Data Mining: Association Analysis

Laura Brown

Some slides adapted from G. Piatetsky-Shapiro;
Han, Kamber, & Pei; Tan, Steinbach, & Kumar; A. Wasilewska

Mining Frequent Patterns w/o Cand. Gen.

- Bottlenecks of Apriori
  - breadth-first (i.e., level-wise) search
  - candidate generation and test
    - may generate huge number of candidates
- FPGrowth Approach (Han, Pei, Yin SIGMOD, 2000)
  - depth-first search
  - avoid explicit candidate generation
- Main Idea – grow long patterns from short ones using local frequent items only
  - “abc” is a frequent pattern
  - get all trans. with “abc”, project DB on abc: DB | abc
  - “d” is local frequent item in DB | abc, then abcd is freq. pattern

Construct FP-tree

- Compress a large database into a compact, Frequent-Pattern tree (FP-tree) structure
  - highly condensed, but complete for frequent pattern mining
  - helps avoid costly database scans
- Develop an efficient, FP-tree based frequent pattern mining method
  - divide and conquer methodology: decompose mining tasks into smaller ones
  - avoid candidate generation: sub-database test only

Construct FP-tree: Overview

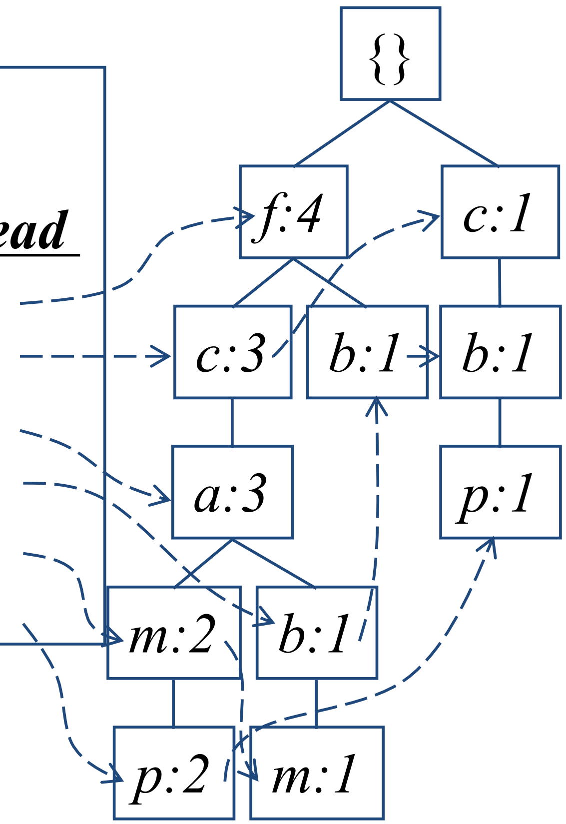
TID	*Items bought*	*(ordered) frequent items*
100	***{f, a, c, d, g, i, m, p}***	***{f, c, a, m, p}***
200	***{a, b, c, f, l, m, o}***	***{f, c, a, b, m}***
300	***{b, f, h, j, o, w}***	***{f, b}***
400	***{b, c, k, s, p}***	***{c, b, p}***
500	***{a, f, c, e, l, p, m, n}***	***{f, c, a, m, p}***

min_support = 3

1. Scan DB once, find frequent 1-itemset (single item pattern)
2. Sort frequent items in frequency descending order, f-list
3. Scan DB again, construct FP-tree

Header Table

Item	*frequency*	*head*
f	*4*	
c	*4*	
a	*3*	
b	*3*	
m	*3*	
p	*3*	



F-list = f-c-a-b-m-p

Partition Patterns and Databases

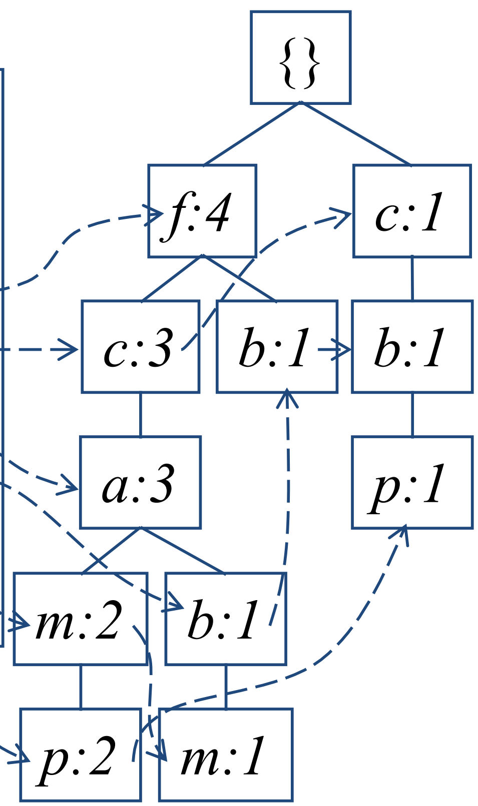
- Frequent patterns can be partitioned into subsets according to f-list
  - F-list = f-c-a-b-m-p
  - Patterns containing p
  - Patterns having m but no p
  - …
  - Patterns having c but no a nor b, m, p
  - Pattern f
- Completeness and non-redundancy

Find Patterns from P-conditional Database

- Starting at the frequent item header table in the FP-tree
- Traverse the FP-tree by following the link of each frequent item *p*
- Accumulate all of the transformed prefix paths of item *p* to form *p*'s conditional pattern base

Header Table

Item	*frequency*	*head*
f	4	
c	4	
a	3	
b	3	
m	3	
p	3	

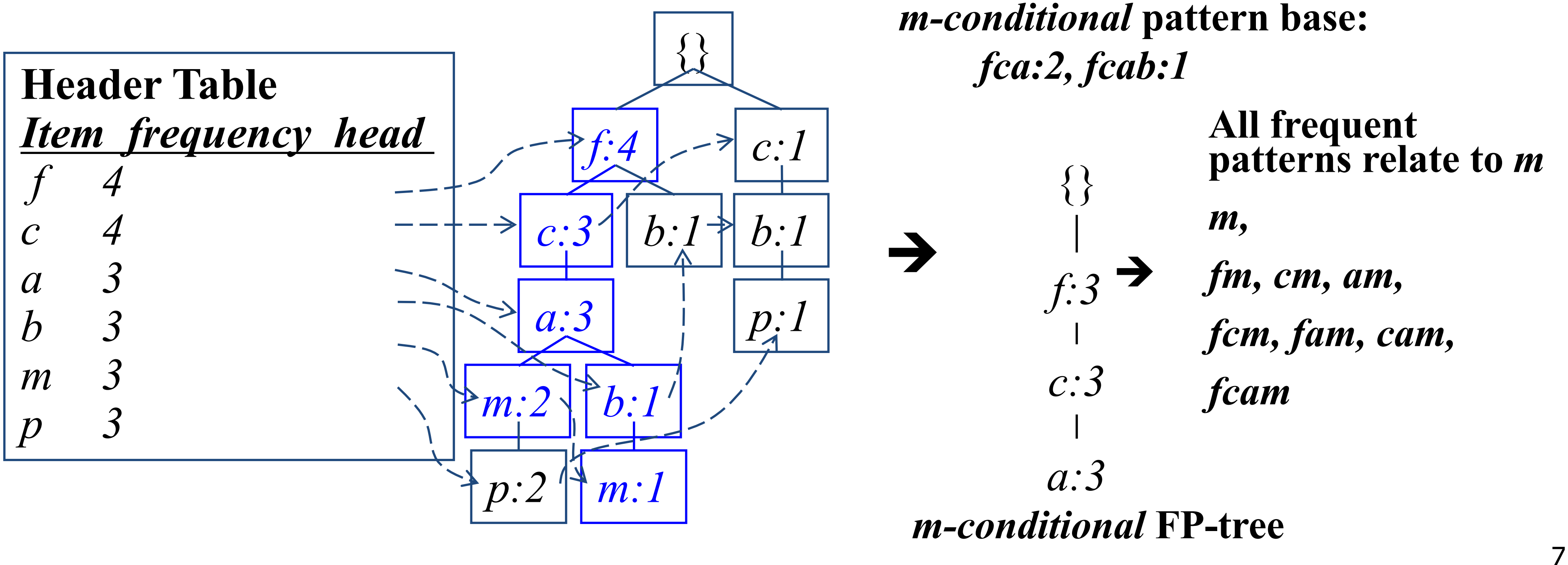


***Conditional* pattern bases**

item	***cond. pattern base***
c	***f:3***
a	***fc:3***
b	***fca:1, f:1, c:1***
m	***fca:2, fcab:1***
p	***fcam:2, cb:1***

From Conditional Pattern-bases to Conditional FP-tree

- For each pattern-base
  - Accumulate the count for each item in the base
  - Construct the FP-tree for the frequent items of the pattern base



FP-Growth Method: An Example

TID	List of Items
T100	I1, I2, I5
T100	I2, I4
T100	I2, I3
T100	I1, I2, I4
T100	I1, I3
T100	I2, I3
T100	I1, I3
T100	I1, I2 ,I3, I5
T100	I1, I2, I3

- Consider database D
- Let *minsup* = 2
- First scan is same as Apriori to derives 1-itemsets and their support counts
- Set of frequent items is sorted in order of descending support count
- L = {I2:7, I1:6, I3:6, I4:2, I5:2}

Construct FP-tree

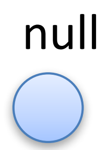
- Create root of tree, labeled “null”
- Scan D a 2nd time (first scan was to create 1-itemsets and L)
- Items are processed in L order (sorted order)
- Branch created for each transaction with items having their support count separated by colon
- Whenever same node is encountered in another transaction just increment support count of common node or prefix
- To facilitate tree traversal, an item header table is built so that each item points to its occurrences in the tree via a chain of node-links
- The problem of mining frequent patterns in D is transformed to mining the FP-tree

Construct the FP-tree: Part 1

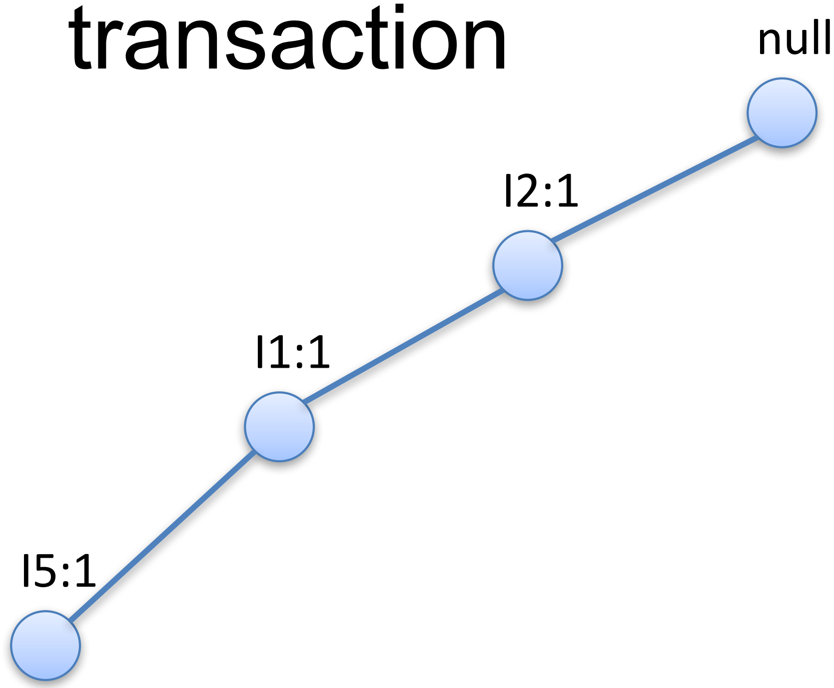
TID	List of Items
T100	I1, I2, I5
T100	I2, I4
T100	I2, I3
T100	I1, I2, I4
T100	I1, I3
T100	I2, I3
T100	I1, I3
T100	I1, I2 ,I3, I5
T100	I1, I2, I3

Header Table	
I2	
I1	
I3	
I4	
I5	

- Start, root = null



- After reading 1st transaction

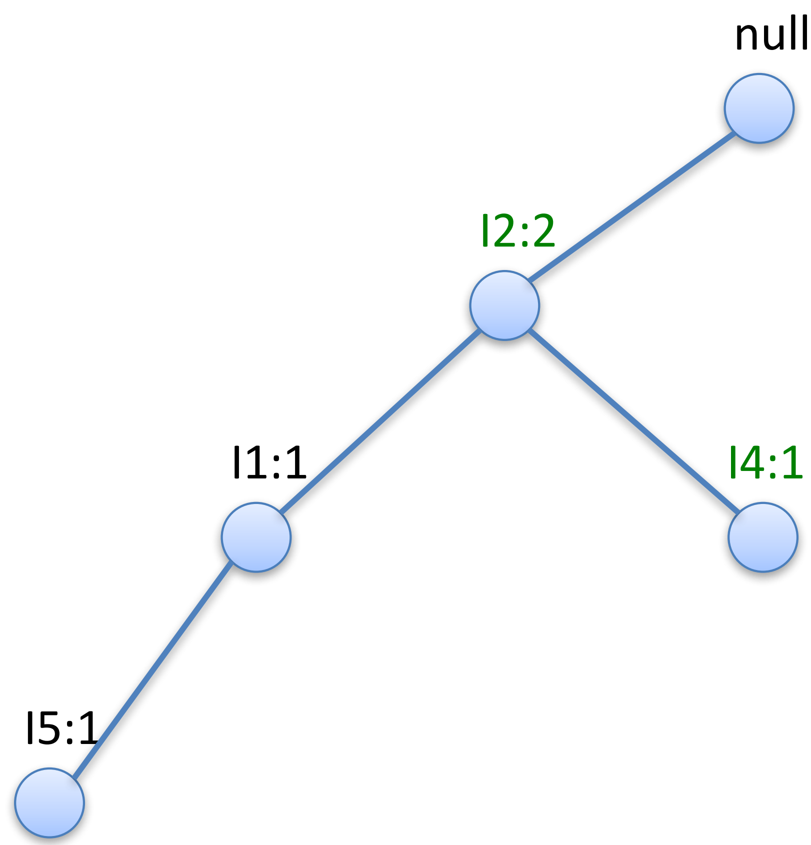


Construct the FP-tree: Part 2

TID	List of Items
T100	I1, I2, I5
T100	I2, I4
T100	I2, I3
T100	I1, I2, I4
T100	I1, I3
T100	I2, I3
T100	I1, I3
T100	I1, I2 ,I3, I5
T100	I1, I2, I3

Header Table	
I2	
I1	
I3	
I4	
I5	

After reading 2nd transaction

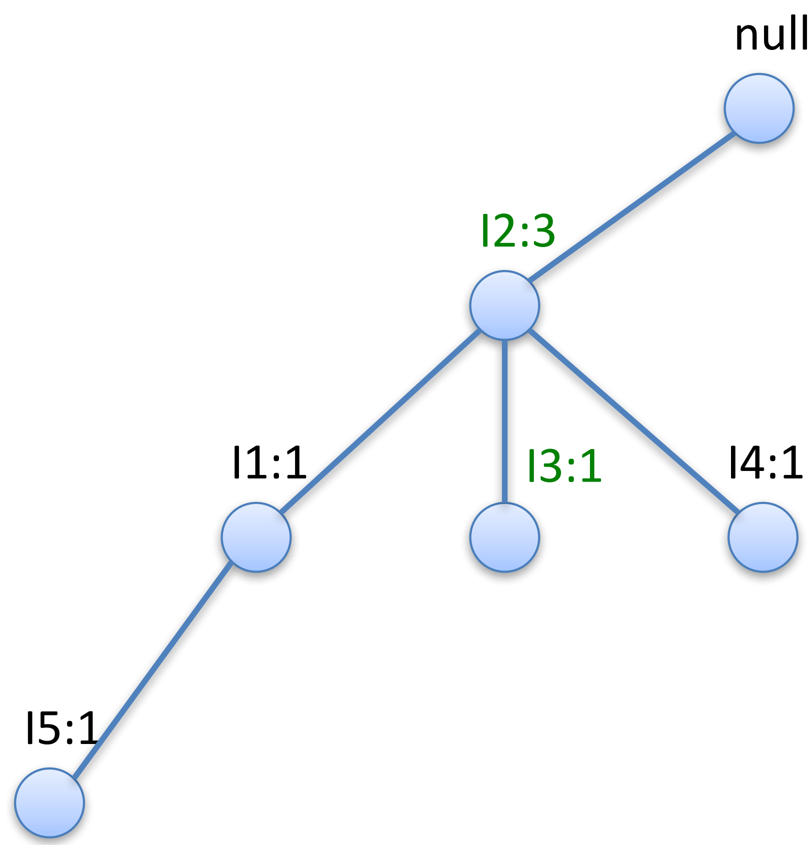


Construct the FP-tree: Part 3

TID	List of Items
T100	I1, I2, I5
T100	I2, I4
T100	I2, I3
T100	I1, I2, I4
T100	I1, I3
T100	I2, I3
T100	I1, I3
T100	I1, I2 ,I3, I5
T100	I1, I2, I3

Header Table	
I2	
I1	
I3	
I4	
I5	

After reading 3rd transaction

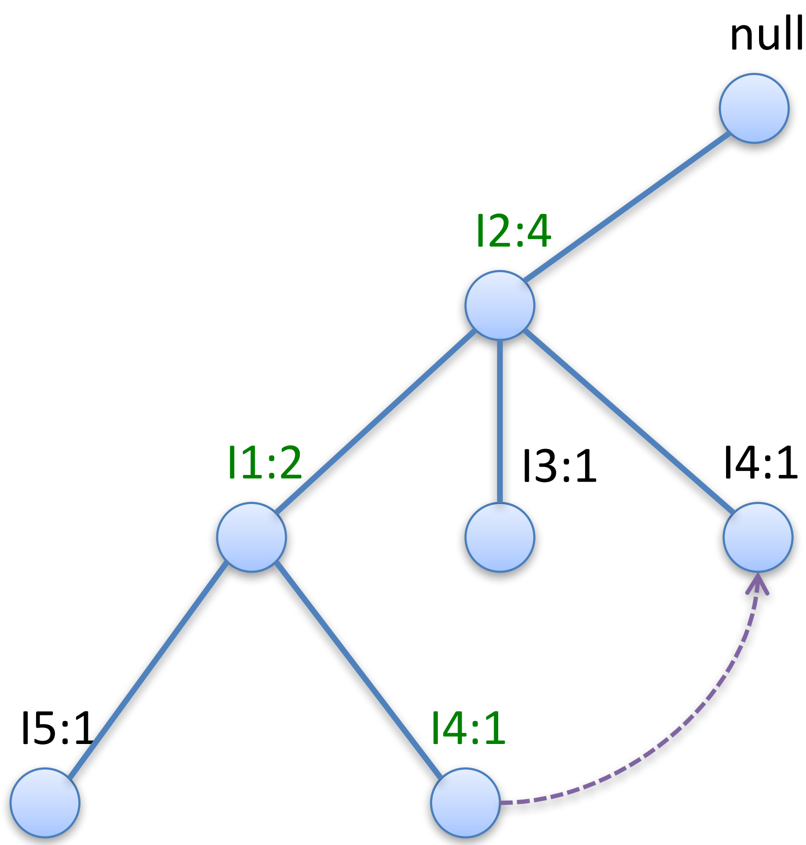


Construct the FP-tree: Part 4

TID	List of Items
T100	I1, I2, I5
T100	I2, I4
T100	I2, I3
T100	I1, I2, I4
T100	I1, I3
T100	I2, I3
T100	I1, I3
T100	I1, I2 ,I3, I5
T100	I1, I2, I3

Header Table	
I2	
I1	
I3	
I4	
I5	

After reading 4th transaction

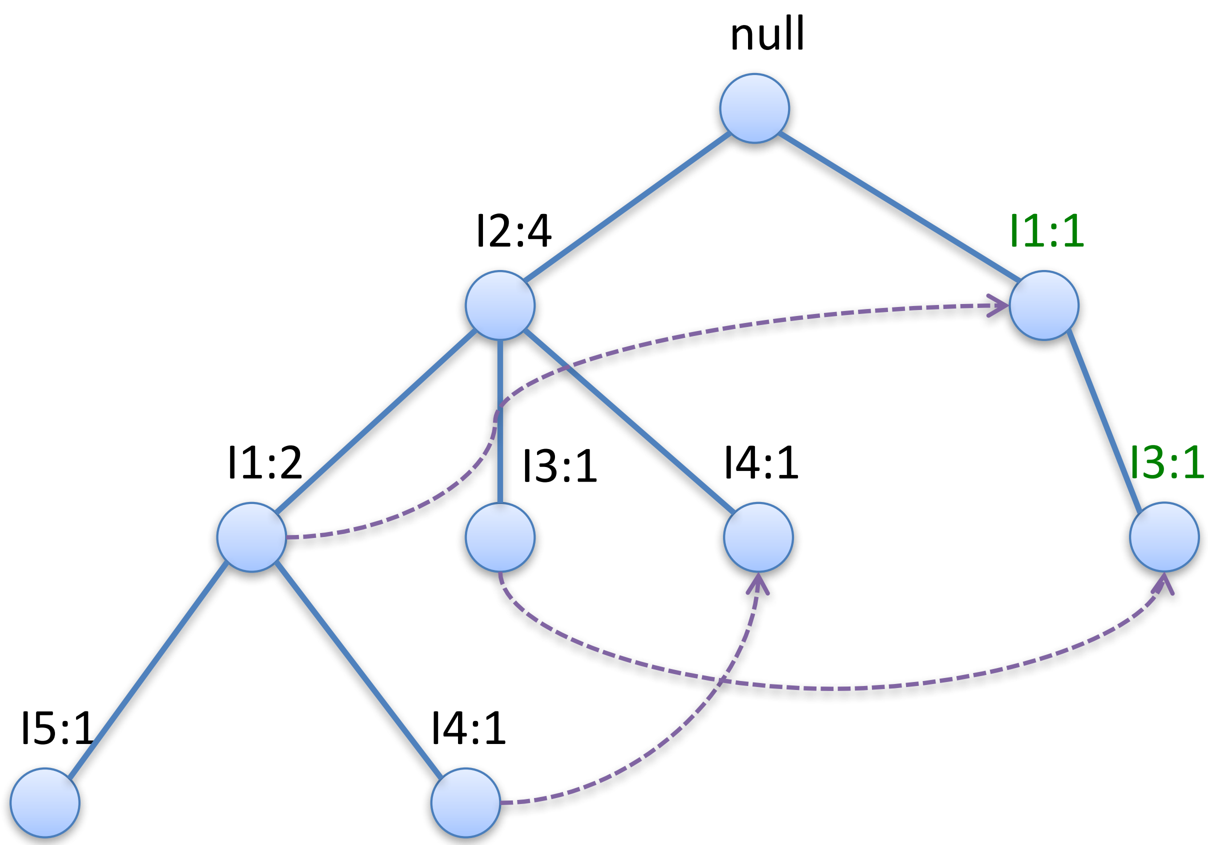


Construct the FP-tree: Part 5

TID	List of Items
T100	I1, I2, I5
T100	I2, I4
T100	I2, I3
T100	I1, I2, I4
T100	I1, I3
T100	I2, I3
T100	I1, I3
T100	I1, I2 ,I3, I5
T100	I1, I2, I3

Header Table	
I2	
I1	
I3	
I4	
I5	

After reading 5th transaction

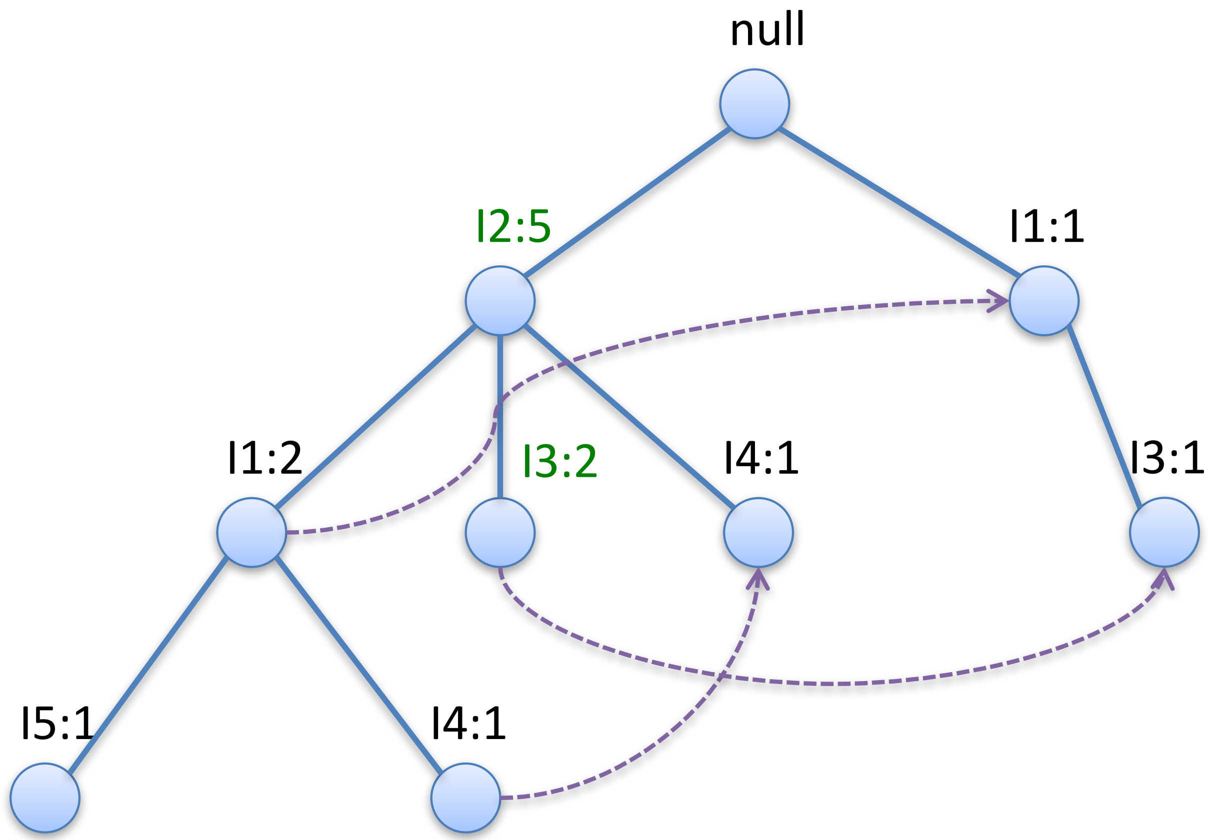


Construct the FP-tree: Part 6

TID	List of Items
T100	I1, I2, I5
T100	I2, I4
T100	I2, I3
T100	I1, I2, I4
T100	I1, I3
T100	I2, I3
T100	I1, I3
T100	I1, I2 ,I3, I5
T100	I1, I2, I3

Header Table	
I2	
I1	
I3	
I4	
I5	

After reading 6th transaction

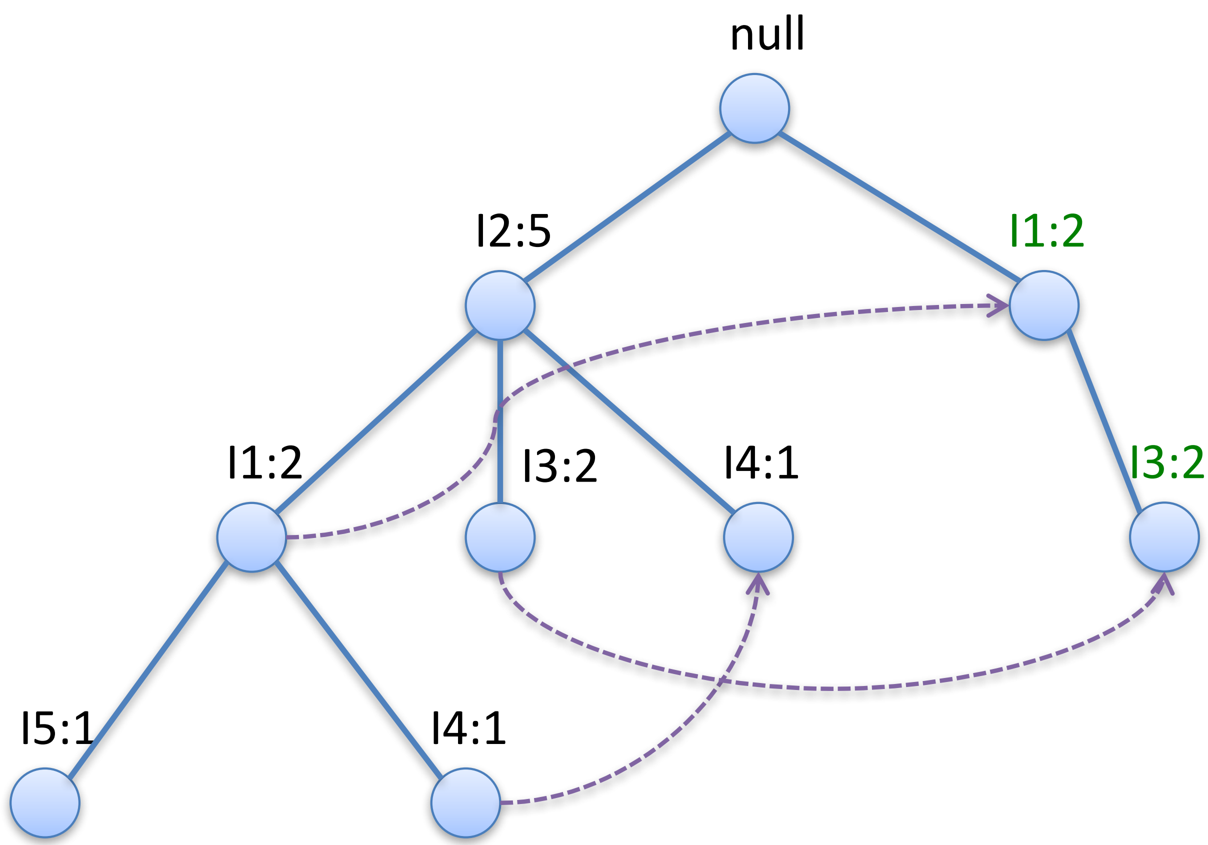


Construct the FP-tree: Part 7

TID	List of Items
T100	I1, I2, I5
T100	I2, I4
T100	I2, I3
T100	I1, I2, I4
T100	I1, I3
T100	I2, I3
T100	I1, I3
T100	I1, I2 ,I3, I5
T100	I1, I2, I3

Header Table	
I2	
I1	
I3	
I4	
I5	

After reading 7th transaction

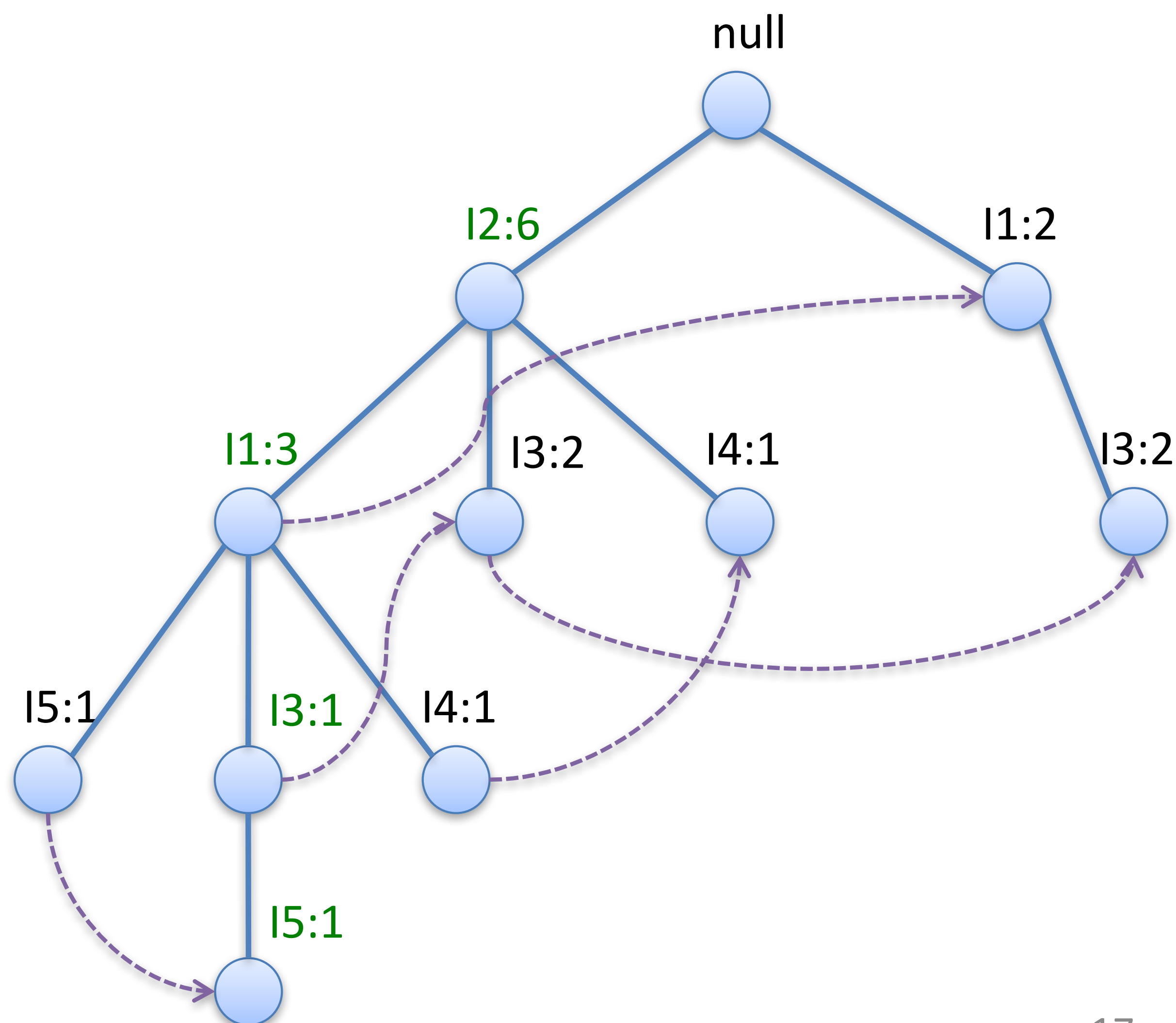


Construct the FP-tree: Part 8

TID	List of Items
T100	I1, I2, I5
T100	I2, I4
T100	I2, I3
T100	I1, I2, I4
T100	I1, I3
T100	I2, I3
T100	I1, I3
T100	I1, I2 ,I3, I5
T100	I1, I2, I3

Header Table	
I2	
I1	
I3	
I4	
I5	

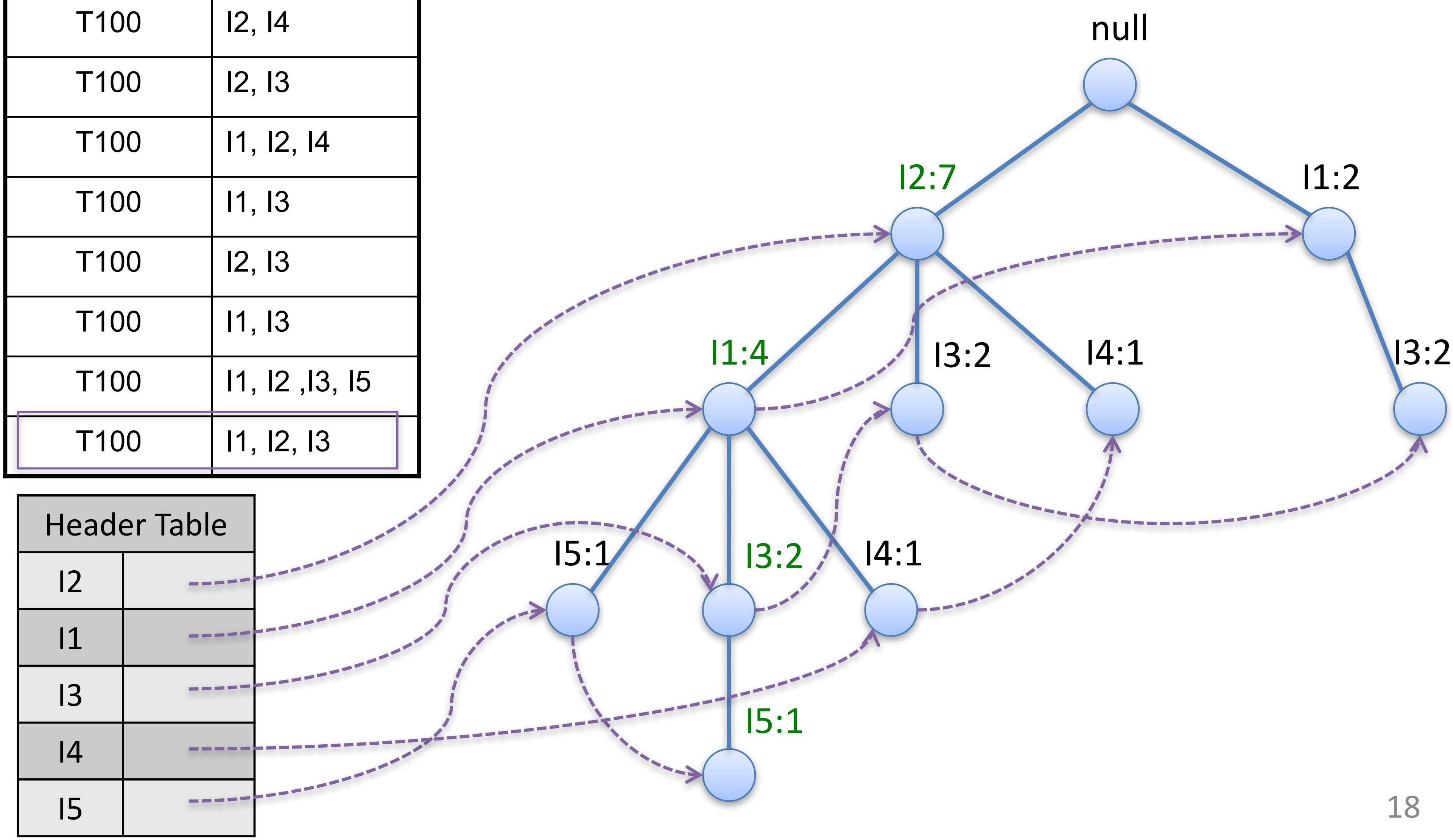
After reading 8th transaction



Construct the FP-tree: Part 9

TID	List of Items
T100	I1, I2, I5
T100	I2, I4
T100	I2, I3
T100	I1, I2, I4
T100	I1, I3
T100	I2, I3
T100	I1, I3
T100	I1, I2 ,I3, I5
T100	I1, I2, I3

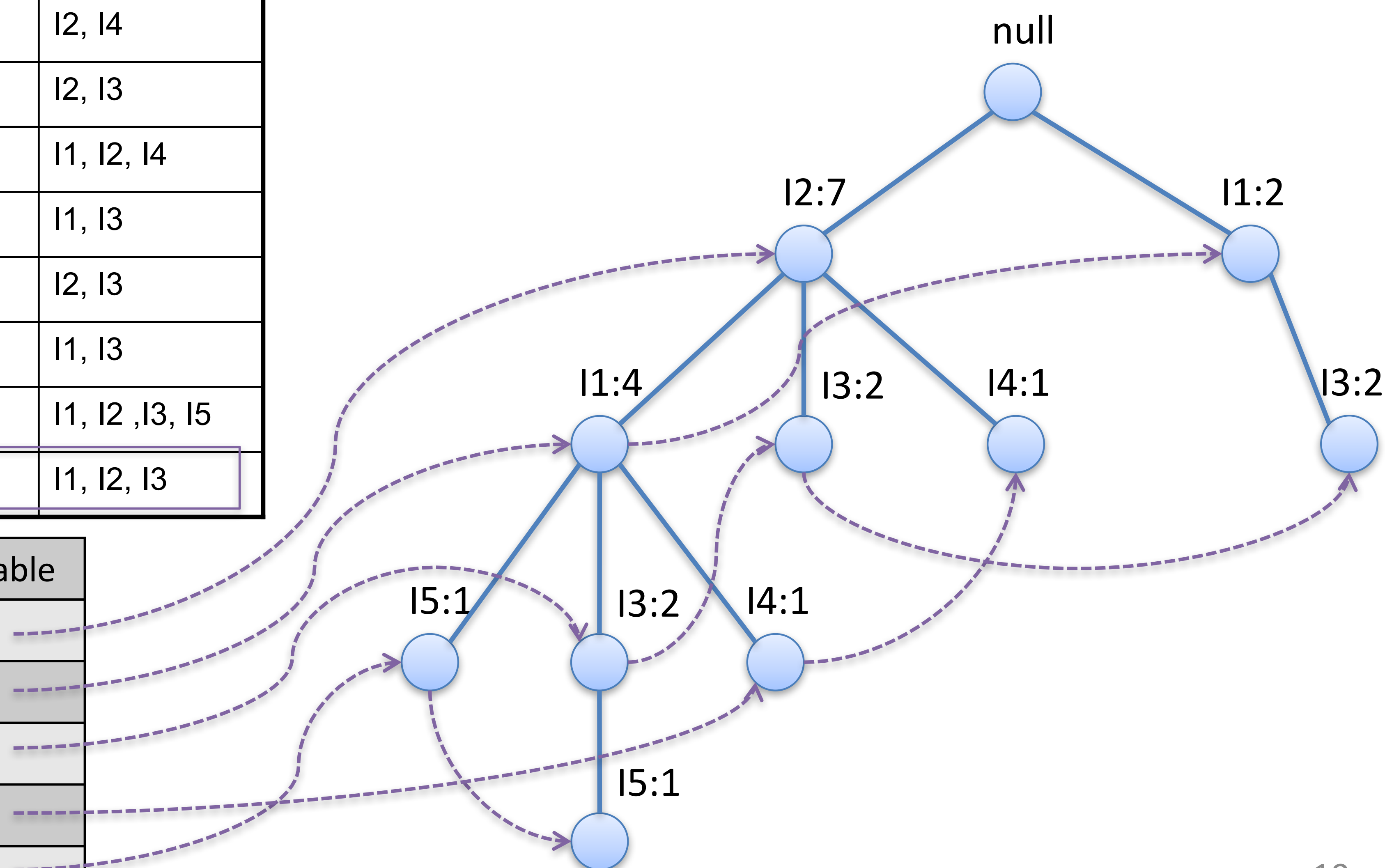
After reading 9th transaction



Construct the FP-tree: complete

TID	List of Items
T100	I1, I2, I5
T100	I2, I4
T100	I2, I3
T100	I1, I2, I4
T100	I1, I3
T100	I2, I3
T100	I1, I3
T100	I1, I2 ,I3, I5
T100	I1, I2, I3

Header Table	
I2	
I1	
I3	
I4	
I5	



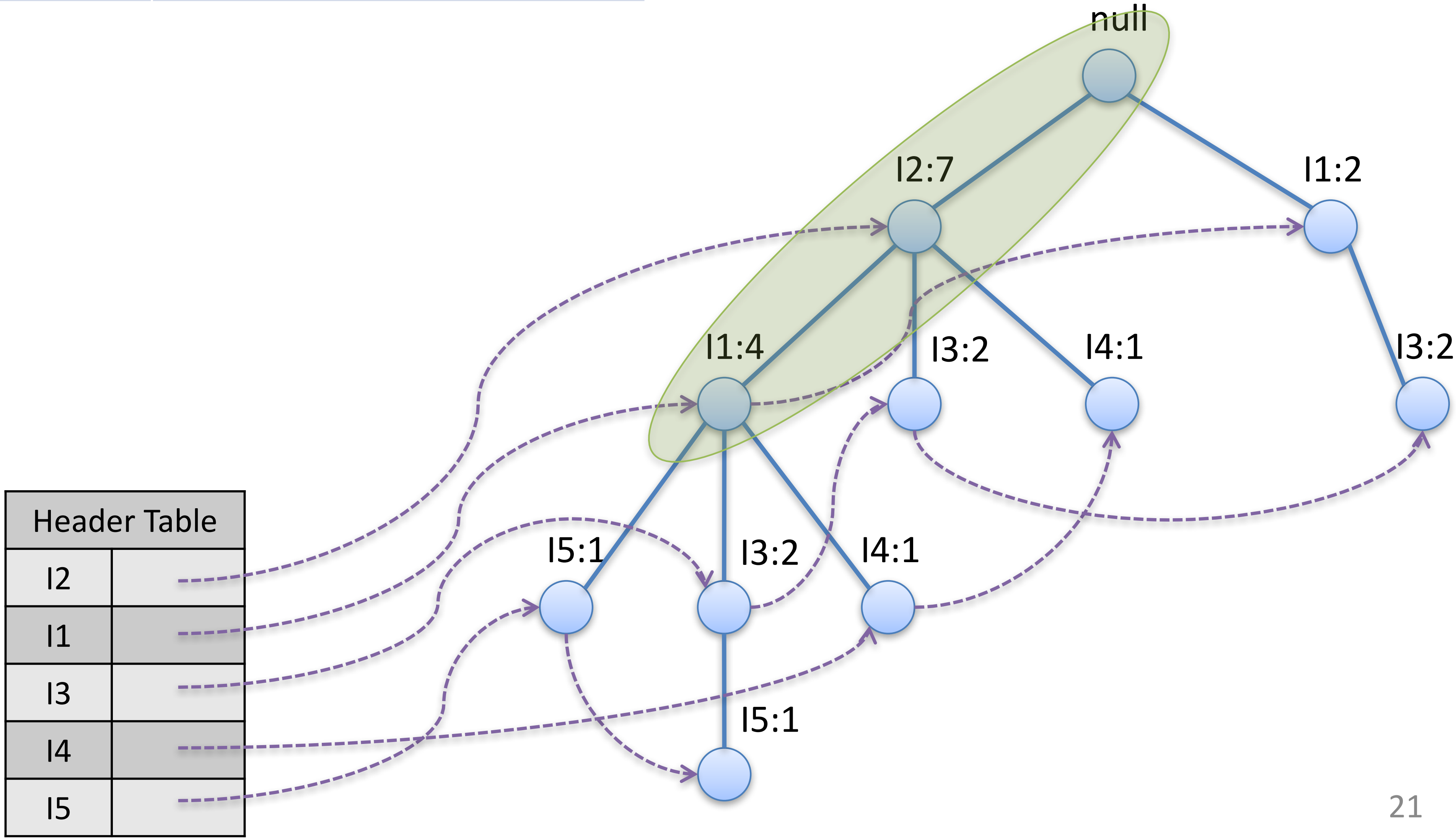
Mine FP-tree: Create Conditional Pattern Base

Steps

1. Start from each frequent length 1-pattern (as an initial suffix pattern)
2. Construct its conditional pattern base which consists of the set of prefix path in the FP-tree co-occurring with suffix pattern
3. Then, construct its conditional FP-tree & perform mining on such a tree
4. The pattern growth is achieved by concatenation of the suffix pattern with the frequent patterns generated from a conditional FP-tree
5. The union of all frequent patterns (generated by step 4) gives the required frequent itemset

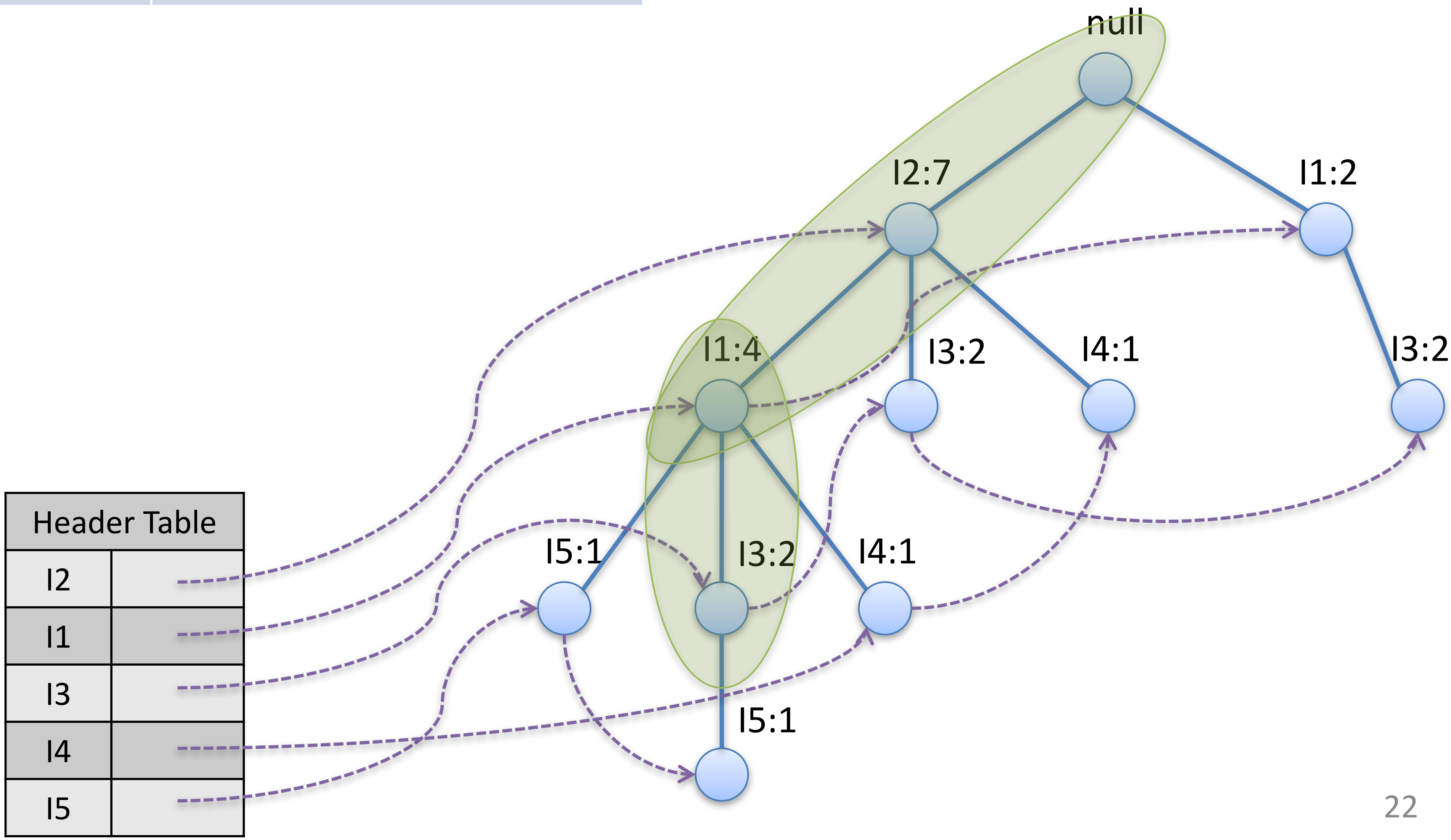
Construct Conditional Pattern Base

Item	Cond. Pattern
I5	{(I2 I1: 1)}



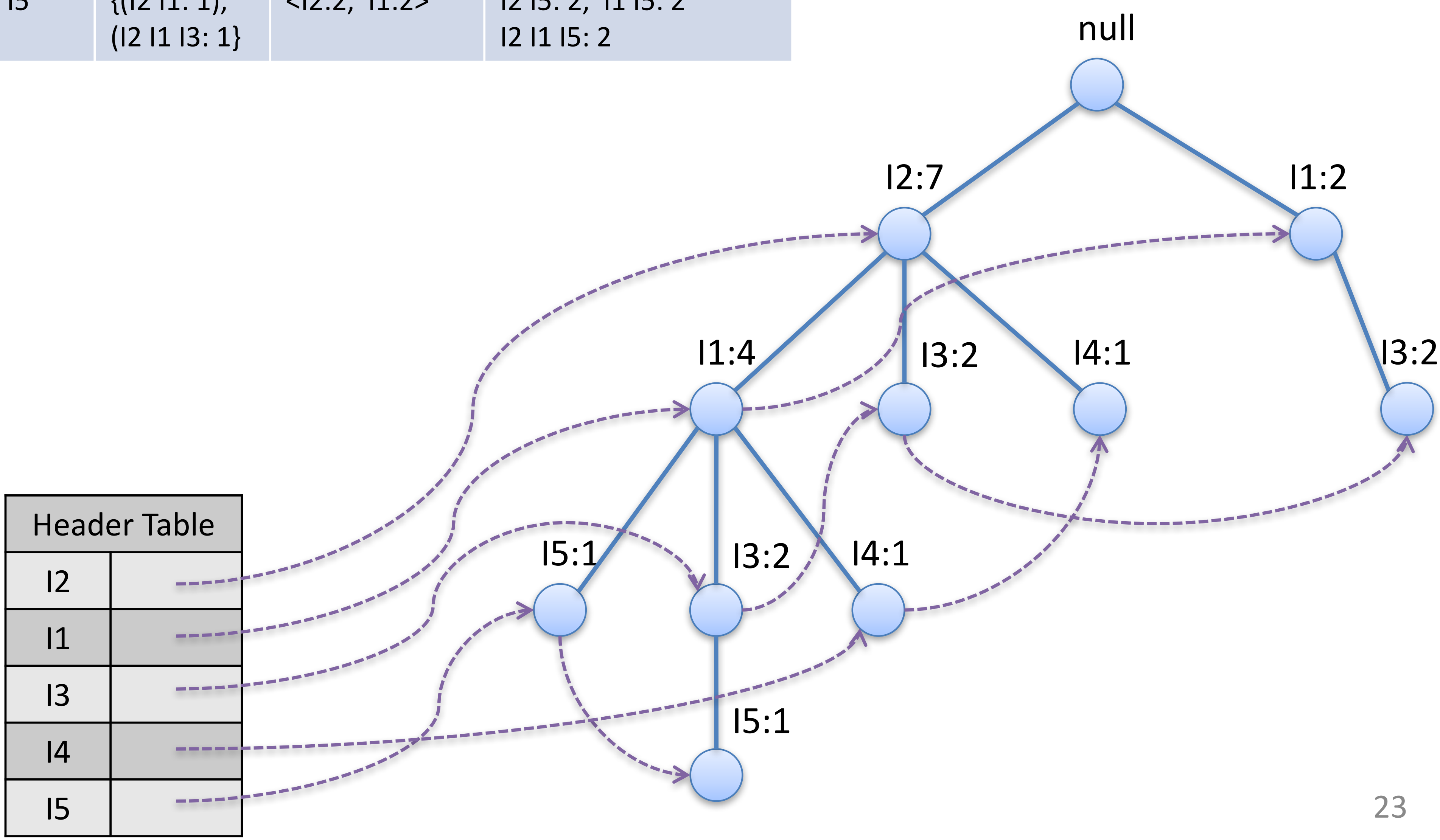
Construct Conditional Pattern Base

Item	Cond. Pattern
I5	{(I2 I1: 1), (I2 I1 I3: 1}



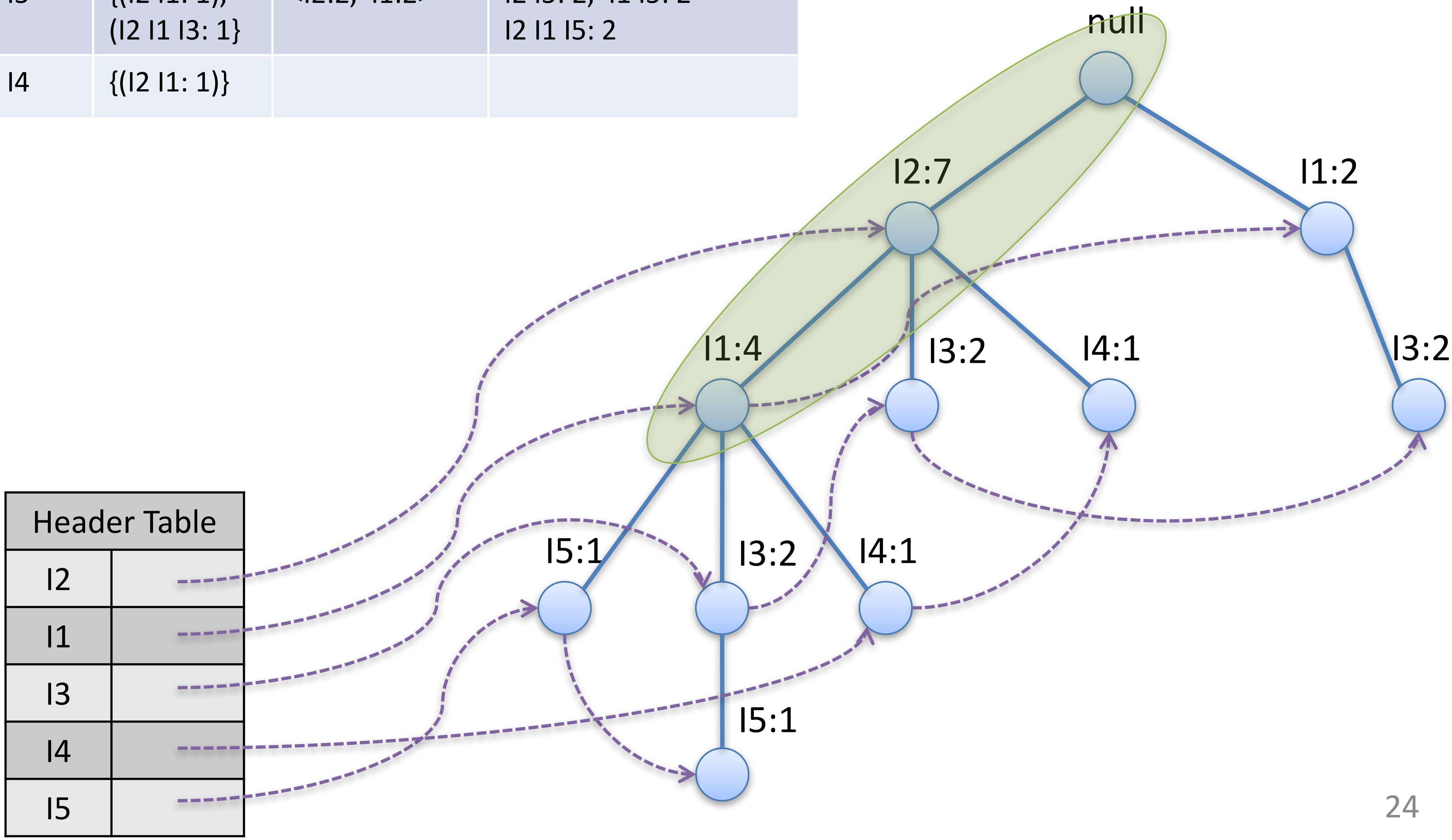
Construct Conditional Pattern Base

Item	Cond. Pattern	Cond. FP-tree	Frequent pattern
I5	{(I2 I1: 1), (I2 I1 I3: 1}	<I2:2, I1:2>	I2 I5: 2, I1 I5: 2 I2 I1 I5: 2



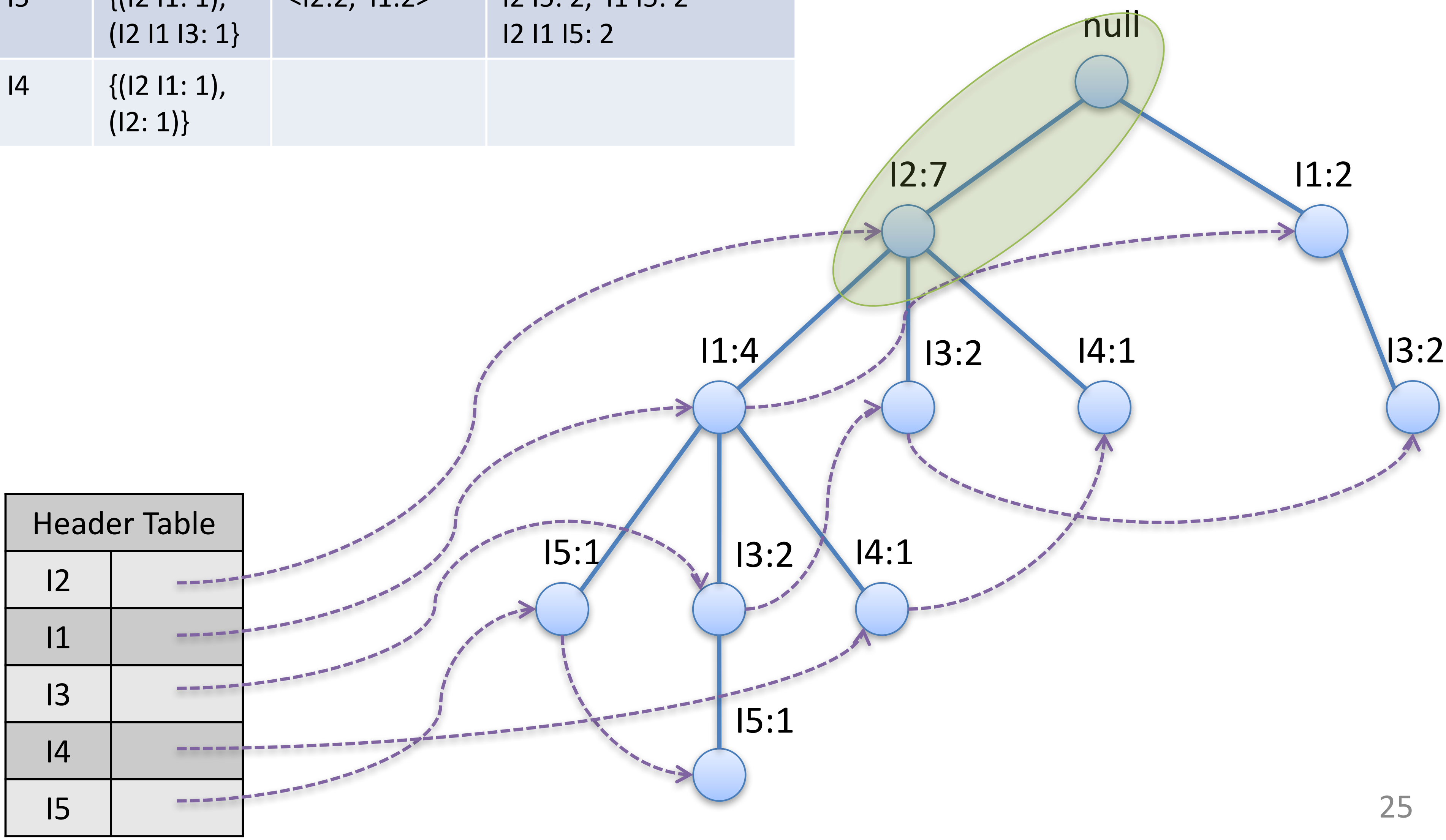
Construct Conditional Pattern Base

Item	Cond. Pattern	Cond. FP-tree	Frequent pattern
I5	{(I2 I1: 1), (I2 I1 I3: 1}	<I2:2, I1:2>	I2 I5: 2, I1 I5: 2 I2 I1 I5: 2
I4	{(I2 I1: 1)}		



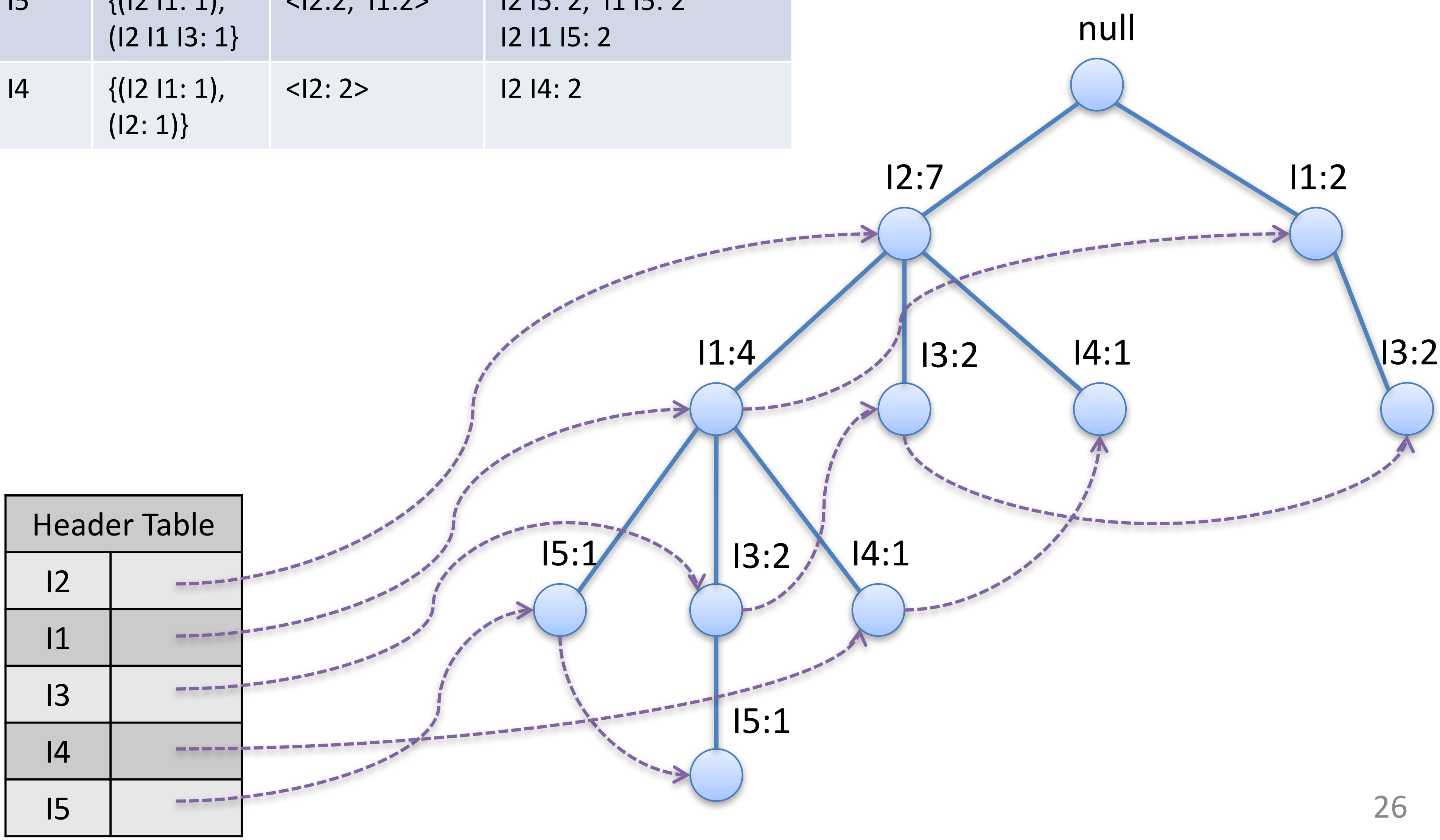
Construct Conditional Pattern Base

Item	Cond. Pattern	Cond. FP-tree	Frequent pattern
I5	{(I2 I1: 1), (I2 I1 I3: 1}	<I2:2, I1:2>	I2 I5: 2, I1 I5: 2 I2 I1 I5: 2
I4	{(I2 I1: 1), (I2: 1)}		



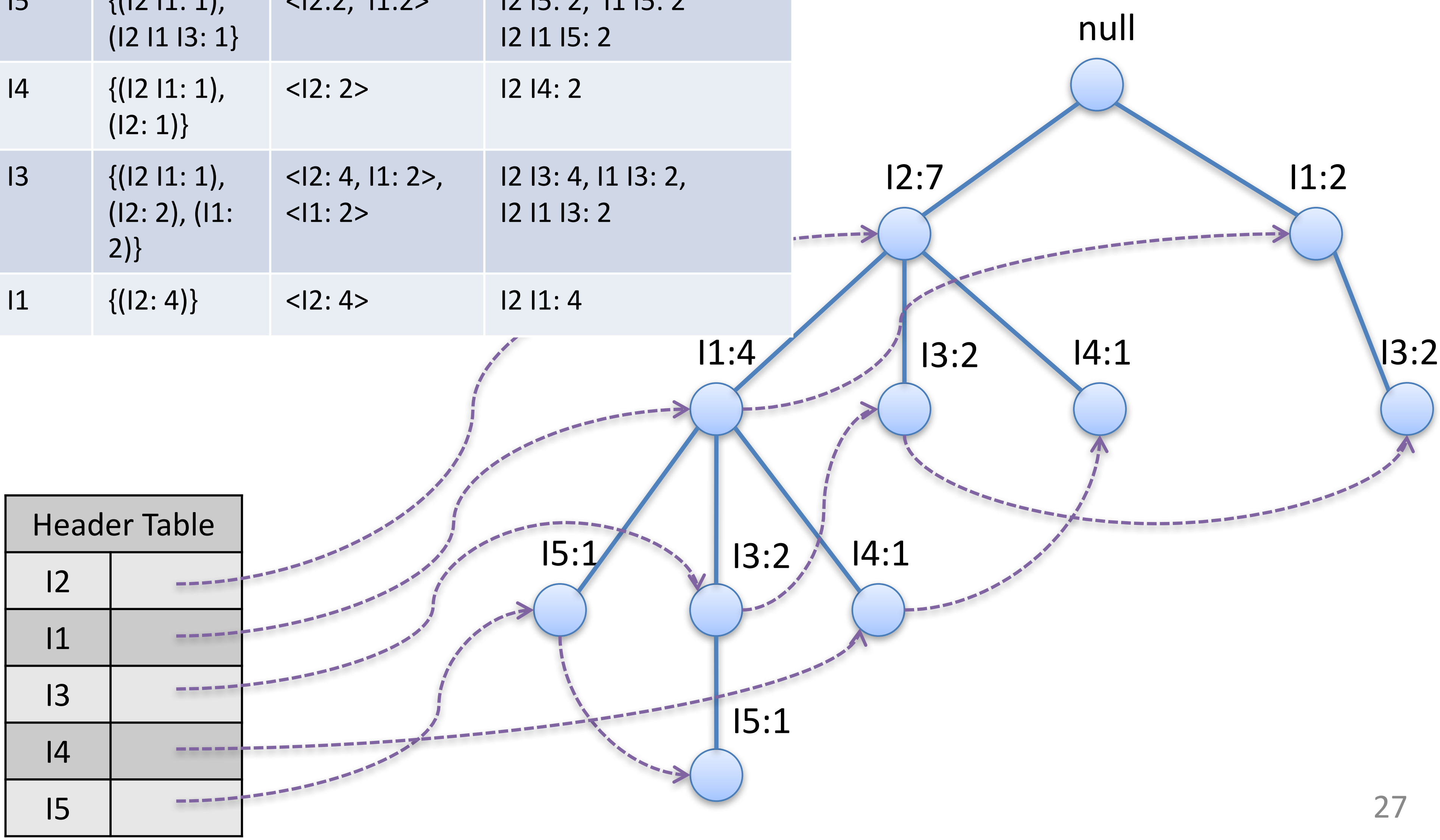
Construct Conditional Pattern Base

Item	Cond. Pattern	Cond. FP-tree	Frequent pattern
I5	{(I2 I1: 1), (I2 I1 I3: 1}	<I2:2, I1:2>	I2 I5: 2, I1 I5: 2 I2 I1 I5: 2
I4	{(I2 I1: 1), (I2: 1)}	<I2: 2>	I2 I4: 2



Construct Conditional Pattern Base

Item	Cond. Pattern	Cond. FP-tree	Frequent pattern
I5	{(I2 I1: 1), (I2 I1 I3: 1}	<I2:2, I1:2>	I2 I5: 2, I1 I5: 2 I2 I1 I5: 2
I4	{(I2 I1: 1), (I2: 1)}	<I2: 2>	I2 I4: 2
I3	{(I2 I1: 1), (I2: 2), (I1: 2)}	<I2: 4, I1: 2>, <I1: 2>	I2 I3: 4, I1 I3: 2, I2 I1 I3: 2
I1	{(I2: 4)}	<I2: 4>	I2 I1: 4



Benefits of FP-tree

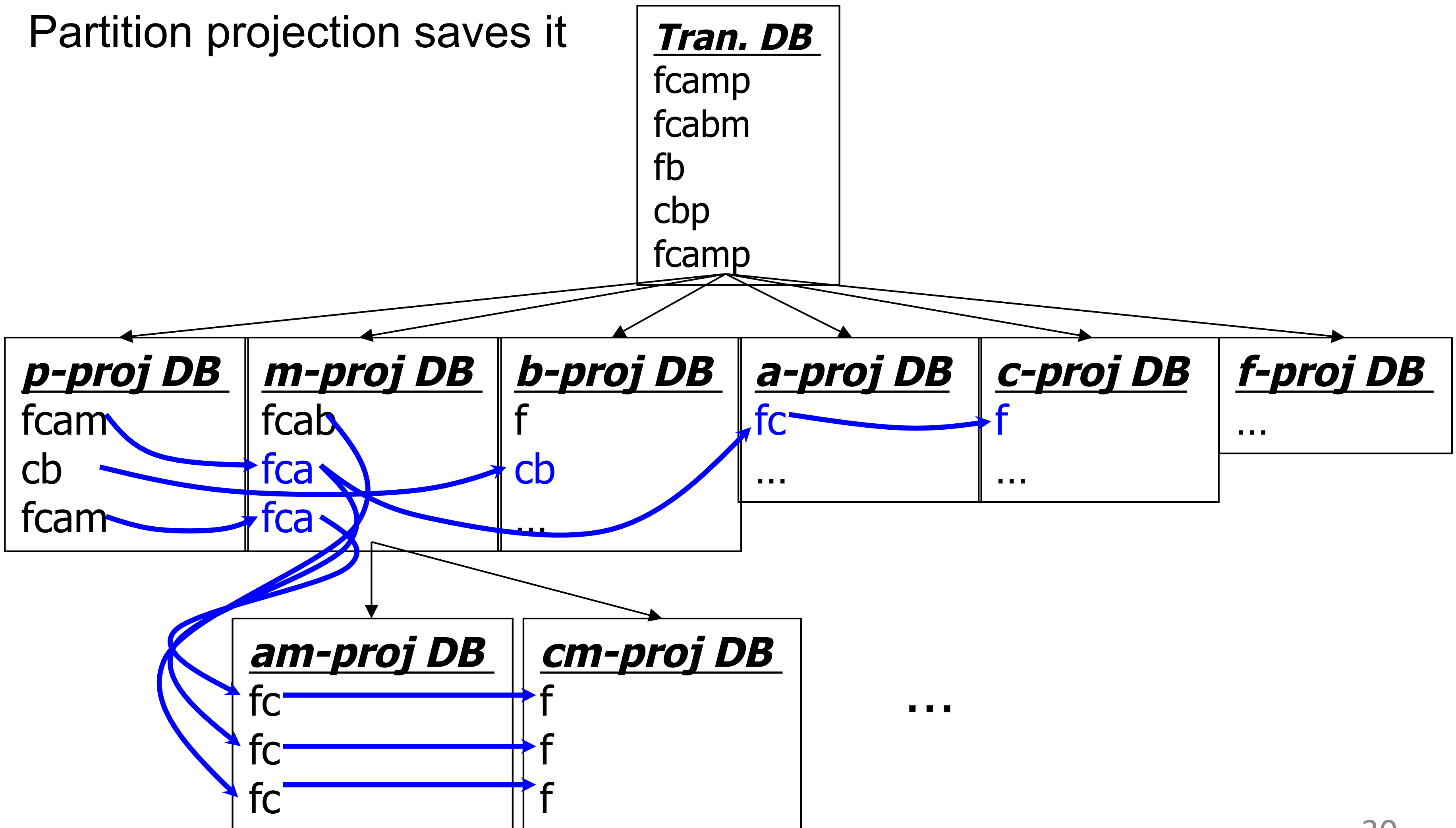
- Completeness
  - Preserve complete information for frequent pattern mining
  - Never break a long pattern of any transaction
- Compactness
  - Reduce irrelevant info – infrequent items are gone
  - Items in frequency descending order: the more frequently occurring, the more likely to be shared
  - Never larger than original database

Scaling FP-Growth w/ DB Projection

- What if FP-tree can not fit in memory?
  - DB projection
- First, partition database into a set of projected DBs
- Then construct and mine FP-trees for each projected DB
- Parallel projection vs. Partition projection methods
  - Parallel projection
    - Project the DB in parallel for each frequent item
    - Parallel projection is space costly
    - All the partitions can be processed in parallel
  - Partition projection
    - Partition the DB based on the ordered frequent items
    - Passing the unprocessed parts to the subsequent partitions

Partition-based Projection

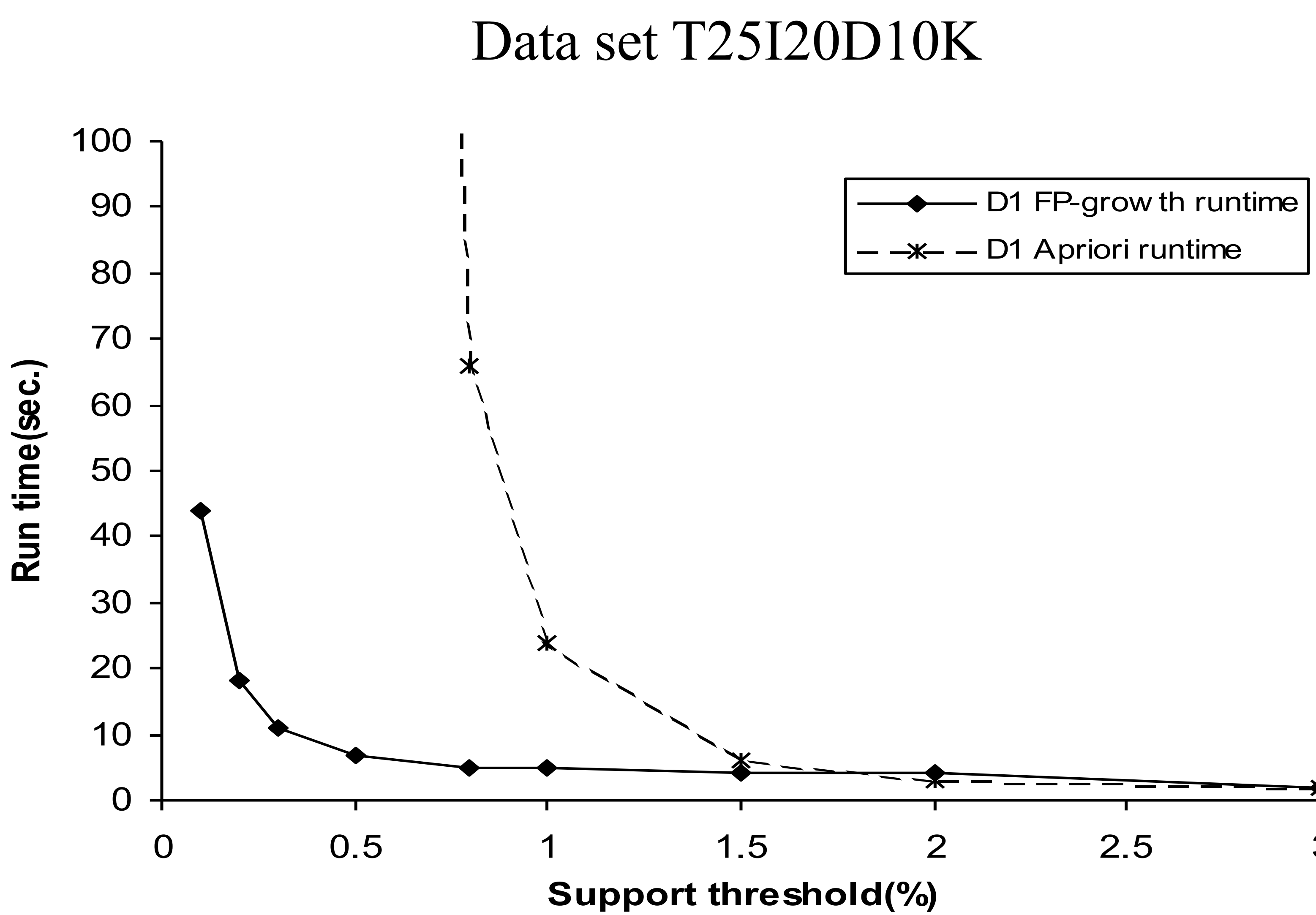
- Parallel projection needs a lot of disk space
- Partition projection saves it



Benefits of FP-Growth

- Performance study shows
  - FP-Growth is an order of magnitude faster than Apriori, also faster than tree-projection
- Reasoning
  - no candidate generation, no candidate test
  - use compact data structure
  - eliminate repeated database scan
  - basic operation is counting and FP-tree building

FP-Growth vs. Apriori

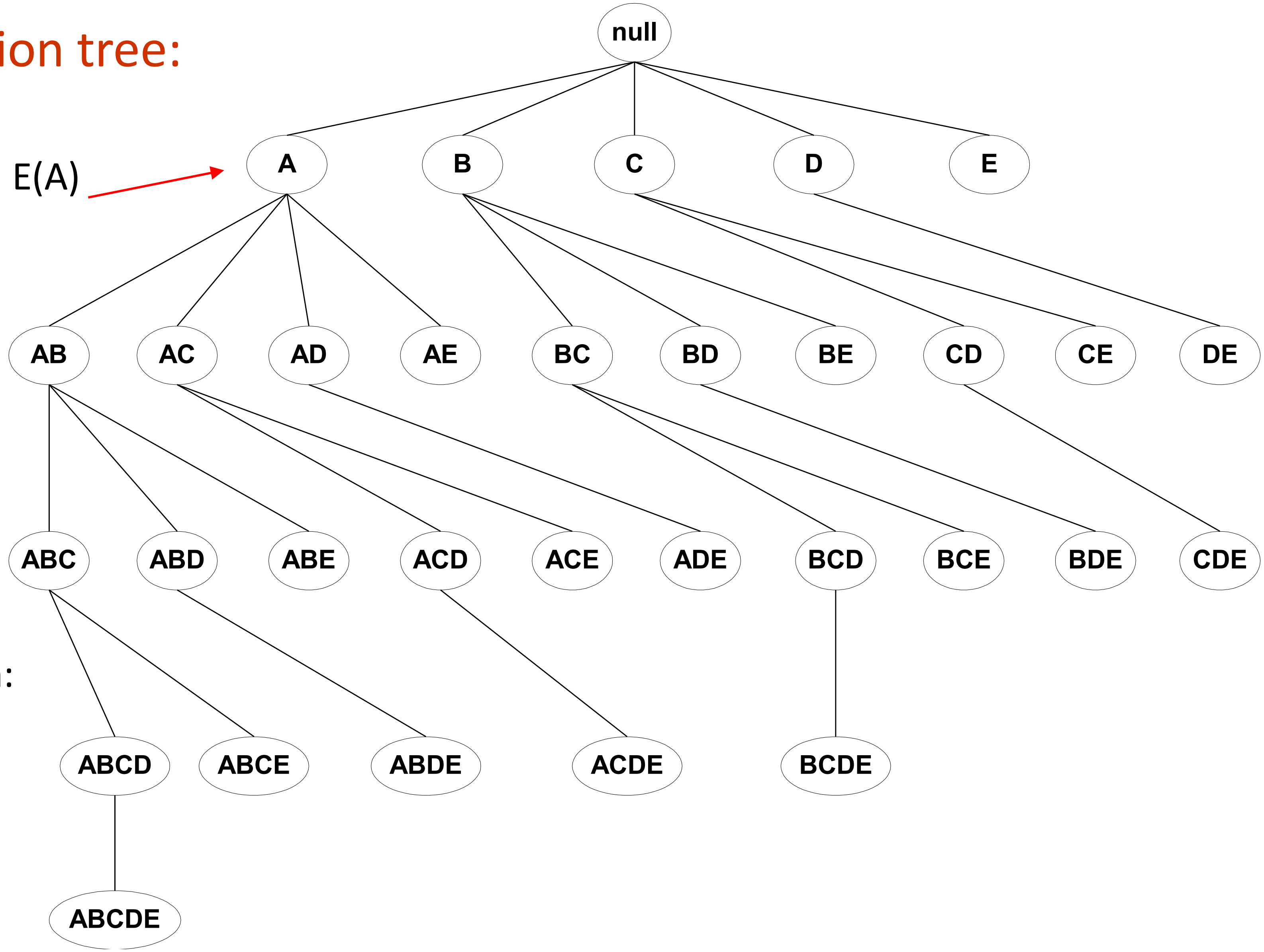


Tree Projection

Set enumeration tree:

Possible Extension: E(A) = {B,C,D,E}

Possible Extension: E(ABC) = {D,E}



Tree Projection

- Items are listed in lexicographic order
- Each node P stores the following information:
  - Itemset for node P
  - List of possible lexicographic extensions of P: E(P)
  - Pointer to projected database of its ancestor node
  - Bitvector containing information about which transactions in the projected database contain the itemset

Projected Database

Original Database:

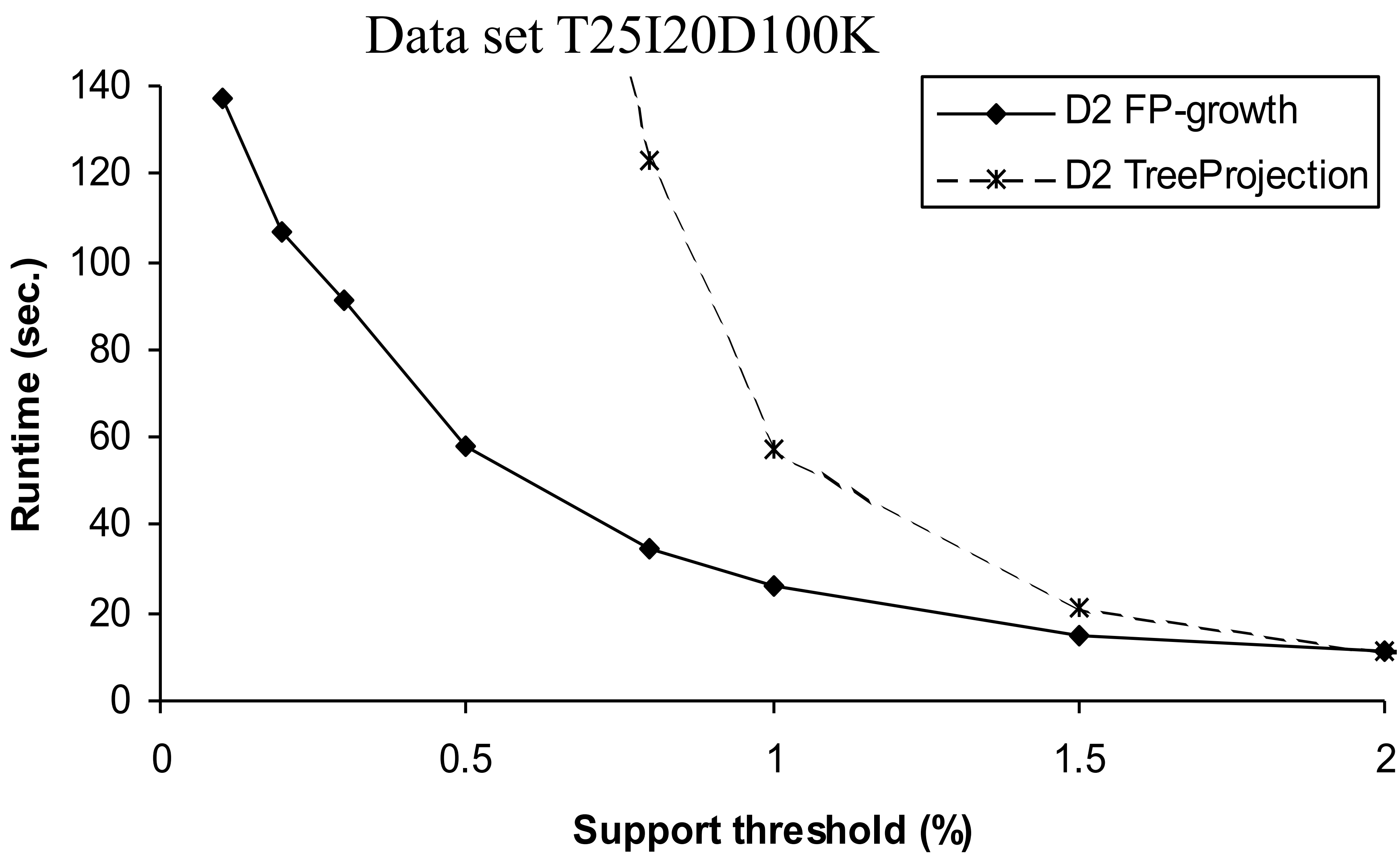
TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{B,C}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

Projected Database for node A:

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

For each transaction T, projected transaction at node A

FP-Growth vs. Tree Projection



Further Improvements of Mining Methods

- AFOPT (Liu et al., KDD 2003)
  - A “push-right” method for mining condensed frequent pattern (CFP) trees
- Carpenter (Pan et al., KDD 2003)
  - Mine data sets with small rows but numerous columns
  - Construct a row-enumeration tree for efficient mining
- Fpgrowth+ (Grahne and Zhu, FIMI 2003)
  - Efficiently using prefix trees, open-source implementation
  - ICDM 2003
- TD-Close (Liu et al., SDM 2006)

Other Extensions

- Mining closed frequent itemsets and max-patterns
  - CLOSET (DMKD'00), FPclose, and FPMax (Grahne & Zhu, Fimi'03)
- Mining sequential patterns
  - PrefixSpan (ICDE'01), CloSpan (SDM'03), BIDE (ICDE'04)
- Mining graph patterns
  - gSpan (ICDM'02), CloseGraph (KDD'03)
- Constraint-based mining of frequent patterns
  - Convertible constraints (ICDE'01), gPrune (PAKDD'03)
- Computing iceberg data cubes with complex measures
  - H-tree, H-cubing, and Star-cubing (SIGMOD'01, VLDB'03)
- Pattern-growth-based Clustering
  - MaPle (Pei, et al., ICDM'03)
- Pattern-Growth-Based Classification
  - Mining frequent and discriminative patterns (Cheng, et al, ICDE'07)

Frequent Mining with Vertical Data

- Vertical format
  - for each item store a list of transaction IDs (tids)
- tid-list: list of tids for itemsets
  - $t(AB) = \{T_{11}, T_{25}, \dots \}$
- Derive frequent patterns based on vertical intersections
  - $t(X) = t(Y)$: X and Y always happen together
  - $t(X) \subseteq t(Y)$: transaction having X always has Y

ECLAT – Equivalence Class Transformation

- For each item, store a list of transaction ids (tids)

Horizontal Data Layout

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

Vertical Data Layout

A	B	C	D	E
1	1	2	2	1
4	2	3	4	3
5	5	4	5	6
6	7	8	9	
7	8	9		
8	10			
9				

TID-list

ECLAT

- Determine support of any k-itemset by intersection

A
1
4
5
6
7
8
9

$\cap$

B
1
2
5
7
8
10

$=$

AB
1
5
7
8

- Use diffset to accelerate mining
  - only keep track of difference of tids
  - Diffset(AB, A) = { 4, 6, 9 }, Diffset(AB, B) = { 2, 10 }

ECLAT

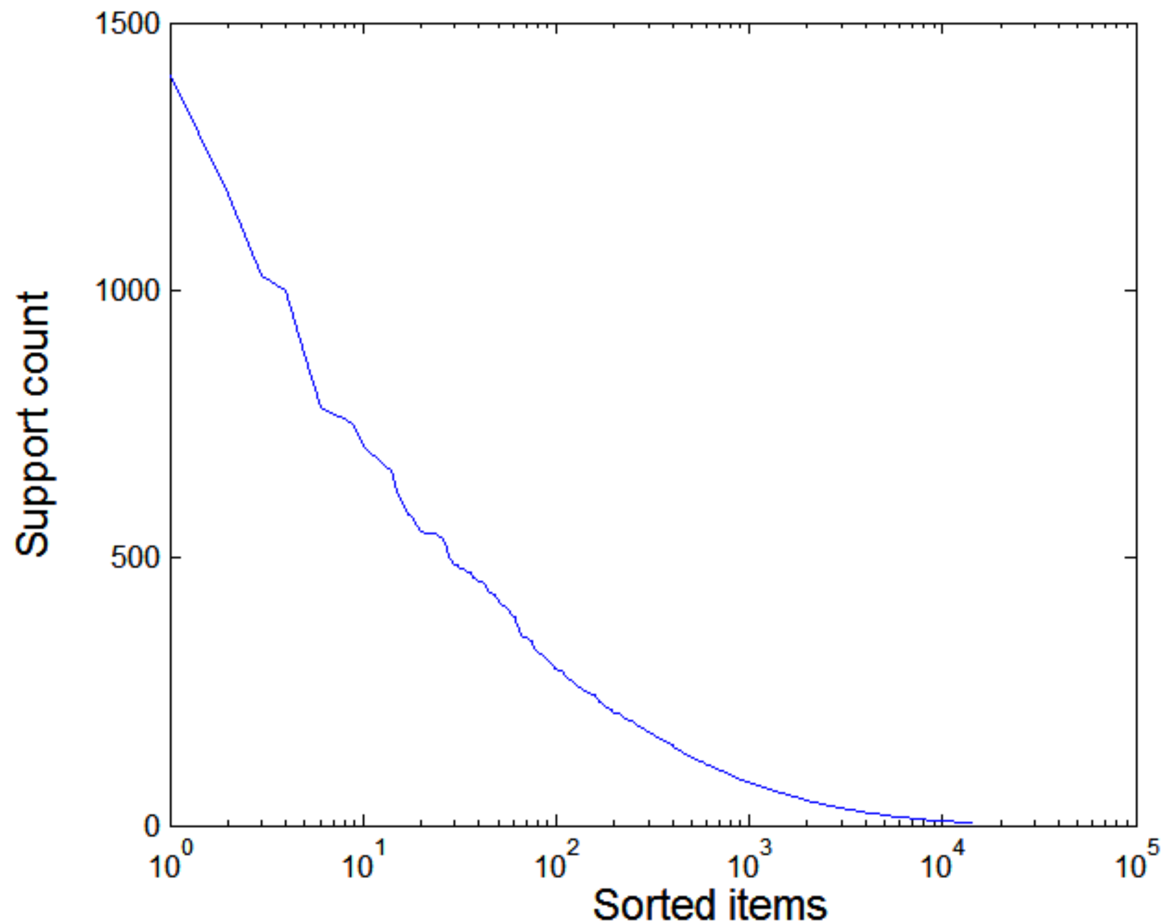
- 3 traversal approaches for itemsets
  - top-down, bottom-up, and hybrid
- Advantages: very fast support counting
- Disadvantages: intermediate tid-lists may become too large for memory

- References:
  - ECLAT – Zaki et al., KDD 1997
  - Mining closed patterns with vertical format: CHARM – Zaki & Hsiao, SDM 2002

Effect of Support Distribution

- Many real data sets have skewed support distributions

Support distribution of a retail data set



Effect of Support Distribution

- How to select appropriate *minsup* threshold?
  - If *minsup* is too high, could miss itemset involving interesting rare items (e.g., expensive products)
  - If *minsup* is too low, computationally expensive and number of itemsets identified grows

- Use of a single minimum support threshold may not be effective

Multiple Minimum Support

- How to apply multiple minimum supports?
  - MS(i) : minimum support for item i
  - Ex. MS(Milk) = 5% MS(Coke) = 3%
    MS(Broccoli) = 0.1% MS(Salmon) = 0.5%
  - MS({Milk, Broccoli}) = min(MS(Milk), MS(Broccoli))
    = 0.1%
- Challenge: Support is no longer anti-monotone
  - Suppose: Support(Milk, Coke) = 1.5%
    Support(Milk, Coke, Broccoli) = 0.5%
  - {Milk, Coke} is infrequent, but {Milk, Coke, Broccoli} is frequent

Multiple Minimum Support

- Order items according to their minimum support (ascending order)
  - Ex. Broccoli, Salmon, Coke, Milk
- Modify Apriori algorithm to support MMS
  - At 1-itemsets create, $F_{1,}$ set of items that pass minimum support levels
  - C_2 is created from join of F_1 rather than L_1
  - Pruning must also be modified to account for itemized support
- Reference: Liu 1999

Evaluation of Patterns

- A number of methods may be used to generate frequent itemsets and association rules
- Methods tend to produce too many rules
  - rules may be redundant or uninteresting
  - Ex. {A, B, C} -> {D} and {A, B} -> {D} are redundant if have same support and confidence
- Other "interestingness" measures can be used to prune or rank association rules

Interestingness Measure

- Using a contingency table many measure may be computed:

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

- Other measure are defined through manipulations of values from the table
  - confidence, lift, Gini, J-measure, …

Interestingness Measure

	Basketball	Not basketball	Sum (row)
Cereal	2000	1750	3750
Not cereal	1000	250	1250
Sum(col.)	3000	2000	5000

- Survey 5000 students
  - 3000 play basketball
  - 3750 eat cereal
  - 2000 both play basketball and eat cereal
- Rule 1: play basketball -> eat cereal [40%, 66.7%]
  - misleading, the overall percentage of students eating cereal is 75% > 66.7%
- Rule 2: play basketball -> not eat cereal [20%, 33.3%]
  - is more accurate, although lower support and confidence

Interestingness Measure

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

- Rule: Tea -> Coffee
  - Conf. = P(Coffee | Tea) = 0.75
  - but P(Coffee) = 0.9

Statistical Independence

- Population of 1000 students
  - 600 students know how to swim (S)
  - 700 students know how to bike (B)
  - 420 students know how to swim and bike (S^B)

  - P(S^B) = 420/1000 = 0.42
  - P(S) * P(B) = 0.6 * 0.7 = 0.42

  - P(S^B) = P(S) * P(B) - statistical independence
  - P(S^B) > P(S) * P(B) - positively correlated
  - P(S^B) < P(S) * P(B) - negatively correlated

Other Interestingness Measures

- Consider Rule X -> Y

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

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

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

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

Interestingness Measures

Many different measures exists

Which measure is best?

- domain dependent

#	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)	$\max\Big(P(A,B)\log(\frac{P(B\|A)}{P(B)})+P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}),$ $P(A,B)\log(\frac{P(A\|B)}{P(A)})+P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})\Big)$
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)	$\max\Big(\frac{NP(A,B)+1}{NP(A)+2},\frac{NP(A,B)+1}{NP(B)+2}\Big)$
13	Conviction (V)	$\max\Big(\frac{P(A)P(\overline{B})}{P(A\overline{B})},\frac{P(B)P(\overline{A})}{P(B\overline{A})}\Big)$
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)	$\max\Big(\frac{P(B\|A)-P(B)}{1-P(B)},\frac{P(A\|B)-P(A)}{1-P(A)}\Big)$
18	Added Value (AV)	$\max(P(B\|A)-P(B),P(A\|B)-P(A))$
19	Collective strength (S)	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})}\times\frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
20	Jaccard (ζ)	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
21	Klosgen (K)	$\sqrt{P(A,B)}\max(P(B\|A)-P(B),P(A\|B)-P(A))$

Properties of a Good Measure

- Piatetsky-Shapiro:
  3 properties a good measure M must satisfy:
  - M(A,B) = 0 if A and B are statistically independent
  - M(A,B) increase monotonically with P(A,B) when P(A) and P(B) remain unchanged
  - M(A,B) decreases monotonically with P(A) [or P(B)] when P(A,B) and P(B) [or P(A)] remain unchanged

Comparison of Measures

10 examples of contingency tables:

Example	f_{11}	f_{10}	f_{01}	f_{00}
E1	8123	83	424	1370
E2	8330	2	622	1046
E3	9481	94	127	298
E4	3954	3080	5	2961
E5	2886	1363	1320	4431
E6	1500	2000	500	6000
E7	4000	2000	1000	3000
E8	4000	2000	2000	2000
E9	1720	7121	5	1154
E10	61	2483	4	7452

Rankings of contingency tables using various measures:

#	ϕ	λ	α	Q	Y	κ	M	J	G	s	c	L	V	I	IS	PS	F	AV	S	ζ	K
E1	1	1	3	3	3	1	2	2	1	3	5	5	4	6	2	2	4	6	1	2	5
E2	2	2	1	1	1	2	1	3	2	2	1	1	1	8	3	5	1	8	2	3	6
E3	3	3	4	4	4	3	3	8	7	1	4	4	6	10	1	8	6	10	3	1	10
E4	4	7	2	2	2	5	4	1	3	6	2	2	2	4	4	1	2	3	4	5	1
E5	5	4	8	8	8	4	7	5	4	7	9	9	9	3	6	3	9	4	5	6	3
E6	6	6	7	7	7	7	6	4	6	9	8	8	7	2	8	6	7	2	7	8	2
E7	7	5	9	9	9	6	8	6	5	4	7	7	8	5	5	4	8	5	6	4	4
E8	8	9	10	10	10	8	10	10	8	4	10	10	10	9	7	7	10	9	8	7	9
E9	9	9	5	5	5	9	9	7	9	8	3	3	3	7	9	9	3	7	9	9	8
E10	10	8	6	6	6	10	5	9	10	10	6	6	5	1	10	10	5	1	10	10	7

Property under Variable Permutation

	B	$\overline{B}$
A	p	q
$\overline{A}$	r	s

$\Rightarrow$

	A	$\overline{A}$
B	p	r
$\overline{B}$	q	s

Does M(A, B) = M(B, A)?

- Symmetric Measures
  - support, lift, collective strength, cosine, Jaccard, …
- Asymmetric Measures
  - confidence, conviction, Laplace, J-measure, …

Property under Row/Column Scaling

- Grade-Gender Example (Mosteller, 1968)

	Male	Female	
High	2	3	5
Low	1	4	5
	3	7	10

	Male	Female	
High	4	30	34
Low	2	40	42
	6	70	76
	2x	10x	

Mosteller:

Underlying association should be independent of the relative number of male and female students in the samples

Property under Null Addition

	B	$\overline{\mathbf{B}}$
A	p	q
$\overline{\mathbf{A}}$	r	s

	B	$\overline{\mathbf{B}}$
A	p	q
$\overline{\mathbf{A}}$	r	s + k

- Invariant measures
  - support, cosine, Jaccard, …
- Non-invariant measures
  - correlation, Gini, mutual information, odds ratio, etc.

Comparison of Measures

Symbol	Measure	Range	P1	P2	P3	O1	O2	O3	O3'	O4
Φ	Correlation	-1 … 0 … 1	Yes	Yes	Yes	Yes	No	Yes	Yes	No
λ	Lambda	0 … 1	Yes	No	No	Yes	No	No*	Yes	No
α	Odds ratio	0 … 1 … ∞	Yes*	Yes	Yes	Yes	Yes	Yes*	Yes	No
Q	Yule's Q	-1 … 0 … 1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Y	Yule's Y	-1 … 0 … 1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
κ	Cohen's	-1 … 0 … 1	Yes	Yes	Yes	Yes	No	No	Yes	No
M	Mutual Information	0 … 1	Yes	Yes	Yes	Yes	No	No*	Yes	No
J	J-Measure	0 … 1	Yes	No	No	No	No	No	No	No
G	Gini Index	0 … 1	Yes	No	No	No	No	No*	Yes	No
s	Support	0 … 1	No	Yes	No	Yes	No	No	No	No
c	Confidence	0 … 1	No	Yes	No	Yes	No	No	No	Yes
L	Laplace	0 … 1	No	Yes	No	Yes	No	No	No	No
V	Conviction	0.5 … 1 … ∞	No	Yes	No	Yes**	No	No	Yes	No
I	Interest	0 … 1 … ∞	Yes*	Yes	Yes	Yes	No	No	No	No
IS	IS (cosine)	0 .. 1	No	Yes	Yes	Yes	No	No	No	Yes
PS	Piatetsky-Shapiro's	-0.25 … 0 … 0.25	Yes	Yes	Yes	Yes	No	Yes	Yes	No
F	Certainty factor	-1 … 0 … 1	Yes	Yes	Yes	No	No	No	Yes	No
AV	Added value	0.5 … 1 … 1	Yes	Yes	Yes	No	No	No	No	No
S	Collective strength	0 … 1 … ∞	No	Yes	Yes	Yes	No	Yes*	Yes	No
ζ	Jaccard	0 .. 1	No	Yes	Yes	Yes	No	No	No	Yes
K	Klosgen's	$\left(\sqrt{\frac{2}{\sqrt{3}}-1}\right)\left(2-\sqrt{3}-\frac{1}{\sqrt{3}}\right)\ldots 0\ldots\frac{2}{3\sqrt{3}}$	Yes	Yes	Yes	No	No	No	No	No

Comparison of Measures

Symbol	Measure	Range	P1	P2	P3	O1	O2	O3	O3'	O4
Φ	Correlation	-1 … 0 … 1	Yes	Yes	Yes	Yes	No	Yes	Yes	No
λ	Lambda	0 … 1	Yes	No	No	Yes	No	No*	Yes	No
α	Odds ratio	0 … 1 … ∞	Yes*	Yes	Yes	Yes	Yes	Yes*	Yes	No
IS	IS (cosine)	0 .. 1	No	Yes	Yes	Yes	No	No	No	Yes
PS	Piatetsky-Shapiro's	-0.25 … 0 … 0.25	Yes	Yes	Yes	Yes	No	Yes	Yes	No
F	Certainty factor	-1 … 0 … 1	Yes	Yes	Yes	No	No	No	Yes	No
AV	Added value	0.5 … 1 … 1	Yes	Yes	Yes	No	No	No	No	No
S	Collective strength	0 … 1 … ∞	No	Yes	Yes	Yes	No	Yes*	Yes	No
ζ	Jaccard	0 .. 1	No	Yes	Yes	Yes	No	No	No	Yes
K	Klosgen's	$\left(\sqrt{\frac{2}{\sqrt{3}}-1}\right)\left(2-\sqrt{3}-\frac{1}{\sqrt{3}}\right)\ldots0\ldots\frac{2}{3\sqrt{3}}$	Yes	Yes	Yes	No	No	No	No	No

where:
- P1: $O(\mathbf{M}) = 0$ if $det(\mathbf{M}) = 0$, *i.e.*, whenever A and B are statistically independent.
- P2: $O(\mathbf{M_2}) > O(\mathbf{M_1})$ if $\mathbf{M_2} = \mathbf{M_1} + [k \ \ -k;\ -k \ \ k]$.
- P3: $O(\mathbf{M_2}) < O(\mathbf{M_1})$ if $\mathbf{M_2} = \mathbf{M_1} + [0 \ k;\ 0 \ -k]$ or $\mathbf{M_2} = \mathbf{M_1} + [0 \ 0;\ k \ -k]$.
- O1: Property 1: Symmetry under variable permutation.
- O2: Property 2: Row and Column scaling invariance.
- O3: Property 3: Antisymmetry under row or column permutation.
- O3': Property 4: Inversion invariance.
- O4: Property 5: Null invariance.
- Yes*: Yes if measure is normalized.
- No*: Symmetry under row or column permutation.
- No**: No unless the measure is symmetrized by taking $\max(M(A,B), M(B,A))$.

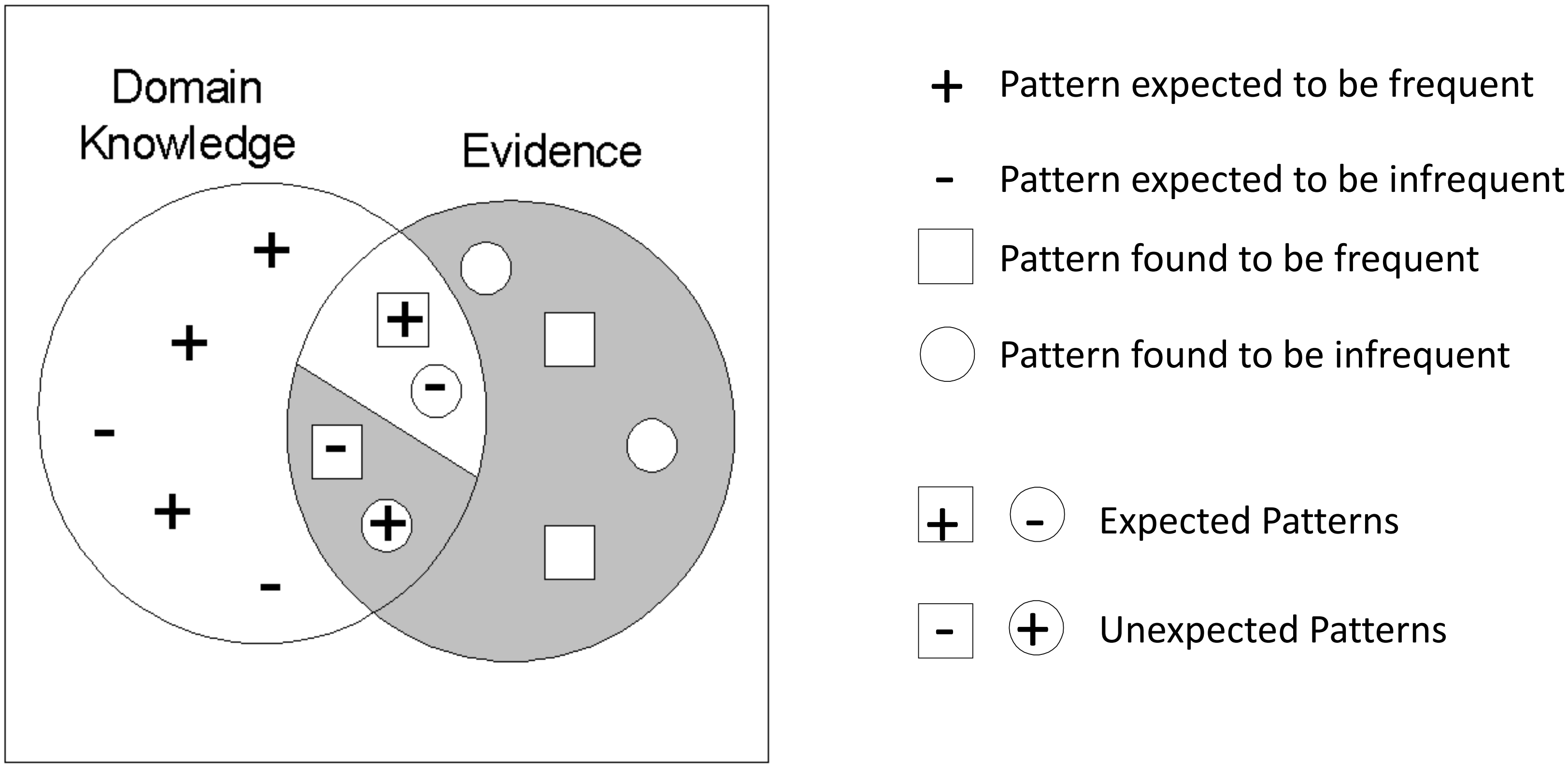
Subjective Interestingness Measures

- Objective measure
  - rank patterns based on statistics computed from data
  - 21 measures reported (support, confidence, Laplace, Gini, mutual information, …)

- Subjective measure
  - Rank patterns according user’s interpretation
    - A pattern is subjectively interesting if it contradicts the expectation of a user (Silberschatz & Tuzhilin)
    - A pattern is subjectively interesting if it is actionable

Interestingness and Unexpectedness

- Need to model expectation of users (domain knowledge)



- Need to combine expectation of users with evidence from data

Summary

- Basic concepts – itemsets, frequent, association rules, support, confidence, closed and max-patterns
- Frequent pattern mining methods
  - Apriori (candidate generation and test)
  - Projection-based (Fpgrowth)
  - Vertical format approach (ECLAT)
- Which patterns are interesting?
  - pattern evaluation