Exercise 4-1 Hash-Tree

(a) **Construction**. Using the hash function

$$h(x) = x \mod 3 \quad (1)$$

construct a hash tree with maximum number of itemsets in inner nodes equal to 4 given the following set of candidates:

(1, 9, 11)	(2, 5, 10)	(3, 6, 8)	(4, 7, 9)	(6, 12, 13)	(9, 12, 14)
(1, 10, 12)	(2, 5, 12)	(3, 7, 10)	(4, 7, 13)	(6, 12, 14)	(10, 11, 15)
(2, 4, 7)	(2, 9, 10)	(3, 12, 14)	(5, 7, 9)	(8, 11, 11)	(12, 12, 15)
(2, 5, 8)	(3, 3, 5)	(4, 5, 8)	(5, 7, 13)	(8, 11, 15)	(14, 14, 15)

In the root node, the itemsets are splitted according to the hash value of the first item in the itemset. Hence, after the root node we have 3 child nodes with content:

N_0	N_1	N_2
(3, 3, 5)	(1, 9, 11)	(2, 4, 7)
(3, 6, 8)	(1, 10, 12)	(2, 5, 8)
(3, 7, 10)	(4, 5, 8)	(2, 5, 10)
(3, 12, 14)	(4, 7, 9)	(2, 5, 12)
(6, 12, 13)	(4, 7, 13)	(2, 9, 10)
(6, 12, 14)	(10, 11, 15)	(5, 7, 9)
(9, 12, 14)		(5, 7, 13)
(12, 12, 15)		(8, 11, 11)
		(8, 11, 15)
		(14, 14, 15)

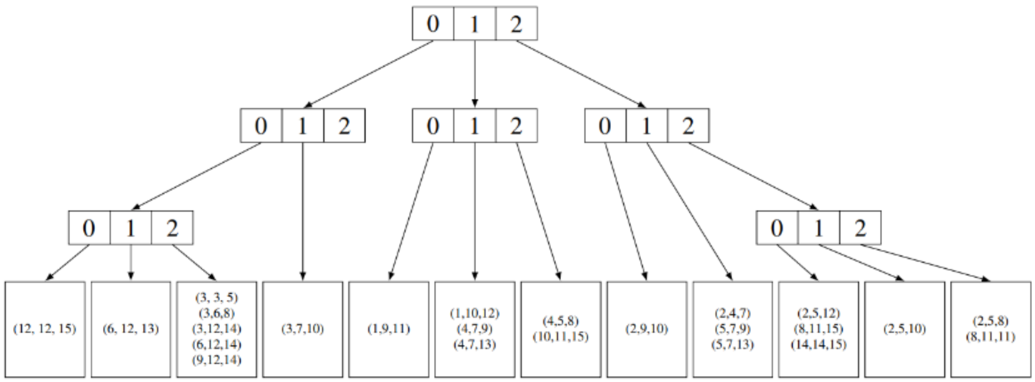
As the fill degree of all nodes is larger 4, all have to be split, now according to the second item.

N_{00}	N_{01}^*	N_{10}^*	N_{11}^*	N_{12}^*	N_{20}^*	N_{21}^*	N_{22}
(3, 3, 5)	(3, 7, 10)	(1, 9, 11)	(1, 10, 12)	(4, 5, 8)	(2, 9, 10)	(2, 4, 7)	(2, 5, 8)
(3, 6, 8)			(4, 7, 9)	(10, 11, 15)		(5, 7, 9)	(2, 5, 10)
(3, 12, 14)			(4, 7, 13)			(5, 7, 13)	(2, 5, 12)
(6, 12, 13)							(8, 11, 11)
(6, 12, 14)							(8, 11, 15)
(9, 12, 14)							(14, 14, 15)
(12, 12, 15)							

Here, only N_{00} and N_{22} have a higher fill degree than allowed (the leaf nodes are marked with *). Hence, they are splitted again, this time using the third item.

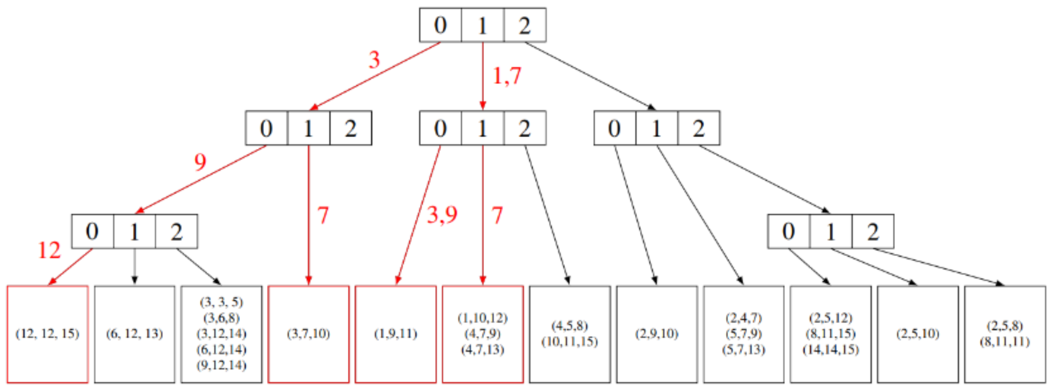
N^*_{000}	N^*_{001}	N^*_{002}	N^*_{220}	N^*_{221}	N^*_{222}
(12, 12, 15)	(6, 12, 13)	(3, 3, 5)	(2, 5, 12)	(2, 5, 10)	(2,5,8)
		(3, 6, 8)	(8, 11, 15)		(8, 11, 11)
		(3, 12, 14)	(14, 14, 15)		
		(6, 12, 14)			
		(9, 12, 14)			

Although N_{002}'s fill degree is larger then 4, there is no remaining item to be used for further splitting. Hence, the hash-tree construction finishes. The final hash-tree is depicted below:



(b) **Counting**. Given the transaction $t = (t_1, \dots, t_5) = (1, 3, 7, 9, 12)$, find all candidates of length $k = 3$ in the previously constructed tree from exercise (a). In absolute and relative numbers: How many candidates need to be refined? How many nodes are visited?

Applying the hash function to the transaction gives $(1, 0, 1, 0, 0)$. The following diagram shows the accessed nodes. A detailed explanation follows below.



(i) Depth $d = 1$. Compute hash values for $t_1, \dots, t_{n-k+d} = t_3$:

$$h(1) = 1 \qquad h(3) = 0 \qquad h(7) = 1 \tag{2}$$

. Continue search in N_0, N_1 (i.e. exclude N_2).

(ii) Depth $d = 2$. Additionally compute $h(t_4) = h(9) = 0$.

- In N_0 reached by item t_2, the nodes for hash values 0 (N_{00} reached by t_4) and 1 (N^*_{01} reached by t_3) are of interest.
- In N_1 reached by item t_1 and t_3, the nodes for hash values 0 (N^*_{10} reached by t_2 and t_4) and 1 (N^*_{11} reached by t_3) are of interest.

(iii) Depth $d = 3$. Additionally compute $h(t_5) = h(12) = 0$.

- In N_{00} reached by $t_2, t_4 = 3, 9$ continue with N^*_{000}.
- In N^*_{01} reached by $t_2, t_3 = 3, 7$ search for $t_2, t_3, t_4 = 3, 7, 9$ and $t_2, t_3, t_5 = 3, 7, 12$. Both are not found.
- In N^*_{10} reached by
  - $t_1 t_2 = 1, 3$,
  - $t_1 t_4 = 1, 9$, or
  - $t_3 t_4 = 7, 9$

  search for
  - $t_1 t_2 t_3 = 1, 3, 7$
  - $t_1 t_2 t_4 = 1, 3, 9$
  - $t_1 t_2 t_5 = 1, 3, 12$
  - $t_1 t_4 t_5 = 1, 9, 12$
  - $t_3 t_4 t_5 = 7, 9, 12$

  None of them is found.
- In N^*_{11} reached by $t_1, t_3 = 1, 7$search for $t_1, t_3, t_4 = 1, 7, 9$ and $t_1, t_3, t_5 = 1, 7, 12$. Both are not found.

(iv) Depth $d = 4$.

- In N^*_{000} reached by $t_2, t_4, t_5 = 3, 9, 12$ search for this transaction. It is not found.

In total, $4/12 \approx 33\%$ of the leaf nodes are visited, $8/18 \approx 44\%$ of the nodes are visited and $6/24 = 25\%$ of the candidates are compared. As result, none of the candidates is supported by the transaction.