University of Colorado
Boulder

CSCI 4502/5502
Data Mining

Fall 2020
Lecture 09 (Sep 22)

Reminders

- Homework 3
  - due at 9:30am, Th, Sep 24
- Temporary 2-week remote instruction
  - Wed Sep 23 to Wed Oct 7
  - Stay healthy! Take good care!

Announcements

- Homework 1
  - grades posted in Canvas, please check
  - contact GSS first with grading questions
- Homework 2 is being graded
- No new homework this Thursday
  - work on course project proposal

Review

- Chapter 6: Mining Frequent Patterns
  - basic concepts, Apriori algorithm, correlation: lift
  - improve the efficiency of Apriori
    - #scans, #candidates, support counting
  - FP-growth: grow patterns w/o generating candidates
    - if c is frequent in DB|ab, then abc is frequent

FP-tree Construction

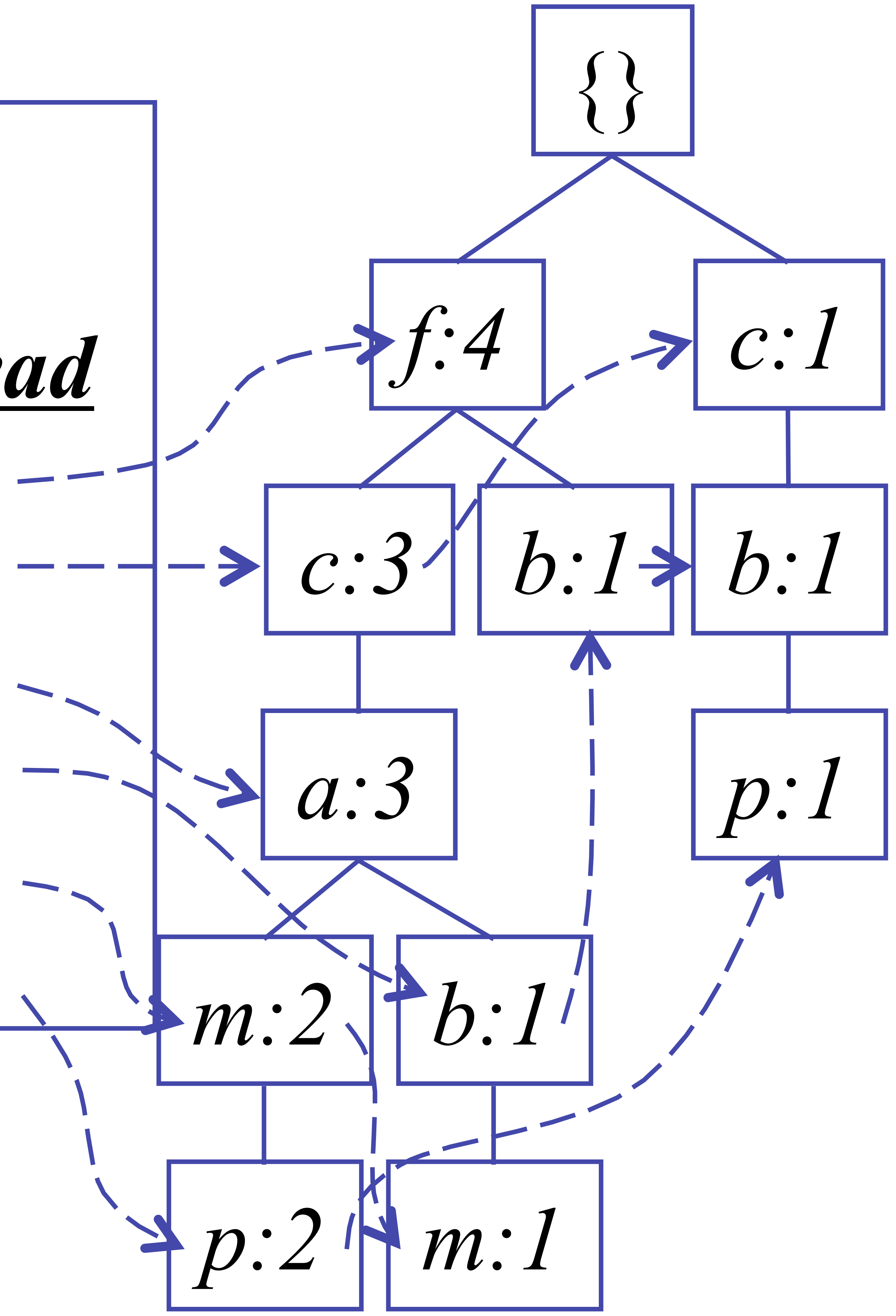
TID	*Items bought*	*(ordered) frequent items*
100	***{f, a, c, d, g, i, m, p}***	***{f, c, a, m, p}***
200	***{a, b, c, f, l, m, o}***	***{f, c, a, b, m}***
300	***{b, f, h, j, o, w}***	***{f, b}***
400	***{b, c, k, s, p}***	***{c, b, p}***
500	***{a, f, c, e, l, p, m, n}***	***{f, c, a, m, p}***

min_sup = 0.6

- Scan, find freq. 1-itemset
- Sort freq. items in descending frequency
- Scan, construct FP-tree

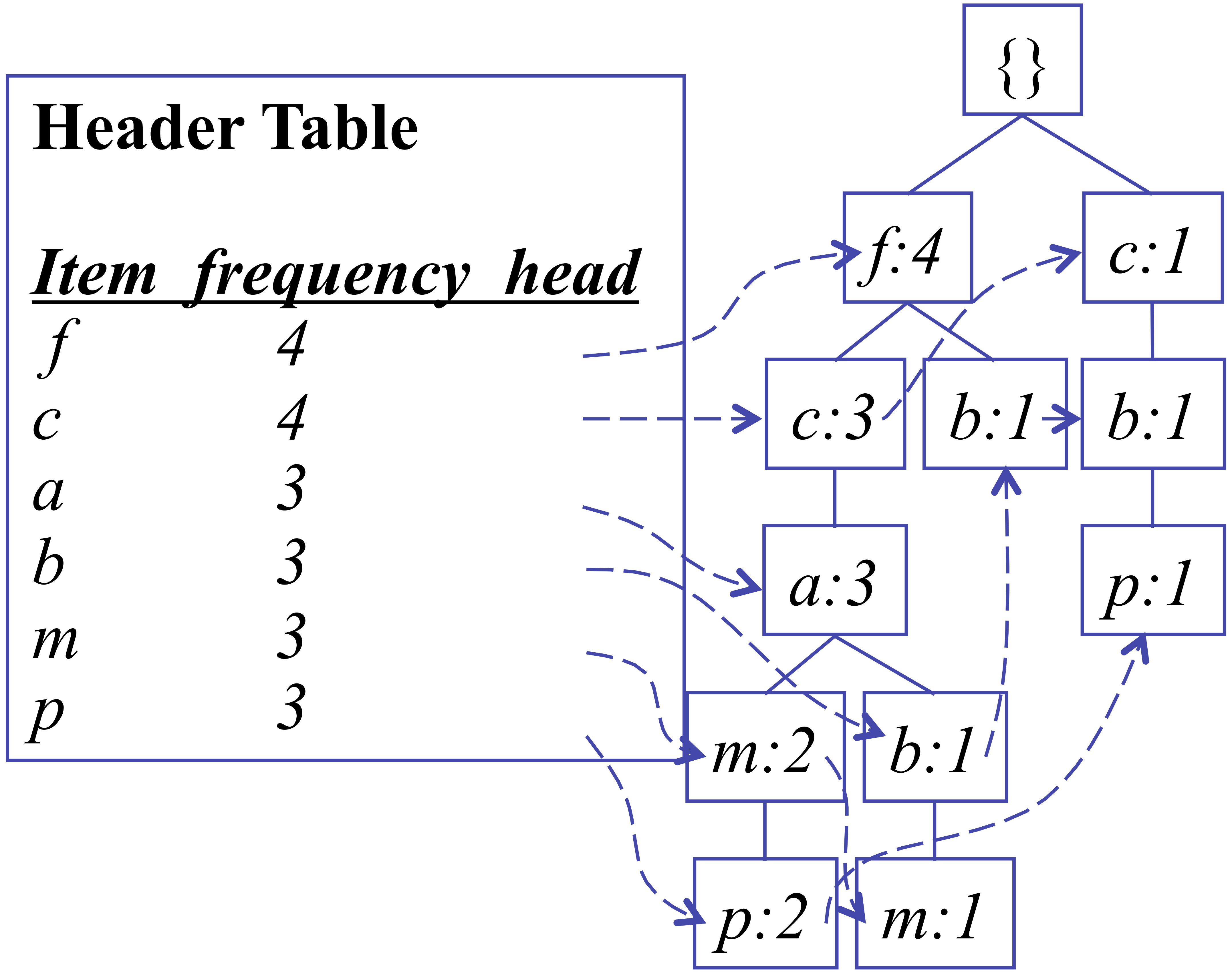
Header Table

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



Conditional Pattern Base

- Traverse links of each frequent item, prefix paths



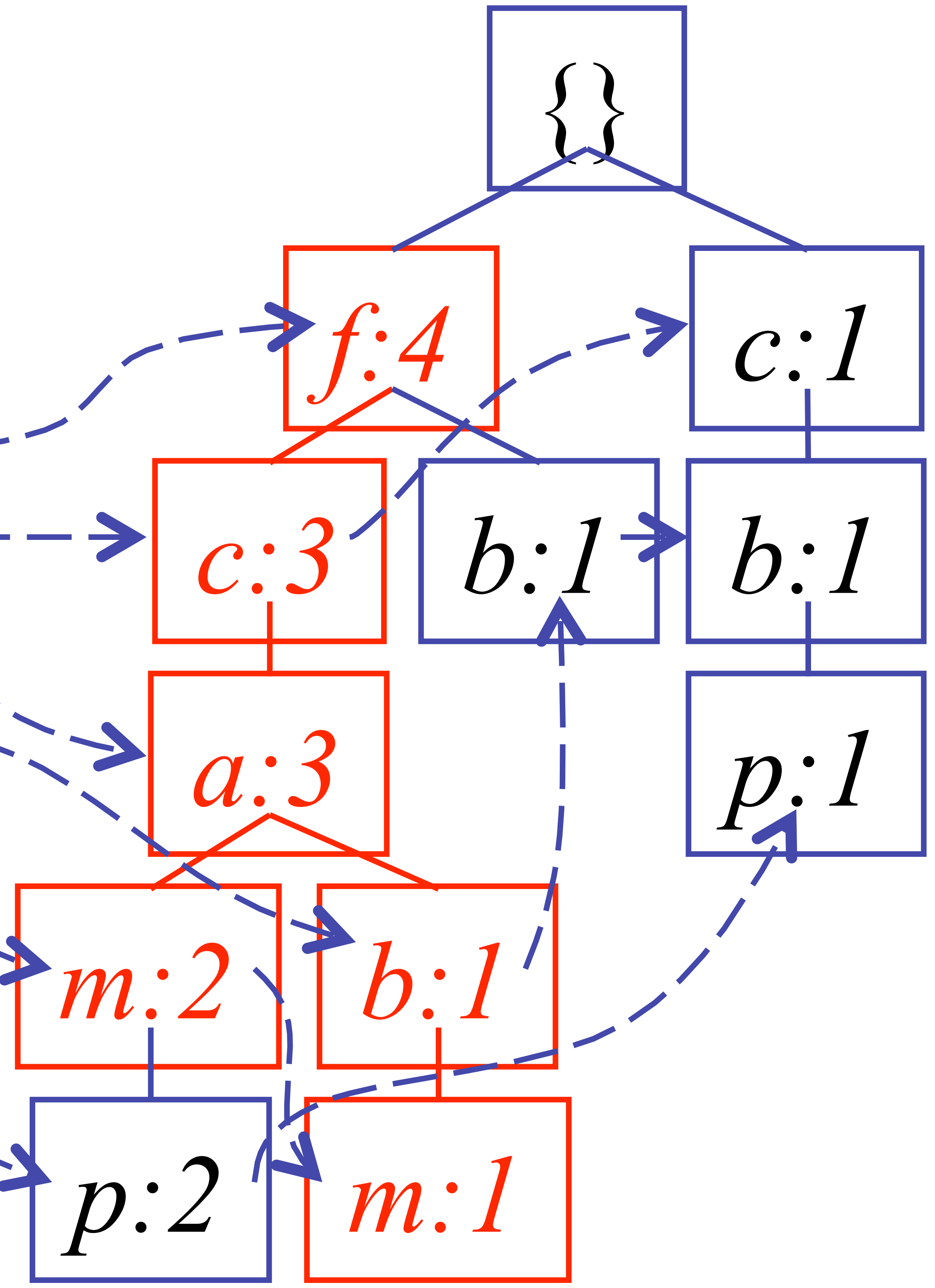
***Conditional* pattern bases**

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***

Conditional FP-trees

Header Table

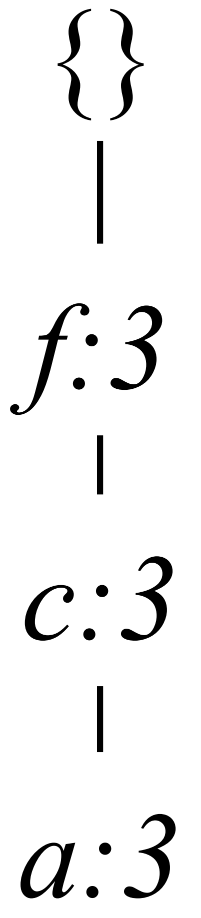
Item	***frequency***	***head***
f	*4*	
c	*4*	
a	*3*	
b	*3*	
m	*3*	
p	*3*	



m-conditional **pattern base:**

fca:2, fcab:1

➔



m-conditional **FP-tree**

➔

All frequent patterns relate to *m*

m,

fm, cm, am,

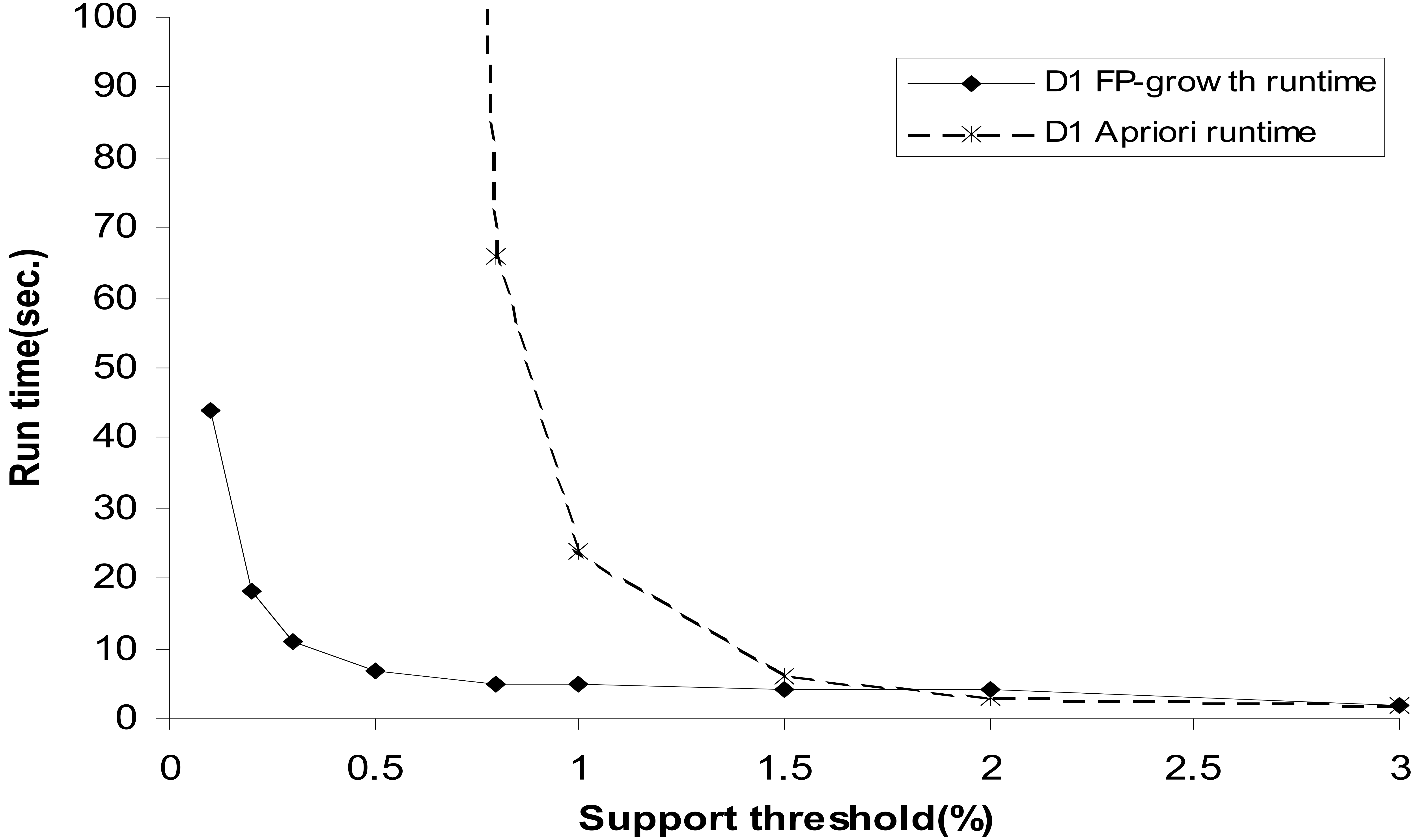
fcm, fam, cam,

fcam

FP-growth

- Idea: Frequent pattern growth
  - recursively grow freq. patterns by pattern and data partition
- Method
  - freq. item => conditional pattern base => conditional FP-tree
  - repeat on each newly created FP-tree
  - until FP-tree is empty or single path

FP-growth vs. Apriori



Correlation Rules

- Correlation rule
  - A ⇒ B [support, confidence, correlation]
- Measure of dependent/correlated events

$$lift(A, B) = \frac{P(A \cup B)}{P(A)P(B)}$$

  - lift = 1? independent
  - lift < 1? negatively dependent
  - lift > 1? positively dependent

$$lift(A, B) = \frac{P(A \cup B)}{P(A)P(B)}$$

	basketball	not basketball	sum (row)
cereal	2000	1750	3750
not cereal	1000	250	1250
sum (col)	3000	2000	5000

$$lift(B, C) = \frac{2000/5000}{(3000/5000) \times (3750/5000)} = 0.89$$

$$lift(B, \overline{C}) = \frac{1000/5000}{(3000/5000) \times (1250/5000)} = 1.33$$

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} \qquad e_{ij} = \frac{count(A=a_i)\times count(B=b_j)}{N}$$

	basketball	not basketball	sum (row)
cereal	2000	1750	3750
not cereal	1000	250	1250
sum (col)	3000	2000	5000

- e_cb = 3750 * 3000 / 5000 = 2250
- X^2 = (2000 - 2250)^2 / 2250 + (1750 - 1500)^2 / 1500 + (1000 - 750)^2 / 750 + (250 - 500)^2 / 500 = 227.78 (correlated)
- o_cb = 2000 < e_cb = 2250 (negative)

Other Correlation Measures

$$all\_conf(A,B) = \frac{sup(A \cup B)}{max\{sup(A), sup(B)\}} = min\{P(A|B), P(B|A)\}$$

$$max\_conf(A, B) = max\{P(A \mid B), P(B \mid A)\}$$

$$Kulc(A, B) = \frac{1}{2}(P(A|B) + P(B|A))$$

$$cosine(A, B) = \frac{P(A \cup B)}{\sqrt{P(A) \times P(B)}} = \frac{sup(A \cup B)}{\sqrt{sup(A) \times sup(B)}}$$

$$= \sqrt{P(A|B) \times P(B|A)}.$$

Comparison (1)

Data Set	mc	$\overline{m}c$	$m\overline{c}$	$\overline{mc}$	χ^2	*lift*	*all_conf.*	*max_conf.*	*Kulc.*	*cosine*
D_1	10,000	1000	1000	100,000	90557	9.26	0.91	0.91	0.91	0.91
D_2	10,000	1000	1000	100	0	1	0.91	0.91	0.91	0.91
D_3	100	1000	1000	100,000	670	8.44	0.09	0.09	0.09	0.09
D_4	1000	1000	1000	100,000	24740	25.75	0.5	0.5	0.5	0.5
D_5	1000	100	10,000	100,000	8173	9.18	0.09	0.91	0.5	0.29
D_6	1000	10	100,000	100,000	965	1.97	0.01	0.99	0.5	0.10

- Null-transaction: e.g., ¬m¬c
- Null-variant: lift and X^2
- Null-invariant: all_conf, max_conf, Kulc, cosine

Comparison (2)

Data Set	mc	$\overline{m}c$	$m\overline{c}$	$\overline{mc}$	χ^2	*lift*	*all_conf.*	*max_conf.*	*Kulc.*	*cosine*
D_1	10,000	1000	1000	100,000	90557	9.26	0.91	0.91	0.91	0.91
D_2	10,000	1000	1000	100	0	1	0.91	0.91	0.91	0.91
D_3	100	1000	1000	100,000	670	8.44	0.09	0.09	0.09	0.09
D_4	1000	1000	1000	100,000	24740	25.75	0.5	0.5	0.5	0.5
D_5	1000	100	10,000	100,000	8173	9.18	0.09	0.91	0.5	0.29
D_6	1000	10	100,000	100,000	965	1.97	0.01	0.99	0.5	0.10

- Imbalance ratio

$$IR(A,B) = \frac{|sup(A) - sup(B)|}{sup(A) + sup(B) - sup(A \cup B)}$$