, E} συχνό



Φάση 2

Κατασκεύασε το υπο-συνθήκη FP-δένδρο για το {C, E}

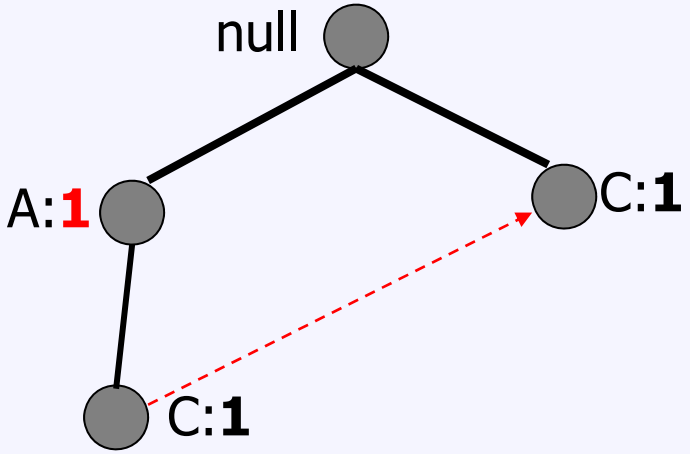
1. Αλλαγή υποστήριξης
2. Περικοπές κόμβων



Φάση 2

Κατασκεύασε το υπο-συνθήκη FP-δένδρο για το {C, E}

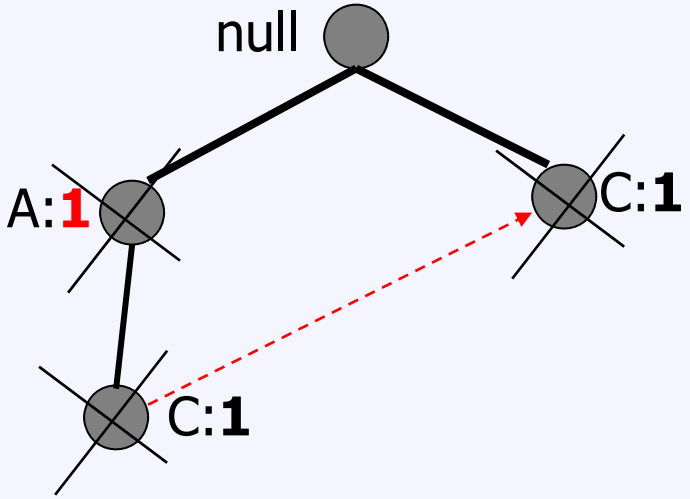
1. Αλλαγή υποστήριξης
2. Περικοπές κόμβων



Φάση 2

Κατασκεύασε το υπο-συνθήκη FP-δένδρο για το {C, E}

1. Αλλαγή υποστήριξης
2. Περικοπές κόμβων



Φάση 2

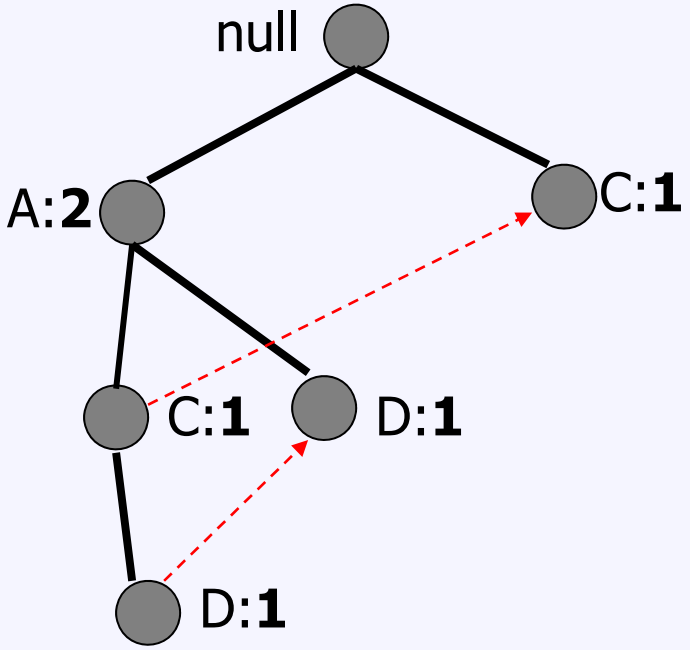
2. Περικοπή Κόμβων

Άρα, επιστροφή στο επόμενο υποπρόβλημα.

Αναδρομή

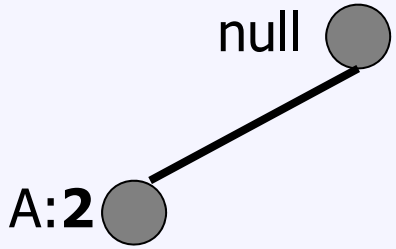
Υπο-συνθήκη FP-δένδρο για το E.

Ο αλγόριθμος επαναλαμβάνεται για το {D, E}, {C, E}, **{A, E}**



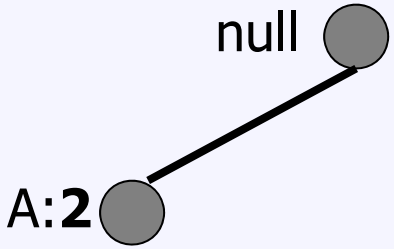
Φάση 1

Όλα τα μονοπάτια που περιέχουν το A (AE)



Υποστήριξη AE

{A, E} συχνό



Δε χρειάζεται να φτιάξουμε υπο-συνθήκη FP-δένδρο για το {A, E}

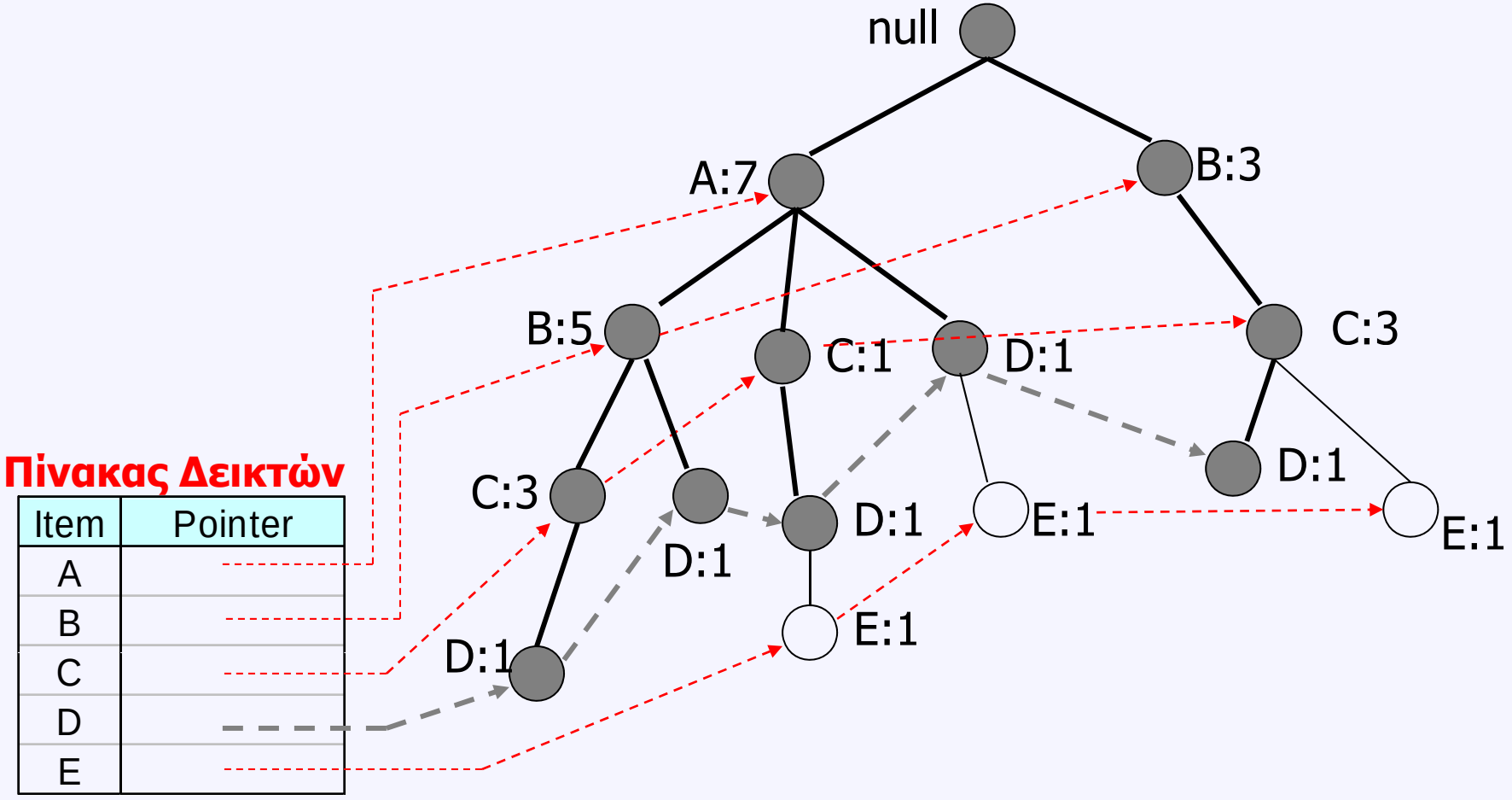
Συνολικά για το E

Έχουμε τα εξής συχνά στοιχειοσύνολα

- {E} {D, E} {A, D, E} {C, E} {A, E}

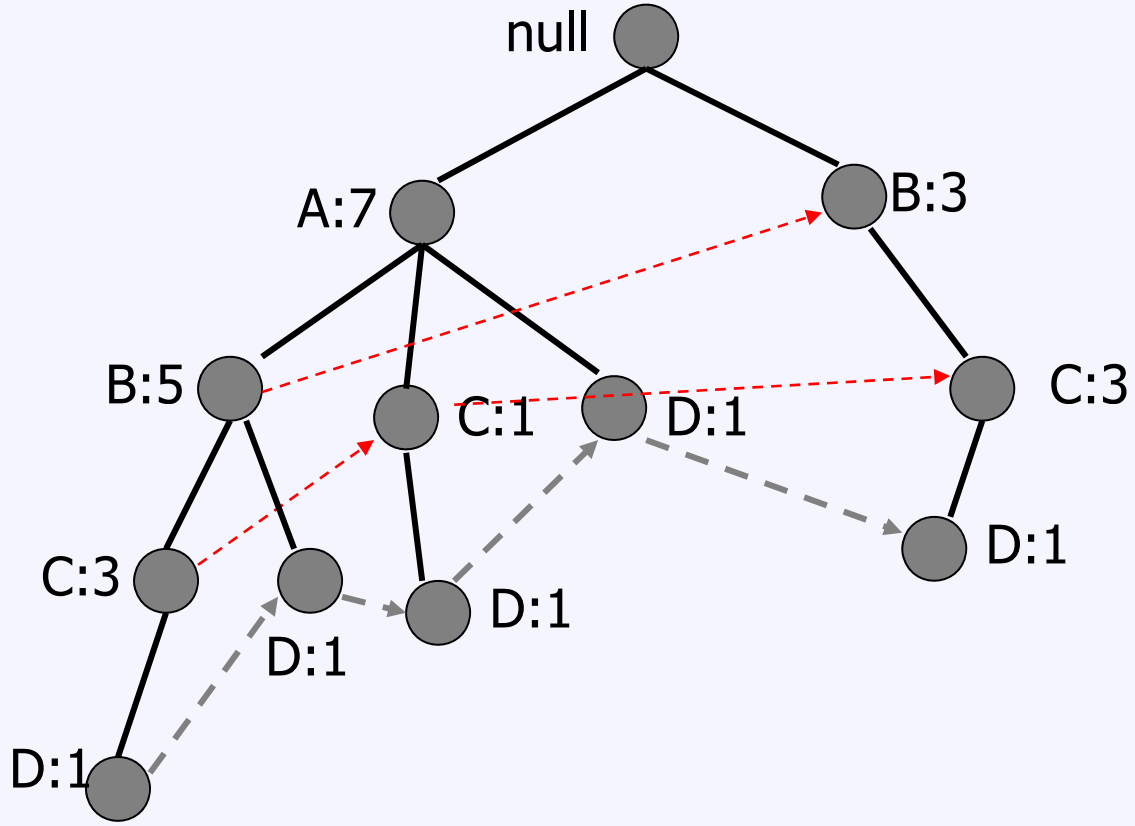
- Συνεχίζουμε για το D

Συχνά στοιχειοσύνολα που λήγουν σε D

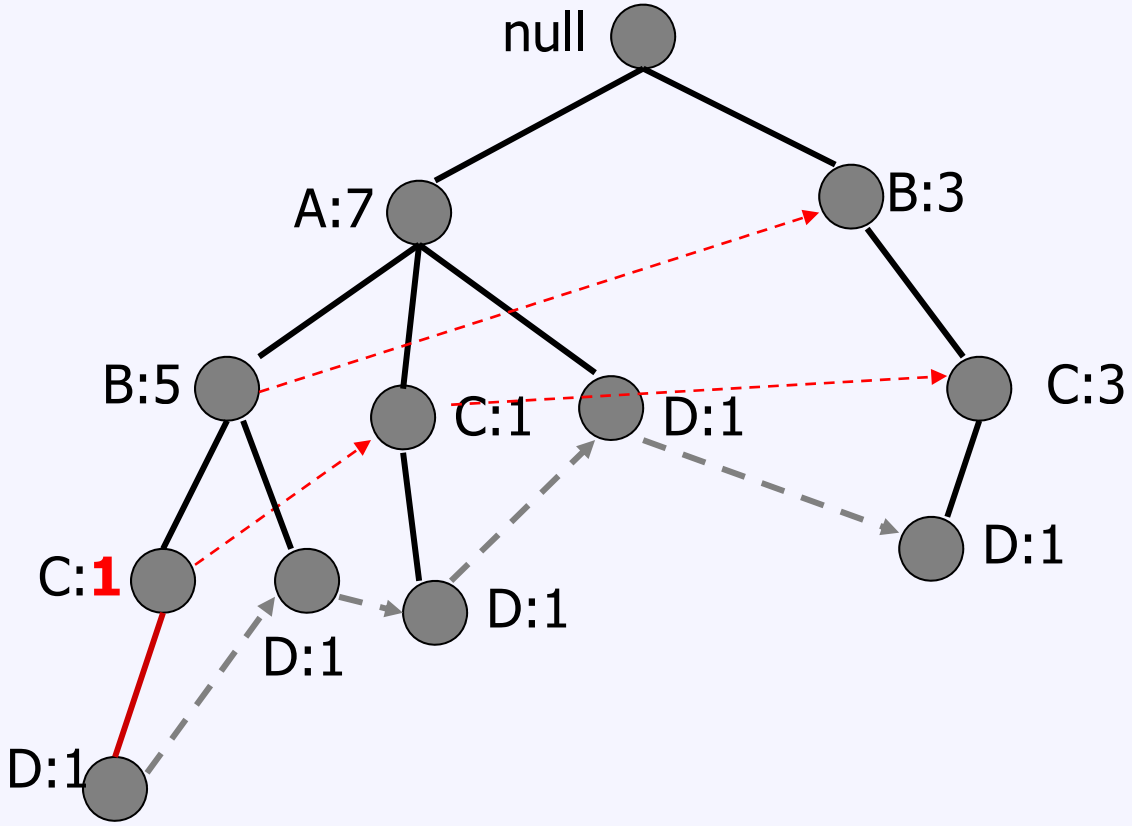


Φάση 1

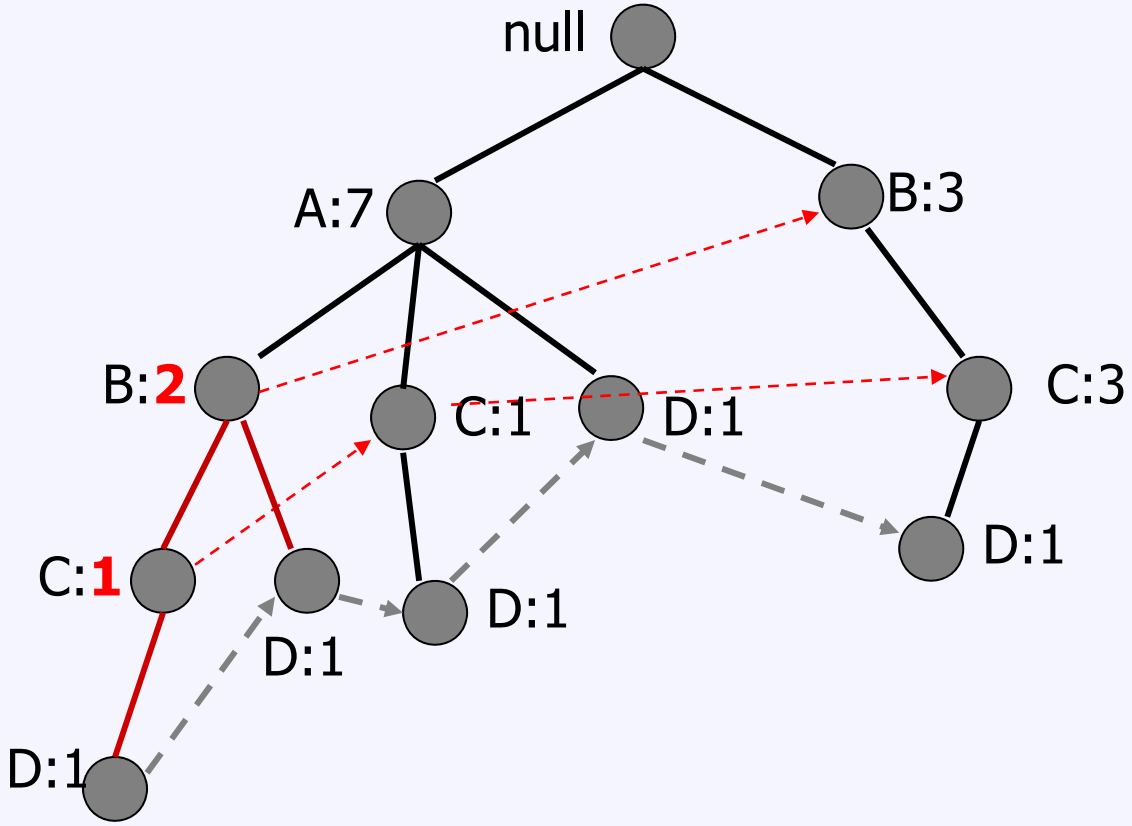
- Κρατάμε όλα τα προθεματικά μονοπάτια που περιέχουν το D
- Υποστήριξη 5>2: άρα συχνό το D
- Μετατροπή του προθεματικού δένδρου σε FP-δένδρο υπό συνθήκη



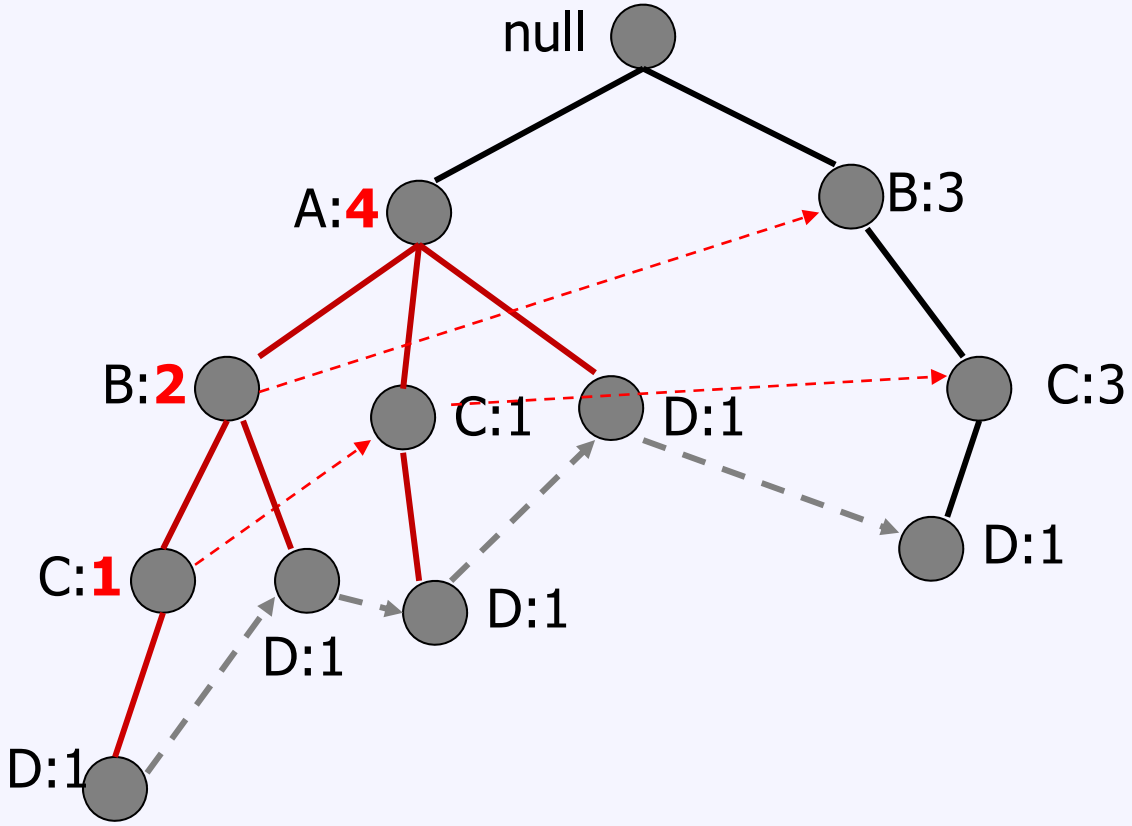
Αλλαγή υποστήριξης



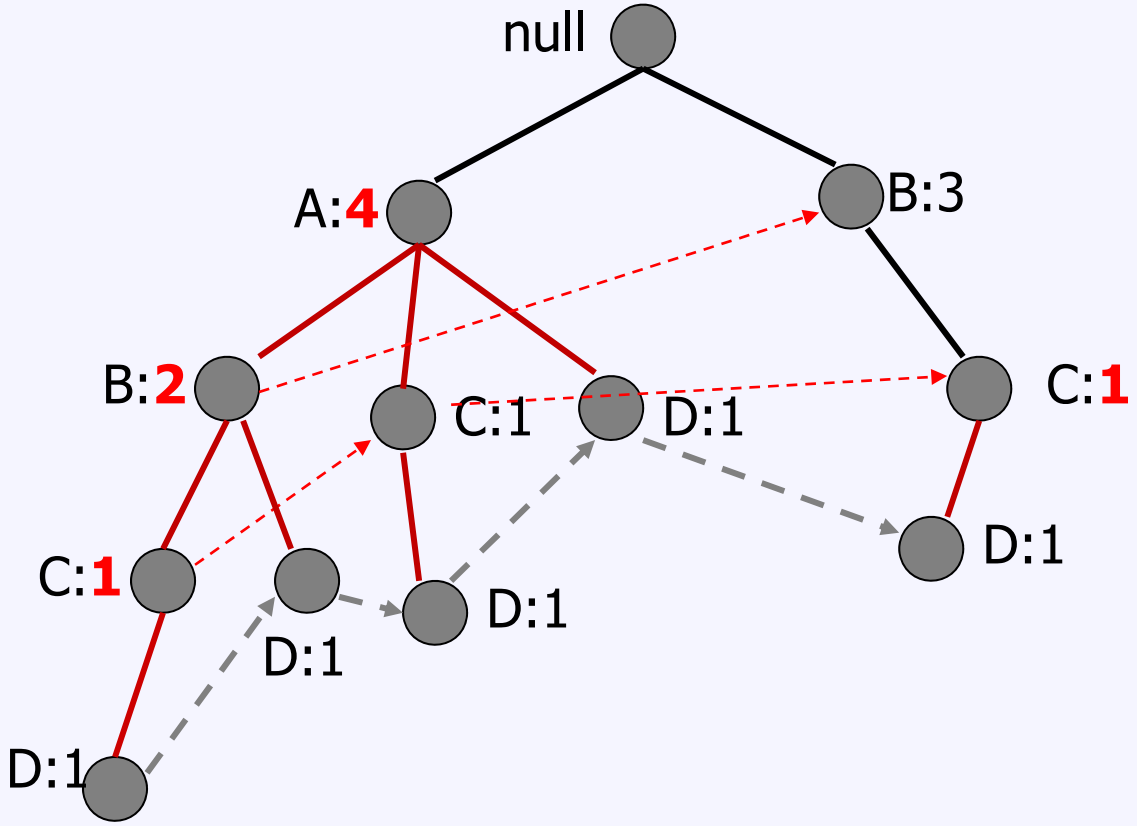
Αλλαγή υποστήριξης



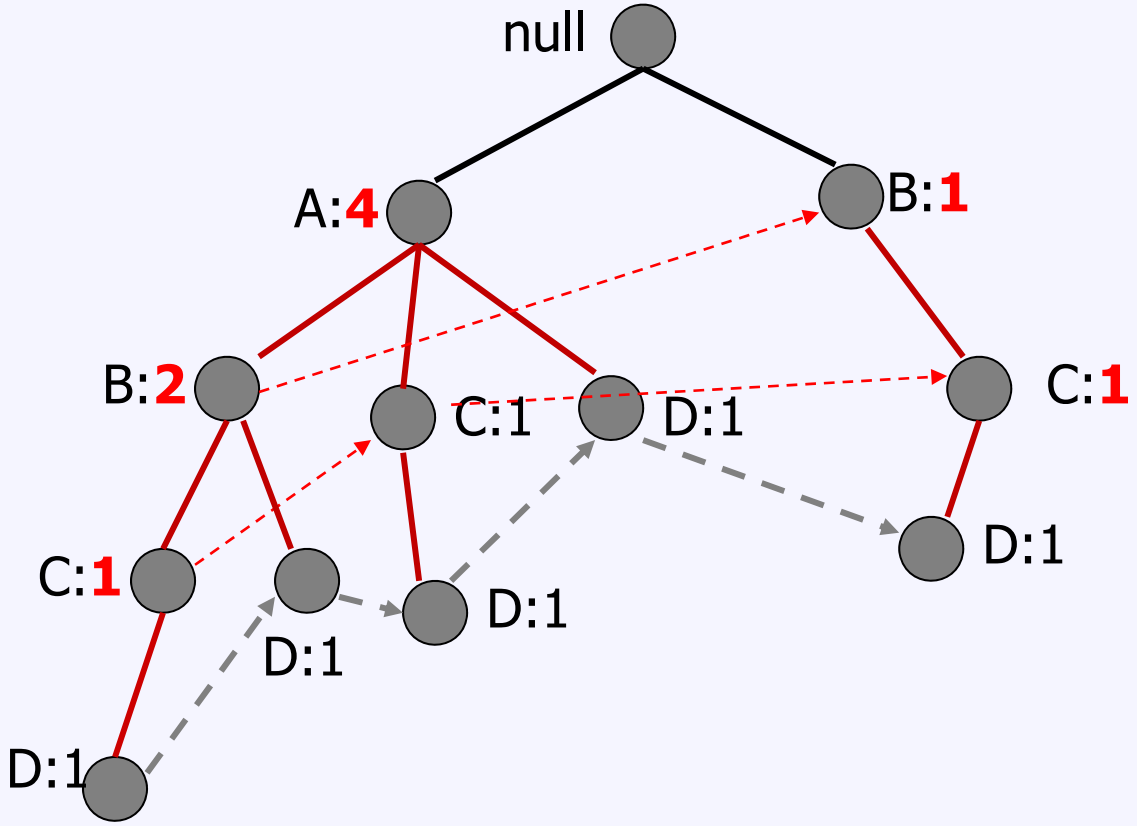
Αλλαγή υποστήριξης



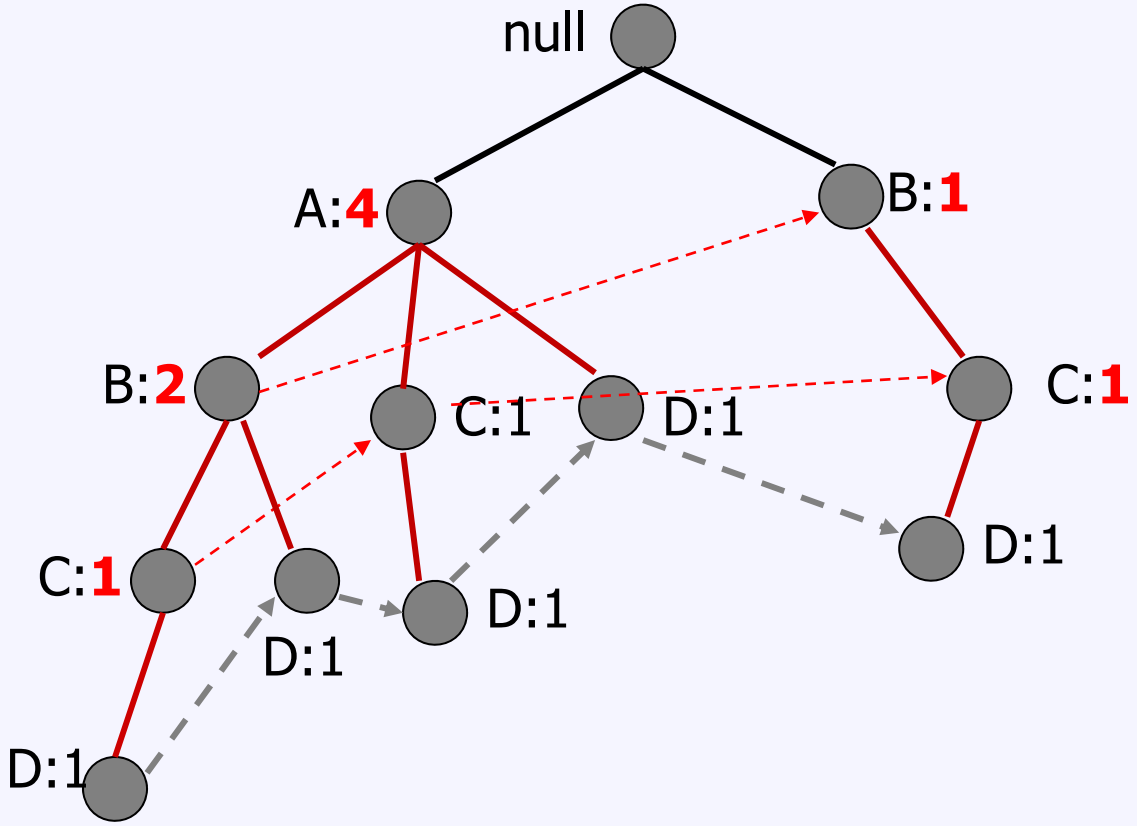
Αλλαγή υποστήριξης



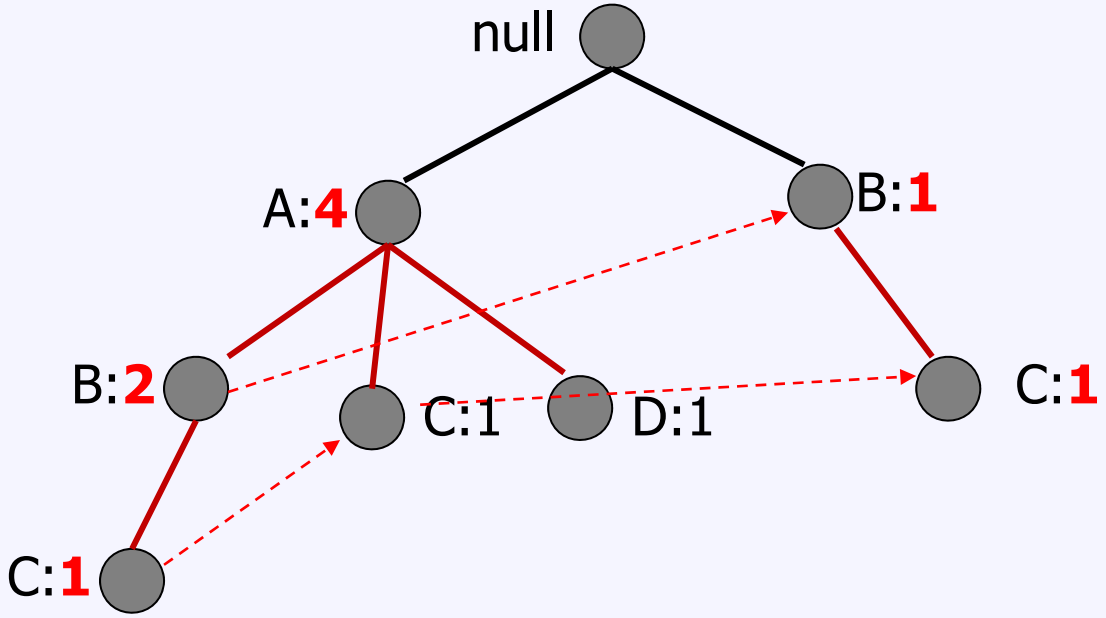
Αλλαγή υποστήριξης



Επόμενο βήμα: Περικοπή κόμβων



Επόμενο βήμα: Περικοπή κόμβων



Κατόπιν συνεχίζουμε για AD, BD, CD

Παρατηρήσεις

- Εφαρμογή τεχνικής διαίρει-και-βασίλευε.
- Σε κάθε αναδρομικό βήμα, λύνεται και ένα υπο-πρόβλημα:
  - Κατασκευάζεται το προθεματικό δένδρο
  - Υπολογίζεται η νέα υποστήριξη για τους κόμβους του
  - Περικόπτονται οι κόμβοι με μικρή υποστήριξη
- Επειδή τα υποπροβλήματα είναι ξένα μεταξύ τους, δεν δημιουργούνται τα ίδια συχνά στοιχειοσύνολα δυο φορές.
- Ο υπολογισμός της υποστήριξης είναι αποδοτικός
  - γίνεται ταυτόχρονα με τη δημιουργία των συχνών στοιχειοσυνόλων
- Η απόδοση του FP-Growth εξαρτάται από τον παράγοντα συμπίεσης του συνόλου των δεδομένων (compaction factor)
  - Βοηθάει η ταξινόμηση αντικειμένων κατά φθίνουσα σειρά υποστήριξης.

Άλλο ένα παράδειγμα

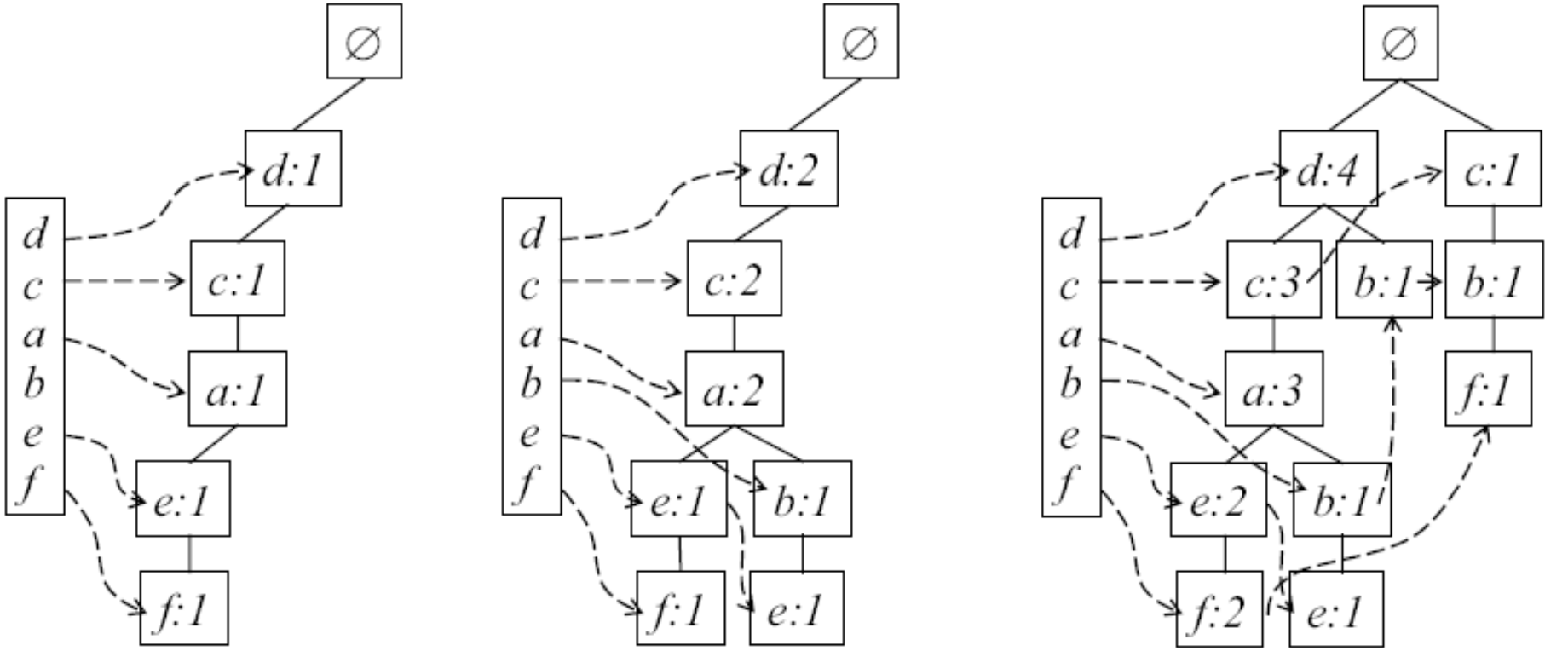
Κωδ	*Εγγραφή*	*Υποστ. Αντικειμένων*	*Αναταξινόμηση*
1	{*a, c, d, e, f*}	*d:4, c:4*	{*d, c, a, e, f*}
2	{*a, b, c, d, e*}	*a:3, b:3, e:3, f:3*	{*d, c, a, b, e*}
3	{*b, d, g*}	*g:1, h:1*	{*d, b*}
4	{*b, c, f*}		{*c, b, f*}
5	{*a, c, d, e, f, h*}		{*d, c, a, e, f*}

Minsup = 2

data engineering

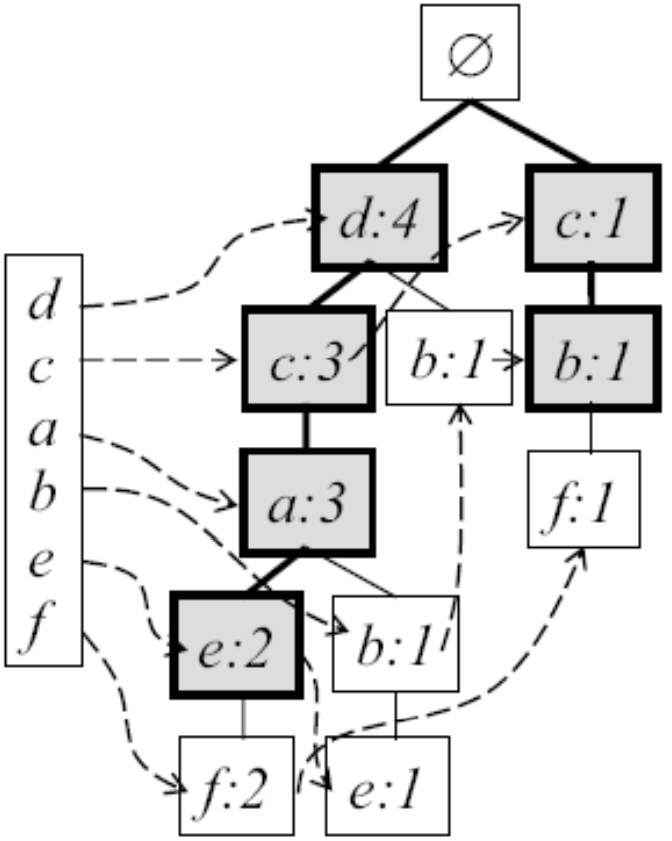
Άλλο ένα παράδειγμα

{d, c, a, e, f}
{d, c, a, b, e}
{d, b}
{c, b, f}
{d, c, a, e, f}



Άλλο ένα παράδειγμα

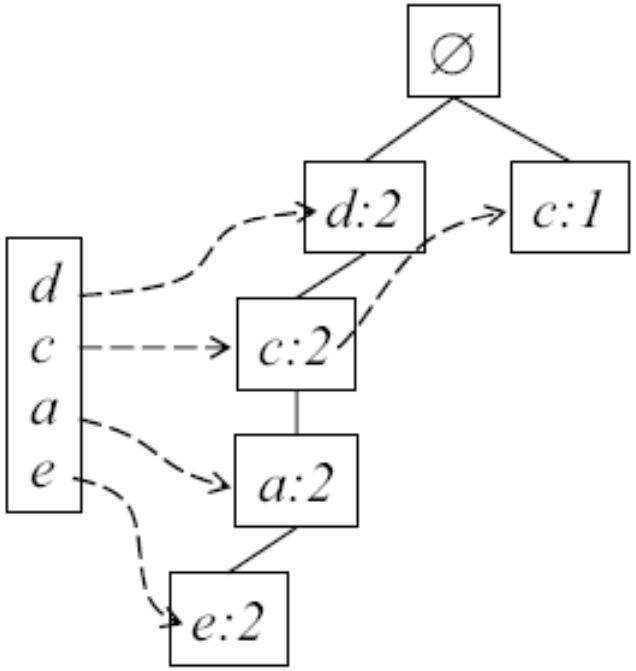
Επίθεμα	Υ.Σ. Μονοπάτια
f	{(dcae:2), (cb:1)}
e	{(dca:2), (dcab:1)}
b	{(dca:1), (d:1), (c:1)}
a	{(dc:3)}
c	{(d:3), ∅:1}
d	∅:4



Υπό Συνθήκη FP-tree | f

Άλλο ένα παράδειγμα

f	{(dcae:2), (cb:1)}



Υ.Σ. FP-tree |f

Επίθεμα	Υ.Σ. Μονοπάτια |f
ef	{(dca:2)}
af	{(dc:2)}
df	{∅:2}
cf	{(d:2), ∅:1}

Υ.Σ. Μονοπάτια |f

Άλλο ένα παράδειγμα

ef	{(dca:2)}

- Συνδυασμός ef με κάθε υποσύνολο του dca (και το κενό):
  - ef, def, cef, aef, dcef, daef, caef, dcaef
  - Όλα με υποστήριξη ίση με 2
  - Δηλ. όταν μένει μόνο ένα υπο συνθήκη μονοπάτι, σταματάμε την αναδρομική διαδικασία.
- (όμοια και για όλα τα άλλα ΥΣ μονοπάτια του f)