Greedy Algorithms

Heejin Park

Hanyang University

Contents

- **Introduction**
- **An activity selection problem**
- **Elements of the greedy strategy**
- **Huffman codes**

Introduction

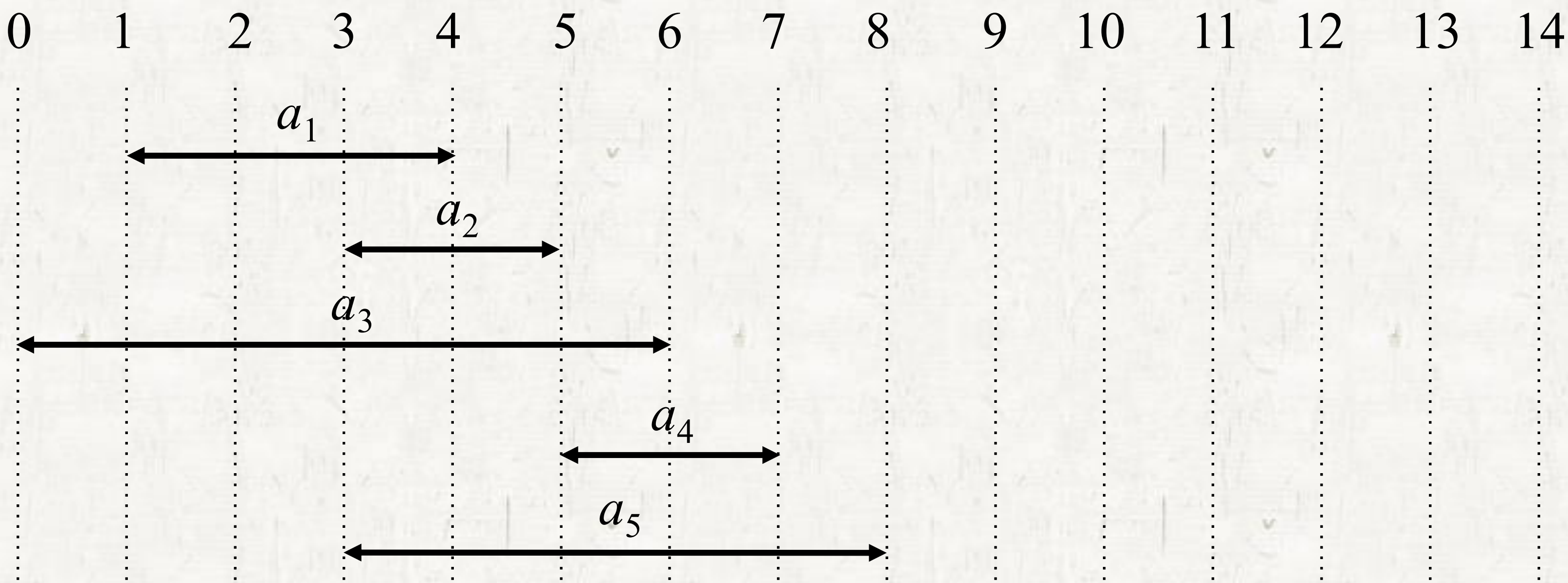
- A ***greedy algorithm*** always makes the choice that looks best at the moment.
- It makes a locally optimal choice in the hope that this choice will lead to a globally optimal solution.
- It makes the choice ***before*** the subproblems are solved.

An activity selection problem

- **An activity selection problem**
  - To select a maximum-size subset of mutually compatible activities.
  - For example
    - Given n classes and 1 lecture room,
    - to select the maximum number of classes

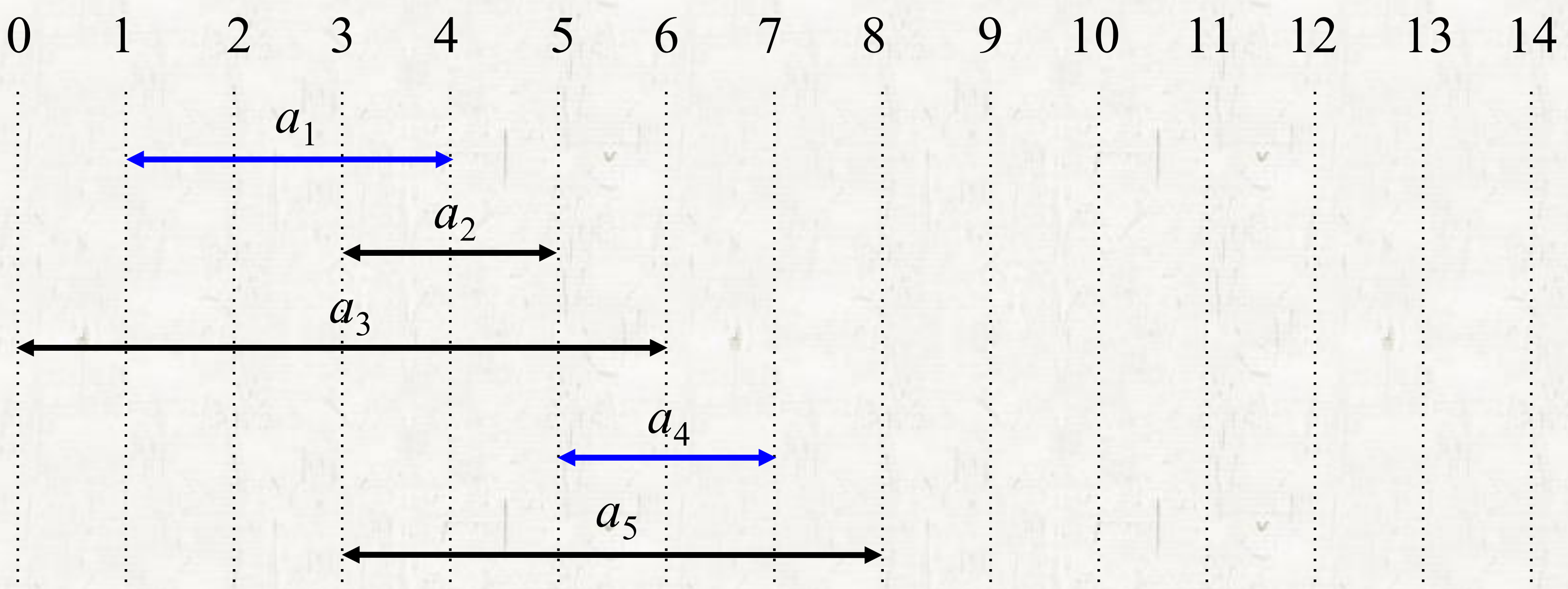
An activity selection problem

- A set of ***activities***: $S = \{a_1, a_2, ..., a_n\}$
- Each activity a_i has its ***start time*** $\boldsymbol{s_i}$ and ***finish time*** $\boldsymbol{f_i}$.
  - $0 \leq s_i < f_i < \infty$

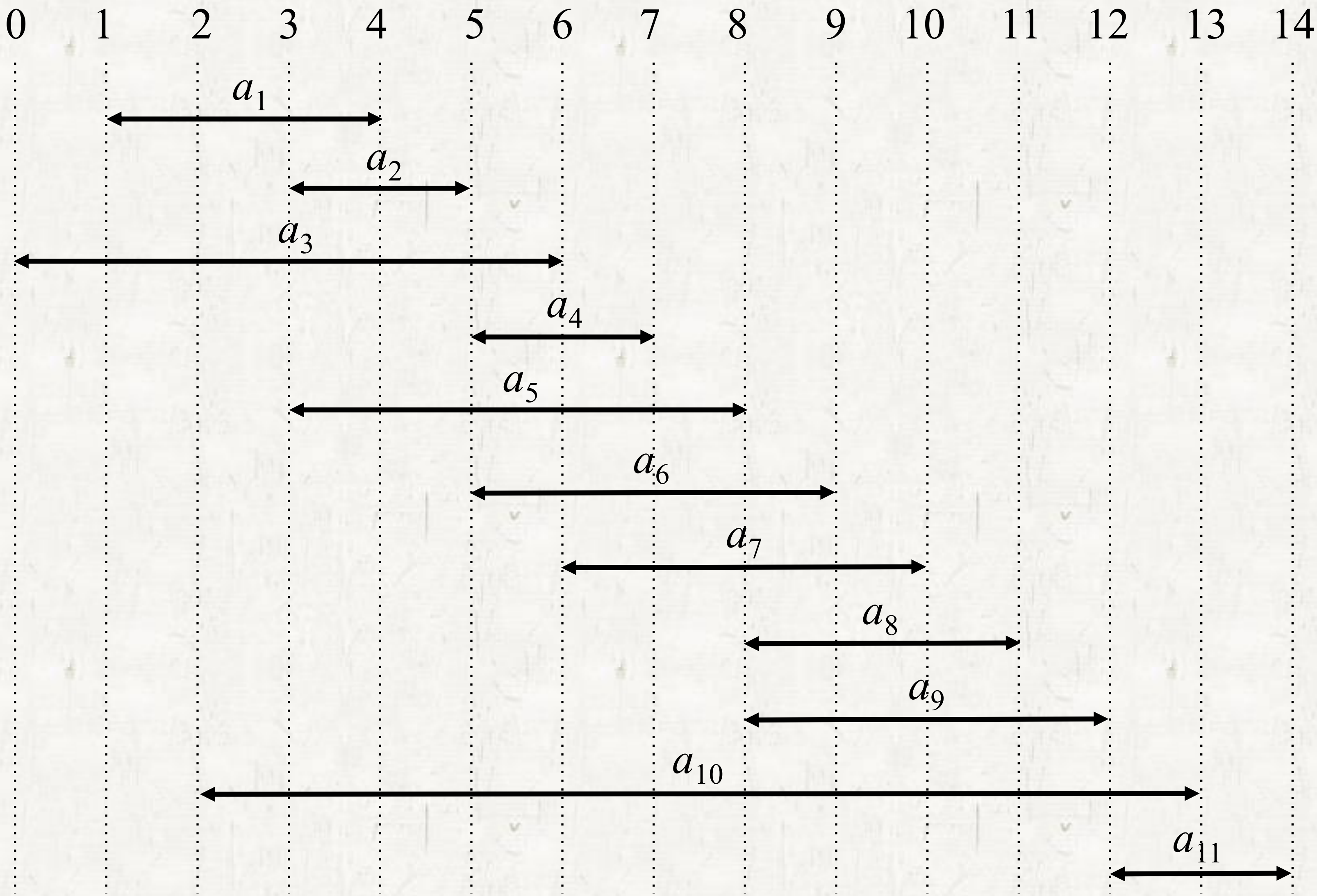


An activity selection problem

- Activity a_i takes place during $[s_i, f_i)$
- Activities a_i and a_j are ***compatible*** if the intervals $[s_i, f_i)$ and $[s_j, f_j)$ do not overlap.

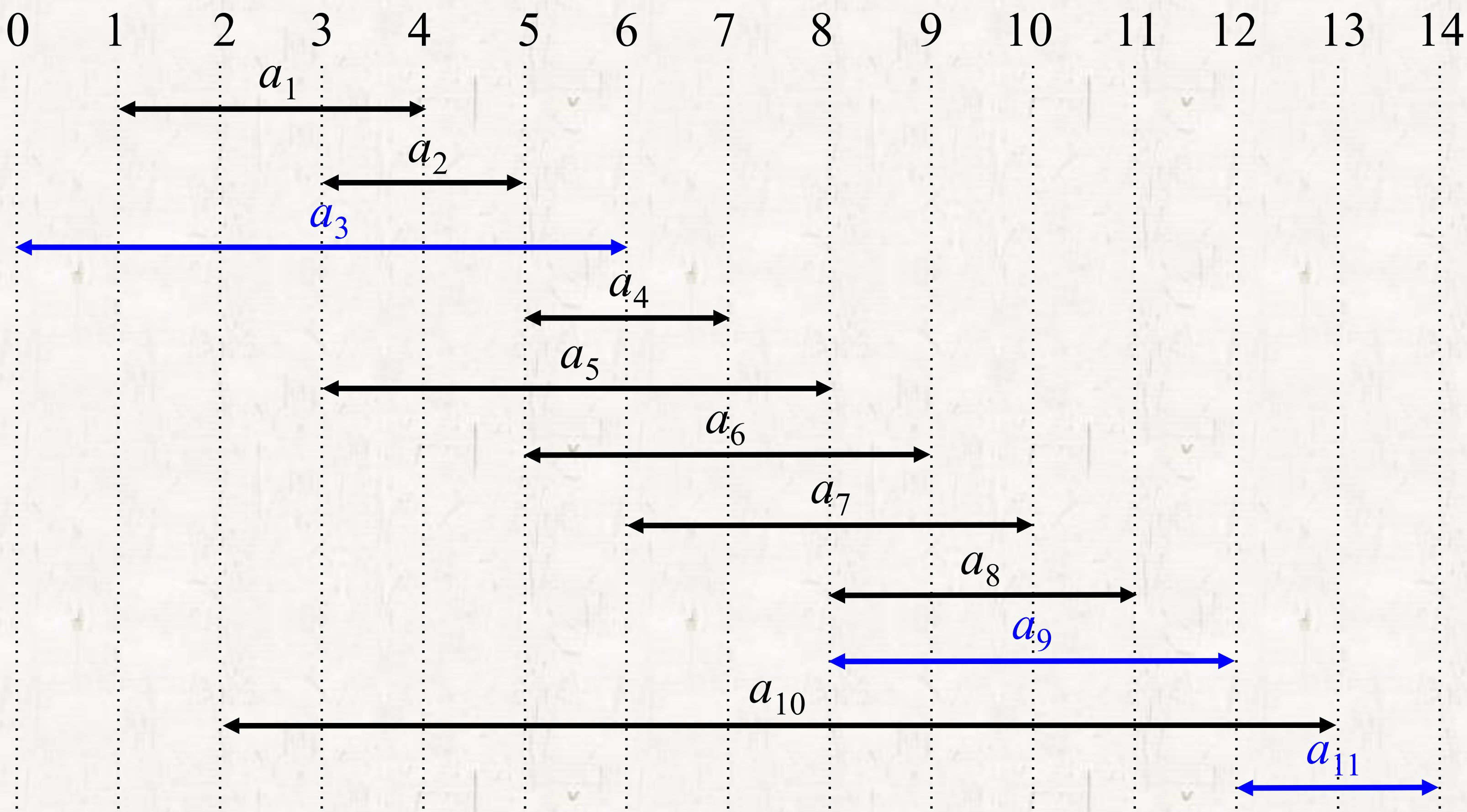


An activity selection problem



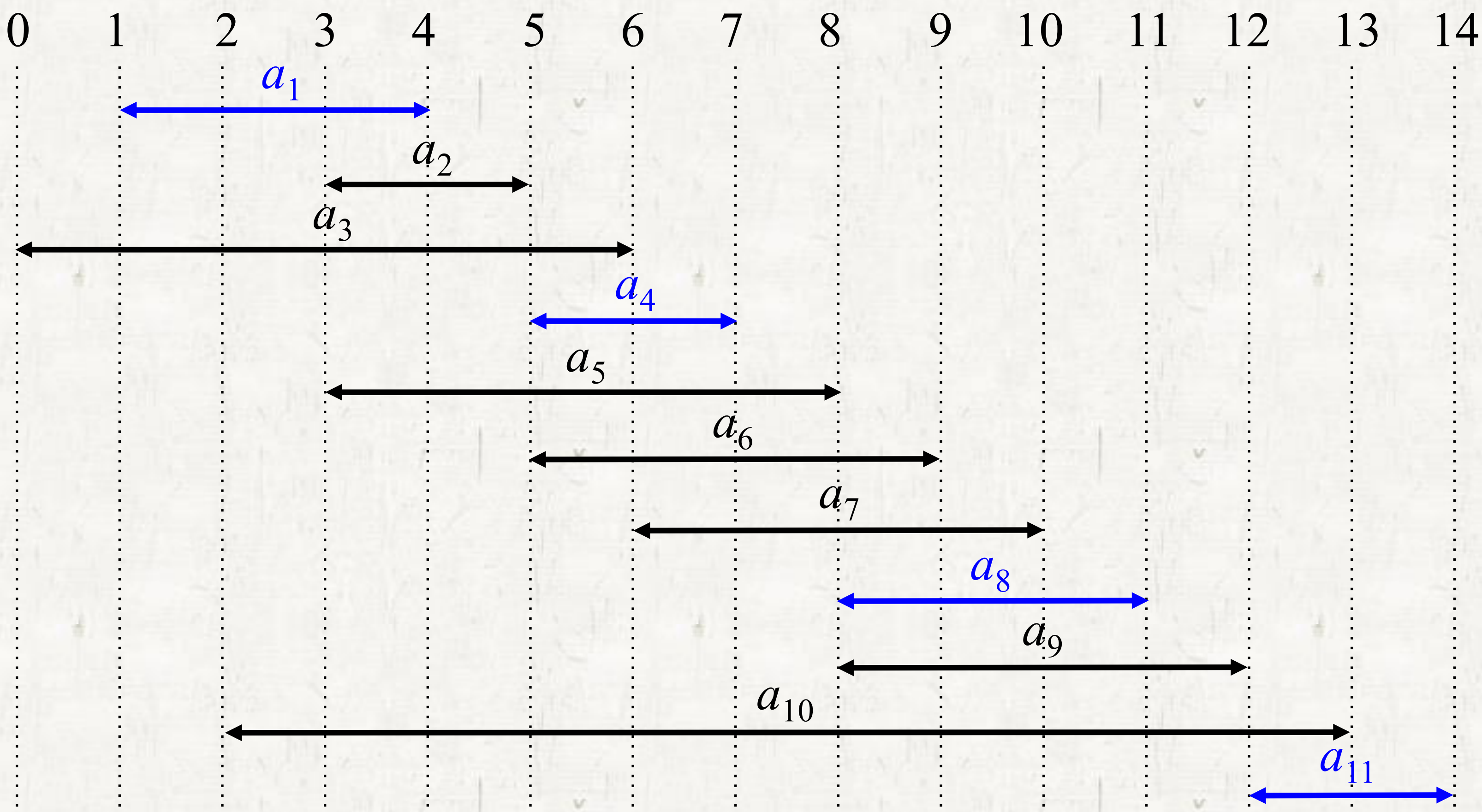
An activity selection problem

- $\{a_3, a_9, a_{11}\}$: mutually compatible activities, not a largest set



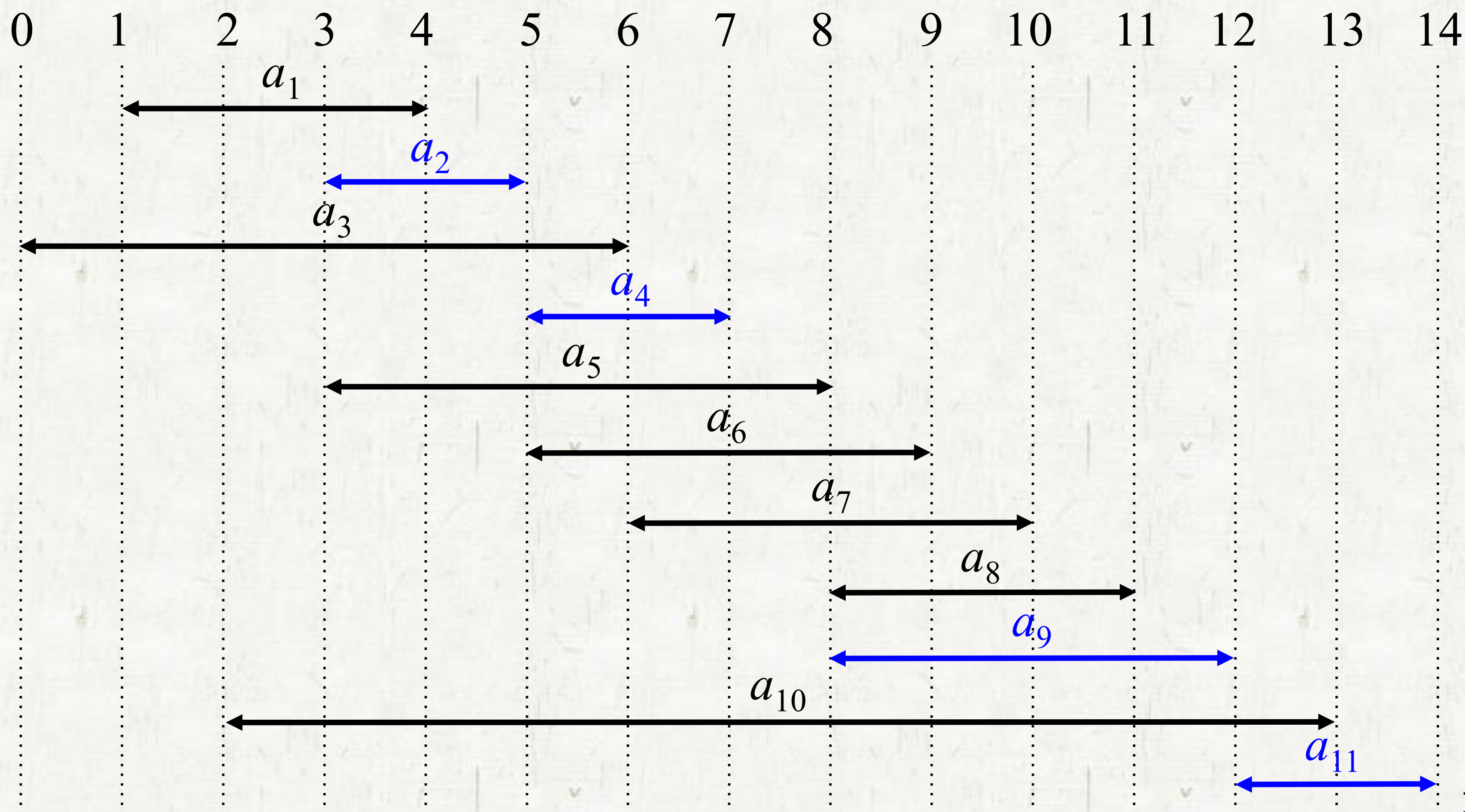
An activity selection problem

- $\{a_1, a_4, a_8, a_{11}\}$: A largest set of mutually compatible activities



An activity selection problem

- $\{a_2, a_4, a_9, a_{11}\}$: Another largest subset



An activity selection problem

- **Optimal substructure**
  - Assume that activities are sorted in increasing order of finish time.

$$f_0 \le f_1 \le f_2 \le \cdots \le f_n < f_{n+1}$$

i	1	2	3	4	5	6	7	8	9	10	11
s_i	1	3	0	5	3	5	6	8	8	2	12
f_i	4	5	6	7	8	9	10	11	12	13	14

An activity selection problem

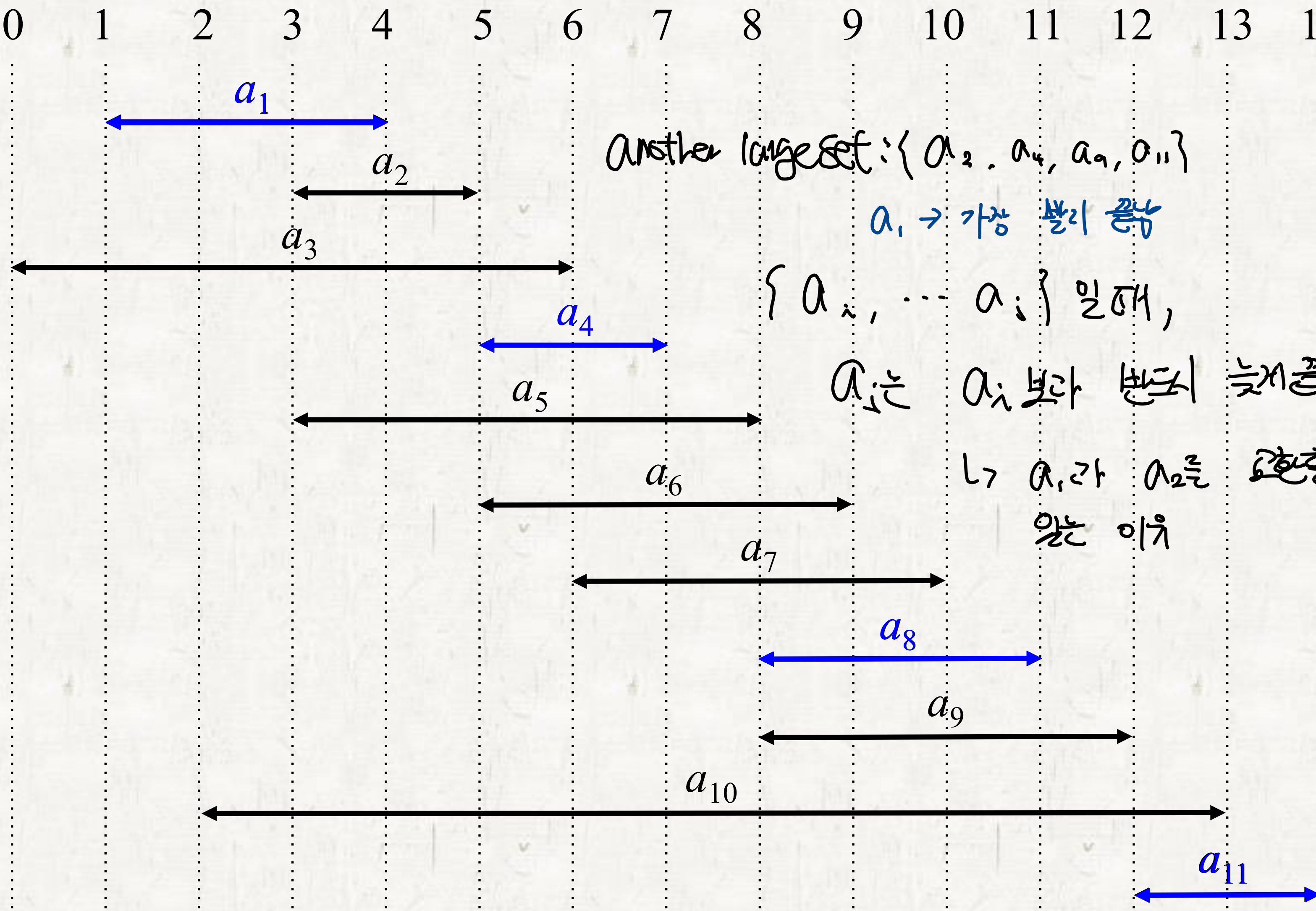
- **Greedy algorithm**
  - Select the earliest finishing activity one by one.

Running Time : $\Theta(n)$

n개의 정렬된 요소 탐색

new activity 와 latest activity를 비교

An activity selection problem



another largeset: $\{a_2, a_4, a_9, a_{11}\}$

a_1 → 가장 빨리 끝남

$\{a_i, \cdots a_j\}$ 일때,

a_j는 a_i보다 반드시 늦게끝난다.

↳ a_1가 a_2를 교환할수 있는 이유

Contents

- *Introduction*
- *An activity selection problem*
- **Elements of the greedy strategy**
- **Huffman codes**

Elements of the greedy strategy

- **Greedy-choice property**
  - Make the choice ***before*** the subproblems are solved.
  - Only one subproblem is generated.
- **Dynamic programming**
  - Make a choice ***after*** the subproblems are solved.
  - Several subproblems may be generated.

Elements of the greedy strategy

- **Greedy vs. Dynamic programming**
  - **0-1 knapsack**
    - A thief robbing a store finds n items.
    - The ith item is worth v_i dollars and weighs w_i pounds.
    - He can carry at most W pounds in his knapsack.
    - The n, v_i, w_i, and W are integers.
    - Which items should he take?
  - **Fractional knapsack**
    - In this case, the thief can take fractions of items.

Item이 나뉘어질 수 있다.

Elements of the greedy strategy

- **Fractional knapsack**
  - The greedy strategy works.
  - Compute the value per pound v_i/w_i for each item.
  - Take as much as possible of the item with the greatest value per pound.

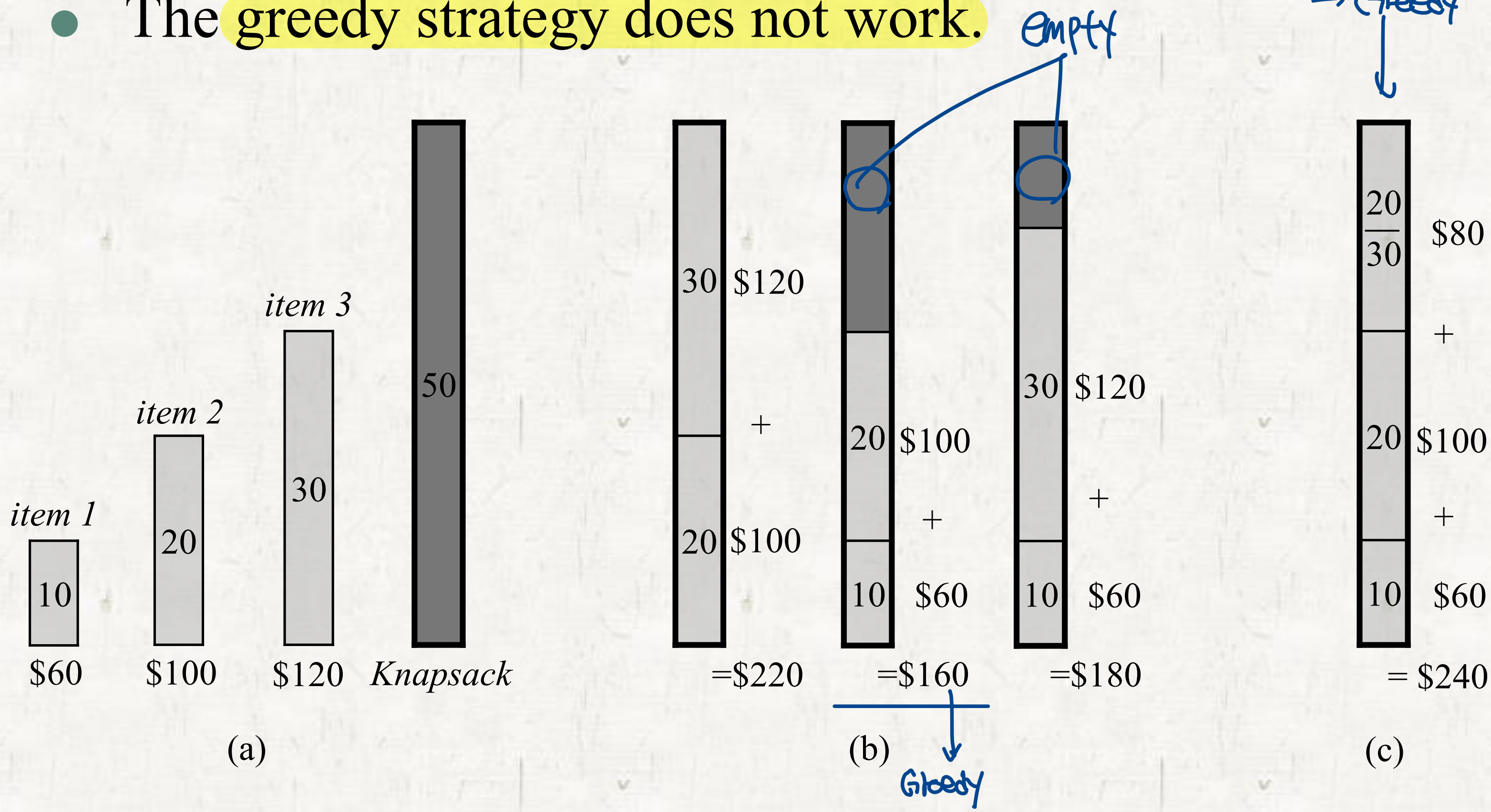
$$\boxed{\frac{V_1}{W_1}} \geq \frac{V_2}{W_2} \geq \frac{V_3}{W_3}$$

1

Elements of the greedy strategy

- **0-1 knapsack**
  - The greedy strategy does not work.

fraction 0-1 knapsack
→ Greedy work!



Self-study

- **Exercise 16.2-1**
- **Exercise 16.2-2**
- **Exercise 16.2-5**
- **Exercise 16.2-7**

Contents

- *Introduction*
- *An activity selection problem*
- *Elements of the greedy strategy*
- **Huffman codes**

Huffman codes

- **Huffman Codes**
  - A widely used technique for compressing data.

$3 \times 100{,}000 = 300{,}000$

- Consider representing 100,000 characters from {a, b, c, d, e, f}.
  - 3-bit ***fixed-length code*** is used in general.
  - It takes 300,000 bits in total

	a	b	c	d	e	f
Frequency (in thousands)	45	13	12	16	9	5
Fixed-length codeword	000	001	010	011	100	101

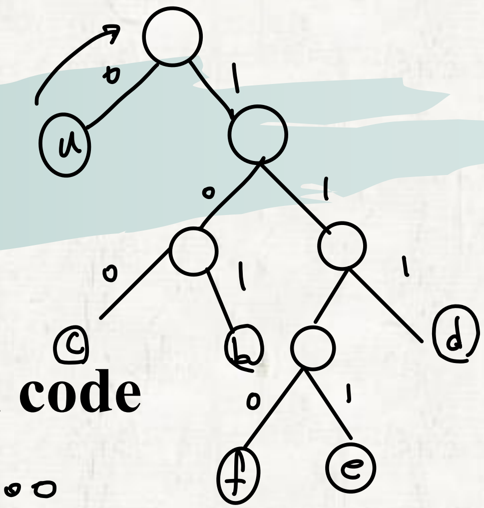
Huffman codes

- We can reduce the space if ***variable-length code*** is used.

	a	b	c	d	e	f
Frequency (in thousands)	45	13	12	16	9	5
Fixed-length codeword	000	001	010	011	100	101
Variable-length codeword	0	101	100	111	1101	1100

- Shorter **codewords** for frequent characters.
- 224,000 bits in total
  - $(45\cdot 1+13\cdot 3+12\cdot 3+16\cdot 3+9\cdot 4+5\cdot 4)\cdot 1000$ bits

radix Tree

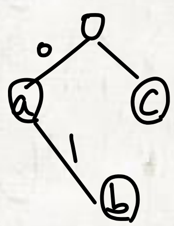


Huffman codes

- **Encoding and decoding of variable-length code**
  - Encoding abc : 0·101·100 : abc → 0101100
  - Decoding 001011101 = aabe!
    - 0·0·101·1101: aabe

prefix 존재 X
→ 중복 decoding 없다

	a	b	c	d	e	f
Variable-length codeword	0	101	100	111	1101	1100



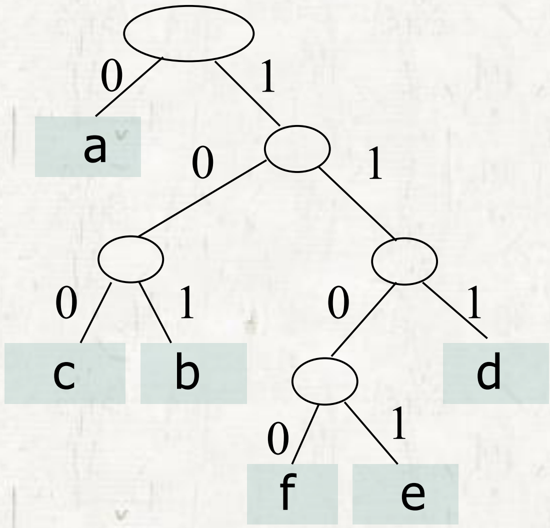
이경우, a가 b의 prefix

  - Decoding 001 when a: 0 b: 01 c: 1
    - 001: aac or ab → 둘 다 가능
    - The codeword 0 for a is a prefix of the codeword 01 for b.

Huffman codes

- **Prefix codes**
  - No codeword is a prefix of some other codeword.

	a	b	c	d	e	f
Variable-length codeword	0	101	100	111	1101	1100

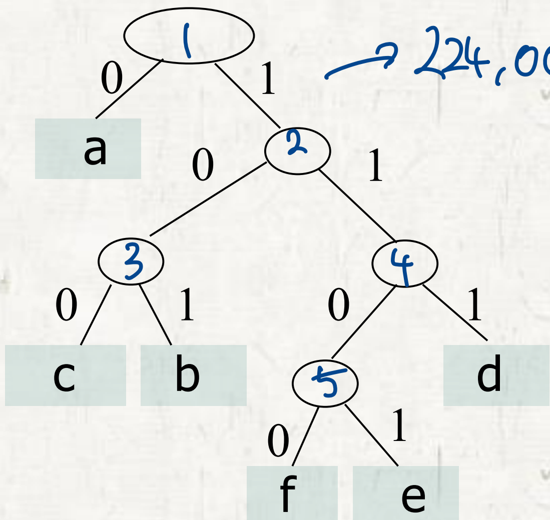


Huffman codes

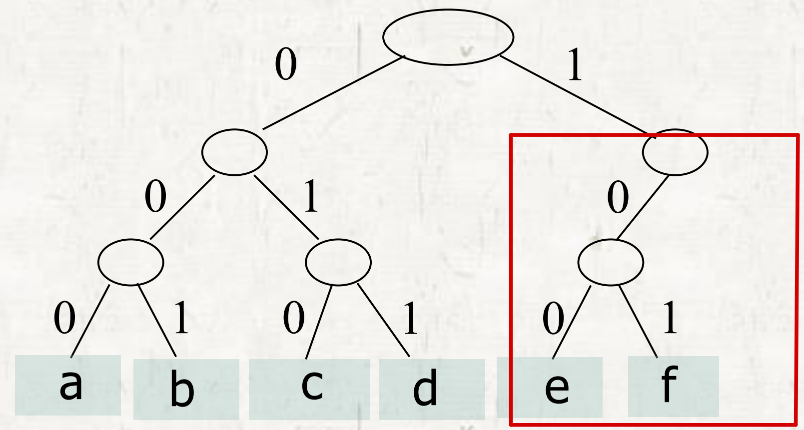
match → 1 winner
unmatch → loser → |C| - 1

- **Prefix codes**
  - 3-bit fixed-length code is also a prefix code.

→ 224,000 bits



we need → 300,000 bits



100 → e
101 → f
구별가능 → fixed-length와
10 → e
11 → f
로 알 수가 → full binary로 안됨 (11-는 필요가 없다)

- The left tree is a ***full binary tree*** *while the right one is not.*
  - Every node is either leaf or has two children
  - A full binary tree for alphabet C has $|C|$ leaves and $|C|$-1 internal nodes.

{a, b, c, d, e, f} 집합
6

Huffman codes

- **The cost of tree *T***
  - $f(c)$: frequency of a character c
  - $d_T(c)$: length of the codeword for c

$$B(T) = \sum_{c \in C} f(c) d_T(c)$$

  - An optimal code is represented by a full binary tree.

Huffman codes

- Huffman invented a greedy algorithm that constructs an optimal prefix code called an ***Huffman code***.

	a	b	c	d	e	f
Frequency (in thousands)	45	13	12	16	9	5

f : 5   e : 9   c : 12   b : 13   d : 16   a : 45

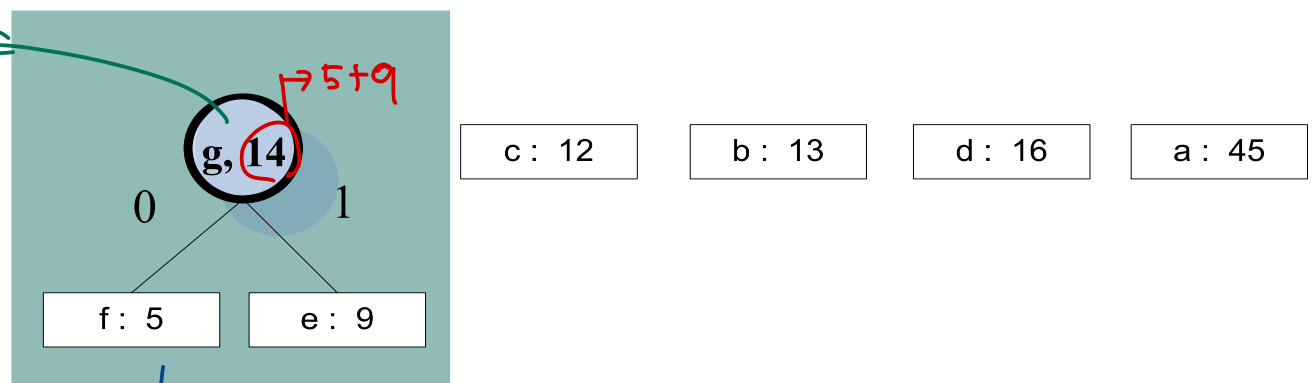
Huffman codes

Choose two least frequency character

f : 5	e : 9	c : 12	b : 13	d : 16	a : 45

최소 2개를 찾기 위해 $O(n)$

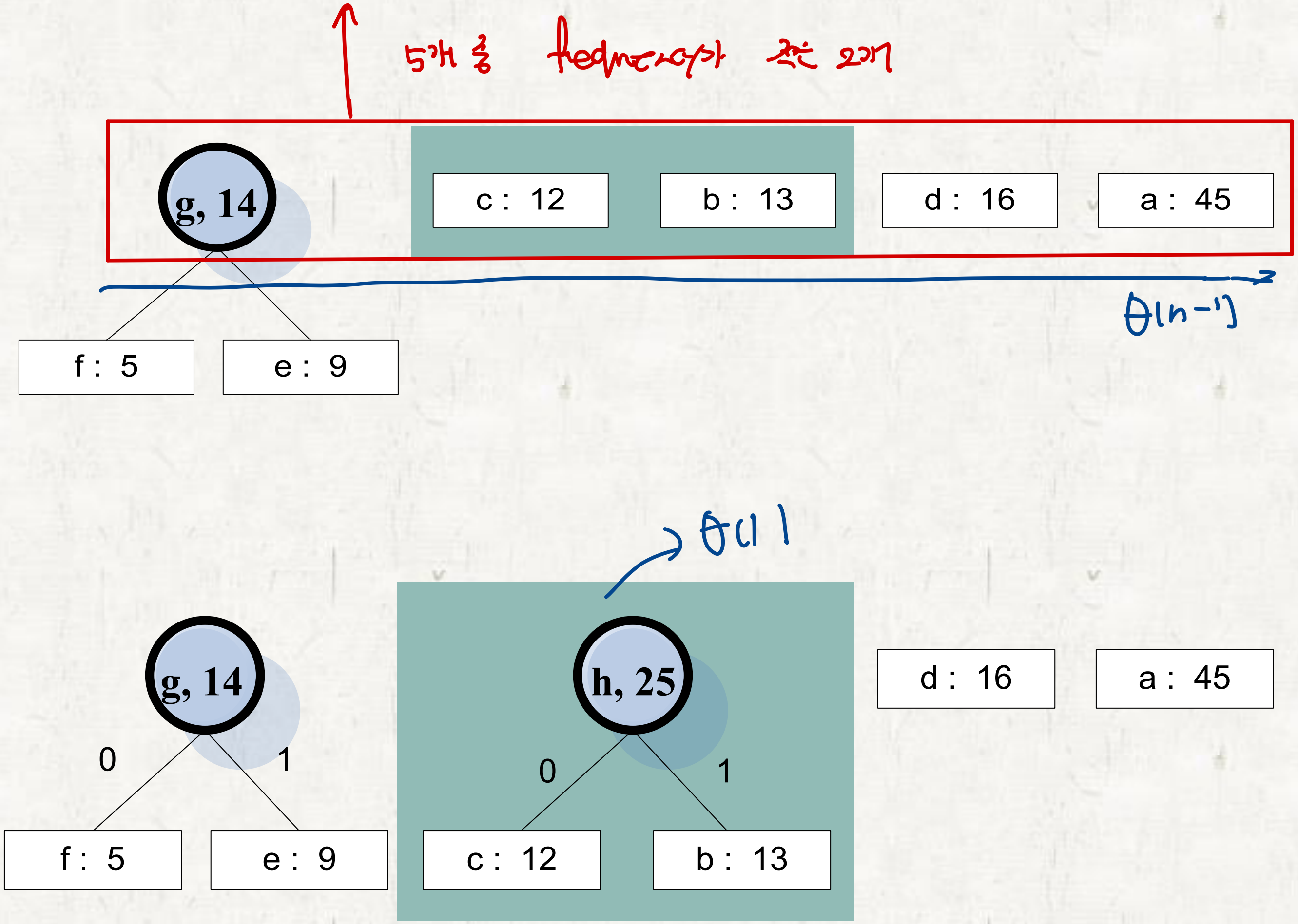
virtual node



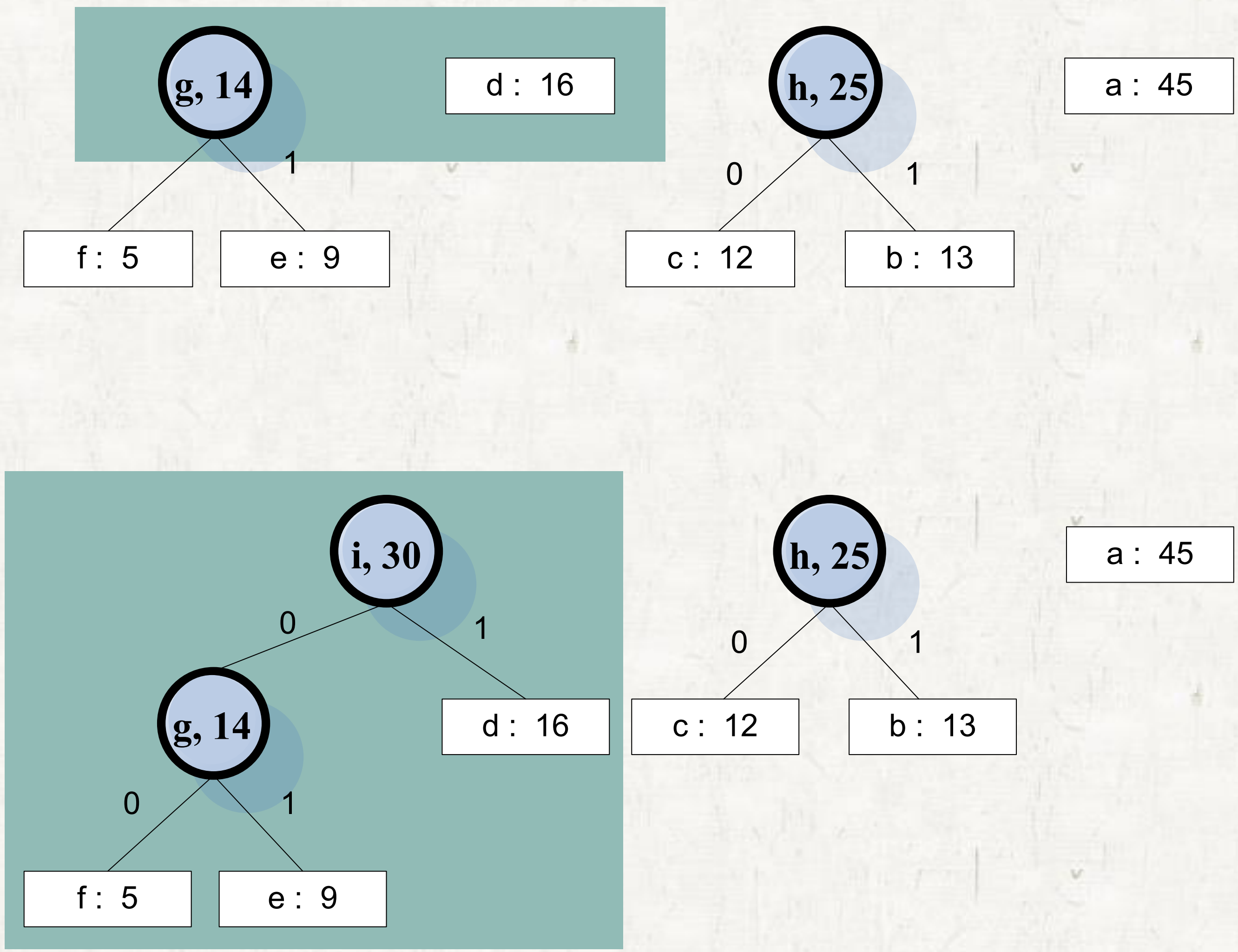
build sub tree $O(1)$

Huffman codes

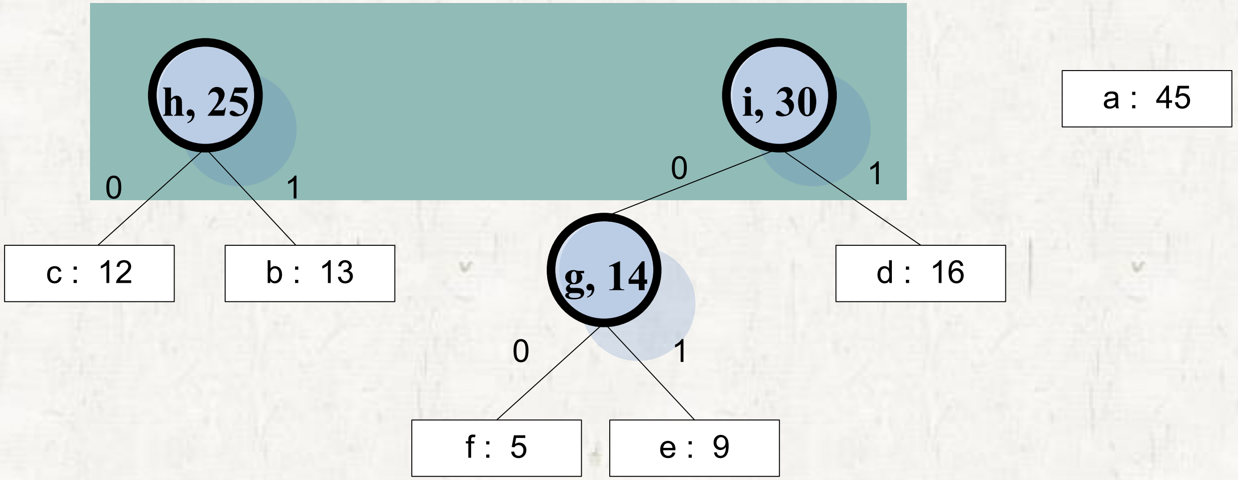
앞장에서 한 줄 반복

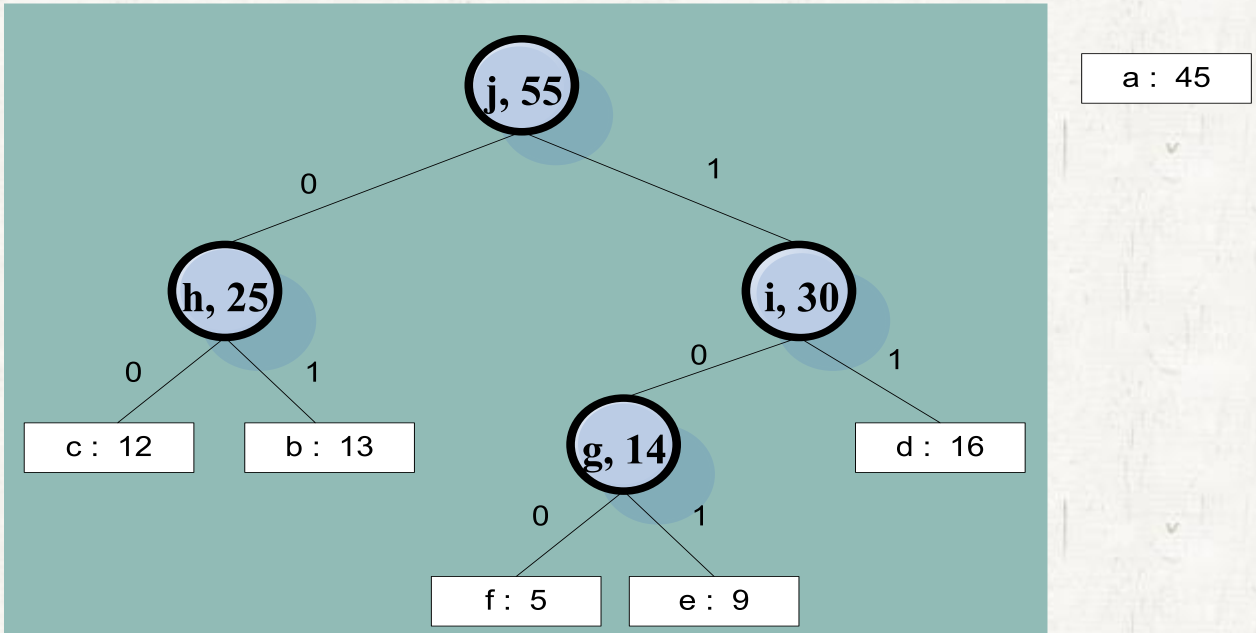


Huffman codes

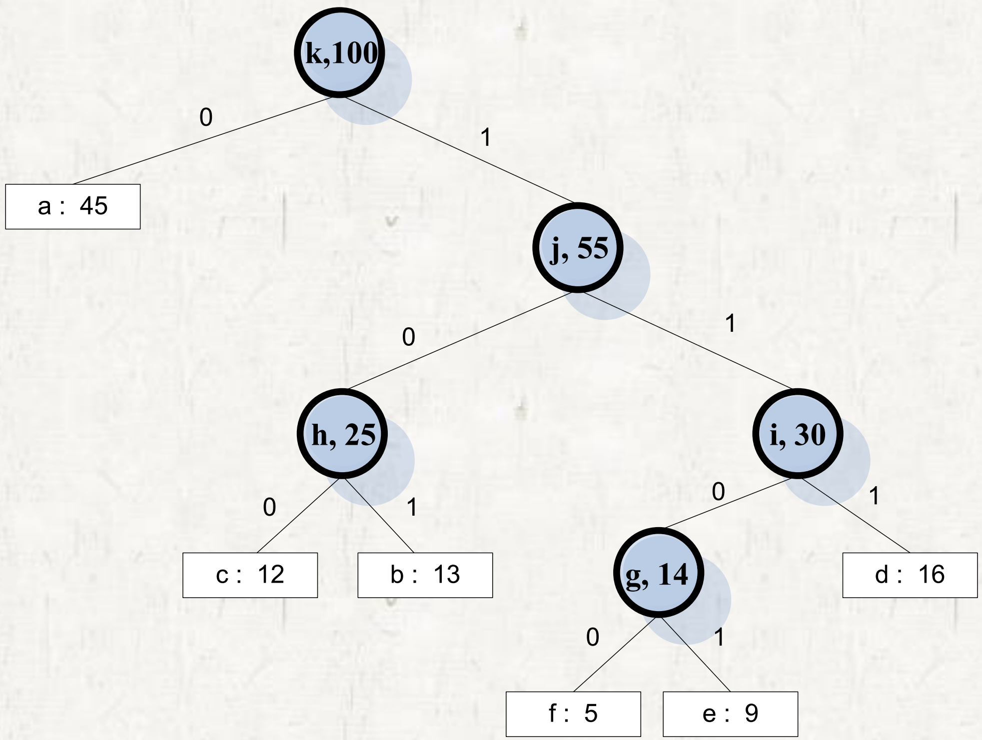


Huffman codes





Huffman codes



$\Theta(n) + \Theta(n-1) + \cdots + \Theta(1)$

$= \Theta(n^2)$

Huffman codes

f : 5	e : 9	c : 12	b : 13	d : 16	a : 45

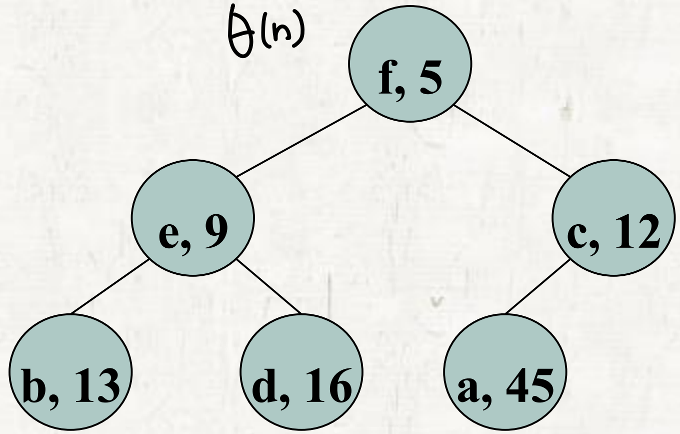
- Min Heap

↳ extract two least node

$\Theta(\log n)$

$\Theta(n\log n)$

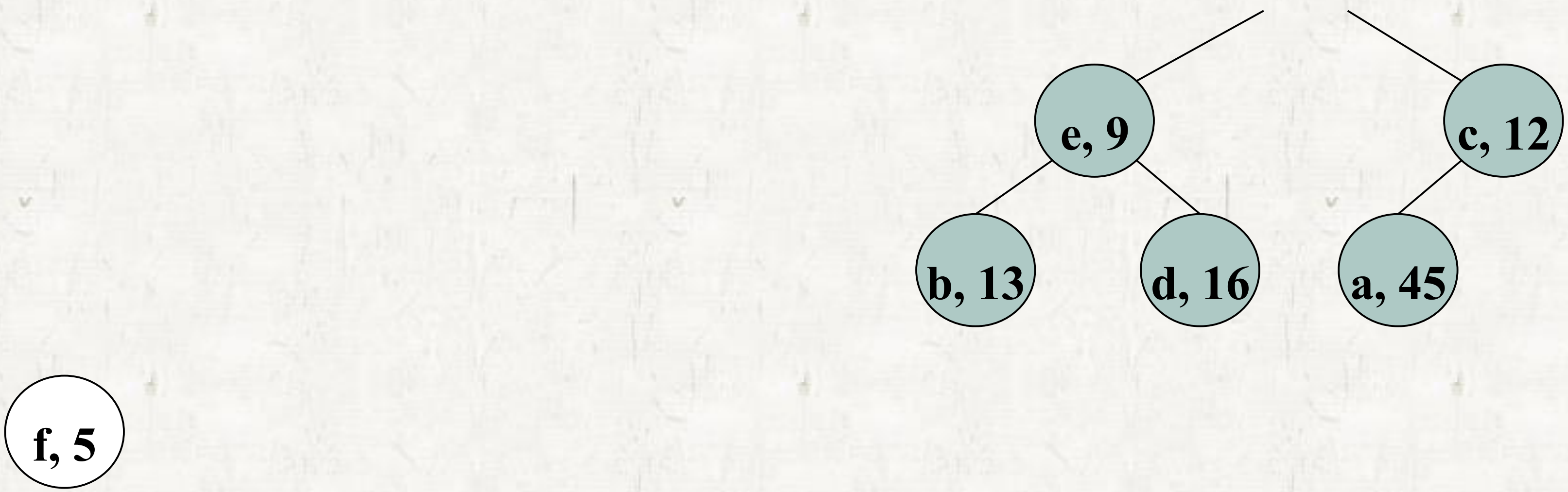
$\Theta(n)$



Huffman codes

f : 5	e : 9	c : 12	b : 13	d : 16	a : 45

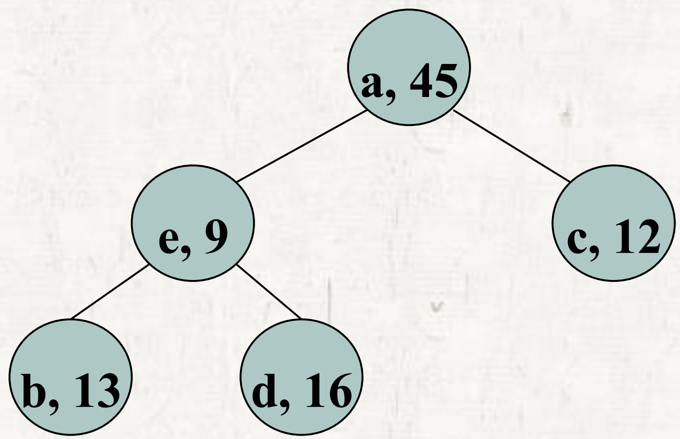
- Min Heap



Huffman codes

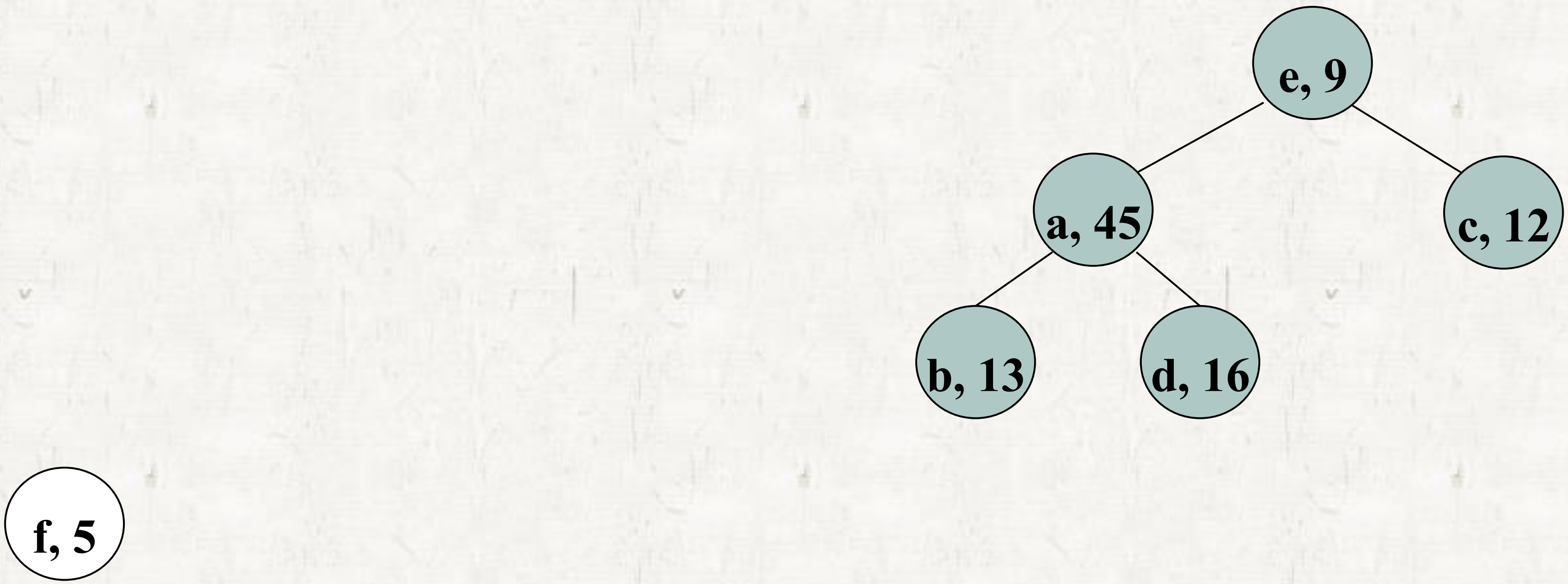
f : 5	e : 9	c : 12	b : 13	d : 16	a : 45

- Min Heap



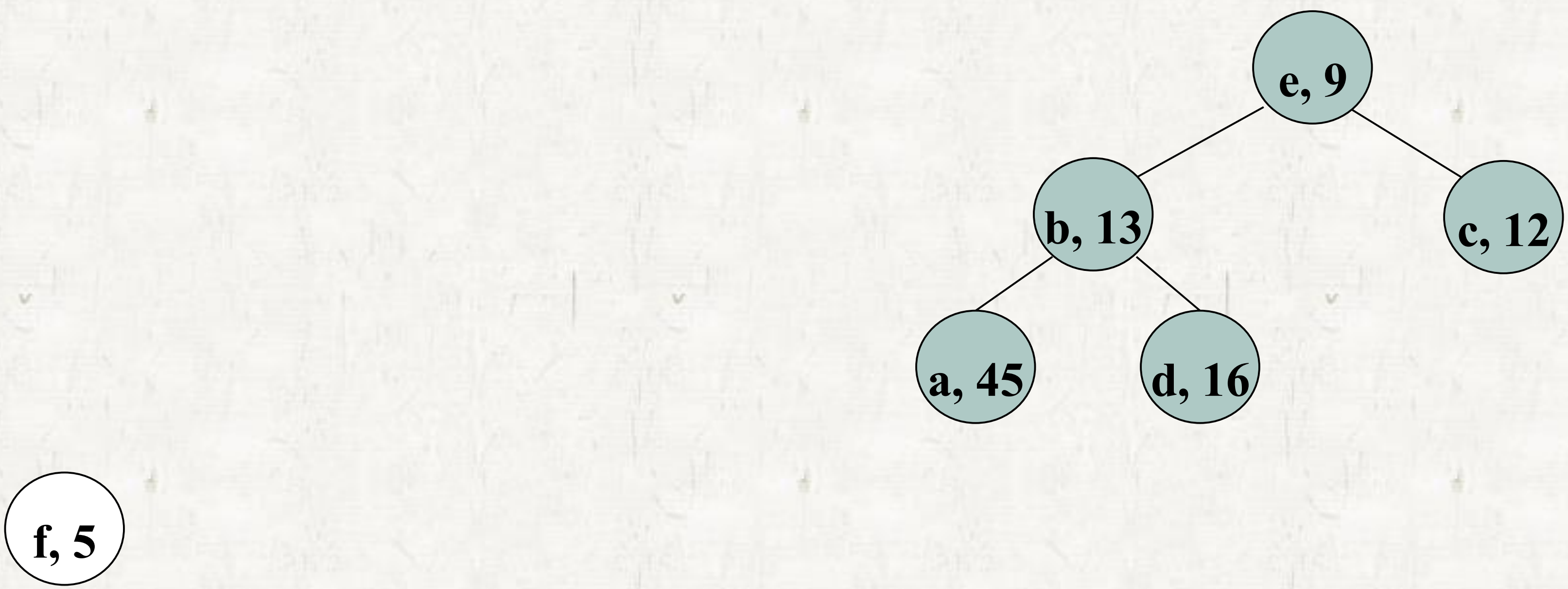
Huffman codes

- Min Heap



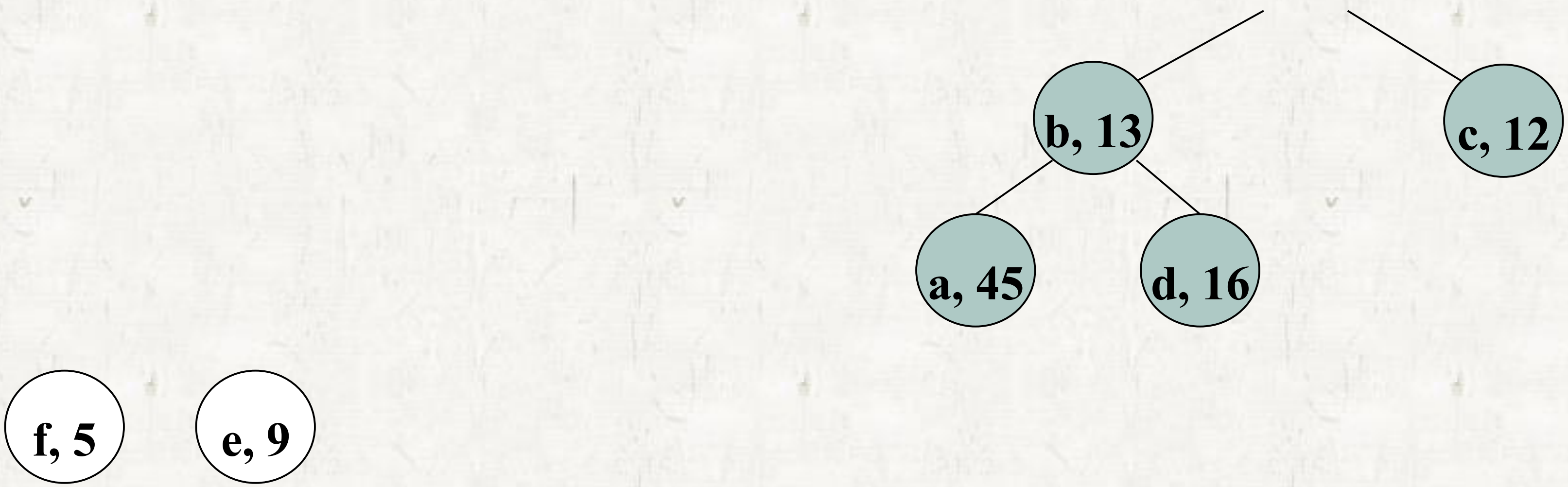
Huffman codes

- Min Heap



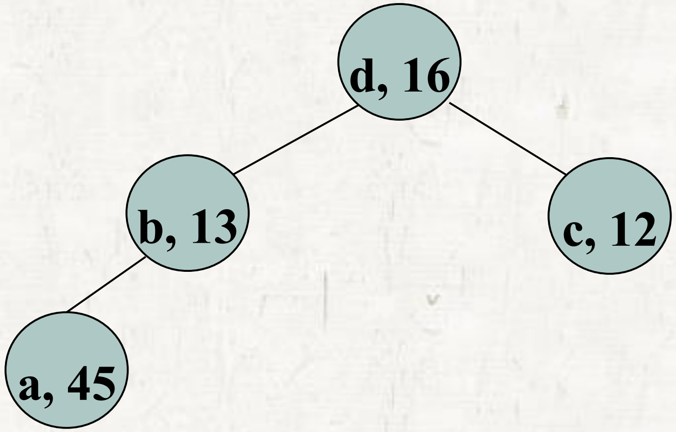
Huffman codes

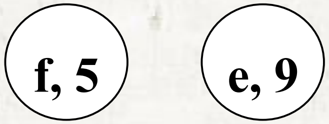
- Min Heap



Huffman codes

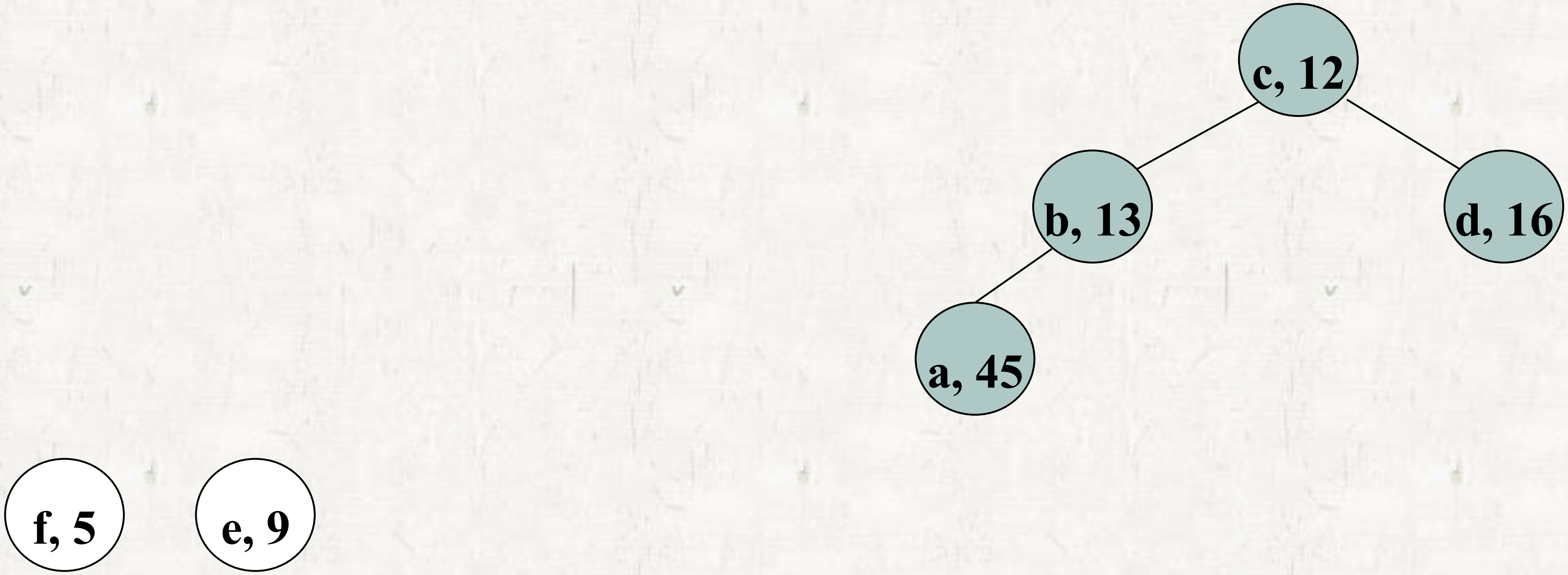
- Min Heap





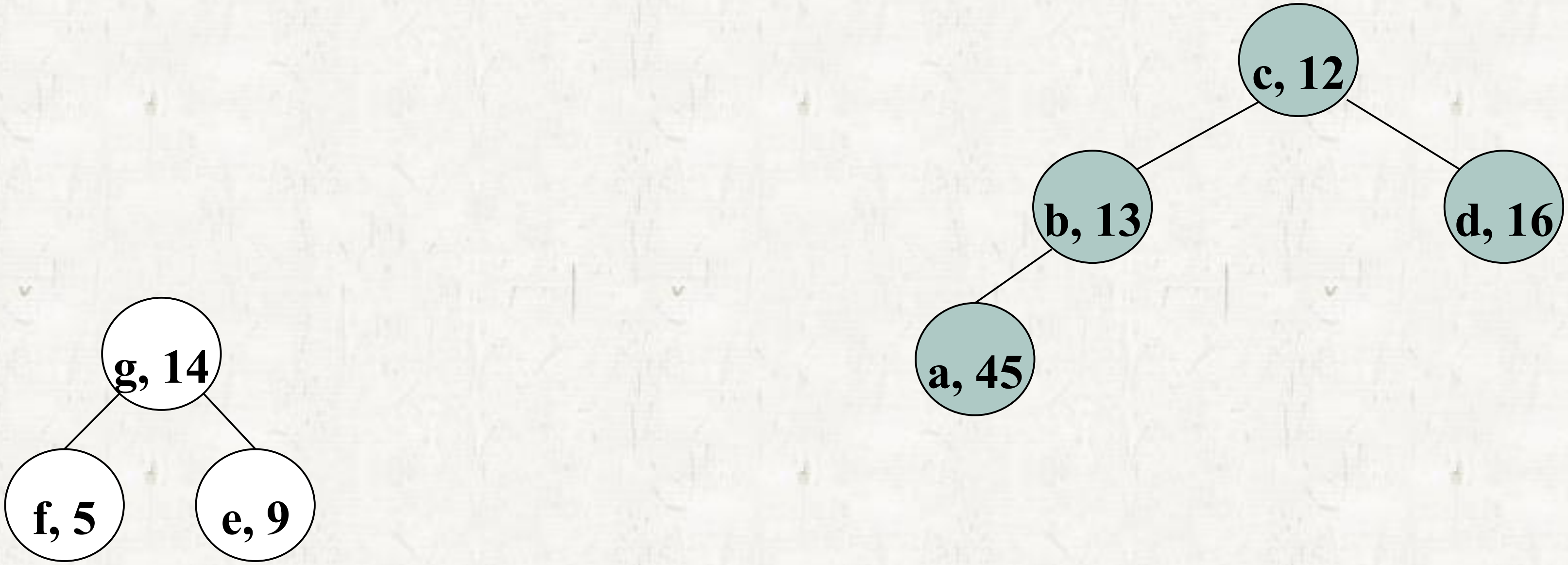
Huffman codes

- Min Heap



Huffman codes

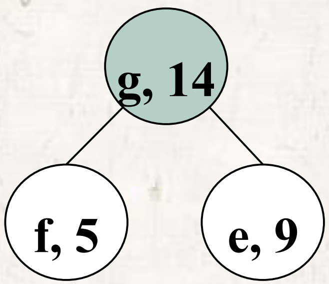
- Min Heap

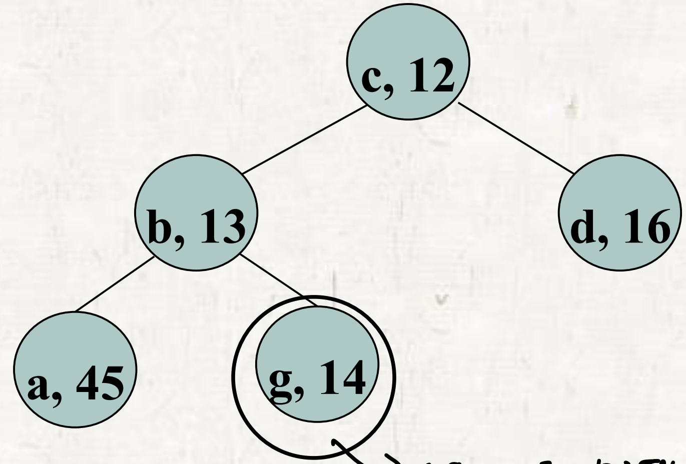


Huffman codes

- Min Heap

: 2개를 min Heap으로 추출하고
2개를 결합한 것을 삽입하는데) $\Theta(\log n)$

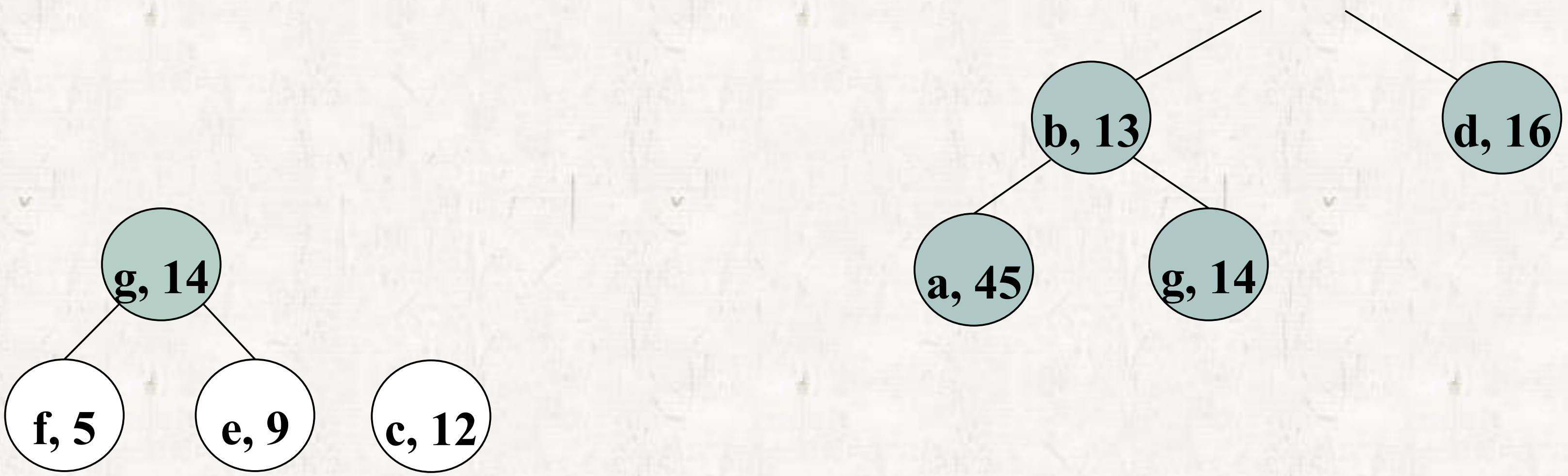




(g,14)를 넣는데
$\Theta(\log n)$

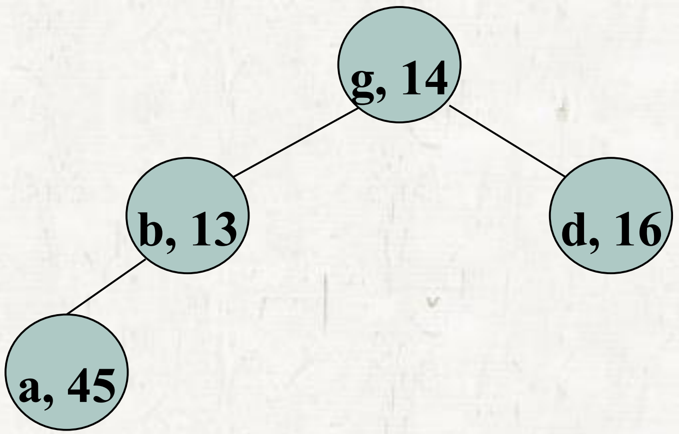
Huffman codes

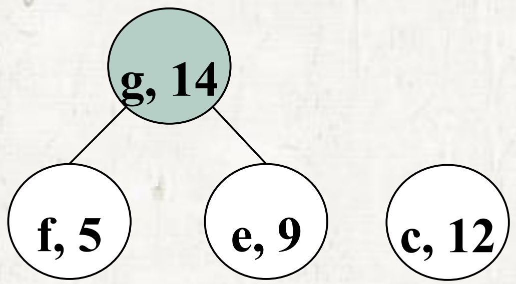
- Min Heap



Huffman codes

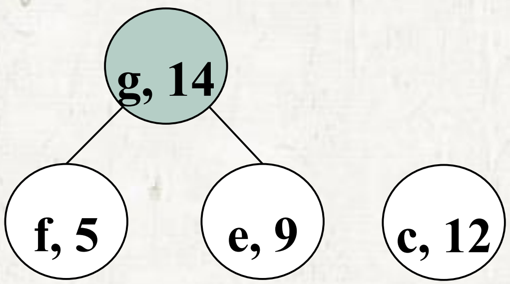
- Min Heap

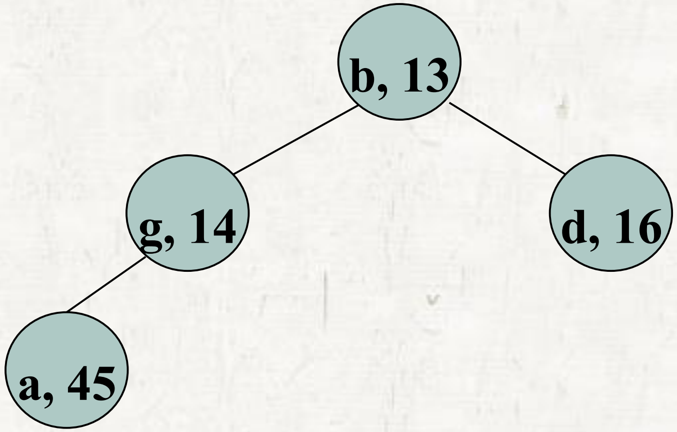




Huffman codes

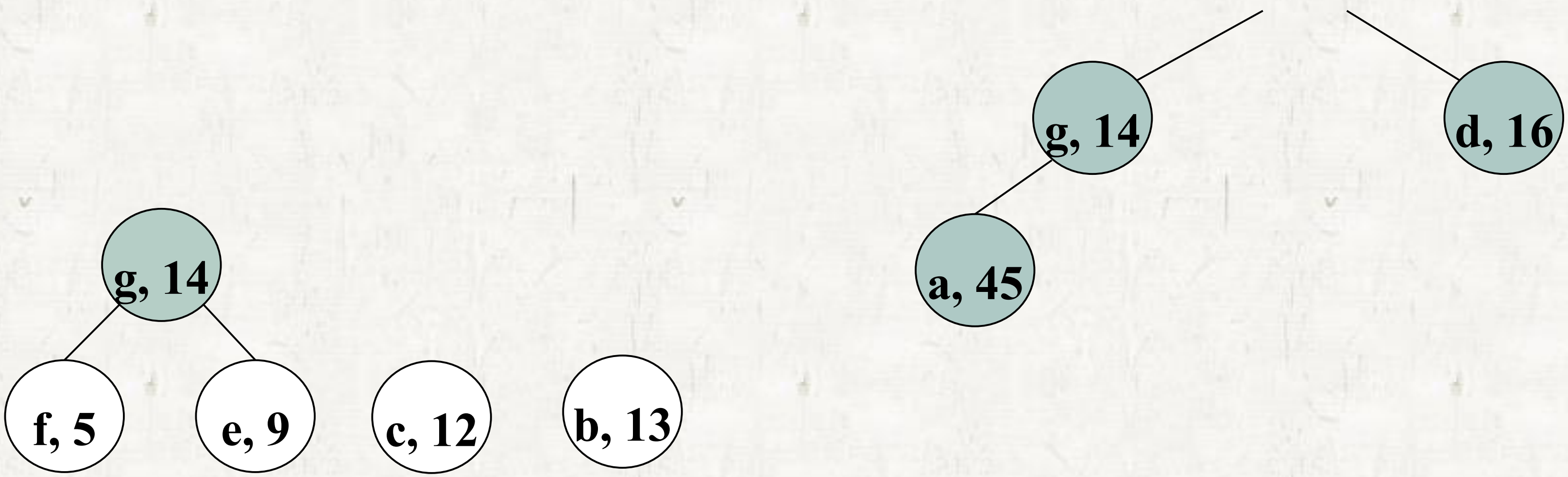
- Min Heap





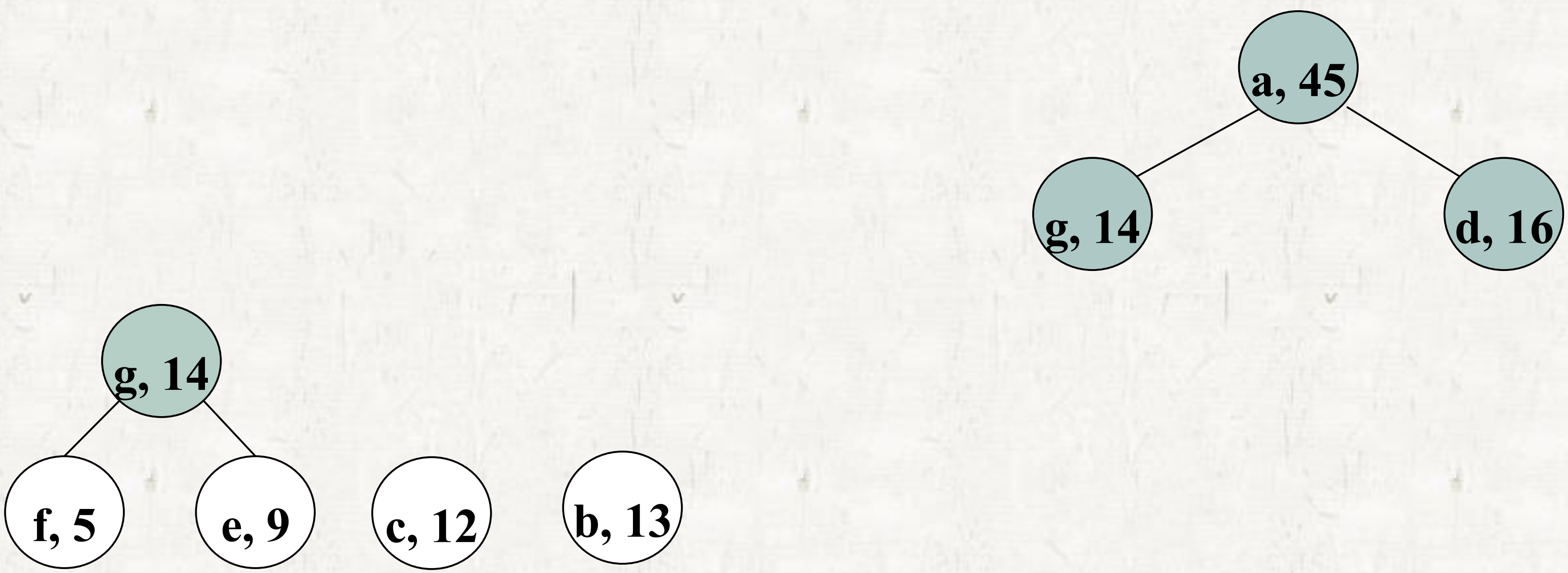
Huffman codes

- Min Heap



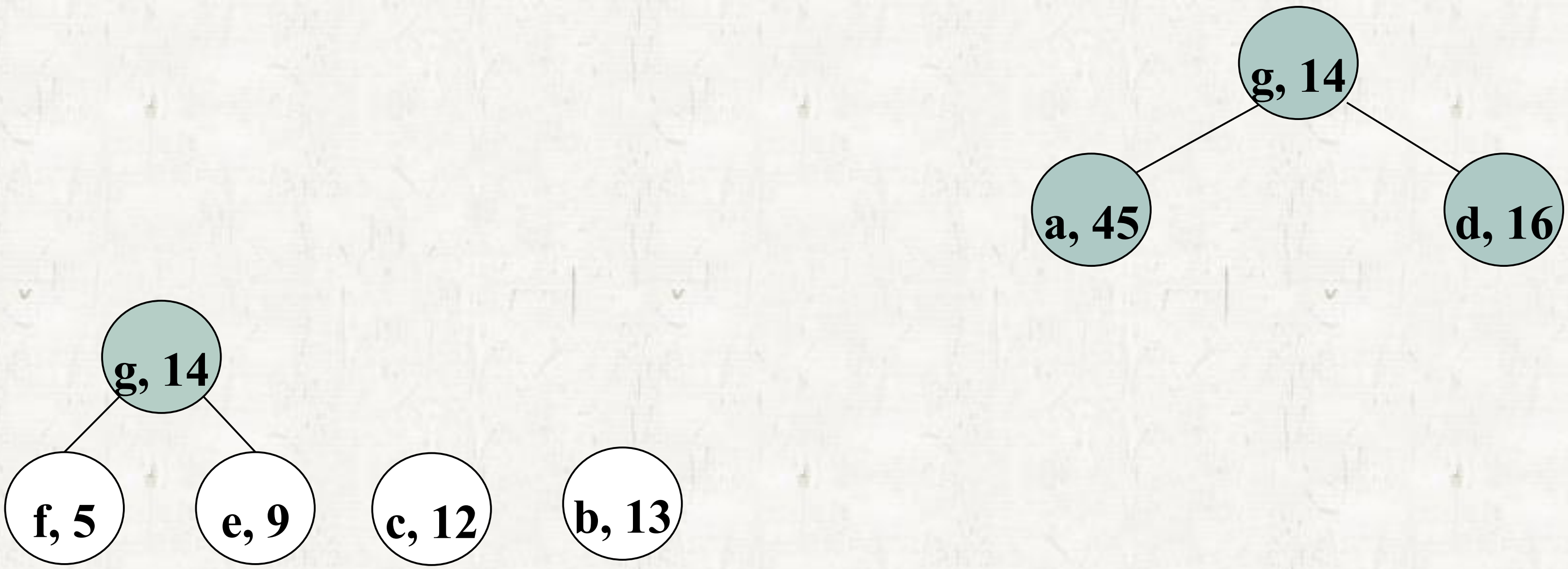
Huffman codes

- Min Heap



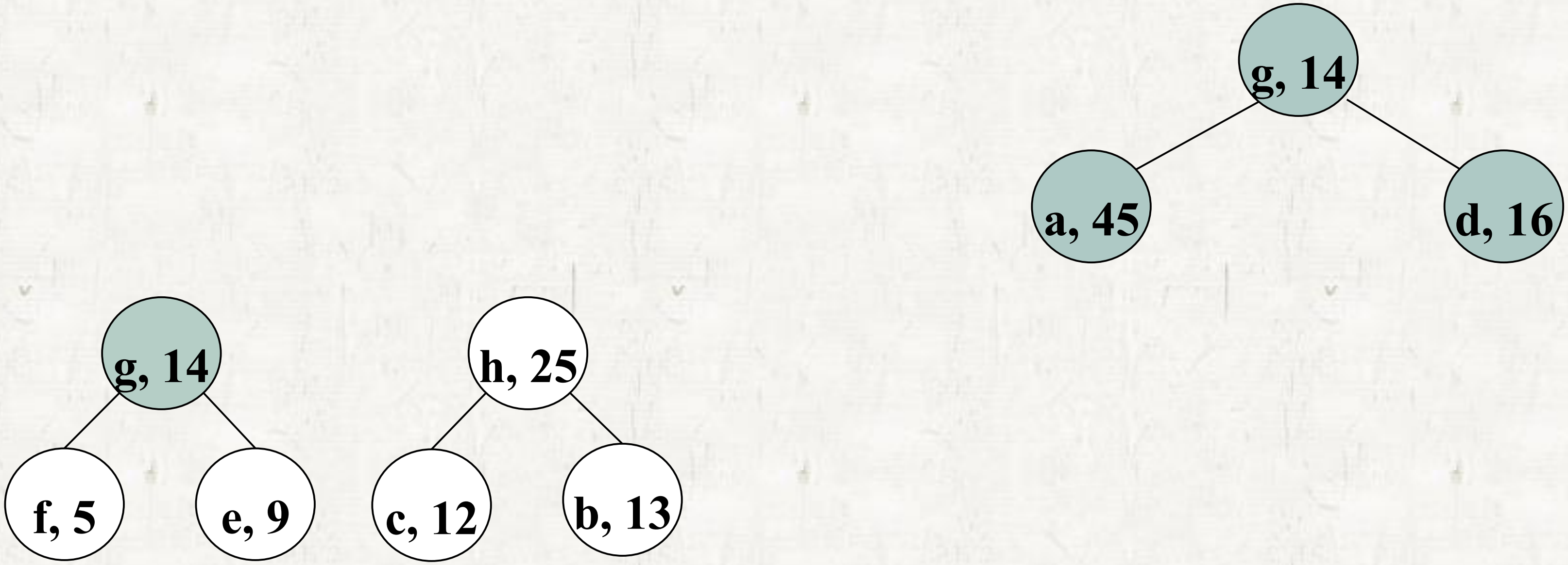
Huffman codes

- Min Heap



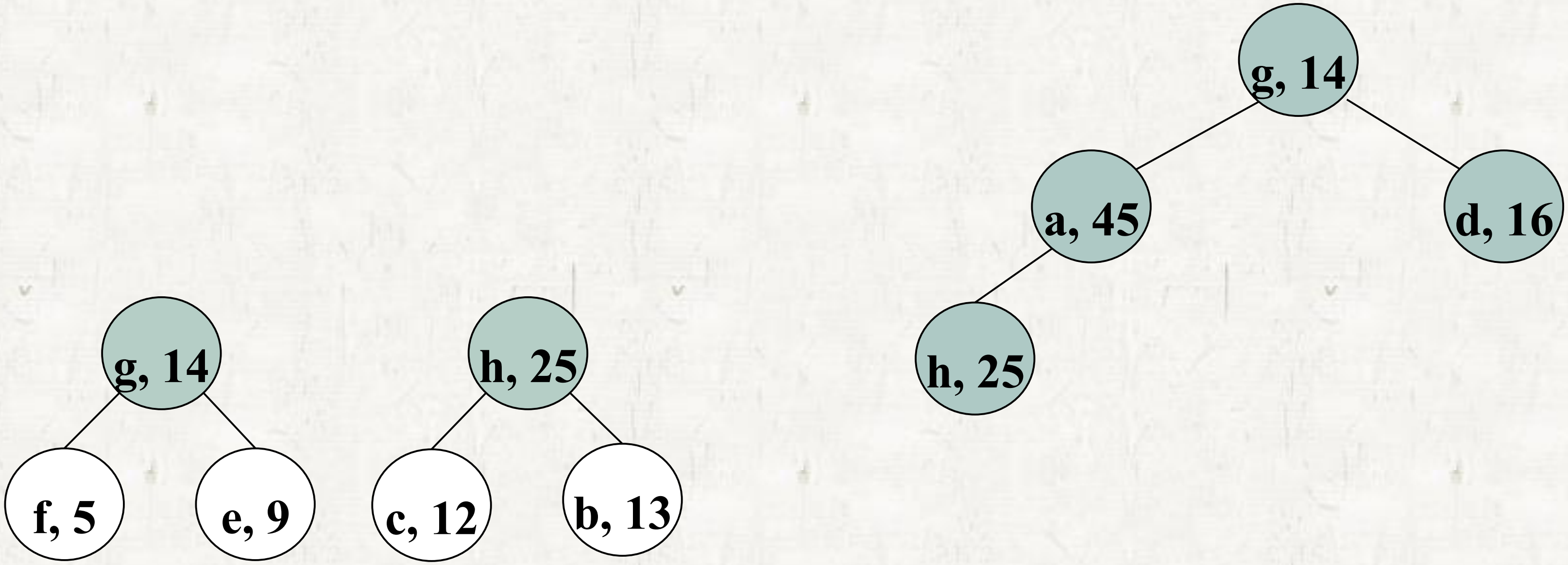
Huffman codes

- Min Heap



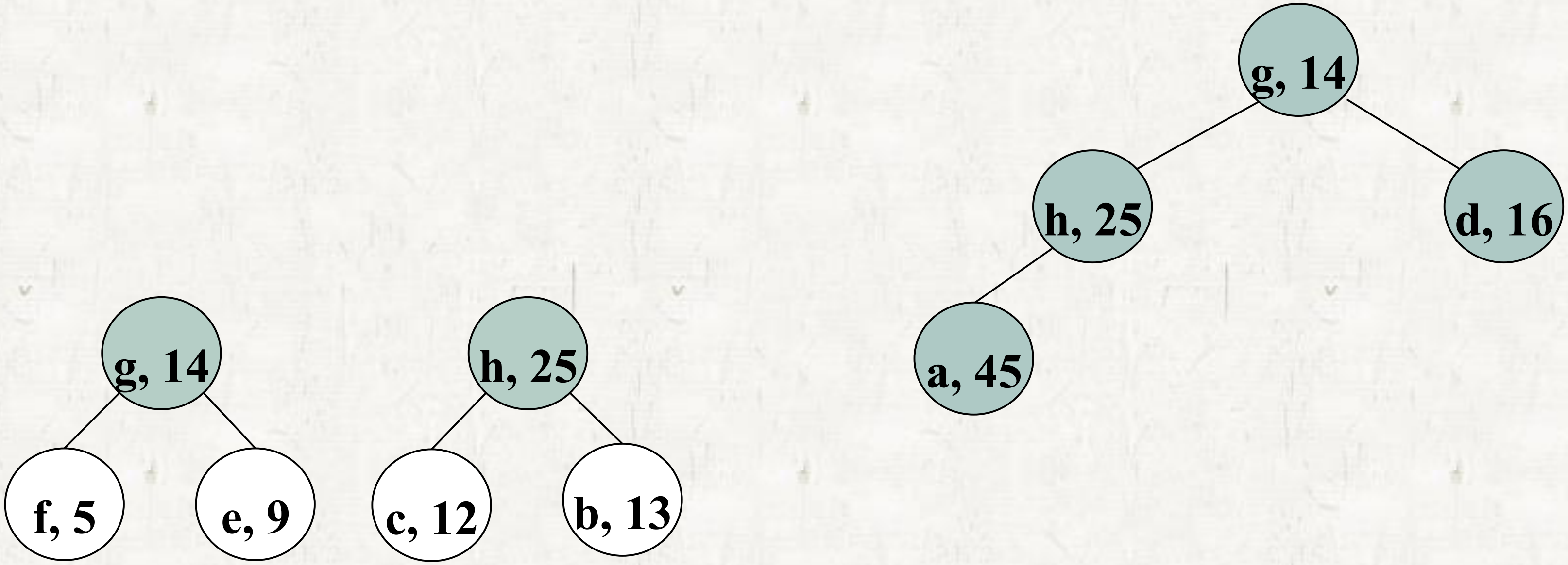
Huffman codes

- Min Heap



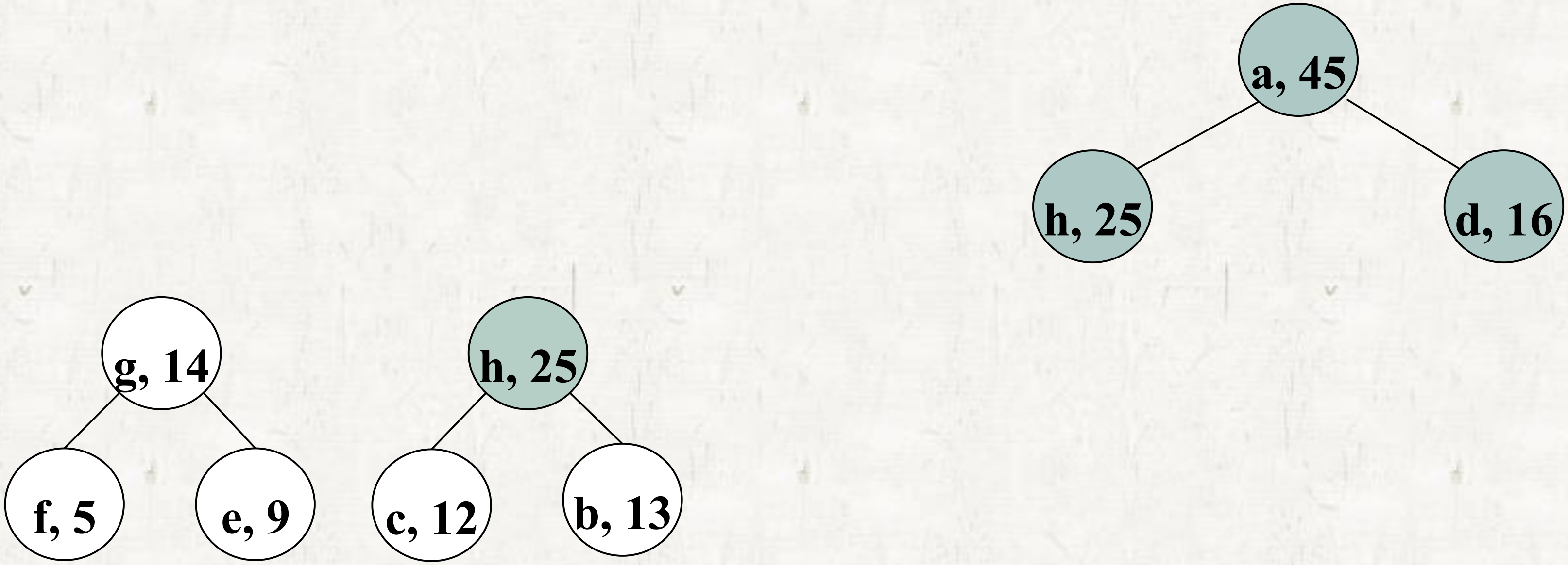
Huffman codes

- Min Heap



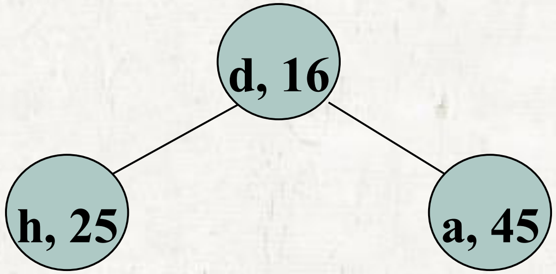
Huffman codes

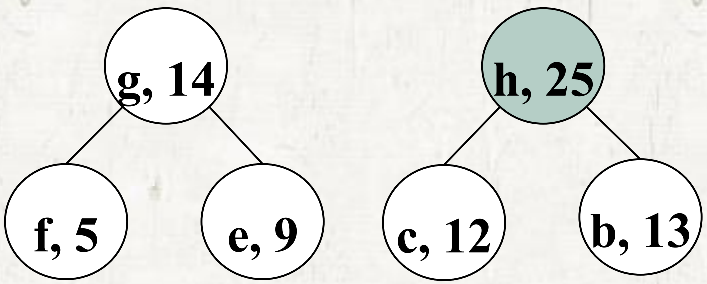
- Min Heap



Huffman codes

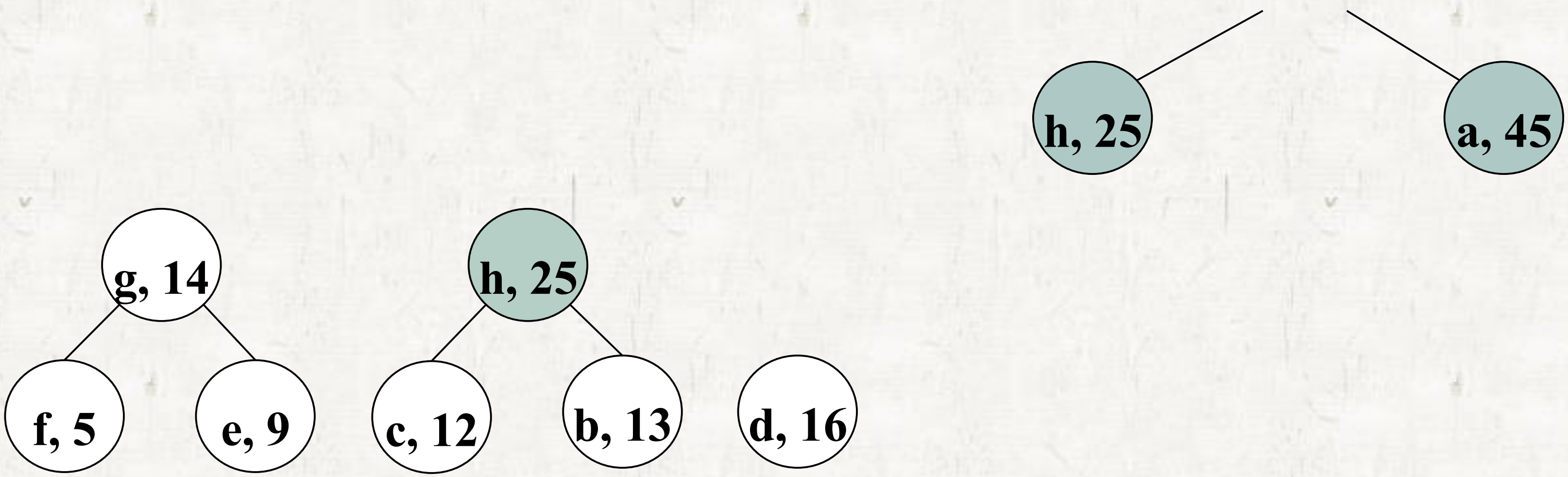
- Min Heap





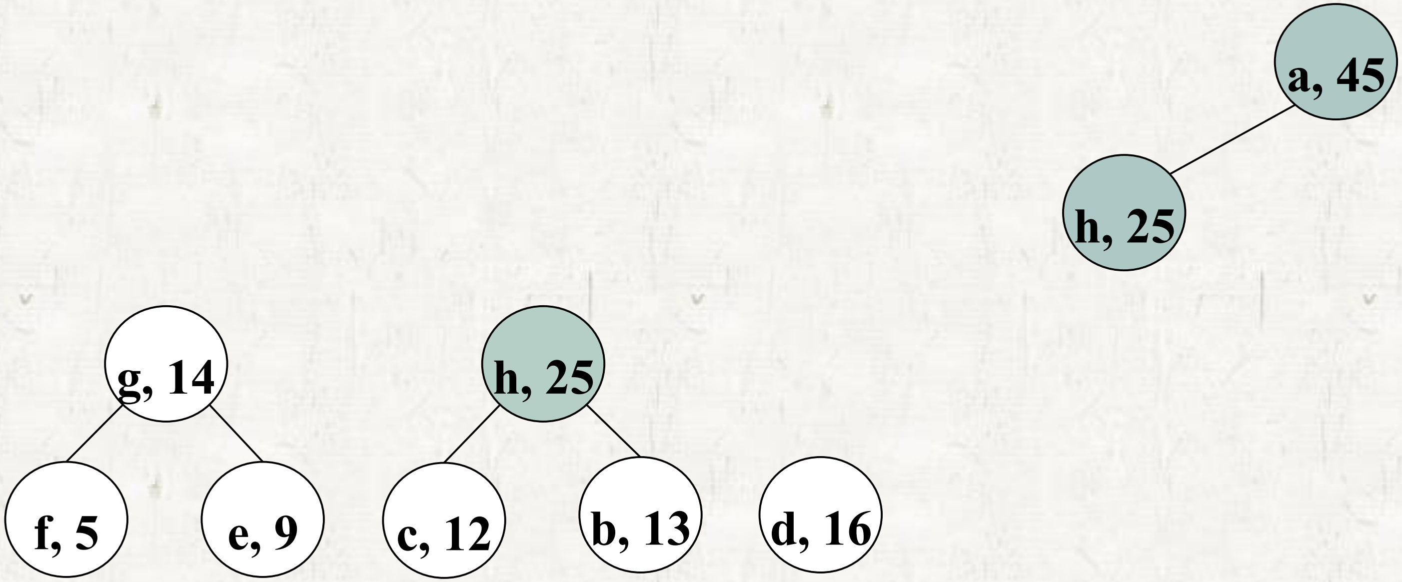
Huffman codes

- Min Heap



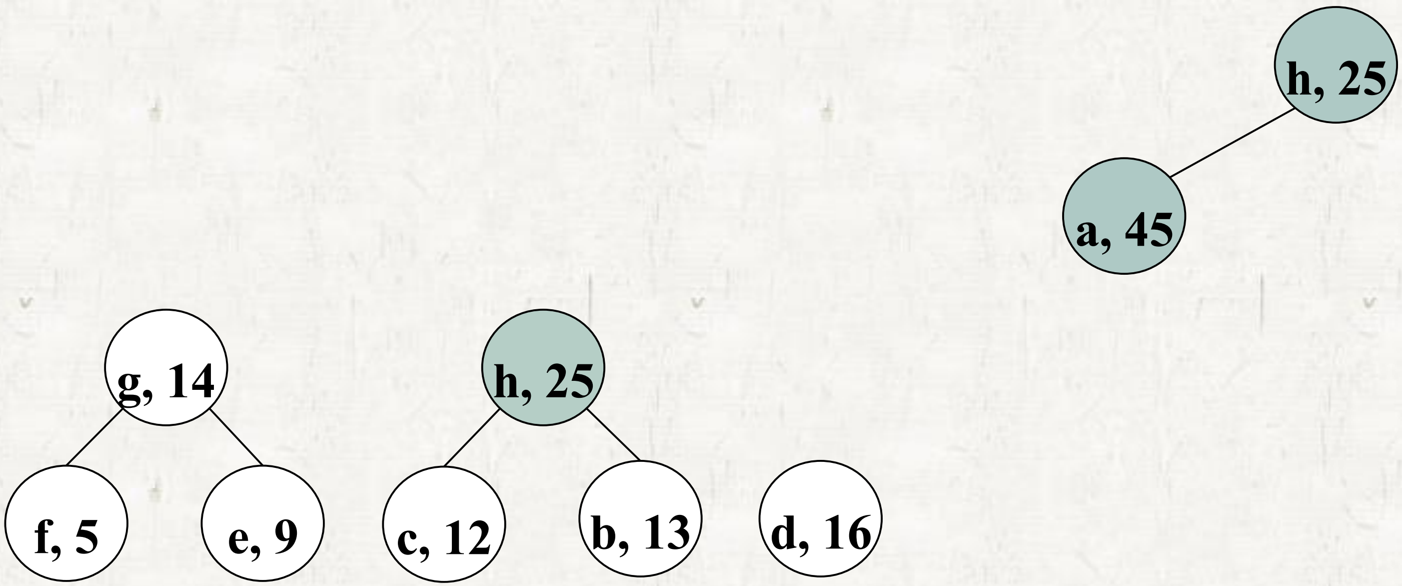
Huffman codes

- Min Heap



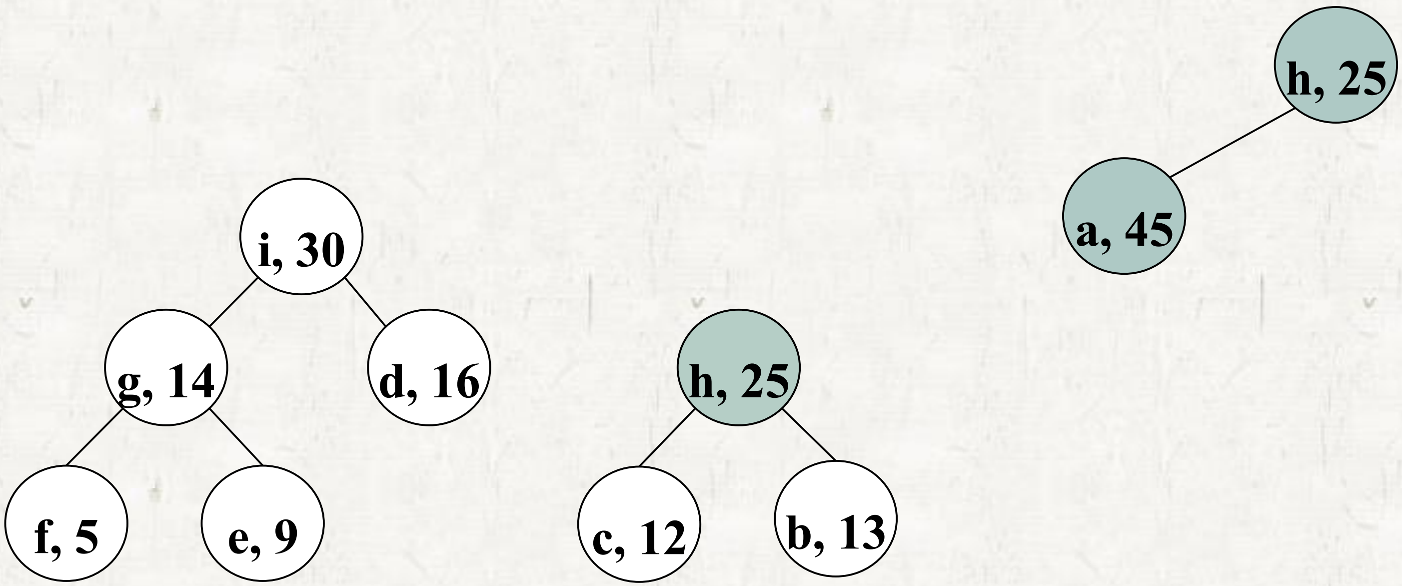
Huffman codes

- Min Heap



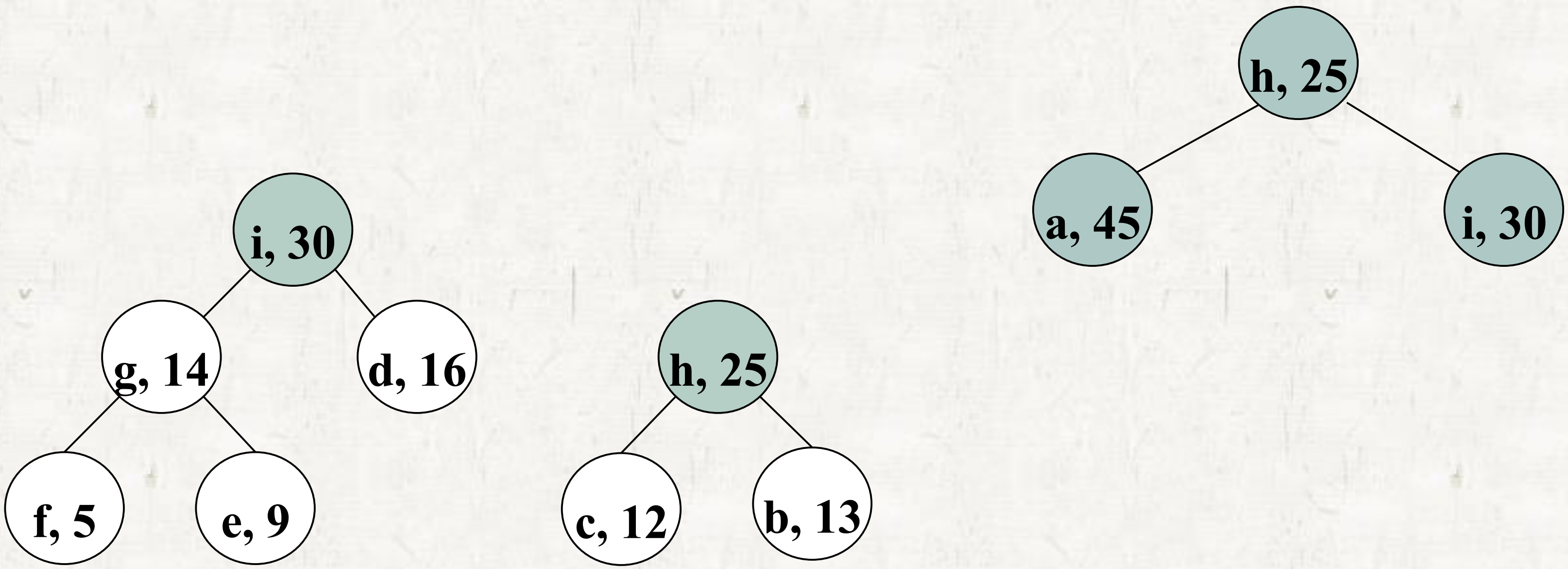
Huffman codes

- Min Heap



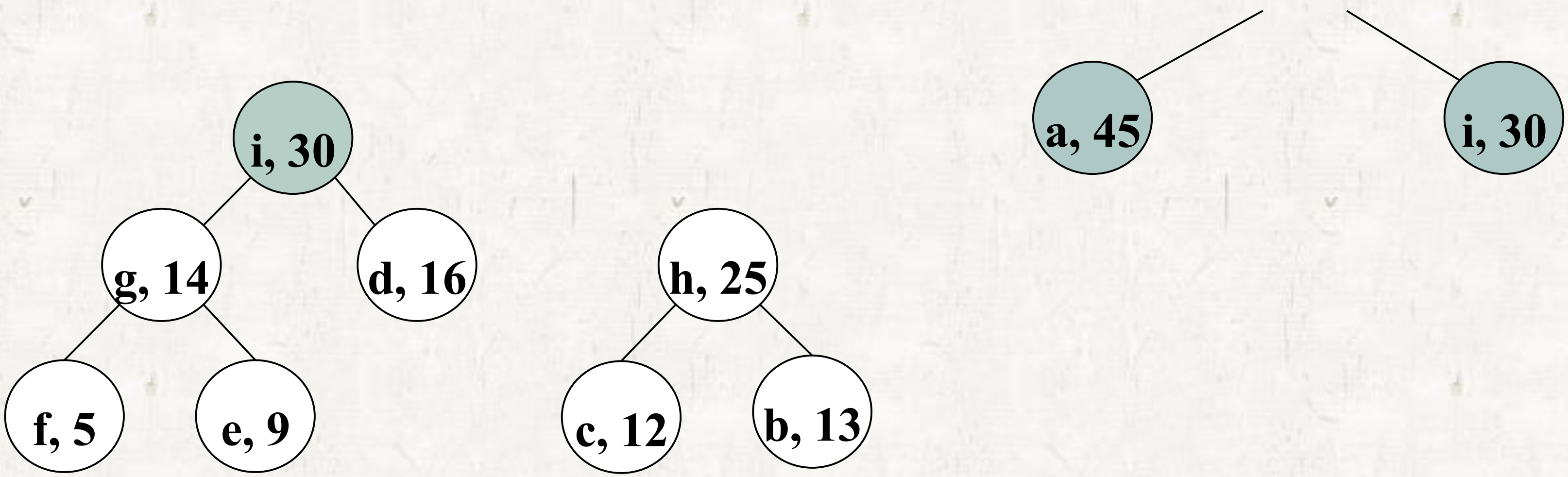
Huffman codes

- Min Heap



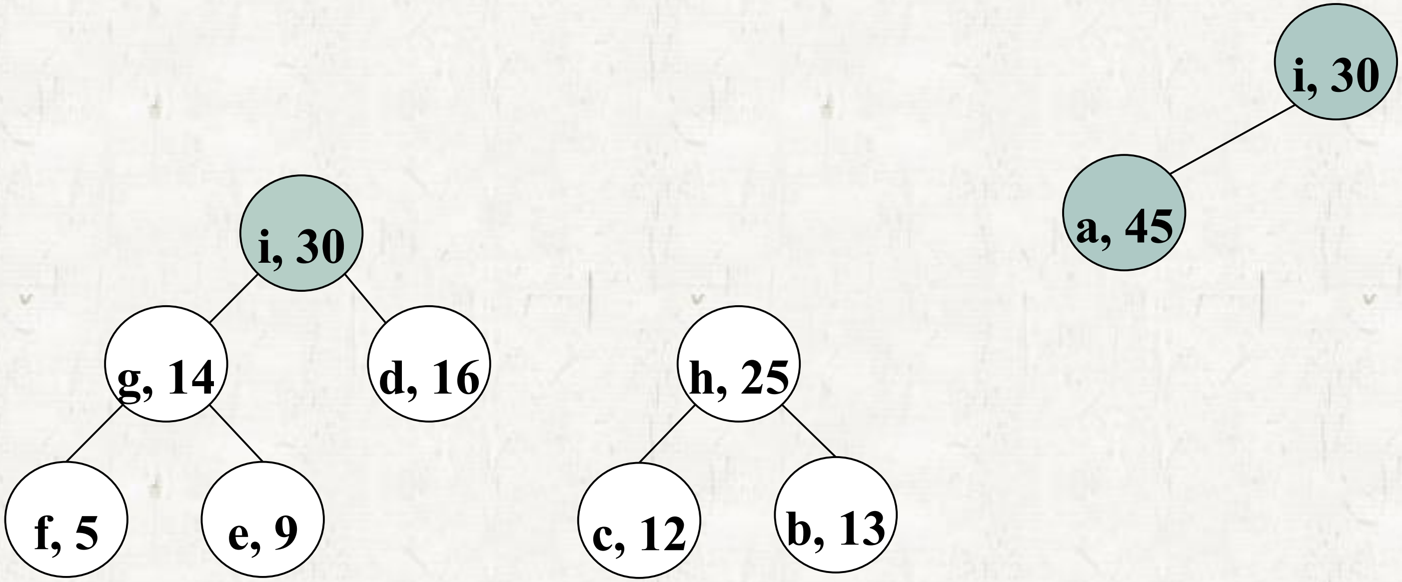
Huffman codes

- Min Heap



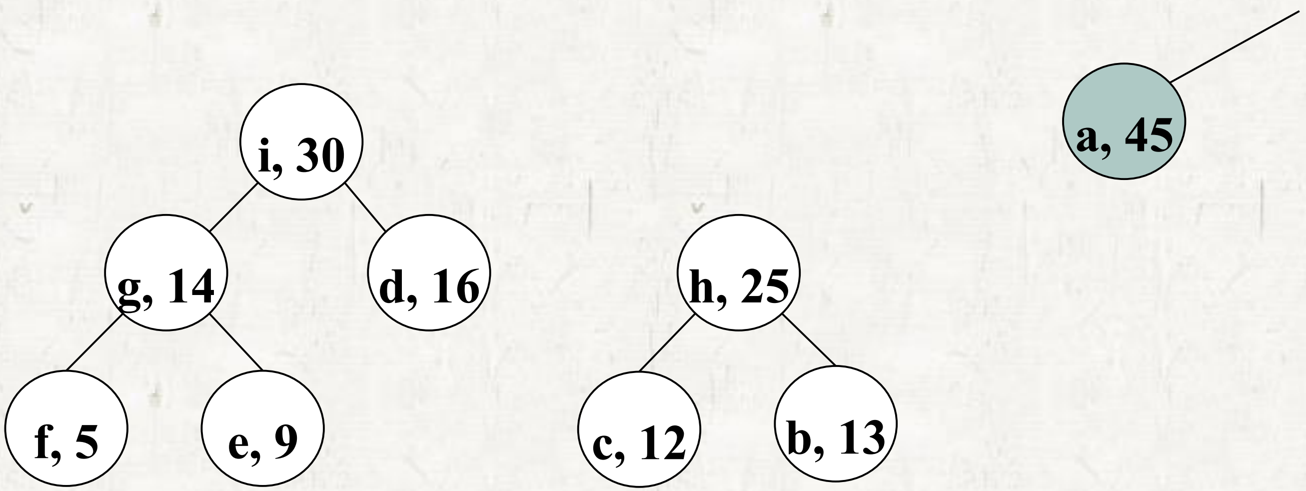
Huffman codes

- Min Heap



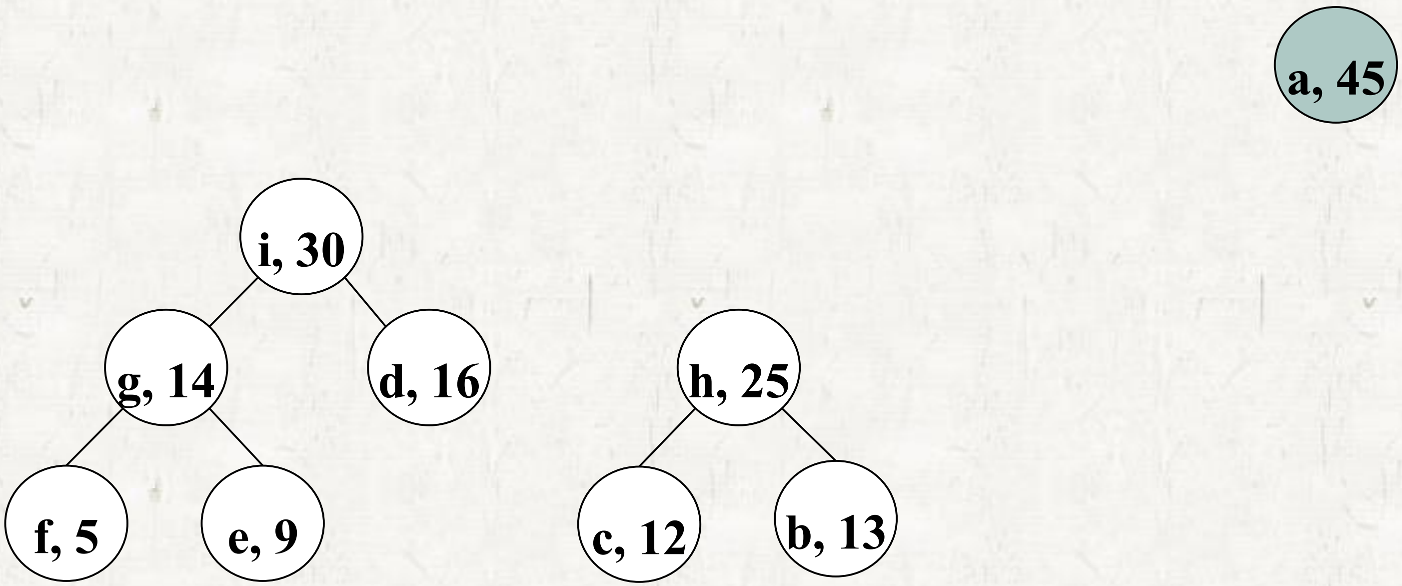
Huffman codes

- Min Heap



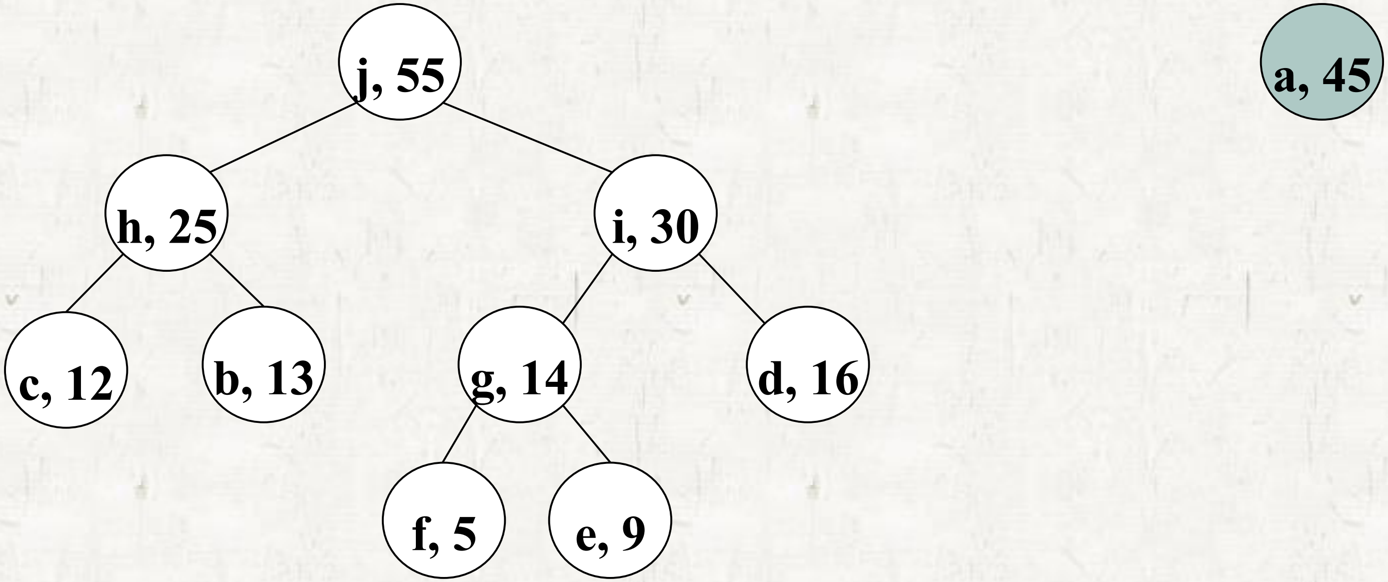
Huffman codes

- Min Heap



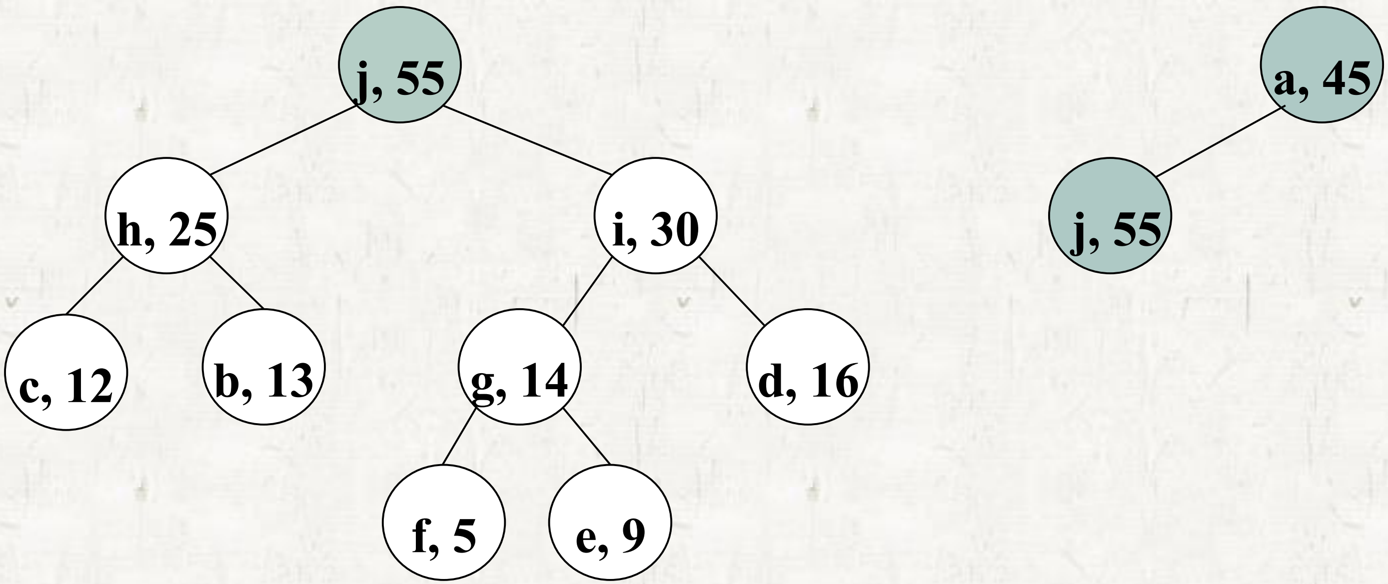
Huffman codes

- Min Heap



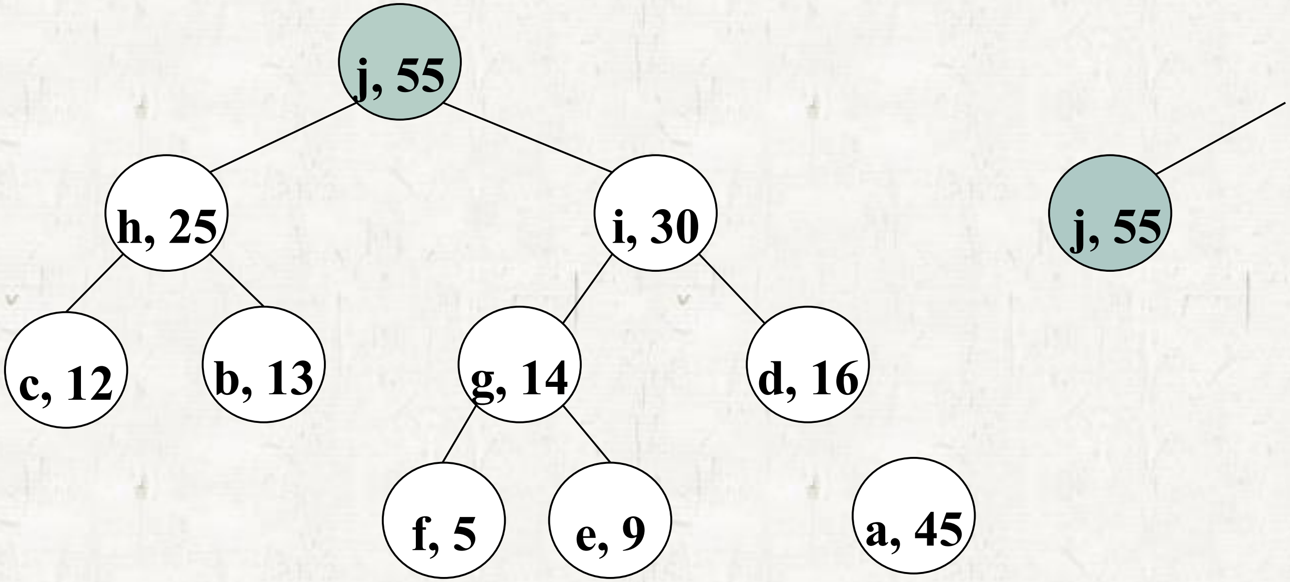
Huffman codes

- Min Heap



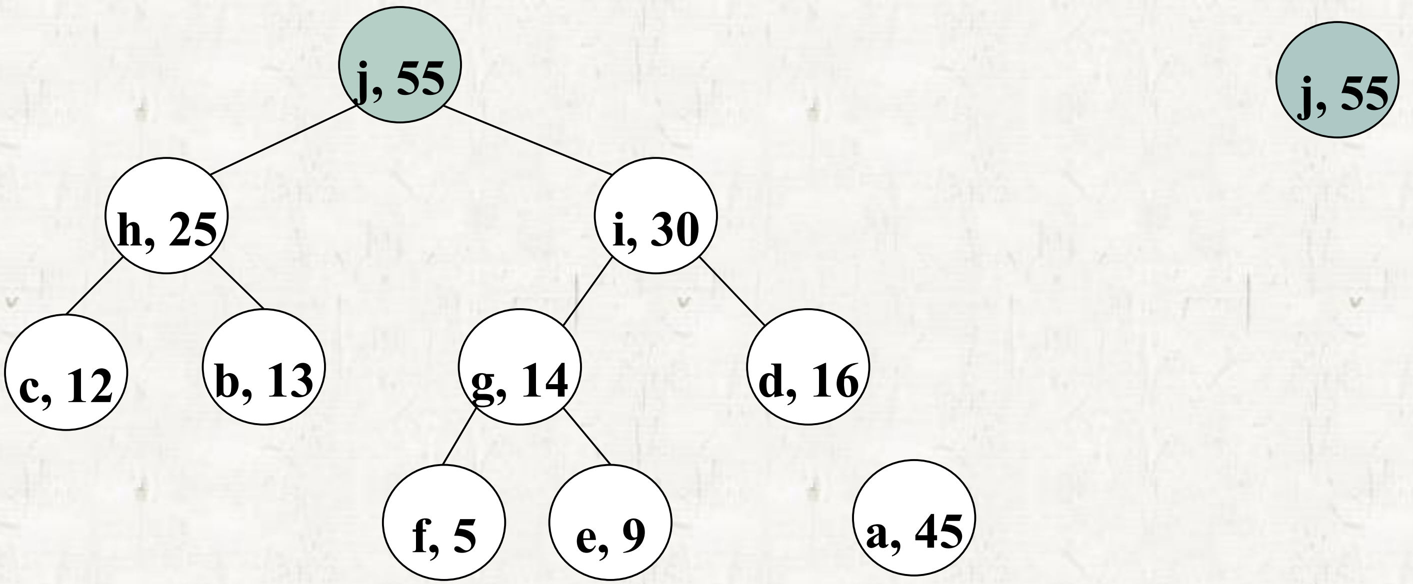
Huffman codes

- Min Heap



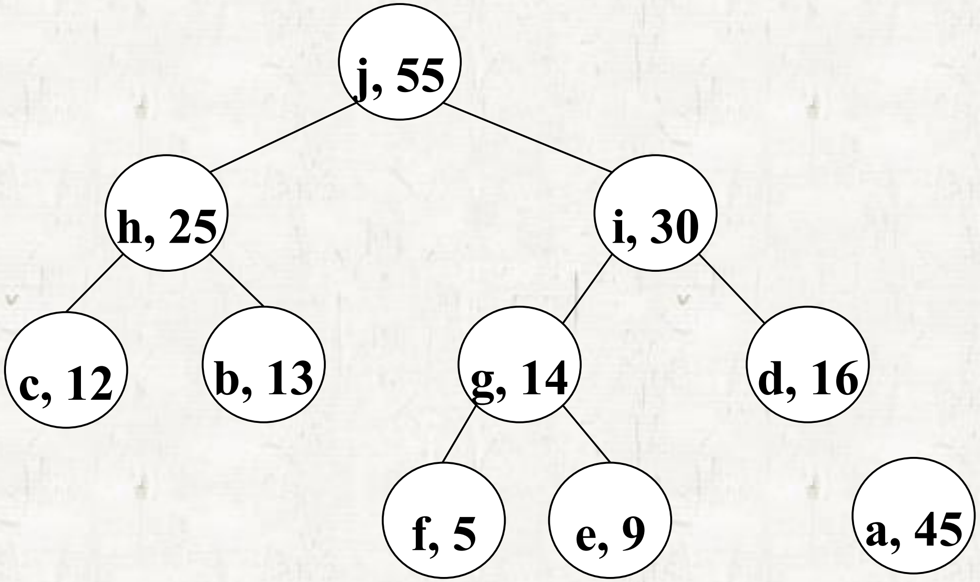
Huffman codes

- Min Heap



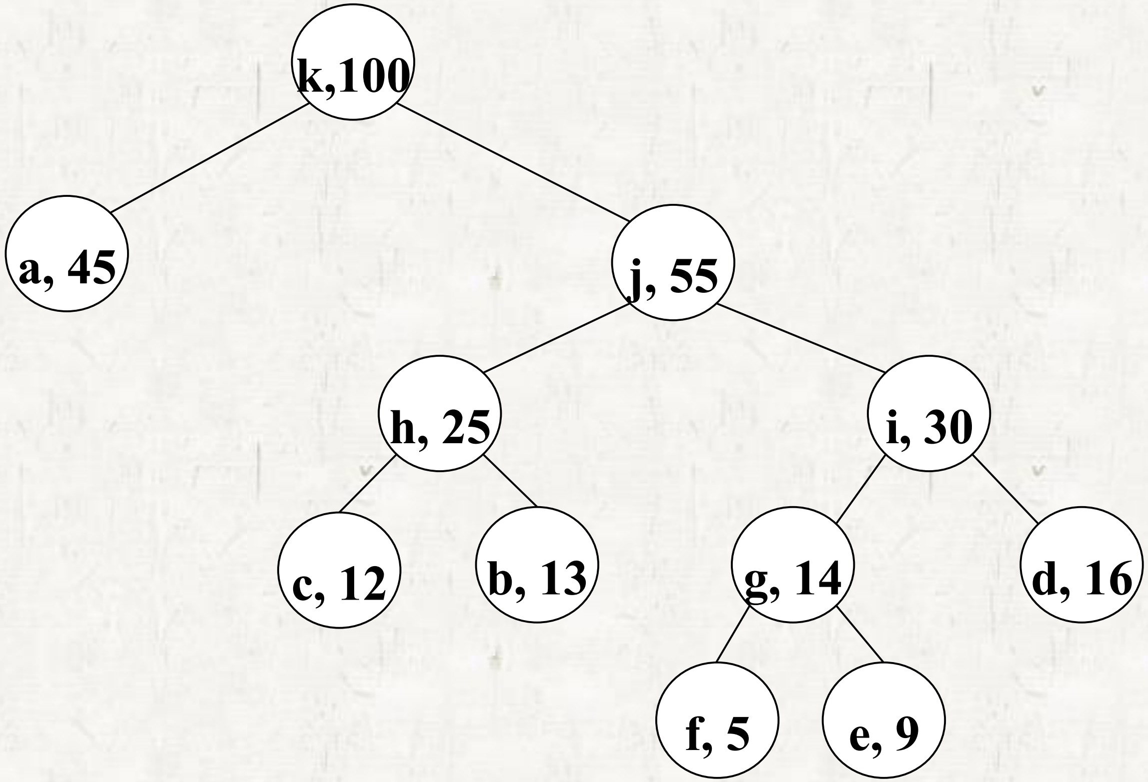
Huffman codes

- Min Heap



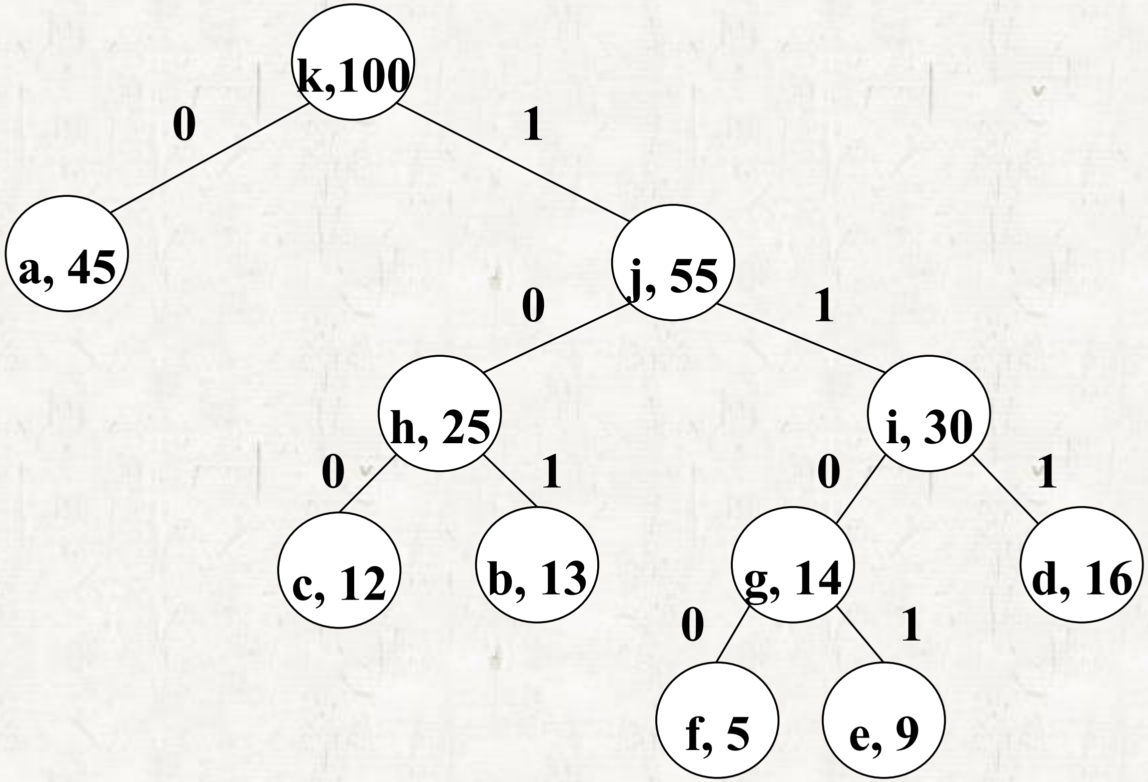
Huffman codes

- Min Heap



Huffman codes

- Min Heap



Huffman codes

- **Running time**: O($n$lg$n$)
  - Build min heap: $O(n)$
  - Merge: n-1 times
    - Each merge requires two minimum selection: $O(\lg n)$

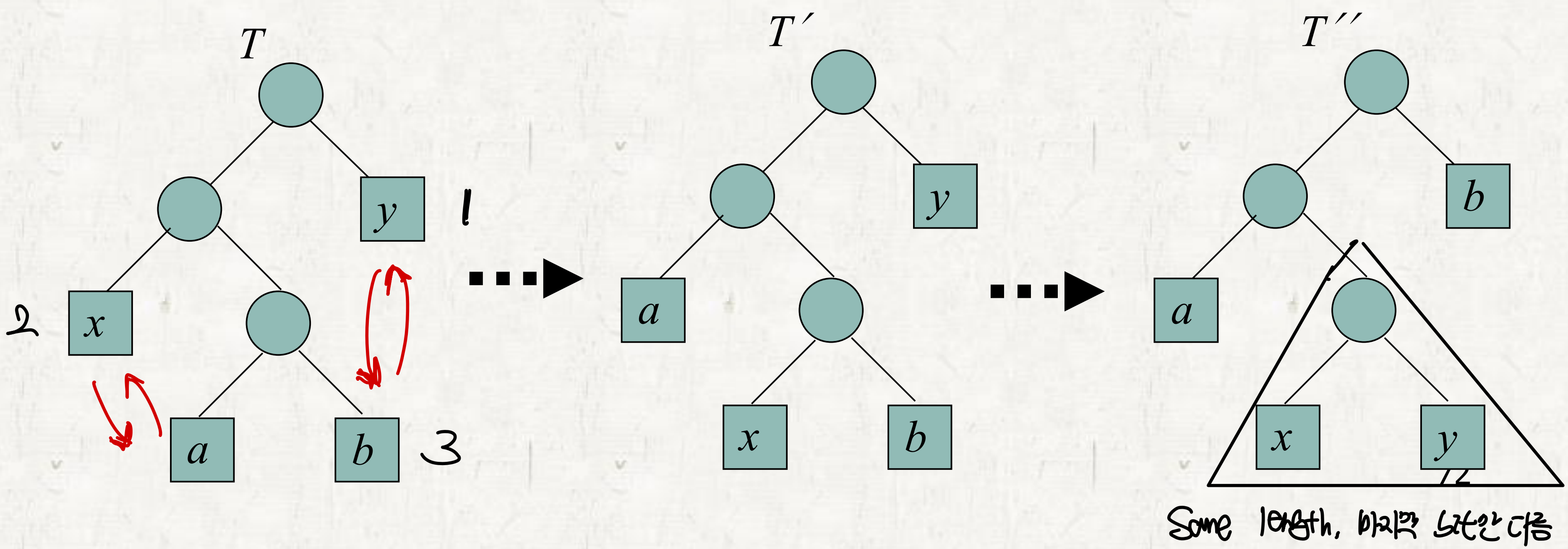
      + one insertion

Huffman codes

- **Correctness**
  - ***Lemma 16.2***
    - Let C be an alphabet in which each character $c \in C$ has frequency $f[c]$.
    - Let x and y be two characters in C having the lowest frequencies.
    - Then there exists an optimal prefix code for C in which the ***codewords for x and y have the same length and differ only in the last bit***.

Huffman codes

- ***Proof***
  - **Idea**: take an arbitrary optimal prefix code tree *T* and modify it and to make a tree representing another optimal prefix code such that the characters *x* and *y* appear as sibling leaves of maximum depth in the new tree.



Huffman codes

- **The cost of tree *T***
  - $f(c)$: frequency of a character c
  - $d_T(c)$: length of the codeword for c

$$B(T) = \sum_{c \in C} f(c) d_T(c)$$

$f(x), f(y) \le f(a), f(b)$

→ 가장 아래 node이기 때문

$$\begin{array}{rl} & B(T) = \cdots 2f(x) + 1 \cdot f(y) + 3f(a) + 3f(b) \\ - & B(T'') \cdots 2f(a) + 1 \cdot f(b) + 3f(x) + 3f(y) \\ \hline \end{array}$$

$$2(f(x) - f(a) + f(y) - f(b) + 3(f(a) - f(x)) + 3(f(b) - f(y))$$

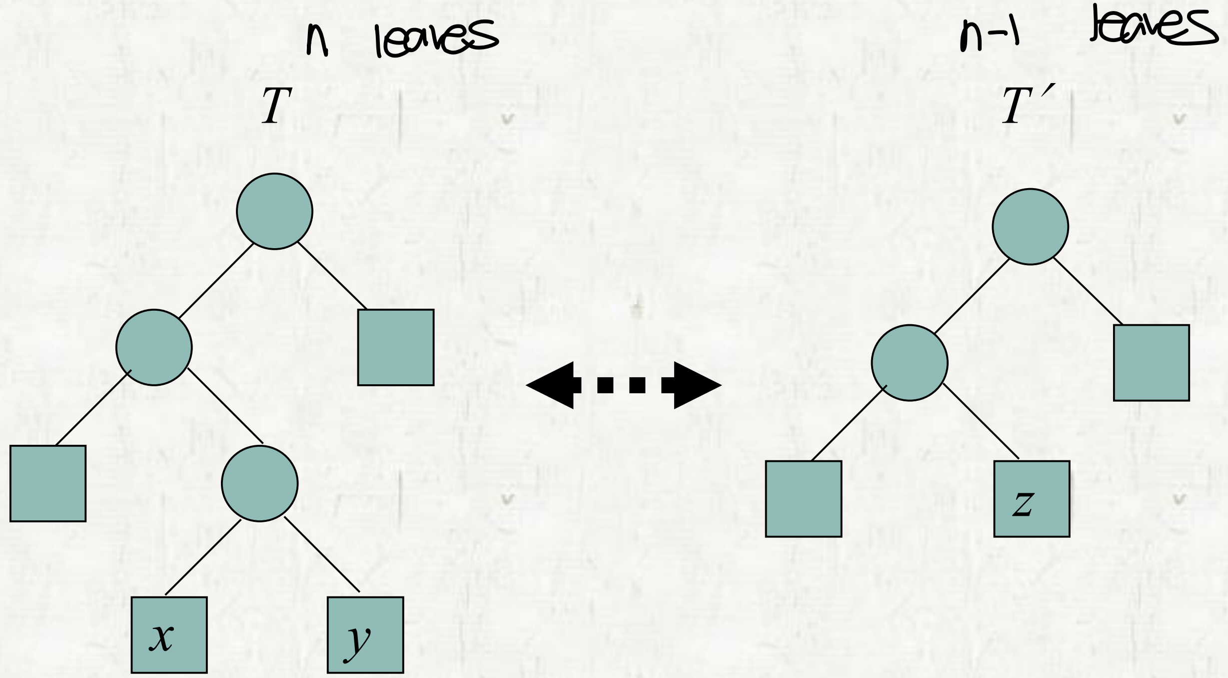
$$= f(a) - f(x) + 2(f(b) - f(y)) \ge 0$$

$\Rightarrow B(T) \ge B(T'')$ T가 optimal → T''도 optimal tree이다.

Huffman codes

- ***Lemma 16.3***
  - Let x and y be two characters in a given alphabet C with minimum frequency.
  - Let C' be the alphabet C with characters x, y removed and character z added, so that $\boxed{C' = C - \{x, y\} \cup \{z\}}$; define f for C' as for C, except that $f[z] = f[x] + f[y]$. z라는 virtual node로 x,y 연결
  - Let T' be any tree representing an optimal prefix code for the alphabet C'.
  - Then the optimal prefix code tree T for C can be obtained from T' by replacing the leaf node for z with an internal node having x and y as children.

Huffman codes



Huffman codes

- ***Proof***
  - Show $B(T) = B(T') + f[x] + f[y]$
    - For each $c \in C$ - $\{x, y\}$, we have $d_T(c) = d_{T'}(c)$, and hence $f[c]d_T(c) = f[c]d_{T'}(c)$.
    - Since $d_T(x) = d_T(y) = d'(z) + 1$, we have

$$f[x]d_T(x) + f[y]d_T(y) = (f[x] + f[y])(d_{T'}(z) + 1)$$
$$= f[z]d_{T'}(z) + (f[x] + f[y])$$

    - From which we conclude that $B(T) = B(T') + f[x] + f[y]$ or, equivalently $B(T') = B(T) - f[x] - f[y]$.

Huffman codes

- ***Proof***
  - Suppose T does not represent an optimal prefix code for C.
  - There exists T'' such that $B(T'') < B(T)$.
  - By Lemma 16.2, there exists T'' having x and y as siblings.
  - Let T''' be the tree T'' with the common parent of x and y replaced by a leaf z with frequency $f[z] = f[x] + f[y]$.
  - Then, $B(T''') = B(T'') - f[x] - f[y]$

    $< B(T) - f[x] - f[y]$

    $= B(T')$

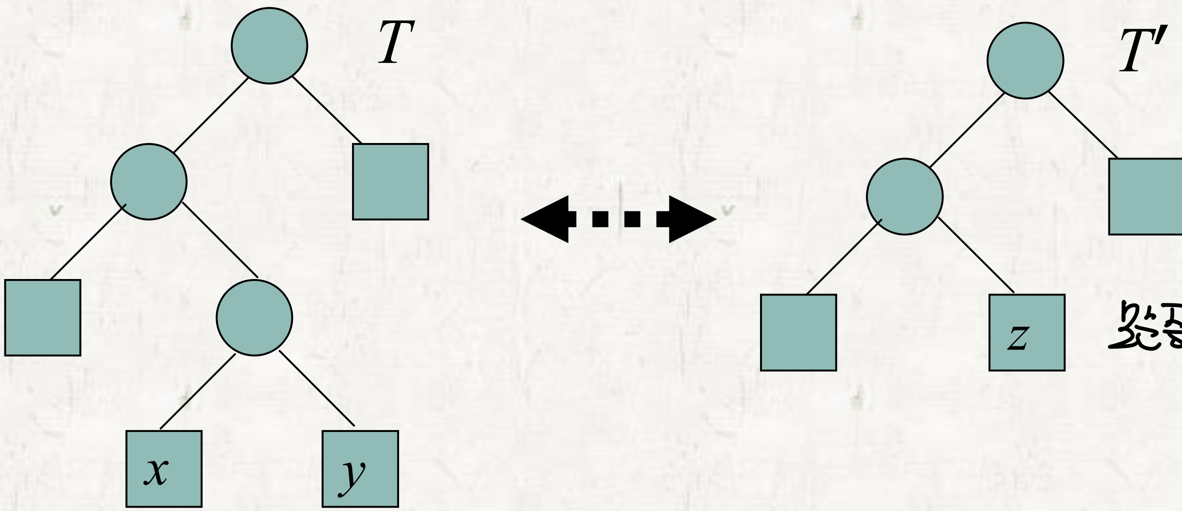
    ➔Contradiction
  - T must represent an optimal prefix code for the alphabet C.

$B(T) = B(T') - d_T(z)f(z) + d_T(x)f(x) + d_T(y)f(y)$

Huffman codes

$f(z) = f(x) + f(y),$
$d_T(z) = d_T(x) - 1$
$d_T(y) = d_T(x)$

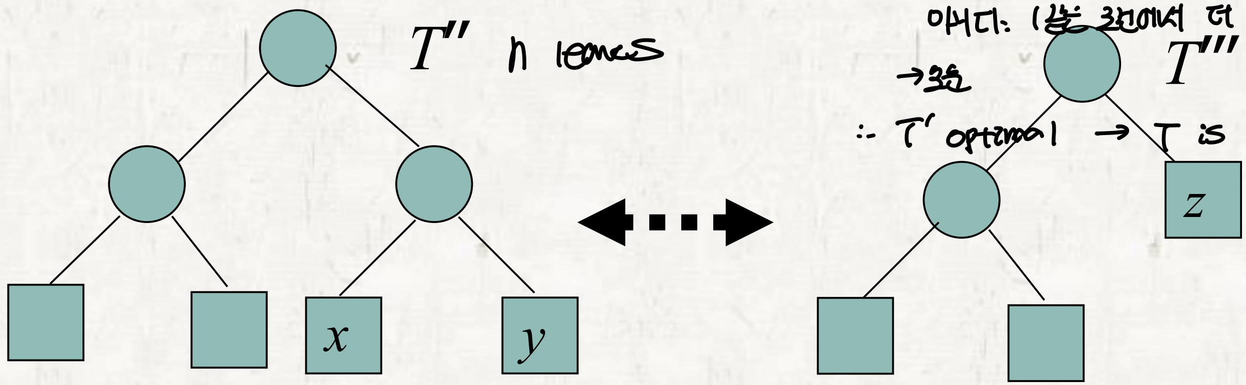
$\boxed{\Rightarrow B(T) = B(T') + f(x) + f(y)}$



T' optimal → T is optimal

대우증명) 가정: T' → T is not optimal
$B(T) > B(T'')$ 라고 하자
$B(T'') = B(T''') + f(x) + f(y)$
$B(T') > B(T''')$ → B(T')는 optimal이 아니다. (같은 조건에서 더 나은 tree 존재)
→ 모순
∴ T' optimal → T is optimal

n leaves



Self-study

- **Exercise 16.3-3 (16.3-2 in the 2nd ed.)**
  - Fibonacci number definition is in p. 59 (p. 56 in the 2nd ed.)

- **Exercise 16.3-7 (16.3-6 in the 2nd ed.)**