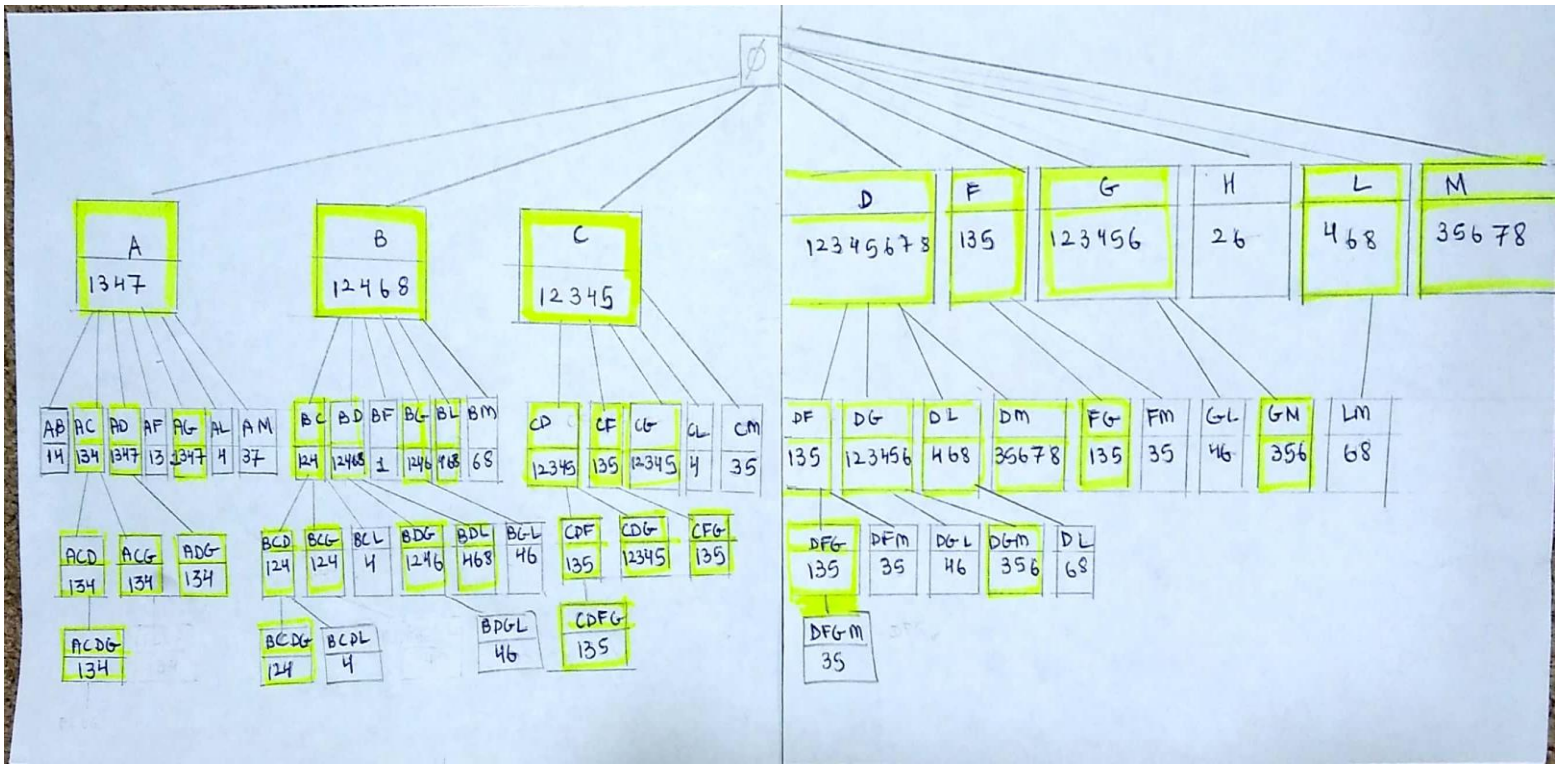
1. **(Association Analysis) Consider the following transactions: (T1: A B C D F G), (T2: B C D G H), (T3: A C D F G K M), (T4: A B C D G L), (T5: C D F G M), (T6: B D G H L M), (T7: A D M), (T8: B D L M). Perform the following tasks using minimum support (minsup) value of 3.**
    a. **(15) Use Eclat algorithm to find all itemsets of size 3 or larger that are frequent. Show the execution of the algorithm and the results obtained.**
    b. **(15) Use GenMax algoritm to find all the maximal frequent itemsets in this dataset. Show the execution of the algorithm and the results obtained.**
    c. **(15) Use Charm algorithm to find all the closed itemsets. Show the execution of the algorithm and the results obtained.**
    d. **(15) Use FP-Tree algorithm to construct the FP Tree for the given set of transactions. Find all the frequent itemsets that contain the item "C". Show the final FP-Tree and main steps in constructing it. Also, show the execution of steps to find the frequent itemsets.**

Answer 1)a)



I have dropped K as it's support is 1. Also drop H as t's support is 2. At each step the itemsets with minimum support less than 3 are eliminated.

All those highlighted are frequent itemset

Results obtained-

The frequent itemsets are:

Minimum support	Itemsets
3	AC, BC, BL, CF, DF, DL, FG, GM, ACD, ACG, ADG, BCD, BCG, BDL, CDF, CFG, DFG, DGM ACDG, BCDG, CDFG, F, L
4	A, AD, AG, BG, BDG, BCDG
5	B, C, M, BD, CD, CG, CDG
6	G, DG
8	D

Answer 1)b)

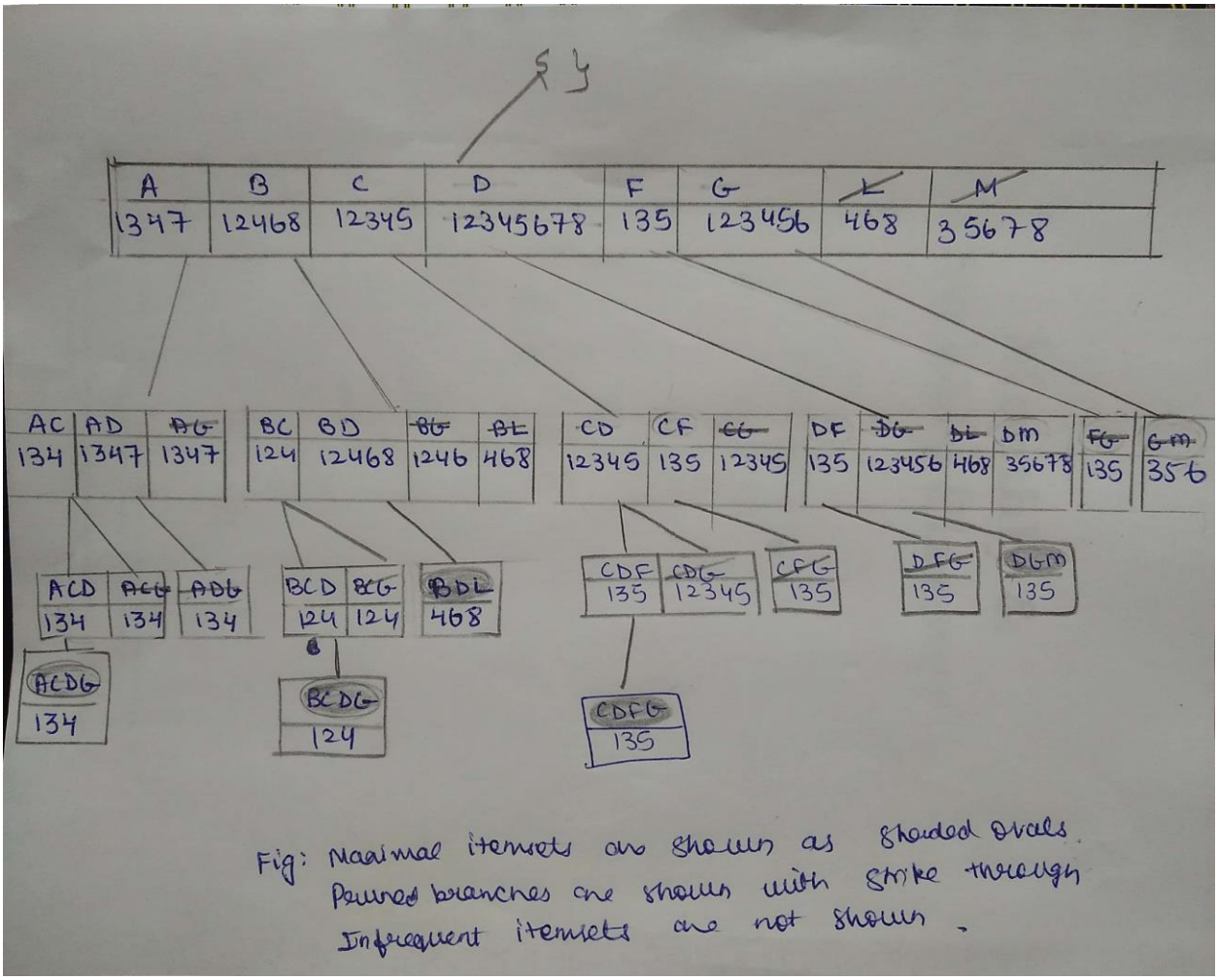


Fig: Maximal itemsets are shown as shaded ovals. Pruned branches are shown with strike through. Infrequent itemsets are not shown.

The maximal frequent itemsets are:

ACDG, BCDG, BDL, CDFG, DGM

Initially, all the itemsets with size 1 are written whose support is greater than 3. Then, itemsets of size 2 are generated and only the frequent itemsets are drawn in the graph. Following the **Depth First Strategy**. Itemset of size 3 and 4 are generated for the first branch.

Now, the maximal frequent itemset is selected. Now, **backtracking** is done. All those itemsets are pruned which are the subset of this maximal frequent itemset.

These steps are repeated until we find all the maximal frequent itemsets.

Answer 1)c)

Initially the set of all closed itemsets, C, is empty.

Given any IT-pair set $P = \{(X_i, t(X_i))\}$, first sort them in the increasing order of their support

F	L	A	B	C	M
135	468	1347	12468	12345	35678

G	D
123456	12345678

Now, the properties of charm's algorithm are applied, we write only the frequent itemset,

As FL, FA, FB, FCM are infrequent, we ignore them (min sup < 3)

Since $c(F) = c(FC)$

$\therefore$ F can be replaced by FC (property 1)

Similarly $c(FC) = c(FCG)$

$\therefore$ FC is replaced by FCG

Similarly, the process is carried on. When we get to A, $c(AC) \neq c(A)$.

$t(AC) \subset t(A)$ is not true

$\therefore \quad t(AC) \neq t(A)$

we apply the property 3 of Charnis Algorithm.

We add new extension P_A,

If set P_A is not empty, we make recursive call to Charnis.

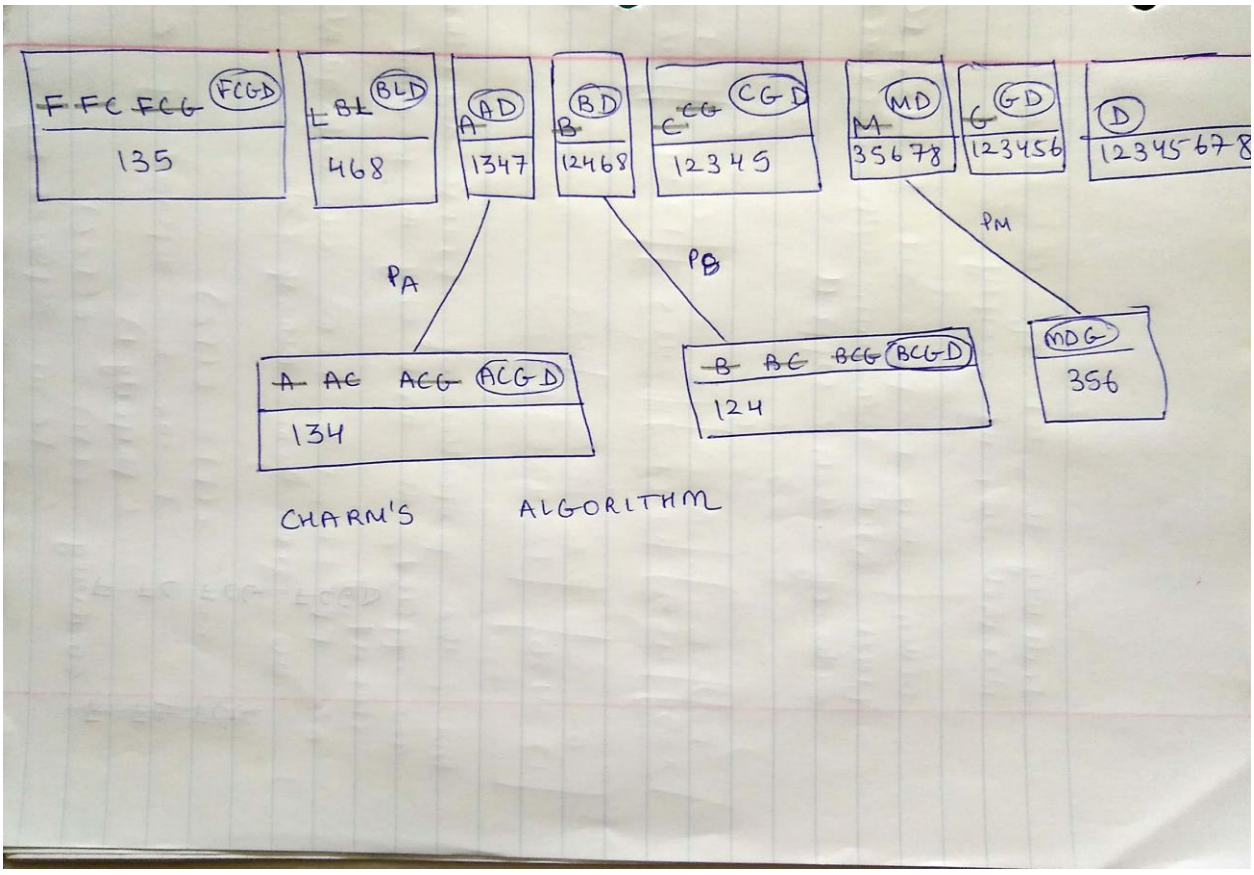
Hence, continue the same procedure.

Property (1) If $\mathbf{t}(X_i) = \mathbf{t}(X_j)$, then $\mathbf{c}(X_i) = \mathbf{c}(X_j) = \mathbf{c}(X_i \cup X_j)$, which implies that we can replace every occurrence of X_i with $X_i \cup X_j$ and prune the branch under X_j because its closure is identical to the closure of $X_i \cup X_j$.

Property (2) If $\mathbf{t}(X_i) \subset \mathbf{t}(X_j)$, then $\mathbf{c}(X_i) \neq \mathbf{c}(X_j)$ but $\mathbf{c}(X_i) = \mathbf{c}(X_i \cup X_j)$, which means that we can replace every occurrence of X_i with $X_i \cup X_j$, but we cannot prune X_j because it generates a different closure. Note that if $\mathbf{t}(X_i) \supset \mathbf{t}(X_j)$ then we simply interchange the role of X_i and X_j.

Property (3) If $\mathbf{t}(X_i) \neq \mathbf{t}(X_j)$, then $\mathbf{c}(X_i) \neq \mathbf{c}(X_j) \neq \mathbf{c}(X_i \cup X_j)$. In this case we cannot remove either X_i or X_j, as each of them generates a different closure.

Thus, using the above properties of Charm's algorithms conclusions were drawn as follows.



The closed itemsets are:

FCDG, BLD, AD, BD, CGD, MD, GD, D, ACGD, BCGD, MDG

Answer 1)d)

FP Tree Algorithm

Step 1 Scan the database for the first time, find the frequent items (single item patterns) and order them into a list L in frequency descending order.

Given.

	Itemsets
T_1	A, B, C, D, F, G
T_2	B, C, D, G, H
T_3	A, C, D, F, G, K, M
T_4	A, B, C, D, G, L
T_5	C, D, F, G, M
T_6	B, D, G, H, L, M
T_7	A, D, M
T_8	B, D, L, M

Performing step 1:

$L = \{ D:8, G:6, B:5, C:5, M:5, A:4, F:3, L:3, H:2, K:1 \}$

Now, drop the items which are not frequent i.e their support is less than min support = 3.

$\therefore L = \{ D:8, G:6, B:5, C:5, M:5, A:4, F:3, L:3 \}$

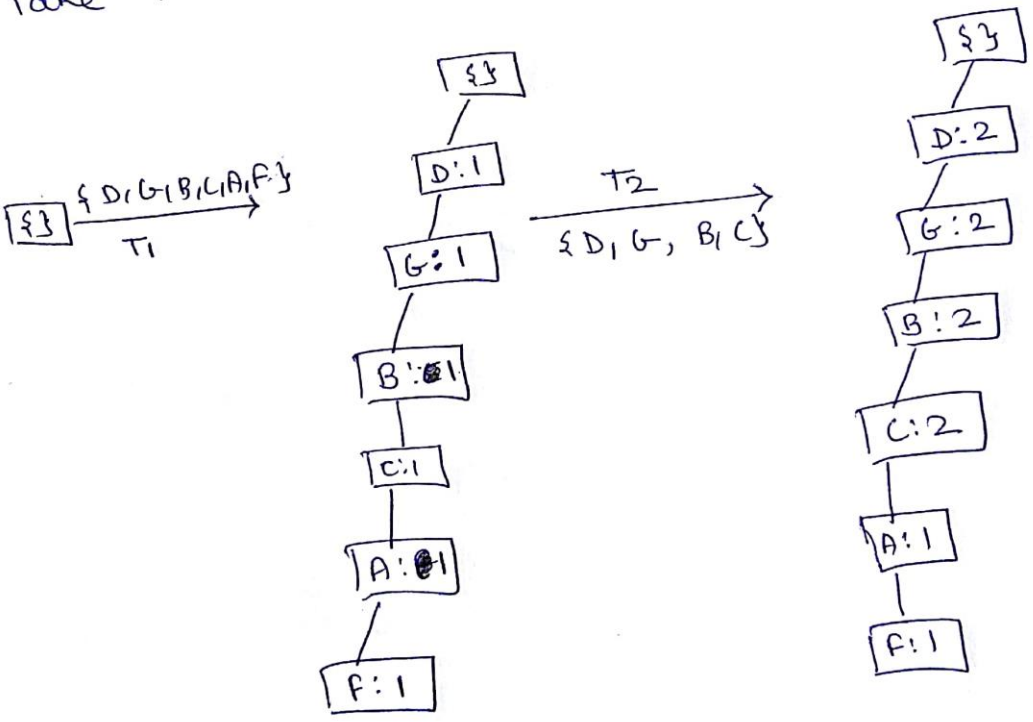
Step 2 For each transaction, order its frequent items according to order in L.

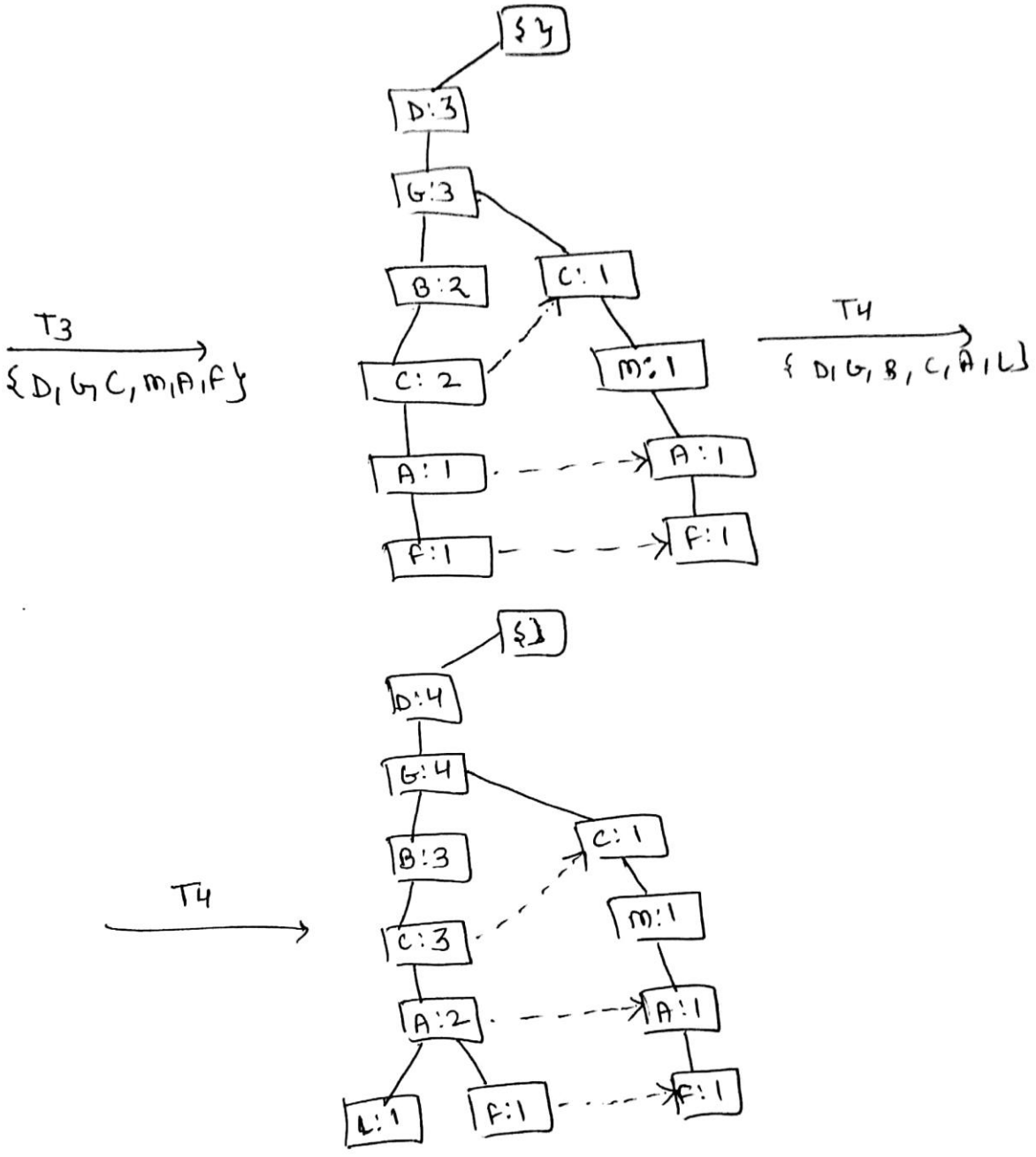
Scan the database second time, construct FP tree by putting each frequency ordered transaction onto it.

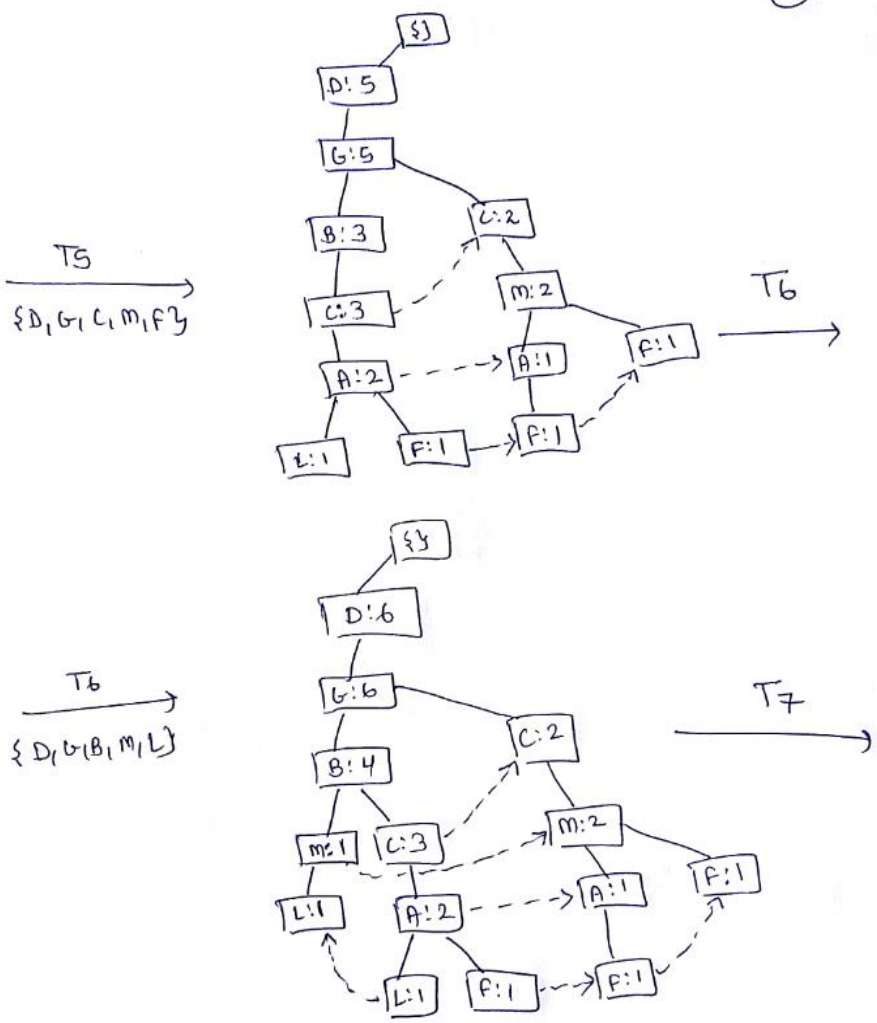
Transaction	Ordered frequent item
T_1	D, G, B, C, A, F
T_2	D, G, B, C
T_3	D, G, C, M, A, F
T_4	D, G, B, C, A, L
T_5	D, G, C, M, F
T_6	D, G, B, M, L
T_7	D, M, A
T_8	D, M, B, L

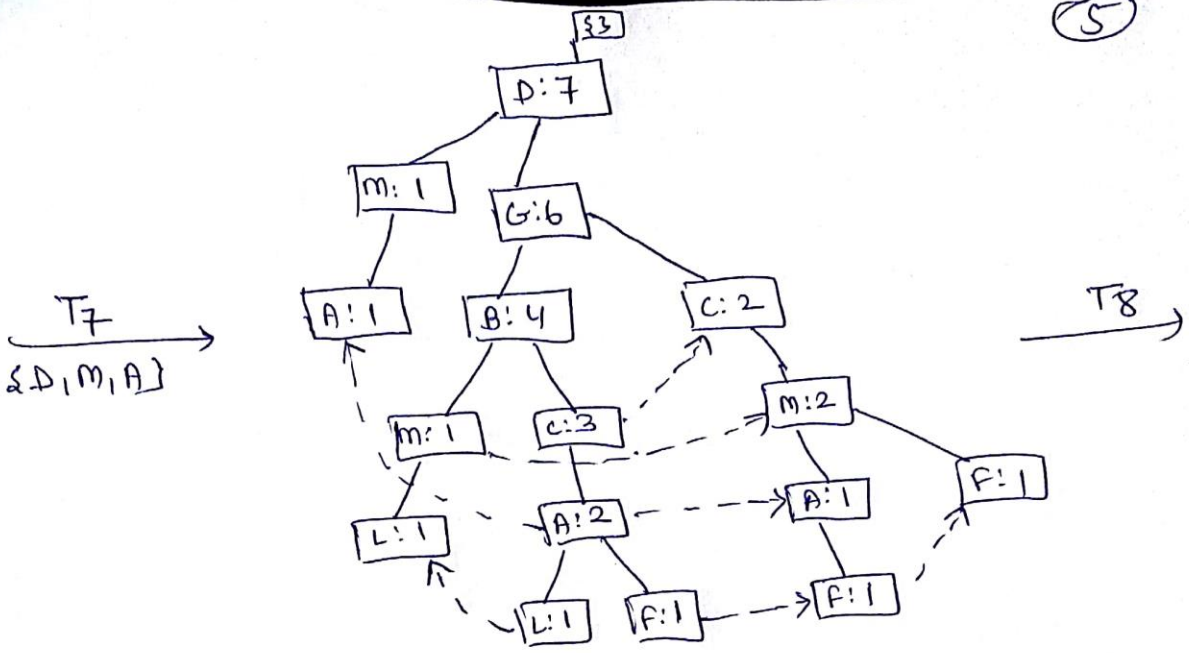
Now, construct the FP tree

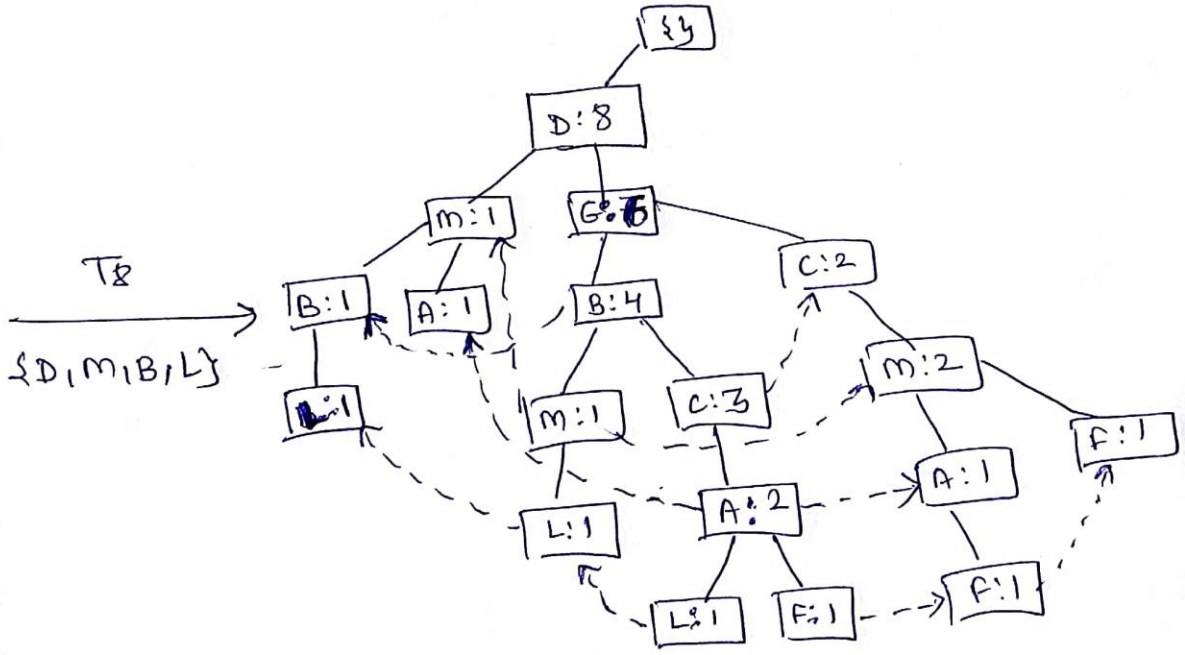
Take transaction 1 (T_1)




T3
{D, G, C, m, A, F}
$ 3
D:3
G:3
B:2
C:1
C:2
m:1
A:1
A:1
F:1
F:1
T4
{D, G, B, C, A, L}
T4
$ 1
D:4
G:4
B:3
C:1
C:3
m:1
A:2
A:1
L:1
F:1
F:1

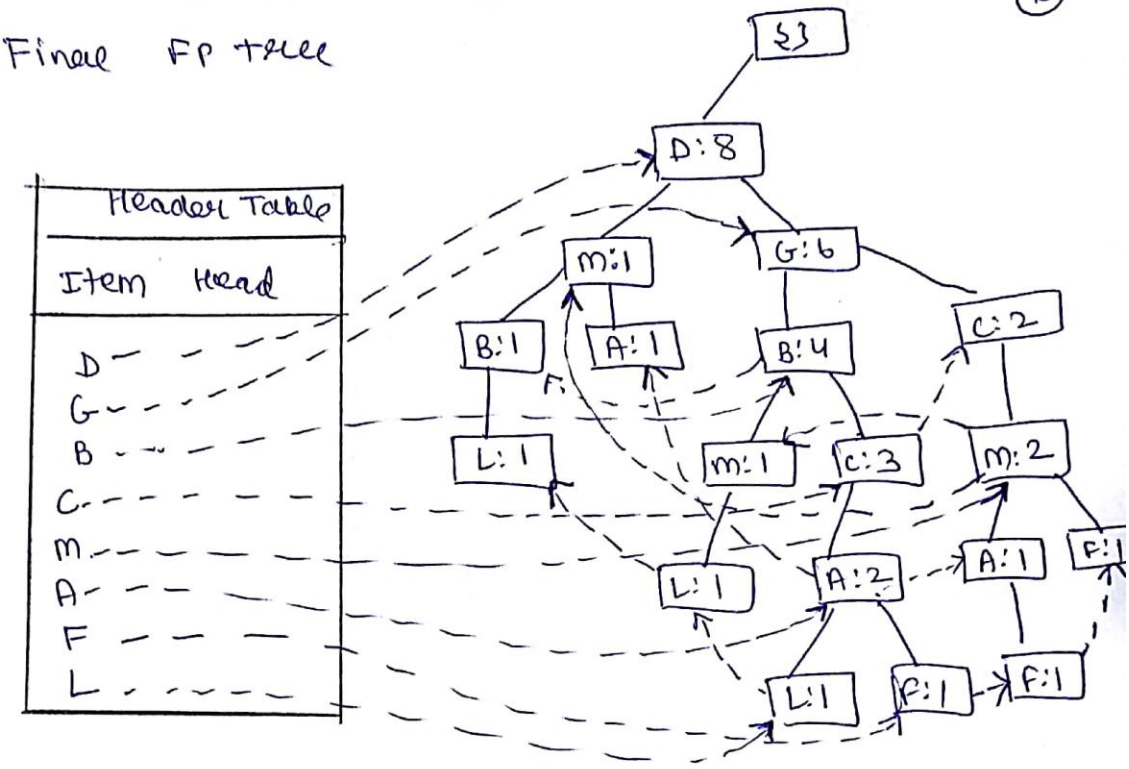

T5
{D, G, C, M, F}
$
D:5
G:5
B:3
C:2
C:3
M:2
A:2
A:1
F:1
L:1
F:1
F:1
T6
T6
{D, G, B, M, L}
$
D:6
G:6
B:4
C:2
M:1
C:3
M:2
L:1
A:2
A:1
F:1
L:1
F:1
F:1
T7


T7
{D, M, A}
T8
{3
D:7
M:1
G:6
A:1
B:4
C:2
M:1
C:3
M:2
L:1
A:2
A:1
F:1
L:1
F:1
F:1


T8
{D, M, B, L}
{4
D:8
M:1
G:6
B:1
A:1
B:4
C:2
L:1
M:1
C:3
M:2
F:1
A:2
A:1
L:1
L:1
F:1
F:1

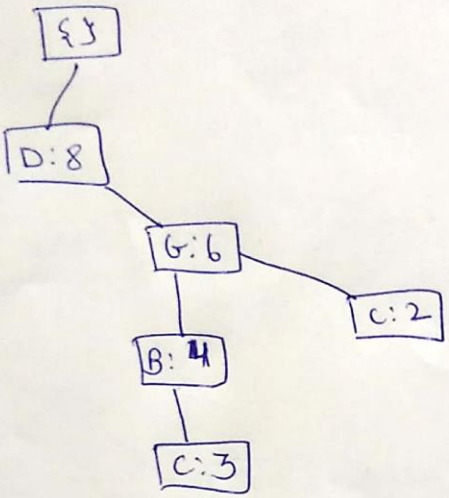
(b)

Final FP tree



Now, the aim is to find out all the frequent itemset that contain C:

(2) Construct Fb tree for the frequent item c



But since occurance of c = 5 ∴, the tree becomes

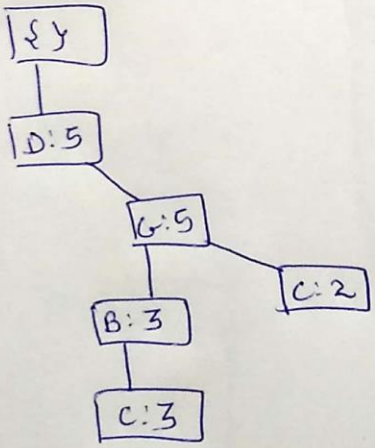
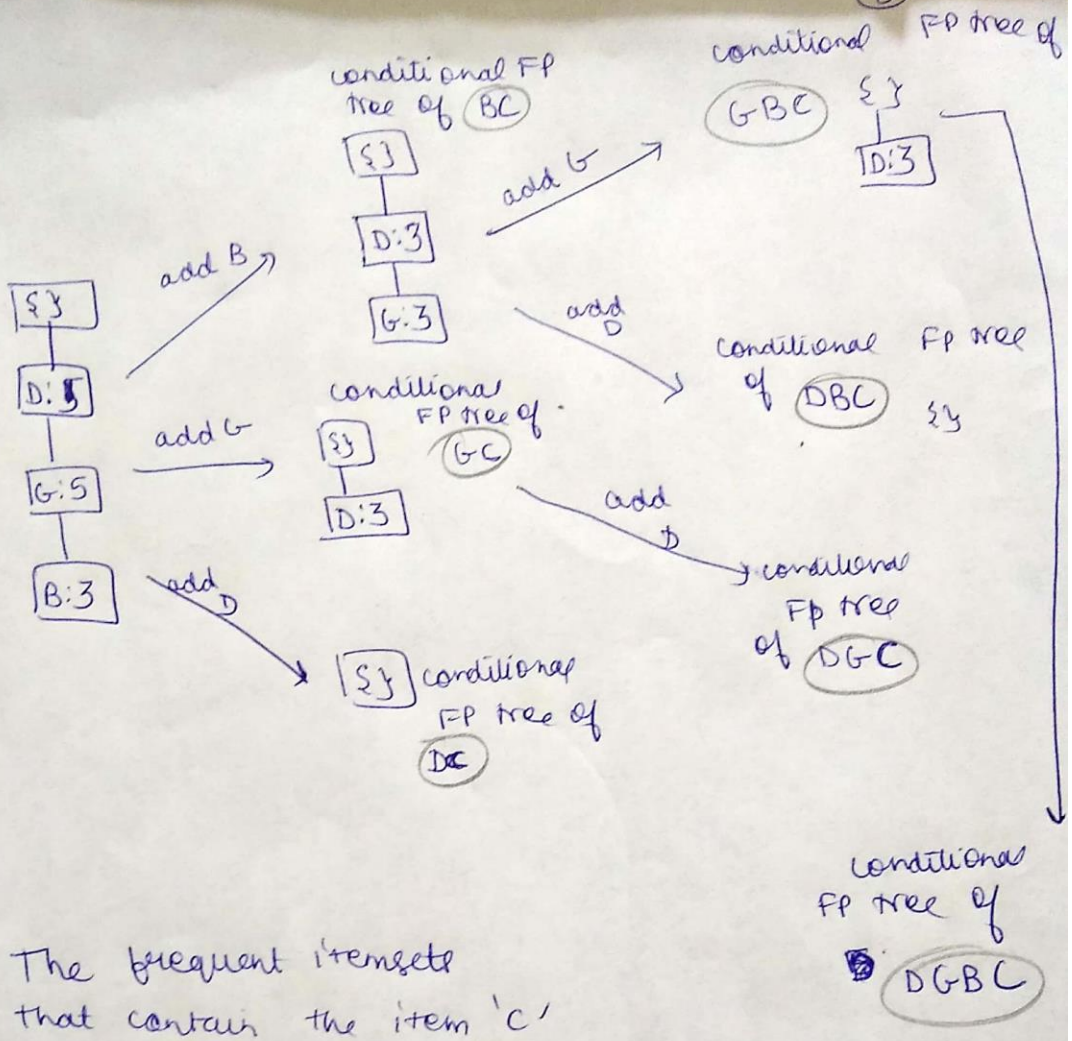


Fig: C-conditional FP tree



The frequent itemsets that contain the item 'C' are:

BC, BCG, BCDG
GC, BCD
DC, CDG

2. Consider a dataset in which each row corresponds to a Plant and each column corresponds to a State in the US. Each entry in the table is marked as "1" when the Plant in the row is known to grow in the State for the column. An example table is shown below.

	S1	S2	S3	S4	S5	S6
Pa	0	0	1	1	1	0
Pb	1	1	0	0	1	1
Pc	0	1	1	1	0	0
Pd	1	0	1	1	1	1
Pe	1	1	0	0	0	1
Pf	0	1	0	0	1	1
Pg	1	1	1	1	1	1
Ph	1	1	0	1	1	1
Pj	0	0	1	1	1	0

   a. (15) Without formally running any algorithm, and using only your intuition, select from this table two different closed itemsets and two different maximal frequent itemsets. Each of these itemsets must contain at least three items and also use minsup=3. Show the supporting sets of rows for each of these itemsets. Explain and show why each itemset has the property of being closed or maximal frequent.
   b. (10) What interpretation can be given to the closed itemsets and what is a potential use for the knowledge of these closed itemsets?
   c. (10) What interpretation can be given to the maximal frequent itemsets and what is a potential use for these maximal frequent itemsets?

Answer 2)a)

Ans 5)
Firstly all the frequent itemsets are found:-

{S1} = 5
{S2} = 6
{S3} = 5
{S4} = 6
{S5} = 7
{S6} = 6

} all of them have support greater than the minsup, ∴ they are frequent.

Now, we find frequent itemsets of size 2.

itemset	support	
{S1, S2}	= 4	
~~{S1, S3} = 2~~		[Not frequent]
{S1, S4}	= 3	
{S1, S5}	= 4	
{S1, S6}	= 5	
~~{S2, S3} = 2~~		[Not frequent]
{S2, S4}	= 3	
{S2, S5}	= 4	
{S2, S6}	= 5	
{S3, S4}	= 5	

{S3, S5} = 4
~~{S3, S6} = 2~~ [Not frequent]

{S4, S5} = 5
{S4, S6} = 3
{S5, S6} = 5

Now, we find frequent itemsets of size 3:

~~{S1, S2, S3} = 1~~ [Not frequent]
~~{S1, S2, S4} = 2~~ [Not frequent]
{S1, S2, S5} = 3
{S1, S2, S6} = 4

~~{S1, S3, S4} = 2~~ [Not frequent]
~~{S1, S3, S5} = 2~~ [Not frequent]
~~{S1, S3, S6} = 2~~ [Not frequent]

{S1, S4, S5} = 3
{S1, S4, S6} = 3

{S1, S5, S6} = 4

~~{S2, S3, S4} = 2~~ [Not frequent]
~~{S2, S3, S5} = 1~~ [Not frequent]

~~{S2, S3, S6} = 1 [Not frequent]~~

~~{S2, S4, S5} = 2 [Not frequent]~~
~~{S2, S4, S6} = 2 [Not frequent]~~

~~{S2, S4, S5}~~
{S2, S5, S6} = 3
{S3, S4, S5} = 4
~~{S3, S4, S6} = 2~~ [Not frequent]

~~{S3, S5, S6} = 2~~ [Not frequent]
{S4, S5, S6} = 3

Itemset which are frequent of size 4:

~~{S1, S2, S3, S4} = 1~~ [Not frequent]
~~{S1, S2, S3, S5} = 1~~ [Not frequent]
~~{S1, S2, S3, S6} = 1~~ [Not frequent]

~~{S1, S2, S4, S5} = 2~~ [Not frequent]
~~{S1, S2, S4, S6} = 2~~ [Not frequent]

{S1, S2, S5, S6} = 3

~~{S1, S3, S4, S5} = 2~~ [Not frequent]
~~{S1, S3, S4, S6} = 2~~ [Not frequent]

{S1, S4, S5, S6} = 3

{S2, S3, S4, S5} = 2 [Not frequent]

{S2, S3, S4, S6} = 1 [Not frequent]

{S2, S4, S5, S6} = 2 [Not frequent]

{S3, S4, S5, S6} = 2 [Not frequent]

{S1, S2, S3, S4, S5} = 1 [Not frequent]

{S1, S2, S3, S4, S6} = 1 [Not frequent]

{S1, S2, S3, S5, S6} = 1 [Not frequent]

{S1, S2, S4, S5, S6} = 2 [Not frequent]

{S1, S3, S4, S5, S6} = 2 [Not frequent]

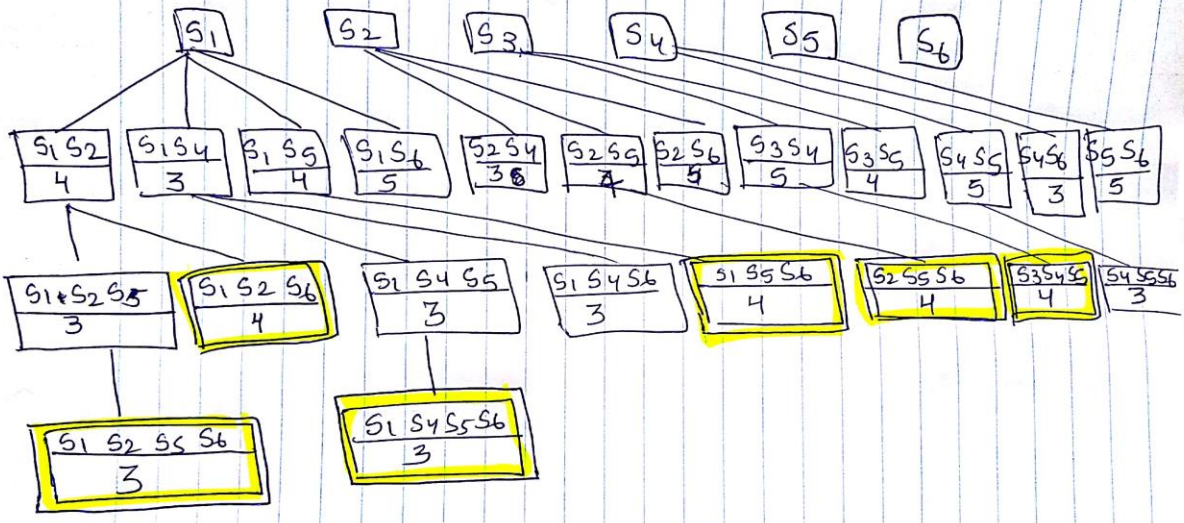
{S2, S3, S4, S5, S6} = 1 [Not frequent]

{S1, S2, S3, S4, S5, S6} = 1 [Not frequent]

Frequent itemsets :

$\{S_1\} = 5$
$\{S_2\} = 6$
$\{S_3\} = 5$
$\{S_4\} = 6$
$\{S_5\} = 7$
$\{S_6\} = 6$
$\{S_1, S_2\} = 4$
$\{S_1, S_4\} = 3$
$\{S_1, S_5\} = 4$
$\{S_1, S_6\} = 5$
$\{S_2, S_4\} = 3$
$\{S_2, S_5\} = 4$
$\{S_2, S_6\} = 5$
$\{S_3, S_4\} = 5$
$\{S_3, S_5\} = 4$
$\{S_4, S_5\} = 5$
$\{S_4, S_6\} = 3$
$\{S_5, S_6\} = 5$
$\{S_1, S_2, S_5\} = 3$
$\{S_1, S_2, S_6\} = 4$
$\{S_1, S_4, S_5\} = 3$
$\{S_1, S_4, S_6\} = 3$
$\{S_1, S_5, S_6\} = 4$
$\{S_2, S_5, S_6\} = 3$

$\{S_3, S_4, S_5\} = 4$
$\{S_4, S_5, S_6\} = 3$
$\{S_1, S_2, S_5, S_6\} = 3$
$\{S_1, S_4, S_5, S_6\} = 3$



The Highlighted boxes contain maximal frequent itemsets. The maximal frequent itemsets are those itemsets which any of its superset becomes infrequent.

$\{S_1\ S_2\ S_5\ S_6\}$ is a maximal frequent itemset.

	S_1	S_2	S_5	S_6
Pb	1	1	1	1
Pg	1	1	1	1
Ph	1	1	1	1

Adding any other item such as S_3

	S_1	S_2	S_5	S_6	S_3
Pb	1	1	1	1	0
Pg	1	1	1	1	0
Ph	1	1	1	1	0

The support goes down from 3 to 0, i.e the superset $\{S_1\ S_2\ S_5\ S_6\ S_3\}$ becomes infrequent.

Adding any other item eg: S_4

	S_1	S_2	S_5	S_6	S_4
Pb	1	1	1	1	0
Pg	1	1	1	1	1
Ph	1	1	1	1	1

The support goes down from 3 to 2, hence superset $\{S_1\ S_2\ S_5\ S_6\ S_4\}$ is infrequent.

Thus, $\{S_1\ S_2\ S_5\ S_6\}$ is a maximal frequent itemset

Similarly $\{S_1, S_4, S_5, S_6\}$ is a maximal frequent itemset

	S_1	S_4	S_5	S_6
Pd	1	1	1	1
Pg	1	1	1	1
Pn	1	1	1	1

Adding any other item suppose S_2 is added

	S_1	S_4	S_5	S_6	S_2
Pd	1	1	1	1	0
Pg	1	1	1	1	1
Pn	1	1	1	1	1

Support goes down from 3 to 2. which is below the minsup Threshold Hence the superset of $\{S_1, S_2, S_5, S_6\}$ which is $\{S_2, S_1, S_4, S_5, S_6\}$ is infrequent.
Thus, $\{S_1, S_4, S_5, S_6\}$ is maximal frequent.
Similarly other items are added and support definately goes below the minsup.

$\{S_1\ S_2\ S_6\}$ is a closed itemset

	S_1	S_2	S_6
P_b	1	1	1
P_e	1	1	1
P_g	1	1	1
P_h	1	1	1

Adding another item for example S_5 to it

	S_1	S_2	S_6	S_5
P_b	1	1	1	1
P_e	1	1	1	0
P_g	1	1	1	1
P_h	1	1	1	1

The support goes down from 4 to 3, but is still above the minsup threshold, i.e the item is still frequent.

∴ $\{S_1\ S_2\ S_6\}$ is a closed itemset as any of its superset has lower support than it.

{S_1 S_5 S_6} is also a closed itemset.

	S_1	S_5	S_6
P_b	1	1	1
P_d	1	1	1
P_g	1	1	1
P_h		1	1

Adding another item such as S_3

	S_1	S_5	S_6	S_3
P_b	1	1	1	0
P_d	1	1	1	1
P_g	1	1	1	1
P_h	1	1	1	1

Support goes down from 4 to 3, but does not go below the minsup.
∴ the {S_1 S_5 S_6} is a closed itemset.
As any of its superset has minimum support than it.
Similarly adding on S_2 ~~S_4~~ or S_4 will also reduce minsupport.

<u>Answer 2)b</u>) An itemset is closed if none of its immediate supersets has the same support as the itemset

Use of knowledge of these closed itemsets:

- Having the closed itemsets we can **generate maximal frequent itemset as well as generate association rules.**
- Since, the datsets in real life are very large and finding all the frequent itemsets will just be resulting in **redundancy** of most of the sets. This redundancy of itemsets needs to be removed. This is made possible by the closed itemsets
- Knowledge of closed itemsets gives us the exact items we need to work on instead of giving the repeated itemsets.
- Another advantage of closed itemset is that the execution of algorithm to find the closed itemset (using CHARMS Algorithm) requires **less execution time** than to find the frequent itemset (using ECLAT algorithm)
- This difference in execution time is noticeable as we move towards larger dataset
- To generate rules, all we require it.

Answer 2)c) An itemset is maximal frequent if none of its immediate supersets is frequent

The potential use for the knowledge of frequent itemsets:

- Using The maximal frequent itemsets, we can find out frequent itemsets.
- Once we find maximal frequent itemset , we can generate all frequent itemset in **a single scan**. Because every subset of frequent itemset is frequent.
- This **saves time and space**, thus increases the efficiency
- It can be interpreted as a compact representation of all the frequent itemsets.
- There are only few maximal frequent itemsets thus they require the least space to store in the memory
- Another advantage of maximal frequent itemset is that the execution of algorithm to find the maximal frequent itemset (using GenMax Algorithm) requires **less execution time** than to find the frequent itemset (using ECLAT algorithm)