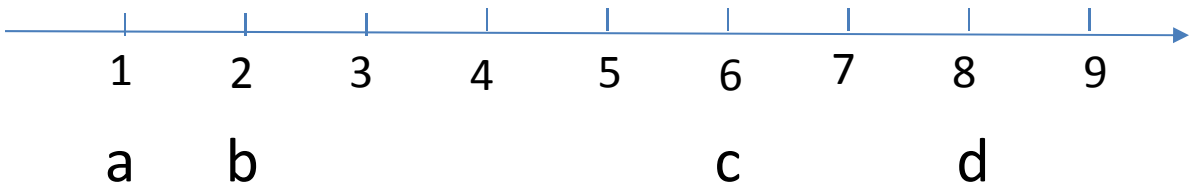
Example of Hierarchical Clustering Using Heap

INF 553

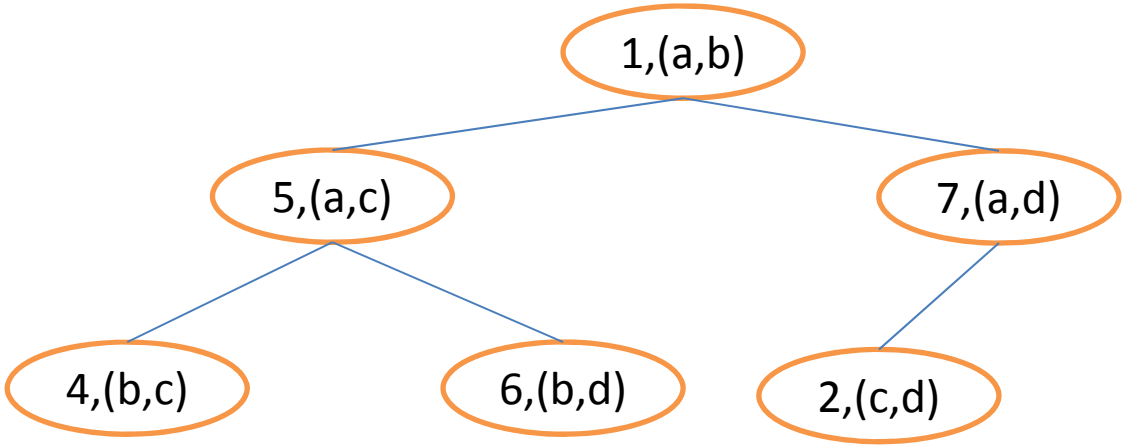
Wensheng Wu

Clustering Problem

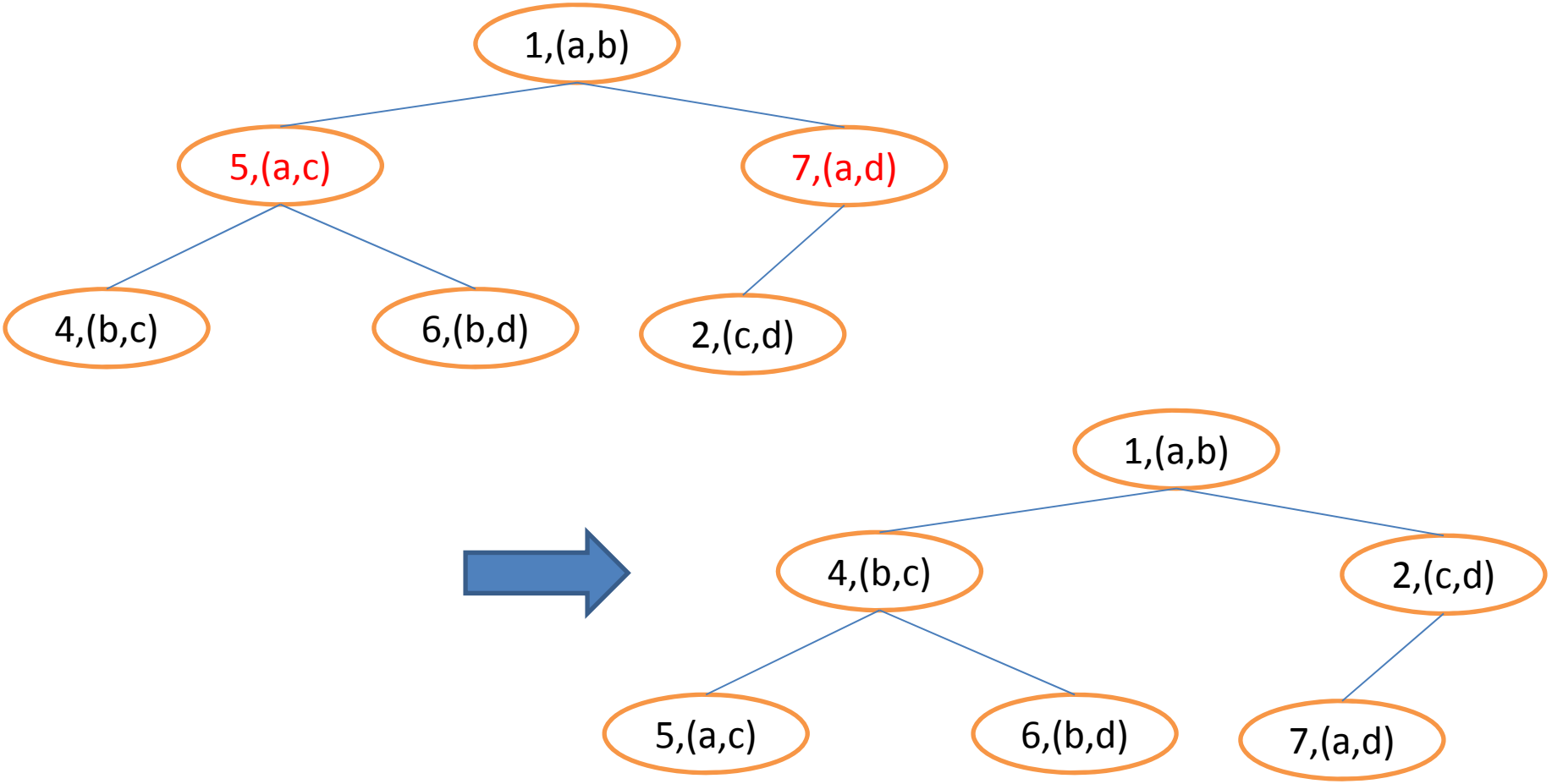


	b	c	d
a	1	5	7
b		4	6
c			2

Initial Heap Before Heapified

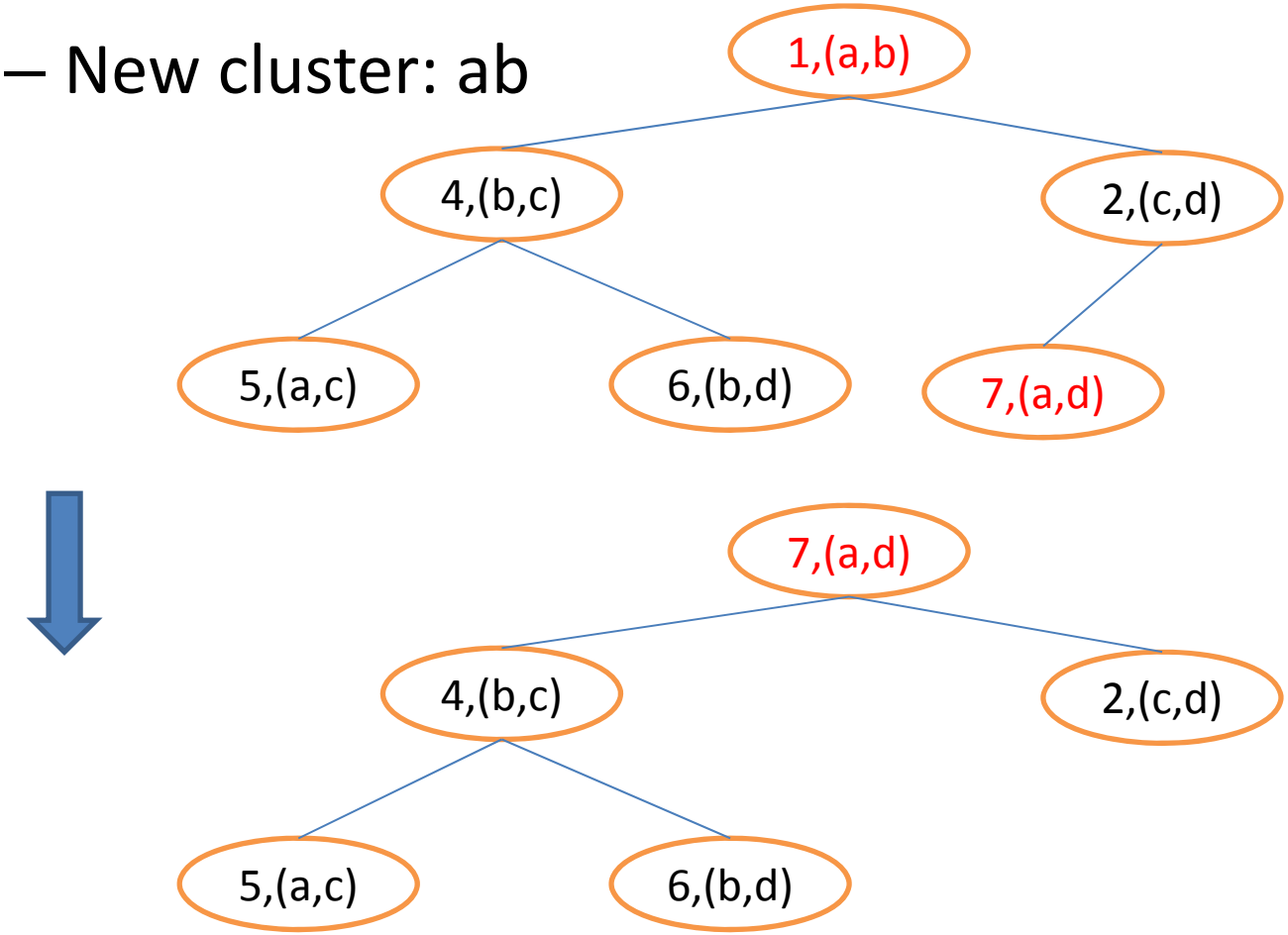


After Heapified

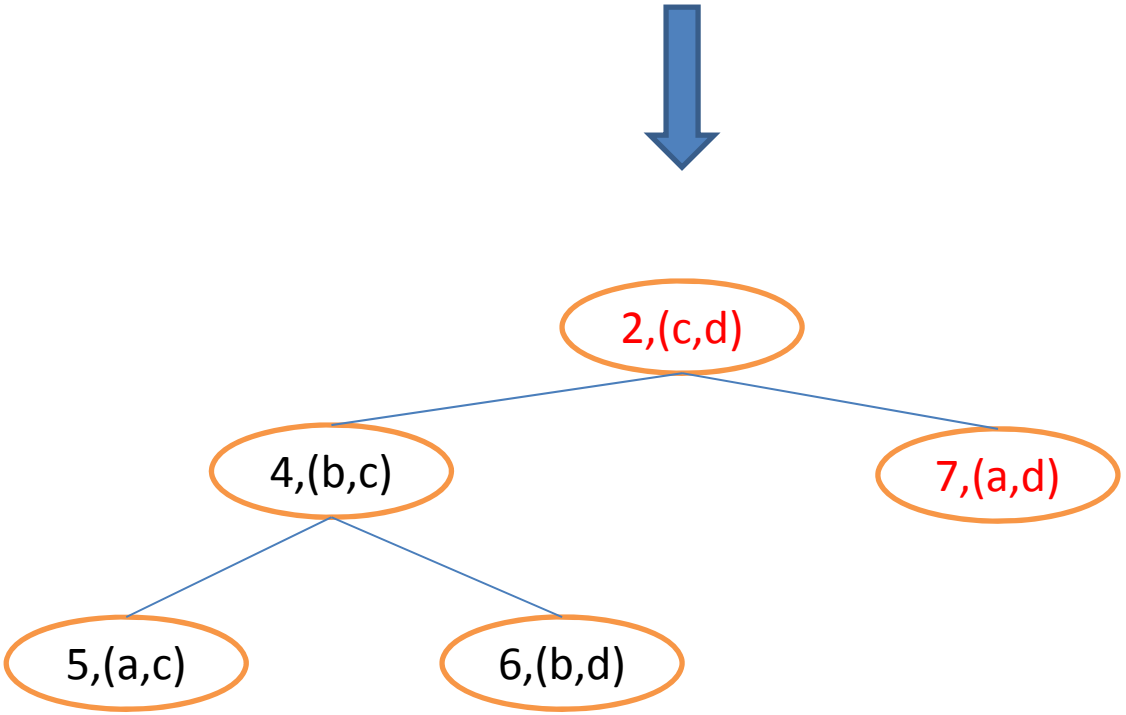


After one merge

- Extract: 1, (a, b)
  - New cluster: ab

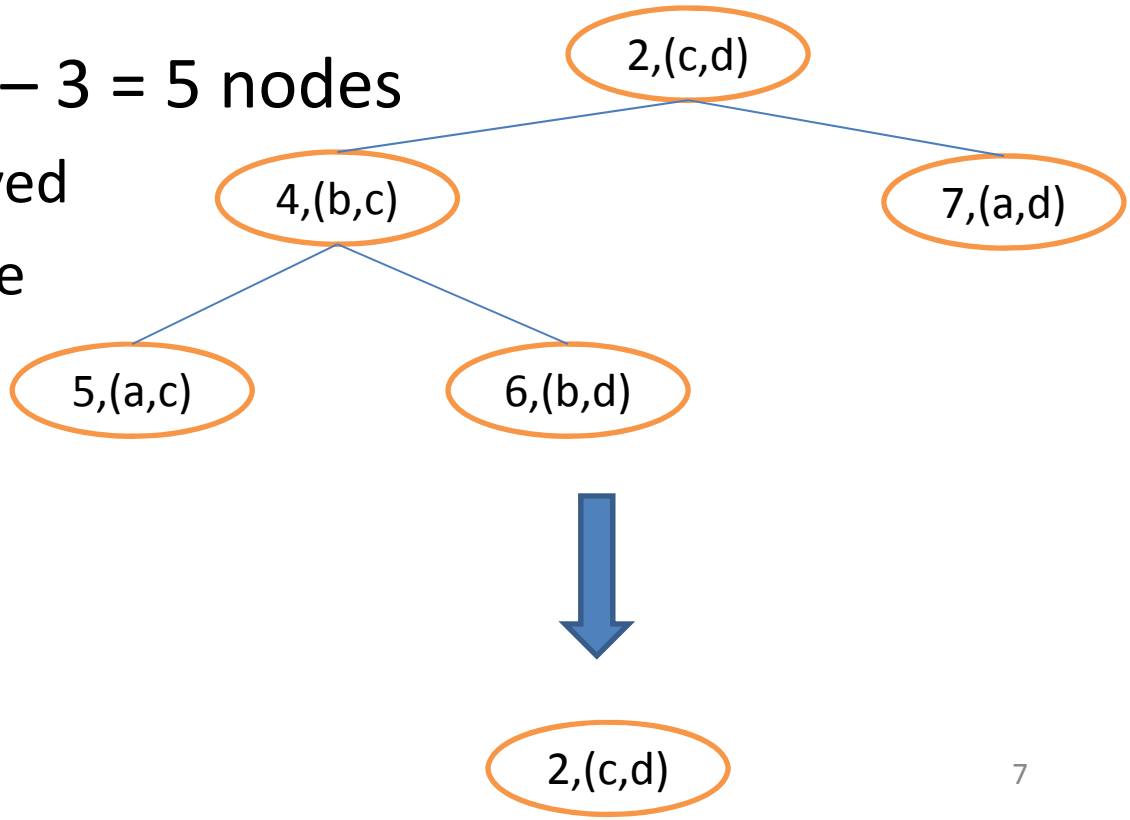


- Sifting root down



Remove Nodes Having Old Clusters

- Need an index to know which nodes having a or b
  - # of clusters = 4
  - Need to remove 2n – 3 = 5 nodes
    - (a, b) already removed
    - Remove 4 more here



Compute Distance of ab with c & d

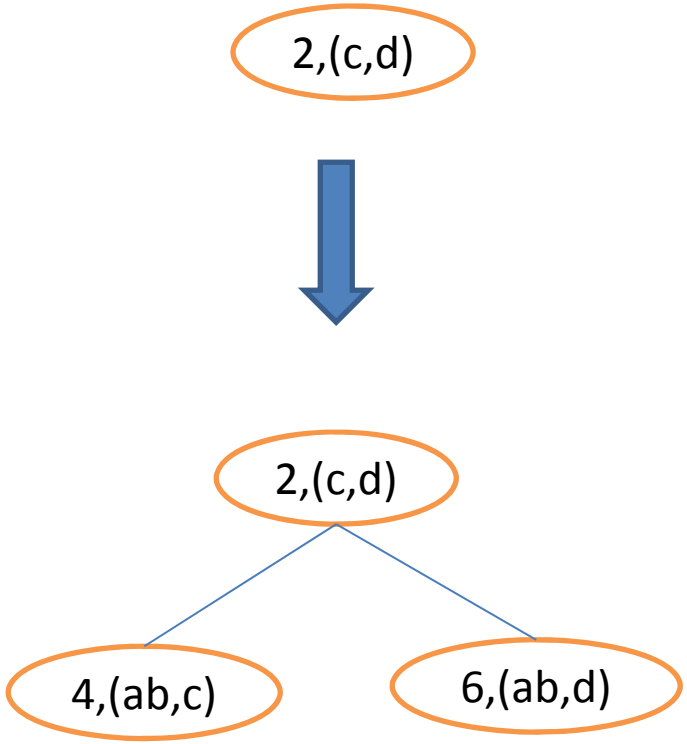
	b	c	d
a	1	5	7
b		4	6
c			2

	c	d
ab	4	6
c		2

New

Adding new pairs to heap

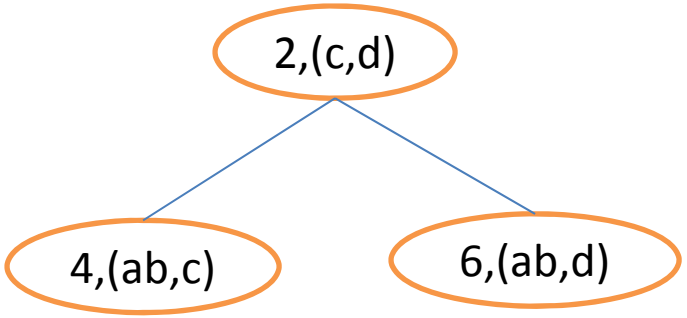
- # of clusters
  - before merge = 4 or n
  - After merge = 3 or n -1
    - ab, c, d
- Add n – 2 pairs
  - Dist btw ab and c
  - Dist btw ab and d



Merging c and d

- Remove 2, (c, d)
- Remove all nodes involving c and d
  - 2k -3 = 3, since k = 3

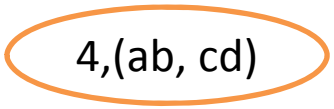
⇒Empty heap



Empty heap

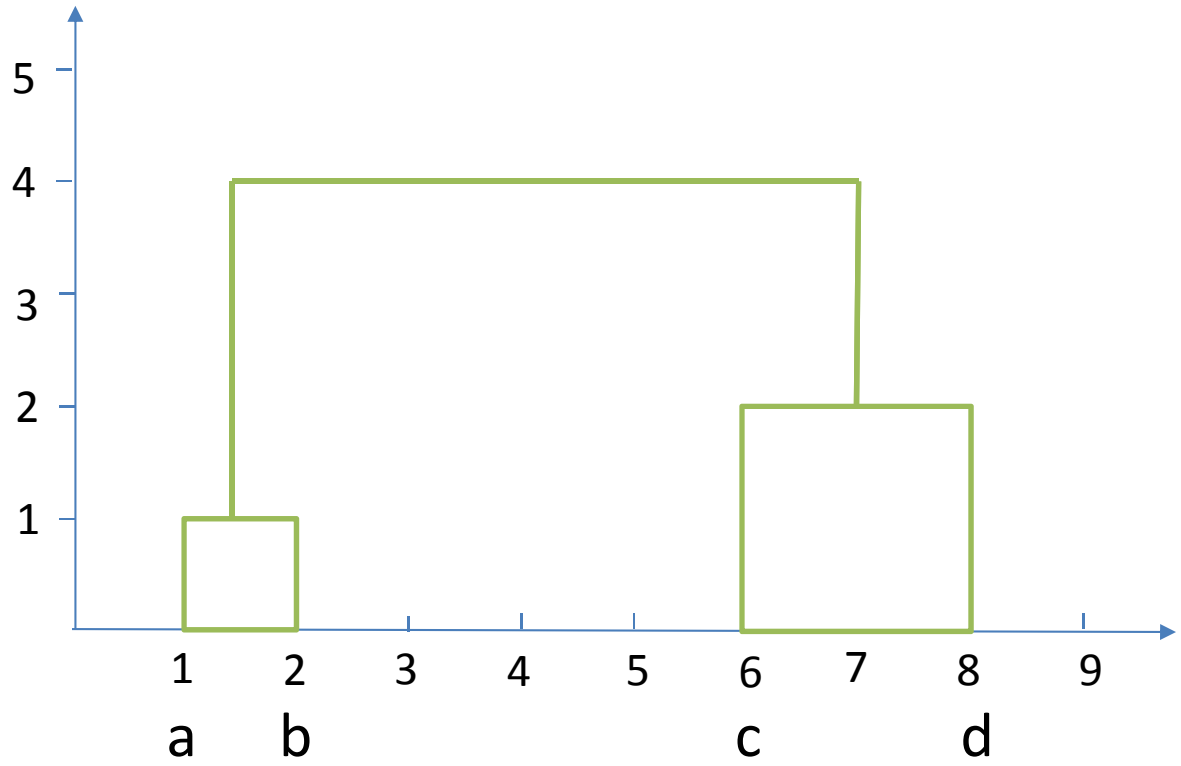
Adding new pairs to heap

- Clusters: (ab) (cd)
  - Distance = dist btw b and c = 4
- Add to heap
  - Add k – 2 = 1 node



- Final merge => (abcd)

Dendrogram

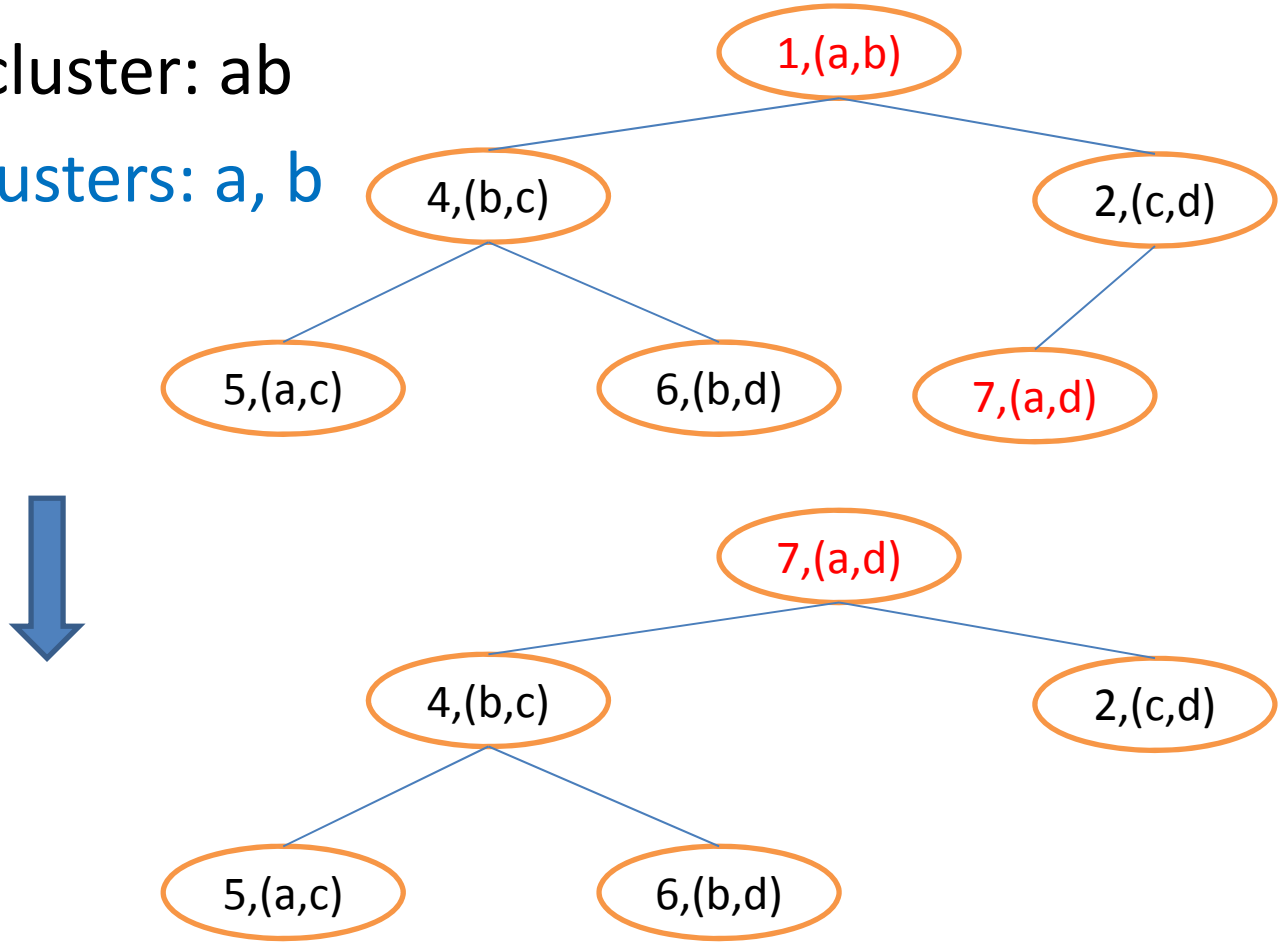


Using Lazy Deletion

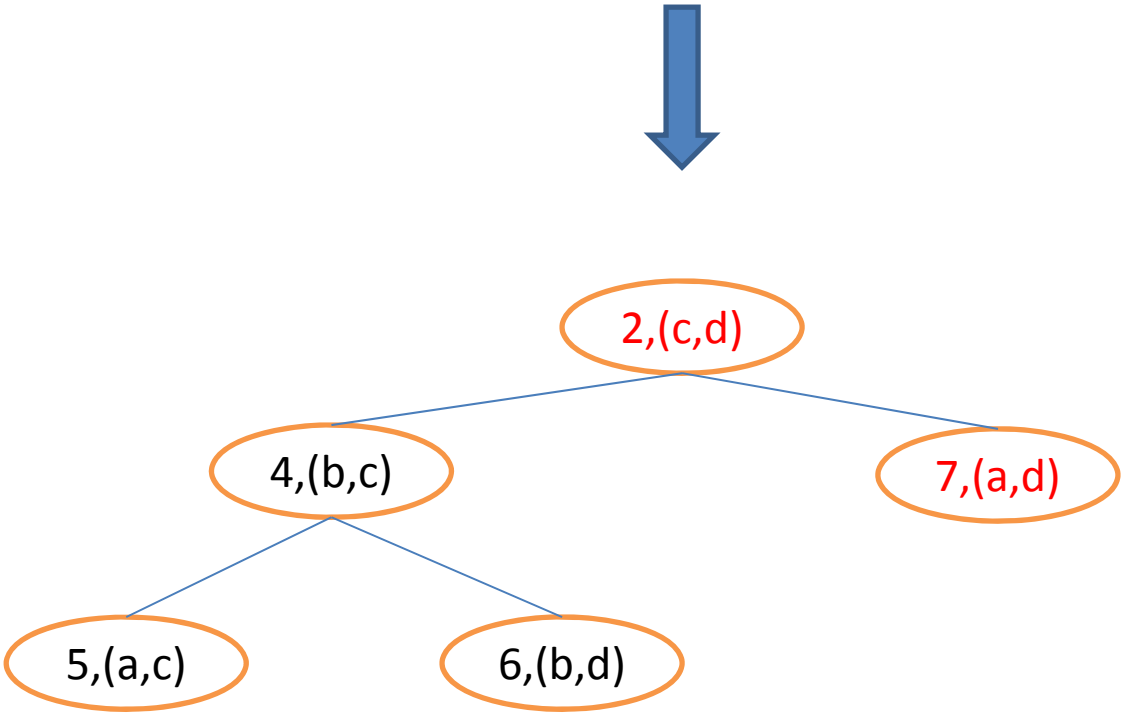
- Do not remove nodes that involve old clusters
- Until they show up at the root
  - When doing ExtractMin

After one merge

- Extract: 1, (a, b)
  - New cluster: ab
  - Old clusters: a, b

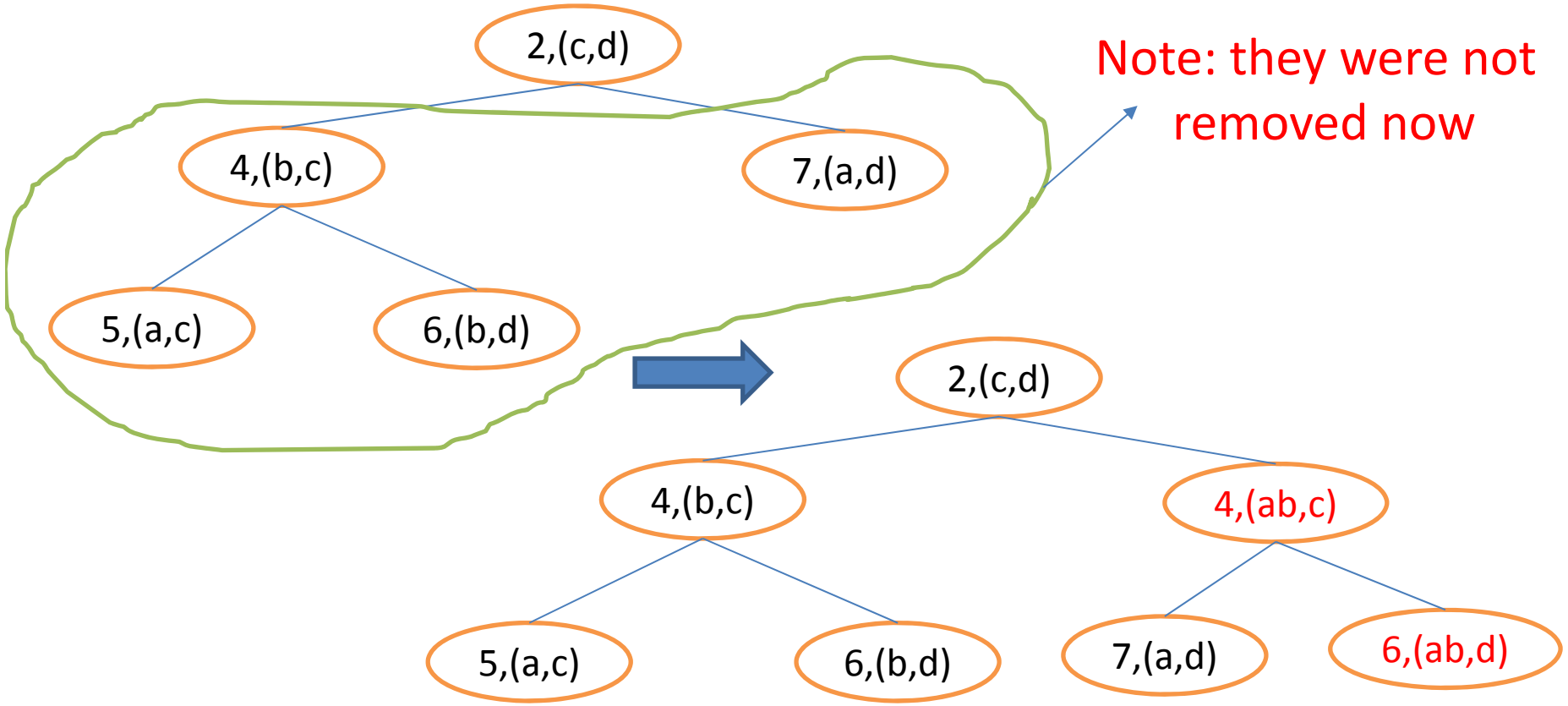


- After sifting root down



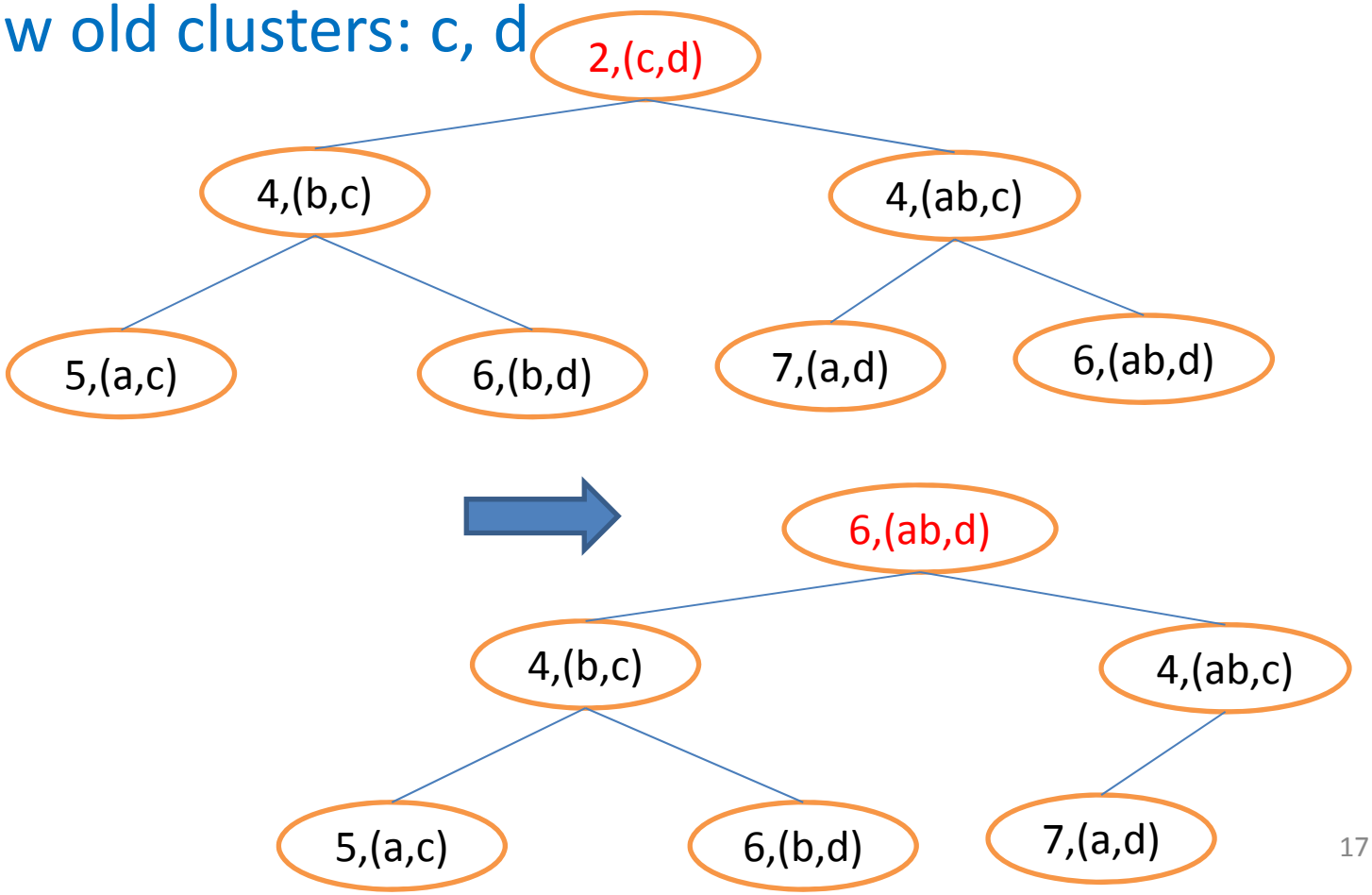
Add Pairs for New Cluster ab

- 4, (ab, c) and 6, (ab, d): sifting up if needed



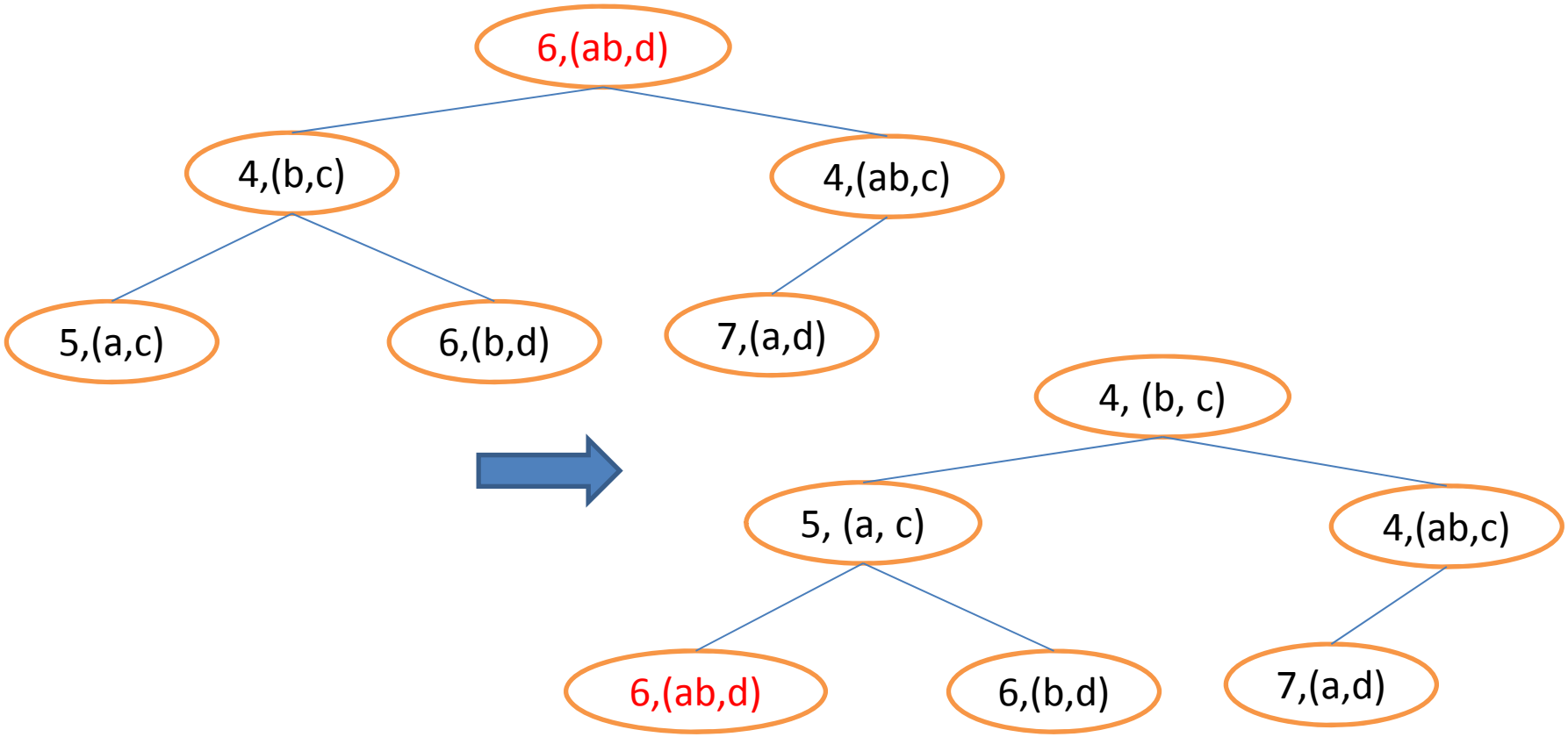
Second Merge

- Extract 2, (c, d), i.e., merge c and d
  - Move last leaf to root and sift down (next slide)
  - New old clusters: c, d



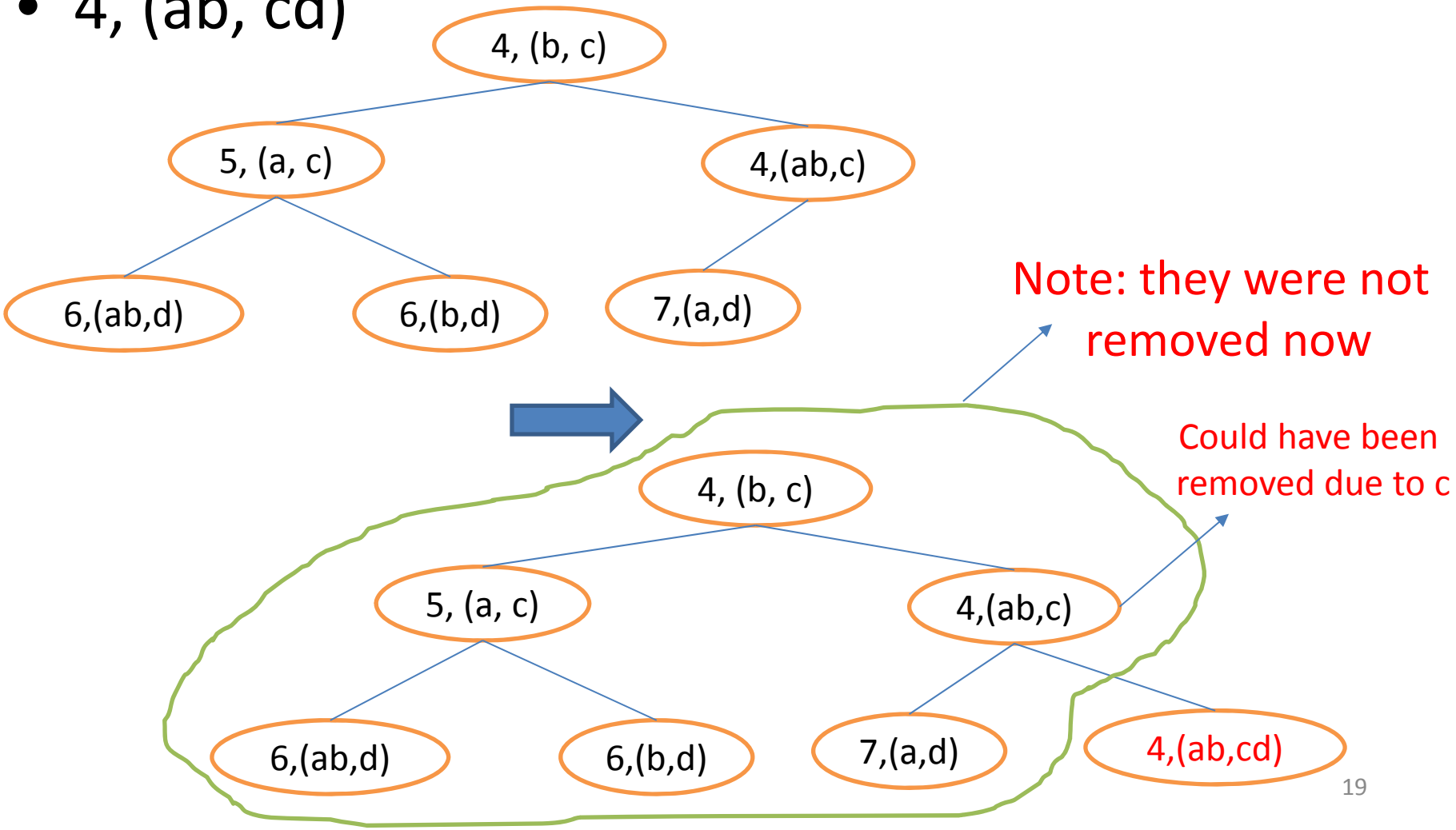
Sift down root

- Can go either way, say left



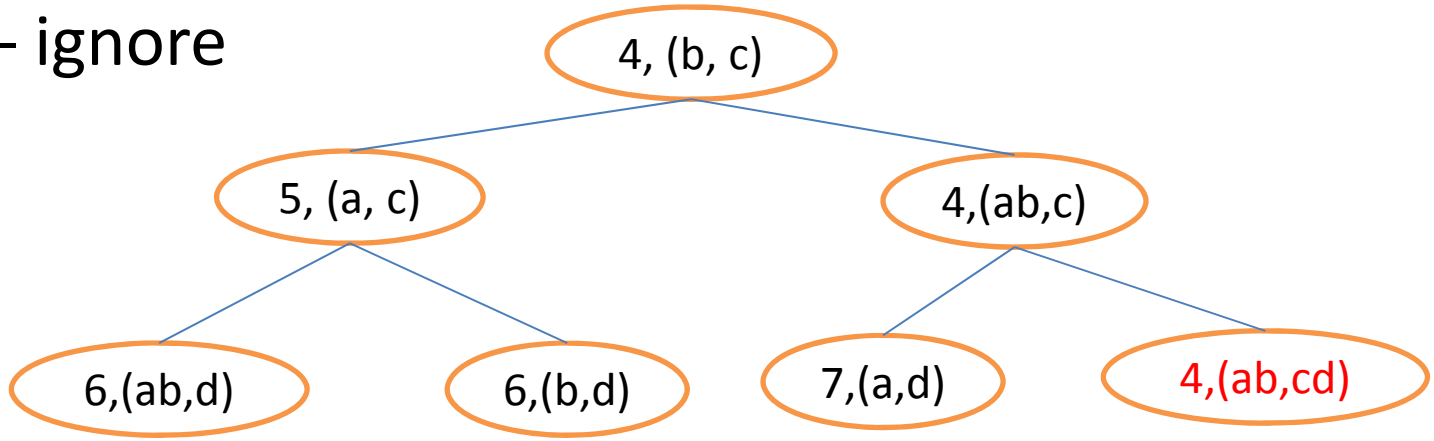
Add Pairs for New Cluster cd

- 4, (ab, cd)

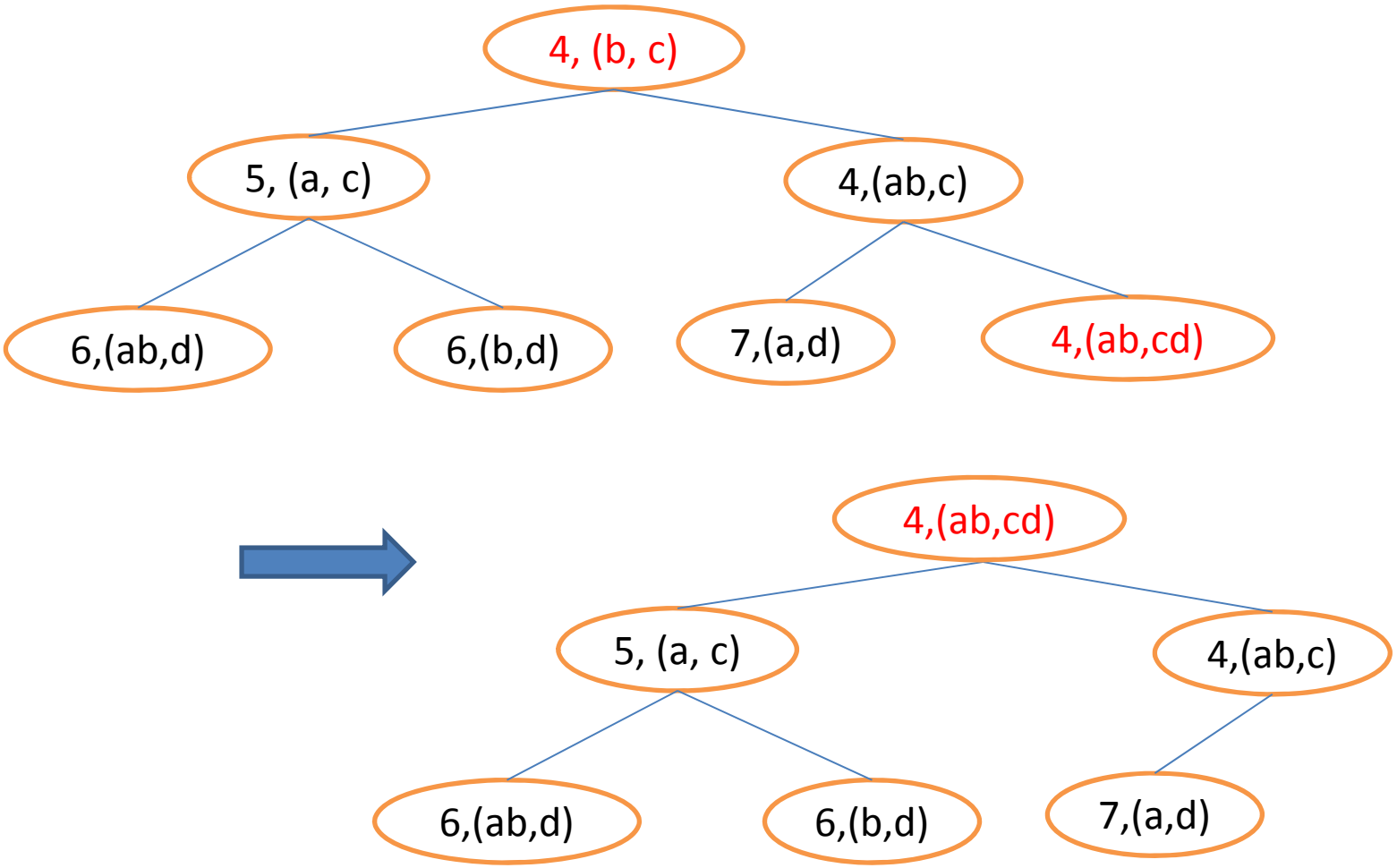


Find Next Two Clusters Merge

- Old clusters: a, b, c, d
  - A & b merged into ab, c & d => cd
- Extract 4, (b, c)
  - found out it is for old cluster
  - ignore



After Exacting 4, (b, c)



Found a Qualified Pair

- Extract 4, (ab, cd)