

Dispense di algoritmi e strutture dati

Leonardo De Faveri

A.A. 2021/2022

Indice

1	Analisi di algoritmi	7
1.1	Studio degli algoritmi	7
1.1.1	Modello di calcolo	8
1.2	Studio di alcuni algoritmi	8
1.2.1	Minimo di un array	8
1.2.2	Ricerca binaria ricorsiva	9
1.2.3	Ordini di complessità	10
1.3	Notazione asintotica	10
1.4	Complessità degli algoritmi e dei problemi	13
1.5	Ordinamento degli array in senso crescente	14
1.5.1	Algoritmo Selection sort	14
1.5.2	Algoritmo Insertion sort	14
1.5.3	Algoritmo Merge sort	15
2	Notazione asintotica	18
2.1	Proprietà della notazione asintotica	19
2.2	Altre notazioni	21
2.3	Classificazione delle funzioni di costo	22
3	Funzioni di ricorrenza	23
3.1	Studio delle equazioni di ricorrenza	24
3.1.1	Metodo dell'albero di ricorsione	24
3.1.2	Metodo della sostituzione	27
3.1.3	Metodo delle ricorrenze comuni	32
4	Strutture dati	39
4.1	Introduzione	39
4.2	Strutture di dati astratte	40
4.2.1	Sequenze	40
4.2.2	Insiemi	41
4.2.3	Dizionari	42
4.2.4	Grafi e alberi	43
4.2.5	Criticità nell'implementazione di strutture dati astratte	43
4.3	Strutture di dati elementari	44
4.3.1	Liste	44
4.3.2	Pile	46
4.3.3	Code	47

5	Alberi binari e alberi generici	51
5.1	Introduzione	51
5.1.1	Terminologia	51
5.2	Alberi binari	52
5.2.1	Specifica	53
5.2.2	Memorizzazione di un albero binario	53
5.2.3	Implementazione di un albero binario	54
5.2.4	Visite di un albero binario	54
5.3	Alberi generici	56
5.3.1	Specifica	56
5.3.2	Memorizzazione di un albero generico	56
5.3.3	Implementazione di un albero generico	57
5.3.4	Visite di un albero generico	58
6	Alberi binari di ricerca e alberi bilanciati	59
6.1	Alberi binari di ricerca	59
6.1.1	Specifica	60
6.1.2	Ricerca	61
6.1.3	Massimo e minimo	62
6.1.4	Successore e predecessore	62
6.1.5	Inserimento	63
6.1.6	Cancellazione	64
6.1.7	Costo computazionale delle operazioni	66
6.2	Alberi binari di ricerca bilanciati	66
6.3	Alberi Red-Black	67
6.3.1	Rotazioni	68
6.3.2	Inserimento	69
6.3.3	Altezza degli Alberi Red-Black	74
6.3.4	Cancellazione	75
7	Grafi	77
7.1	Introduzione	77
7.1.1	Dimensioni dei grafi	78
7.1.2	Casi speciali	79
7.1.3	Cammini e gradi dei nodi	79
7.1.4	Specifica	80
7.1.5	Memorizzare un grafo	81
7.2	Visite di un grafo	84
7.2.1	Visita in ampiezza	84
7.2.2	Visita in profondità	86
7.2.3	Complessità delle visite	87
7.3	Problemi sui grafi risolubili con visite in ampiezza	87
7.3.1	Calcolo della distanza tra nodi	87
7.3.2	Ricerca del cammino più breve fra due nodi	87
7.4	Problemi sui grafi risolubili con visite in profondità	88
7.4.1	Ricerca delle componenti connesse	88
7.4.2	Verifica di esistenza di cicli nei grafi non orientati	90
7.4.3	Verifica di esistenza di cicli nei grafi orientati	91
7.4.4	Realizzare un ordinamento topologico di un grafo orientato	94
7.4.5	Ricerca delle componenti fortemente connesse	97

8	Hashing	102
8.1	Caratteristiche e implementazioni di funzioni hash	103
8.1.1	Realizzare una funzione hash	104
8.1.2	Metodo dell'estrazione	104
8.1.3	Metodo dello XOR	105
8.1.4	Metodo della divisione	105
8.1.5	Metodo della moltiplicazione	106
8.2	Gestione delle collisioni	107
8.2.1	Liste di trabocco	107
8.2.2	Indirizzamento aperto	108
8.2.3	Complessità delle diverse implementazioni	112
9	Analisi ammortizzata	113
9.1	Metodi per l'analisi ammortizzata	113
9.2	Esempio di analisi ammortizzata	115
9.3	Vettori dinamici	117
9.3.1	Inserimento	117
9.3.2	Cancellazione	119
9.4	Discussione sugli insiemi	121
9.4.1	Insieme come vettore di booleani	121
9.4.2	Insieme come lista non ordinata	122
9.4.3	Insieme come lista ordinata	123
9.4.4	Confronto generale	124
10	Divide-et-impera	125
10.1	Introduzione	125
10.2	Algoritmo Quicksort	127
10.2.1	Principi di funzionamento	127
10.2.2	Implementazione	128
10.2.3	Costo computazionale	128
10.3	Esercizio di applicazione del Divide-et-impera	129
11	Strutture dati specializzate	131
11.1	Code a priorità	131
11.1.1	Heap	132
11.1.2	Algoritmo HeapSort	134
11.1.3	Implementazione di code a priorità	139
11.2	Insiemi disgiunti	141
11.2.1	Implementazione basata su insiemi di liste	143
11.2.2	Implementazione basata su insiemi di alberi	143
11.2.3	Implementazioni con tecniche euristiche	143
11.2.4	Complessità	146
12	Programmazione dinamica	147
12.1	Introduzione	147
12.1.1	Approccio generale	147
12.2	Gioco del domino	148
12.2.1	Approccio basato su divide-et-impera	148
12.2.2	Approccio basato su programmazione dinamica	150
12.3	Problema di Hateville	151
12.3.1	Approccio basato su divide-et-impera	151

12.3.2	Approccio basato su programmazione dinamica	152
12.4	Problema dello zaino	154
12.4.1	Approccio basato su programmazione dinamica	155
12.4.2	Approccio basato su memoization	157
12.5	Problema dello zaino senza limiti	158
12.6	Ricerca della sottosequenza comune massimale	160
12.6.1	Approccio basato su programmazione dinamica	161
12.7	Problema dello string matching approssimato	164
12.8	Problema del prodotto a catena di matrici	166
12.8.1	Approccio basato su divide-et-impera	169
12.8.2	Approccio basato su programmazione dinamica	169
12.9	Ricerca dell'insieme indipendente di peso massimo	170
13	Scelta della struttura dati	174
13.1	Ricerca dei cammini minimi da sorgente singola	174
13.1.1	Prototipo di algoritmo	177
13.1.2	Algoritmo di Dijkstra	177
13.1.3	Algoritmo di Bellman-Ford-Moore	180
13.1.4	Ricerca dei cammini minimi sui grafi DAG	181
13.1.5	Casi d'uso delle soluzioni	182
13.2	Ricerca dei cammini minimi da sorgente multipla	182
13.2.1	Algoritmo di Floyd-Warshall	182
14	Programmazione greedy	185
14.1	Problema del resto	185
14.1.1	Approccio basato su programmazione dinamica	185
14.1.2	Approccio basato su programmazione greedy	186
14.2	Insieme indipendente massimale di intervalli	187
14.2.1	Approccio basato su programmazione dinamica	188
14.3	Problema dello scheduling	190
14.4	Problema dello zaino frazionario	192
14.5	Problema della compressione	193
14.5.1	Algoritmo di Huffman	194
14.6	Ricerca degli alberi di copertura di peso minimo	197
14.6.1	Approccio generico	197
14.6.2	Algoritmo di Kruskal	199
14.6.3	Algoritmo di Prim	200
15	Ricerca locale	202
15.1	Introduzione	202
15.2	Problema della ricerca del flusso massimo	203
15.2.1	Metodo delle reti residue	204
15.2.2	Metodo dei cammini aumentanti	207
15.2.3	Complessità	213
16	Backtracking	215
16.1	Introduzione	215
16.2	Problema dell'enumerazione dei sottoinsiemi	215
16.2.1	Approccio generale	216
16.3	Problema dell'enumerazione delle permutazioni	217
16.4	Problema dell'enumerazione dei k-sottoinsiemi	219

16.5	Problema del subset sum	221
16.5.1	Approccio generale	222
16.6	Problema dell'involuppo convesso	223
16.6.1	Algoritmo di Jarvis	223
16.6.2	Algoritmo di Graham	224
17	Problemi intrattabili e teoria dell'NP-completezza	226
17.1	Introduzione	226
17.2	Riduzioni	227
17.2.1	Colorazione di grafi	227
17.2.2	Sudoku	228
17.2.3	Insieme indipendente	228
17.2.4	Copertura di vertici	229
17.2.5	Soddisfacibilità di formule booleane	230
17.3	Classi di problemi	232
17.4	Alcuni problemi NP-completi	234
18	Soluzioni per problemi intrattabili	237
18.1	Algoritmi pseudo-polinomiali	237
18.1.1	Problema del subset sum	237
18.1.2	Problemi fortemente e debolmente NP-completi	239
18.2	Algoritmi di approssimazione	240
18.2.1	Problema del bin packing approssimato	241
18.2.2	Problema del commesso viaggiatore modificato	242
18.3	Algoritmi euristici	244
18.3.1	Prima soluzione greedy - Arco di costo minimo	245
18.3.2	Seconda soluzione greedy - Vicino più vicino	247
18.3.3	Soluzione con ricerca locale	248
18.4	Algoritmi branch-&-bound	249
18.4.1	Schema generale	249
18.4.2	Problema del commesso viaggiatore	250
19	Algoritmi di ordinamento	254
19.1	Algoritmi basati su confronti	254
19.2	Algoritmi non basati su confronti	255
19.2.1	Spaghetti sort	255
19.2.2	Counting sort	255
19.2.3	Pigeonhole sort	256
19.2.4	Bucket sort	256
19.3	Proprietà degli algoritmi di ordinamento	257
19.4	Confronto tra algoritmi di ordinamento	257
20	Algoritmi probabilistici	258
20.1	Algoritmi Montecarlo	258
20.1.1	Test di primalità	258
20.1.2	Identificazione di espressioni polinomiali nulle	260
20.2	Algoritmi Las Vegas	260
20.2.1	Problema della selezione	260
20.3	Bloom filter	263

A	Appendice A	266
A.1	Proprietà dei logaritmi	266
A.2	Serie matematiche convergenti	266

Capitolo Nr.1

Analisi di algoritmi

1.1 Studio degli algoritmi

Studiare gli algoritmi significa studiarne la *complessità* e l'obiettivo di quest'analisi è quello di realizzare nuove versioni di quegli stessi algoritmi che godano di una *complessità* minore e siano quindi più efficienti.

Definizione 1 - Complessità di un algoritmo.

Studiare la complessità di un algoritmo significa fare l'analisi delle risorse da esso impiegate per risolvere un problema, in funzione della dimensione e della tipologia di input.

La *complessità* si misura con una funzione del tipo:

$$\text{Complessità} : \text{"Dimensione dell'input"} \rightarrow \text{"Tempo di esecuzione"}$$

Ma cosa si intende con *dimensione dell'input*?

Dimensione dell'input Ci sono due criteri per stimare la dimensione:

- *Criterio di costo uniforme*: la taglia dell'input è il numero di elementi di cui è costituito;
- *Criterio di costo logaritmico*: la taglia dell'input è il numero di bit necessari per rappresentarlo;

E con *tipologia di input*? Questo aspetto non impatta su tutti gli algoritmi allo stesso modo. Il caso in cui è più facile comprenderne le conseguenze è quello dell'ordinamento di array. L'algoritmo *insertion sort*, per esempio, si adatta più efficientemente a situazioni in cui gli elementi vengono forniti uno per volta, mentre l'algoritmo *selection sort* è preferibile quando si hanno già tutti gli elementi da ordinare¹.

Tempo di esecuzione Il *tempo di esecuzione* coincide con il numero di istruzioni elementari.

Definizione 2 - Istruzione elementare.

Un'istruzione si considera elementare se può essere eseguita in tempo "costante" dal processore.

¹La complessità di questi algoritmi verrà discussa più nel dettaglio nel corso di questo capitolo

1.1.1 Modello di calcolo

Nello studio degli algoritmi, un parametro da tenere in considerazione è il *modello di calcolo* utilizzato, in quanto, modelli diversi potrebbero cambiare, in meglio o in peggio, la *complessità* e quindi l'efficienza di un algoritmo.

Definizione 3 - Modello di calcolo.

Un modello di calcolo è una rappresentazione astratta di un calcolatore.

Da questa definizione deriviamo le caratteristiche che un generico *modello di calcolo* deve soddisfare:

- *Astrazione*: deve nascondere i dettagli tecnico-operativi;
- *Realismo*: deve riflettere la situazione reale del sistema di calcolo;
- *Potenza matematica*: deve consentire di trarre conclusioni *formali* sul costo;

Esistono molteplici *modelli di calcolo* e il modello della *Macchina di Turing* ne è un esempio. Nel corso di questa trattazione faremo riferimento al *Modello RAM*² caratterizzato da:

- *Memoria*: è costituita da infinite celle di dimensione finita
- *Processore*: ha un set di istruzioni simili a quelli reali:
 - Operazioni logico-aritmetiche;
 - Istruzioni di controllo;
- *Costo uniforme*: il costo delle istruzioni elementari è uniforme;

1.2 Studio di alcuni algoritmi

Proviamo a vedere la logica dietro l'analisi della *complessità* con degli esempi.

1.2.1 Minimo di un array

Di seguito viene riportato l'algoritmo per l'estrazione del minimo elemento di un array. La funzione prende in input un array di n elementi e la relativa dimensione n .

Frammento 1 - Implementazione funzione min per la ricerca del minimo.

ITEM	Costo	N. esecuzioni
min(Item[] A, int n)		
min = A[1]	c_1	1
for (i = 2 to n) do	c_2	n
if (A[i] < min) then	c_3	$n - 1$
min = A[i]	c_4	$n - 1$
return min	c_5	1

Analisi del costo Analizziamo il costo di questo algoritmo. Ogni operazione elementare ha un tempo di esecuzione costante che indicheremo con c_i e ogni istruzione viene eseguita un certo numero di volte.

²RAM: Random Access Machine

Indichiamo con $T(n)$ la *funzione di costo* dell'algoritmo, ovvero la funzione che ne traccia il tempo di esecuzione al variare del valore n . Calcoliamo $T(n)$ sommando il tempo di esecuzione di ogni istruzione considerando anche il numero di volte che ogni istruzione viene eseguita.

$$\begin{aligned} T(n) &= c_1 + c_2n + c_3(n-1) + c_4(n-1) + c_5 \\ &= (c_2 + c_3 + c_4)n + (c_1 + c_5 - c_3 - c_4) \\ &= an + b \end{aligned}$$

1.2.2 Ricerca binaria ricorsiva

Consideriamo l'algoritmo per la ricerca binaria.

Frammento 2 - Implementazione funzione `binarySearch` per la ricerca binaria.

	Costo	N. ($i > j$)	N. ($i \leq j$)
<code>int binarySearch(ITEM[] A, ITEM v, int i, int j)</code>			
<code>if (i > j) then</code>	c_1	1	1
<code>return 0</code>	c_2	1	0
<code>else</code>			
<code>int m = [(i + j) / 2]</code>	c_3	0	1
<code>if (A[m] = v) then</code>	c_4	0	1
<code>return m</code>	c_5	0	0
<code>else if (A[m] < v) then</code>	c_6	0	1
<code>return binarySearch(A, v, m + 1, j)</code>	$c_7 + T(\lfloor (n-1)/2 \rfloor)$	0	0/1
<code>else</code>			
<code>return binarySearch(A, v, i, m - 1)</code>	$c_7 + T(\lfloor n/2 \rfloor)$	0	1/0

La funzione prende in input un array A ordinato in senso crescente, l'elemento v da cercare e due indici i e j che indicano rispettivamente l'estremo destro e sinistro della porzione di array all'interno della quale ricercare l'elemento³.

Nell'algoritmo sono presenti delle selezioni che “modificano” il costo in termini di tempo dell'algoritmo. Viene selezionato un elemento “centrale” nell'array e, se non è l'elemento cercato, si applica ricorsivamente l'algoritmo alla parte di sinistra, se l'elemento cercato è minore di quello analizzato, oppure alla parte destra. Per questa caratteristica dell'algoritmo si ha che la parte di sinistra ha dimensione $(n-1)/2$, mentre quella di destra $n/2$.

Analizziamo il caso pessimo, cioè quello in cui l'elemento cercato non è presente e quindi viene controllato l'intero array. Scegliamo di agire in questo modo così da poter dare una stima del tempo massimo d'esecuzione. Per lo stesso motivo, ipotizziamo che l'elemento cercato sia maggiore di ogni elemento presente nell'array e che quindi la chiamata ricorsiva venga effettuata sempre sulla parte di destra che ha una dimensione maggiore.

Inoltre, per semplicità assumiamo che l'array contenga un numero 2^k di elementi e assegniamo ad ogni istruzione elementare un costo c_i .

Nella stima del costo dell'algoritmo abbiamo due casi:

- $i > j$ per $n = 0$ quindi $T(n) = c_1 + c_2 = c$;
- $i \leq j$ per $n > 0$ quindi $T(n) = T(n/2) + c_1 + \dots + c_7 = T(n/2) + d$;

Possiamo vedere il tutto come una funzione definita per casi:

$$T(n) = \begin{cases} T(n/2) & n \geq 1 \\ c & n = 0 \end{cases}$$

³All'inizio indicano rispettivamente l'indice del primo e dell'ultimo elemento

Poiché $T(n/2)$ rappresenta una chiamata ricorsiva, possiamo analizzare le varie chiamate:

$$\begin{aligned}
T(n) &= T(n/2) + d \\
&= T(n/4) + 2d \\
&= T(n/8) + 3d \\
&\dots \\
&= T(1) + kd \quad n = 2^k \Rightarrow k = \log n \\
&= T(0) + (k+1)d \\
&= kd + (c+d) \\
&= d \log(n) + e
\end{aligned}$$

1.2.3 Ordini di complessità

Dall'analisi dei due algoritmi abbiamo ottenuto due espressioni matematiche: $an+b$ e $d \log(n) + e$ ⁴. Viste quelle espressioni, dette *funzioni di complessità*, possiamo affermare che i due algoritmi hanno rispettivamente un costo *polinomiale* e *logaritmico*.

Queste due tipologie di costo sono anche dette *ordini* o *classi di complessità* e la tabella sottostante ne riassume i principali.

$f(n)$	10^1	10^2	10^3	10^4	Tipo
$\log n$	3	6	9	13	Logaritmico
\sqrt{n}	3	10	31	100	Sublineare
n	10	100	1000	10000	Lineare
$n \log n$	30	664	9965	132877	Loglineare
n^2	10^2	10^4	10^6	10^8	Quadratico
n^3	10^3	10^6	10^9	10^{12}	Cubico
2^n	1024	10^{30}	10^{300}	10^{3000}	Esponenziale

1.3 Notazione asintotica

Abbiamo quindi introdotto il concetto di *ordine di complessità* come metro di misura del costo di un algoritmo. Ora andremo ad introdurre lo strumento della *notazione asintotica*, che ci permetterà di descrivere quel costo in maniera più formale.

Quelle che andremo ad introdurre in realtà sono tre notazioni, indicate rispettivamente dalle lettere O , Ω , Θ .

Definizione 4 - Notazione O .

Sia $g(n)$ una funzione di costo. Si indica con $O(g(n))$ l'insieme delle funzioni $f(n)$ tali per cui:

$$\exists c > 0, \exists m \geq 0 : f(n) \leq c \cdot g(n) \quad \forall n \geq m$$

Le funzioni $f(n)$ che rispettano questa disuguaglianza sono dette essere *O -grandi* di $g(n)$ e si scrive in simboli $f(n) = O(g(n))$. $g(n)$ è un *limite asintotico superiore* di $f(n)$, ciò significa che $f(n)$ cresce al più come $g(n)$.

⁴Il logaritmo è sottinteso essere in base 2

Definizione 5 - Notazione Ω .

Sia $g(n)$ una funzione di costo. Si indica con $\Omega(g(n))$ l'insieme delle funzioni $f(n)$ tali per cui:

$$\exists d > 0, \exists m \geq 0 : f(n) \geq d \cdot g(n) \quad \forall n \geq m$$

Le funzioni $f(n)$ che rispettano questa disuguaglianza sono dette essere Ω -grandi di $g(n)$ e si scrive in simboli $f(n) = \Omega(g(n))$. $g(n)$ è un *limite asintotico inferiore* di $f(n)$, ciò significa che $f(n)$ cresce almeno come $g(n)$.

Definizione 6 - Notazione Θ .

Sia $g(n)$ una funzione di costo. Si indica con $\Theta(g(n))$ l'insieme delle funzioni $f(n)$ tali per cui:

$$\exists c_1 > 0, \exists c_2 > 0, \exists m \geq 0 : c_1 \cdot g(n) \leq f(n) \leq c_2 \cdot g(n) \quad \forall n \geq m$$

Le funzioni $f(n)$ che rispettano quella disuguaglianza sono dette essere Θ di $g(n)$ e si scrive in simboli $f(n) = \Theta(g(n))$. $f(n)$ è un Θ di $g(n)$, ovvero $f(n) = \Theta(g(n))$, se e solo se $f(n) = O(g(n))$ e $f(n) = \Omega(g(n))$, da cui deriva che $f(n) = \Theta(g(n))$ significa che $f(n)$ cresce esattamente come $g(n)$.

Graficamente, se $f(n) = \Theta(g(n))$, si osserva un comportamento di questo tipo:

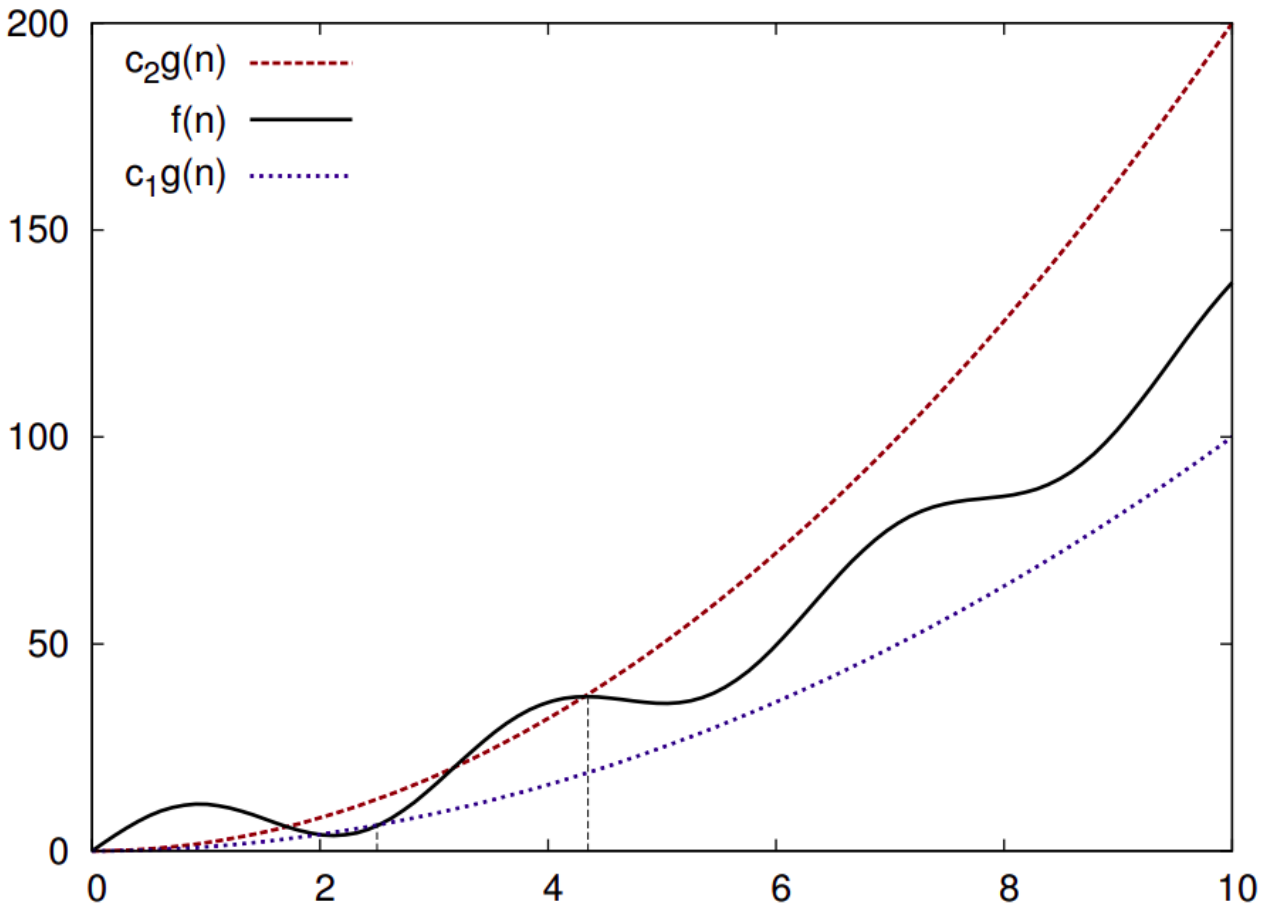


Fig. 1.1: Curva di crescita di $f(n) = \Theta(g(n))$

Si noti come all'inizio la curva di $f(n)$ sembri disattendere le stime, ma essendo *stime asintotiche* esse valgono per $n \rightarrow +\infty$ e infatti, superata una certa soglia, la curva di $f(n)$ rimane costantemente all'interno della porzione di grafico disegnata dalle due curve di $g(n)$.

Esempio 1 - Verifica delle proprietà di un O -grande.

Si dimostri che $f(n) = 10n^3 + 2n^2 + 7$ è un O -grande di n^3 .

Riprendendo la definizione di **O -grande**, dobbiamo dimostrare che:

$$\exists c > 0, \exists m \geq 0 : f(n) \leq c \cdot n^3 \quad \forall n \geq m$$

Procediamo:

$$\begin{aligned} f(n) &= 10n^3 + 2n^2 + 7 \\ &\leq 10n^3 + 2n^3 + 7 & \forall n \geq 1 \\ &\leq 10n^3 + 2n^3 + n^3 & \forall n \geq \sqrt[3]{7} \\ &= 13n^3 \end{aligned}$$

A questo punto ci chiediamo se esistono un valore $c > 0$ e un valore $m \geq 0$ tali per cui:

$$13n^3 \leq c \cdot n^3$$

La risposta è banale poiché basta prendere $c \geq 13$ e $m \geq 1 \geq \sqrt[3]{7}$, ad esempio, $m = 2$.

NB. In questo esempio abbiamo potuto moltiplicare $2n^2$ per n e sostituire il 7 con n^3 , perché $n \geq 0$. In realtà, nelle funzioni di costo n sarà sempre non negativo, quindi lo potremo sempre fare.

Esempio 2 - Verifica delle proprietà di un Θ .

Si dimostri che $f(n) = 3n^2 + 7n$ è un Θ di n^2 .

Procediamo verificando singolarmente i due estremi, inferiore e superiore. Iniziamo con lo studio del limite inferiore, cioè Ω -grande. Ricordando la definizione di **Ω -grande** dobbiamo dimostrare che:

$$\exists c_1 > 0, \exists m_1 \geq 0 : f(n) \geq c_1 \cdot n^2 \quad \forall n \geq m_1$$

Quindi:

$$\begin{aligned} f(n) &= 3n^2 + 7n \\ &\geq 3n^2 & \forall n \geq 0 \\ &\geq c_1 \cdot n^2 & \forall c_1 \leq 3 \wedge \forall n \geq 0 \end{aligned}$$

Pongo quindi $m_1 = 0$.

Passiamo ora allo studio del limite superiore. Come nell'esempio precedente dobbiamo dimostrare che:

$$\exists c_2 > 0, \exists m_2 \geq 0 : f(n) \leq c_2 \cdot n^2 \quad \forall n \geq m_2$$

Quindi:

$$\begin{aligned} f(n) &= 3n^2 + 7n \\ &\leq 3n^2 + 7n^2 & \forall n \geq 1 \\ &= 10n^2 \\ &\leq c_2 \cdot n^2 & \forall c_2 \geq 10 \wedge \forall n \geq 1 \end{aligned}$$

Pongo quindi $m_2 = 1$.

Torniamo ora al problema originario e riprendiamo la definizione di **Θ** :

$$\exists c_1 > 0, \exists c_2 > 0, \exists m \geq 0 : c_1 \cdot g(n) \leq f(n) \leq c_2 \cdot g(n) \quad \forall n \geq m$$

Chiaramente, possiamo prendere $c_1 = 3$ e $c_2 = 10$, mentre, poiché l'ultima disequazione dev'essere vera $\forall n \geq m$, pongo $m = \max\{m_1, m_2\} = \max\{0, 1\} = 1$.

Nel grafico seguente possiamo verificare la correttezza dei nostri calcoli:

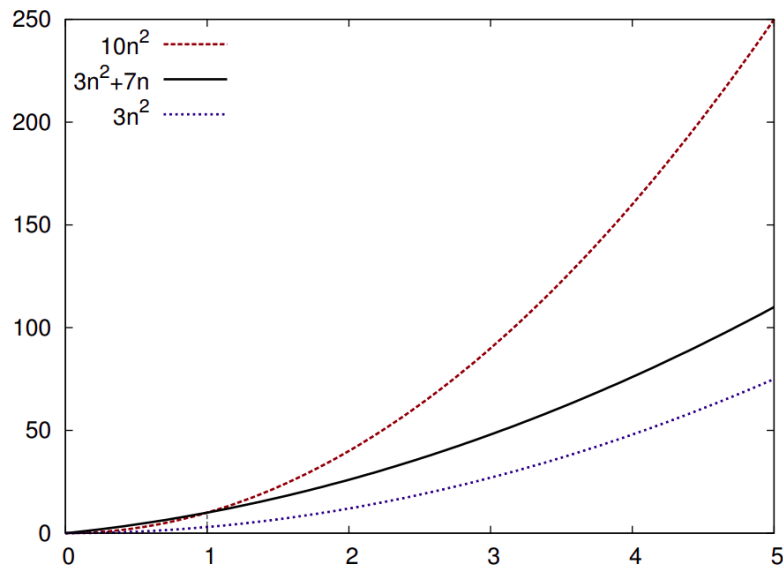


Fig. 1.2: Curva di $f(n) = 3n^2 + 7n$

1.4 Complessità degli algoritmi e dei problemi

Solitamente, per risolvere uno stesso problema esistono una caterva di algoritmi diversi e inevitabilmente alcuni sono più efficienti di altri. In che modo lo studio della *complessità* di questi algoritmi può aiutarci a capire la *complessità del problema*?

Per rispondere a questa domanda sfruttiamo di nuovo la *notazione asintotica*.

Definizione 7 - Notazione O .

Un problema ha complessità $O(f(n))$ se esiste almeno un algoritmo in grado di risolverlo che ha complessità $O(f(n))$.

Definizione 8 - Notazione Ω .

Un problema ha complessità $\Omega(f(n))$ se tutti i possibili algoritmi che lo risolvono hanno complessità $\Omega(f(n))$.

Notazione asintotica per gli algoritmi e per i problemi Nell'analisi della *complessità degli algoritmi*, possiamo riassumere come segue il significato delle *notazioni asintotiche*:

- $O(f(n))$: per tutti gli input n , l'algoritmo costa al più $f(n)$;
- $\Omega(f(n))$: per tutti gli input n , l'algoritmo costa almeno $f(n)$;
- $\Theta(f(n))$: per tutti gli input n , l'algoritmo costa $f(n)$;

Nell'analisi della *complessità dei problemi*, vale invece:

- $O(f(n))$: $O(f(n))$ è la complessità del miglior algoritmo in grado di risolvere il problema;
- $\Omega(f(n))$: vale se si riesce a dimostrare che nessun algoritmo può risolvere il problema in un tempo inferiore a $\Omega(f(n))$;

Se riusciamo a dimostrare che un problema ha complessità $O(f(n))$ e $\Omega(f(n))$, un algoritmo con *complessità* $\Theta(f(n))$ in grado di risolvere quel problema, è il miglior algoritmo possibile.

1.5 Ordinamento degli array in senso crescente⁵

Consideriamo ora tre diversi approcci al problema dell'ordinamento di array e analizziamone la *complessità*.

1.5.1 Algoritmo Selection sort

Questo algoritmo segue un approccio molto intuitivo: cerca il minimo tra tutti gli elementi, lo ordina e poi ripete il tutto per i restanti elementi.

Frammento 3 - Implementazione algoritmo Selection Sort.

```
SelectionSort(ITEM[] A, int n)
    for (i = 1 to n - 1) do
        int min = min(A, i, n)
        A[i] ↔ A[min]                % Scambia il minimo attuale e l'elemento A[i]
```

Frammento 4 - Implementazione funzione min con indice di partenza arbitrario.

```
int min(ITEM[] A, int i, n)
    int min = i                        % Posizione del minimo parziale
    for (j = i + 1 to n) do
        if (A[j] < A[min]) then
            min = j                    % Nuovo minimo parziale
    return min
```

Quali sono le *complessità* nei casi pessimo, medio e migliore?

Il caso pessimo è quello in cui l'array è ordinato in senso decrescente, per cui ad ogni iterazione il minimo si trova nell'ultima posizione. Il caso migliore è invece quello in cui l'array è già ordinato in senso crescente. Nonostante questo, possiamo intuire che il costo dell'algoritmo non cambi, perché in ogni caso si dovrà ricercare il minimo $n - 1$ volte.

Provando a definire la *funzione di costo* di questo algoritmo abbiamo:

$$\sum_{i=1}^{n-1} (n-i) = \sum_{i=1}^{n-1} i = {}^6 \frac{n(n-1)}{2} = n^2 - \frac{n}{2} = O(n^2)$$

Questo vale per tutti i casi, quindi l'algoritmo *insertion sort* ha un costo $\Theta(n^2)$.

1.5.2 Algoritmo Insertion sort

Vediamo ora un approccio nel quale proviamo a prendere un elemento per volta e a metterlo nella posizione giusta.

In particolare, ad ogni iterazione viene preso un valore e viene confrontato progressivamente con i valori precedenti. Se viene trovato un valore minore di quello preso, quest'ultimo viene salvato nella penultima cella controllata.

⁵Il ragionamento è analogo per ordinamenti in senso decrescente

⁶*Formula di Gauss*

Frammento 5 - Implementazione algoritmo Insertion Sort.

```
insertionSort(ITEM[] A, int n)
    for (i = 2 to n) do
        ITEM temp = A[i]
        int j = i
        while (j > 1 and A[j - 1] > temp) do
            A[j] = A[j - 1]
            j = j - 1
        A[j] = temp
```

Cosa succede nei casi pessimo, medio e migliore?

Come prima, il caso pessimo è quello in cui l'array è ordinato in senso contrario. In una situazione del genere, per ogni i devono essere fatti $i - 1$ confronti. Si ha quindi una funzione di questo tipo:

$$\sum_{i=2}^n (i - 1) = \sum_{i=1}^{n-1} (n - i) = O(n^2)$$

Nel caso migliore invece, il vettore è già ordinato nel senso corretto col risultato che ad ogni iterazione il controllo $A[j-1] > \text{temp}$ è sempre falso e quindi la funzione di costo è:

$$\sum_{i=2}^n 1 = n - 1 = O(n)$$

1.5.3 Algoritmo Merge sort

Questo algoritmo è basato sulla tecnica del *divide-et-impera*, cioè procede suddividendo l'array in due metà che vengono ordinate separatamente e quindi ricombinate per ottenere l'array di partenza ordinato.

Come sarà facile intuire, questo algoritmo sfrutta la ricorsione per suddividere l'array in due componenti e una qualche funzione per l'unione dei due sottoarray ordinati.

Frammento 6 - Implementazione algoritmo Merge sort.

```
mergeSort(ITEM[] A, int first, int last)
    if (first < last) then
        int mid = [(first + last) / 2]
        mergeSort(A, first, mid)                % Ordina la parte sinistra
        mergeSort(A, mid + 1, last)              % Ordina la parte destra
        merge(A, first, last, mid)               % Combina i due sottoarray ordinati
```

Nella funzione `merge` ipotizziamo di avere a disposizione un array di appoggio B nel quale andiamo ad inserire, in modo ordinato, i valori dei due sottoarray di A . Una volta riempito, tutti i valori di B vengono ricopiati di nuovo su A .

Frammento 7 - Implementazione funzione merge.

```
merge(ITEM[] A, int first, int last, int mid)
    int i, j, k
```



```

i = first
j = mid + 1
k = first                                % Indice delle posizioni nell'array di appoggio
while (i ≤ mid and j ≤ last) do          % Popolamento di B
    if (A[i] ≤ A[j]) then
        B[k] = A[i]
        i = i + 1
    else
        B[k] = A[j]
        j = j + 1
    k = k + 1
j = last
for (h = mid downto i) do                % Sposta alla fine di A il resto di A[i...mid]
    A[j] = A[h]
    j = j - 1
for (j = first to k - 1) do              % Copia di B su A
    A[j] = B[j]

```

Qual è la complessità del *merge sort*?

Studiamo per prima la sola funzione *merge*. Nel ciclo **while** e nel secondo ciclo **for**, vengono fatte n assegnazioni, dunque la funzione ha complessità $O(n)$.

Passando a considerare l'algoritmo nel suo insieme vediamo che l'array viene diviso in due metà e poi viene invocata *merge*. Ciò significa che se si tracciano i partizionamenti dell'array durante l'esecuzione dell'algoritmo, si ottiene un albero binario di altezza $k = \log n^7$ nel quale, ad ogni livello i , la funzione *merge* viene invocata 2^i volte.

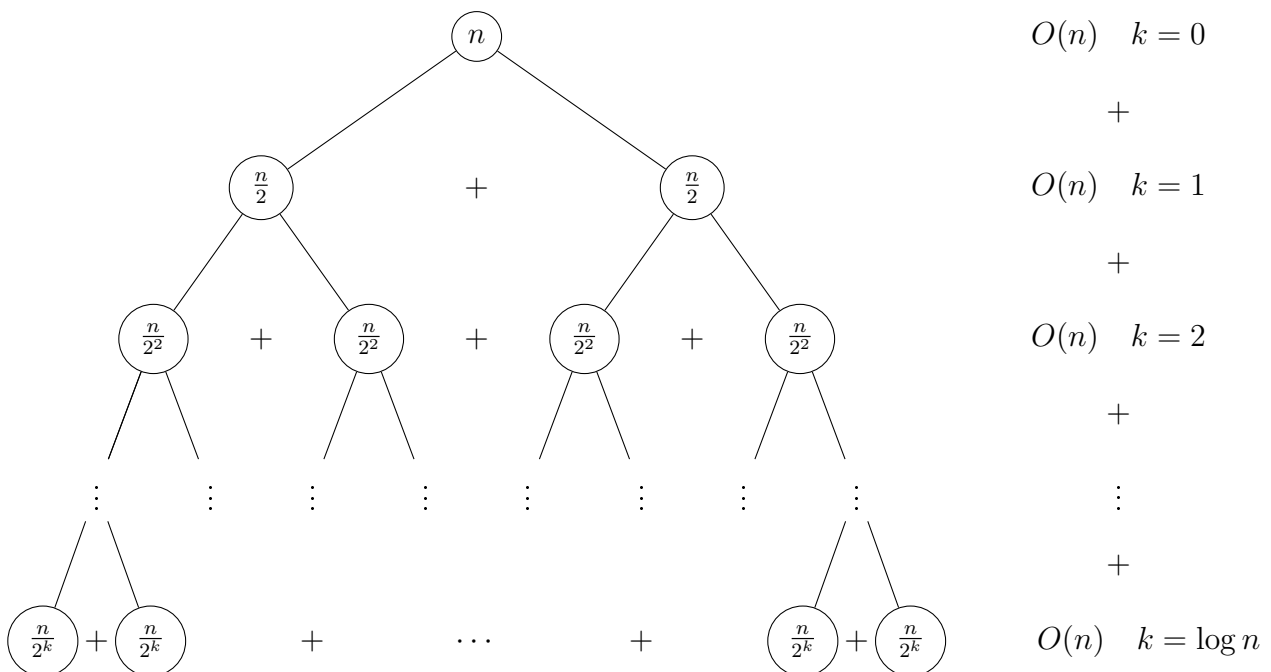


Fig. 1.3: Schema d'esecuzione del *Merge sort*

Ogni livello ha un costo $O(n)$, quindi, essendoci $k = \log n$ livelli, il costo dell'algoritmo è

⁷Per semplicità ipotizziamo che il numero di elementi dell'array sia una potenza di 2

$O(n \log n)$. Possiamo verificare il nostro risultato svolgendo direttamente i calcoli:

$$O\left(\sum_{i=0}^k 2^i \cdot \frac{n}{2^i}\right) = O\left(\sum_{i=0}^k n\right) = O(k \cdot n) = O(n \log n)$$

Nel primo passaggio, l'argomento della sommatoria è quello perché per ogni livello sommiamo il costo necessario ad eseguire il merge su ogni sottoarray di dimensione $\frac{n}{2^i}$ (e.g. il sottoarray al livello 1 avrà dimensione $\frac{n}{2^1}$), ma consideriamo anche la numerosità dei sottoarray: 2^i per ogni livello i .

Ragionando anche sulla *funzione di costo* constatiamo che, poiché ad ogni livello vengono effettuate due invocazioni della **mergeSort** e un'invocazione della **merge**, che ha costo $O(n)$, vale la seguente:

$$T(n) = \begin{cases} 2T(n/2) + dn & n > 1 \\ c & n = 1 \end{cases}$$

Capitolo Nr.2

Notazione asintotica

Nel precedente capitolo abbiamo introdotto velocemente la *notazione asintotica*, ma ora la riprendiamo per approfondirne le caratteristiche e iniziamo con l'esplicitare una regola generale che finora abbiamo sfruttato senza pensarci troppo.

Definizione 9 - Regola generale per le espressioni polinomiali.

Per ogni espressione polinomiale di grado k del tipo

$$f(n) = a_k \cdot n^k + a_{k-1} \cdot n^{k-1} + \dots + a_1 \cdot n + a_0 \quad a_k > 0$$

vale:

$$f(n) = \Theta(n^k)$$

Dimostrazione. La condizione necessaria per poter dire che $f(n) = \Theta(n^k)$ è:

$$f(n) = O(n^k) \quad \wedge \quad f(n) = \Omega(n^k)$$

Iniziamo quindi con la verifica del *limite superiore*, ovvero dimostriamo che vale la seguente:

$$\exists c > 0, \exists m \geq 0 : f(n) \leq c \cdot n^k \quad \forall n \geq m$$

Procediamo:

$$\begin{aligned} f(n) &= a_k \cdot n^k + a_{k-1} \cdot n^{k-1} + \dots + a_1 \cdot n + a_0 \\ &\leq a_k \cdot n^k + |a_{k-1}| \cdot n^{k-1} + \dots + |a_1| \cdot n + |a_0| \\ &\leq a_k \cdot n^k + |a_{k-1}| \cdot n^k + \dots + |a_1| \cdot n^k + |a_0| \cdot n^k \quad \forall n \geq 1 \\ &= (a_k + |a_{k-1}| + \dots + |a_1| + |a_0|) \cdot n^k \\ &\leq c \cdot n^k \end{aligned}$$

L'ultima disequazione è vera per:

$$c \geq (a_k + |a_{k-1}| + \dots + |a_1| + |a_0|) \quad \wedge \quad m = 1$$

Ora passiamo alla verifica del *limite inferiore* e quindi dimostriamo che vale anche la seguente:

$$\exists d > 0, \exists m \geq 0 : f(n) \geq d \cdot n^k \quad \forall n \geq m$$

Procediamo:

$$\begin{aligned}
f(n) &= a_k \cdot n^k + a_{k-1} \cdot n^{k-1} + \dots + a_1 \cdot n + a_0 \\
&\geq a_k \cdot n^k - |a_{k-1}| \cdot n^{k-1} - \dots - |a_1| \cdot n - |a_0| \\
&\geq a_k \cdot n^k - |a_{k-1}| \cdot n^{k-1} - \dots - |a_1| \cdot n^{k-1} - |a_0| \cdot n^{k-1} \quad \forall n \geq 1 \\
&\geq d \cdot n^k
\end{aligned}$$

L'ultima disequazione è vera per:

$$d \leq a_k - \frac{|a_{k-1}|}{n} - \frac{|a_{k-2}|}{n} - \dots - \frac{|a_1|}{n} - \frac{|a_0|}{n}$$

Poiché $d > 0$, anche il termine destro della disequazione dev'essere > 0 , e quindi deve valere:

$$a_k - \frac{|a_{k-1}|}{n} - \frac{|a_{k-2}|}{n} - \dots - \frac{|a_1|}{n} - \frac{|a_0|}{n} > 0 \Leftrightarrow n \geq \frac{|a_{k-1}| + \dots + |a_1| + |a_0|}{a_k} = m$$

□

2.1 Proprietà della notazione asintotica

Definizione 10 - Proprietà di dualità.

Per ogni coppia di funzioni di costo $f(n)$ e $g(n)$ vale:

$$f(n) = O(g(n)) \Leftrightarrow g(n) = \Omega(f(n))$$

Dimostrazione.

$$\begin{aligned}
f(n) = O(g(n)) &\Leftrightarrow f(n) \leq c \cdot g(n) \quad \forall n \geq m \\
&\Leftrightarrow g(n) \geq \frac{1}{c} \cdot f(n) \quad \forall n \geq m \\
&\Leftrightarrow g(n) \geq c' \cdot f(n) \quad \forall n \geq m, c' = \frac{1}{c} \\
&\Leftrightarrow g(n) = \Omega(f(n))
\end{aligned}$$

□

Questa proprietà stabilisce che se $f(n)$ è un $O(g(n))$, allora $g(n)$ è un $\Omega(f(n))$.

Definizione 11 - Proprietà di eliminazione delle costanti.

Per ogni funzione di costo $f(n)$ vale:

$$f(n) = O(g(n)) \Leftrightarrow a \cdot f(n) = O(g(n)) \quad \forall a > 0$$

$$f(n) = \Omega(g(n)) \Leftrightarrow a \cdot f(n) = \Omega(g(n)) \quad \forall a > 0$$

Dimostrazione.

$$\begin{aligned}
f(n) = O(g(n)) &\Leftrightarrow f(n) \leq c \cdot g(n) \quad \forall n \geq m \\
&\Leftrightarrow a \cdot f(n) \leq a \cdot c \cdot g(n) \quad \forall n \geq m, \forall a > 0 \\
&\Leftrightarrow a \cdot f(n) \leq c' \cdot g(n) \quad \forall n \geq m, c' = a \cdot c > 0 \\
&\Leftrightarrow a \cdot f(n) = O(g(n))
\end{aligned}$$

□

NB. La dimostrazione è analoga per $\Omega(g(n))$.

Questa proprietà permette di ignorare le costanti moltiplicative per le *funzioni di costo*.

Definizione 12 - Proprietà di massimo costo.

Per ogni coppia di funzioni di costo $f_1(n)$ e $f_2(n)$ vale:

$$f_1(n) = O(g_1(n)), f_2(n) = O(g_2(n)) \Rightarrow f_1(n) + f_2(n) = O(\max\{g_1(n), g_2(n)\})$$

$$f_1(n) = \Omega(g_1(n)), f_2(n) = \Omega(g_2(n)) \Rightarrow f_1(n) + f_2(n) = \Omega(\max\{g_1(n), g_2(n)\})$$

Dimostrazione.

$$\begin{aligned} f_1(n) = O(g_1(n)) \wedge f_2(n) = O(g_2(n)) &\Rightarrow \\ f_1(n) \leq c_1 \cdot g_1(n) \wedge f_2(n) \leq c_2 \cdot g_2(n) &\Rightarrow \\ f_1(n) + f_2(n) \leq c_1 \cdot g_1(n) + c_2 \cdot g_2(n) &\Rightarrow \\ f_1(n) + f_2(n) \leq \max\{c_1, c_2\} \cdot (2 \max\{g_1(n), g_2(n)\}) &\Rightarrow \\ f_1(n) + f_2(n) = O(\max\{g_1(n), g_2(n)\}) &\Rightarrow \end{aligned}$$

□

NB. La dimostrazione è analoga per $f_1(n) + f_2(n) = \Omega(\max\{g_1(n), g_2(n)\})$.

Questa proprietà stabilisce che nel caso in cui si vada a sommare più *funzioni di costo* (e.g. sequenze di algoritmi), si può considerare solo la funzione di costo maggiore.

Definizione 13 - Proprietà del prodotto dei costi.

Per ogni coppia di funzioni di costo $f_1(n)$ e $f_2(n)$ vale:

$$f_1(n) = O(g_1(n)), f_2(n) = O(g_2(n)) \Rightarrow f_1(n) \cdot f_2(n) = O(g_1(n) \cdot g_2(n))$$

$$f_1(n) = \Omega(g_1(n)), f_2(n) = \Omega(g_2(n)) \Rightarrow f_1(n) \cdot f_2(n) = \Omega(g_1(n) \cdot g_2(n))$$

Dimostrazione.

$$\begin{aligned} f_1(n) = O(g_1(n)) \wedge f_2(n) = O(g_2(n)) &\Rightarrow \\ f_1(n) \leq c_1 \cdot g_1(n) \wedge f_2(n) \leq c_2 \cdot g_2(n) &\Rightarrow \\ f_1(n) \cdot f_2(n) \leq c_1 \cdot c_2 \cdot g_1(n) \cdot g_2(n) &\Rightarrow \\ f_1(n) \cdot f_2(n) = O(g_1(n) \cdot g_2(n)) &\Rightarrow \end{aligned}$$

□

NB. La dimostrazione è analoga per $f_1(n) \cdot f_2(n) = \Omega(g_1(n) \cdot g_2(n))$.

Questa proprietà stabilisce che nel caso in cui si vada a moltiplicare tra loro più *funzioni di costo* (e.g. cicli annidati), il costo totale è proprio il prodotto dei costi delle singole funzioni.

Definizione 14 - Proprietà di simmetria.

Per ogni coppia di funzioni di costo $f(n)$ e $g(n)$ vale:

$$f(n) = \Theta(g(n)) \Leftrightarrow g(n) = \Theta(f(n))$$

Dimostrazione. Grazie alla *proprietà di dualità* vale:

$$\begin{aligned} f(n) = \Theta(g(n)) &\Rightarrow f(n) = O(g(n)) \Rightarrow g(n) = \Omega(f(n)) \\ f(n) = \Theta(g(n)) &\Rightarrow f(n) = \Omega(g(n)) \Rightarrow g(n) = O(f(n)) \end{aligned}$$

□

Questa proprietà stabilisce che se $f(n)$ è un $\Theta(g(n))$ allora anche $g(n)$ è un $\Theta(f(n))$.

Definizione 15 - Proprietà di transitività.

Prese tre funzioni di costo $f(n)$, $g(n)$ e $h(n)$ tali che:

$$f(n) = O(g(n)) \quad \wedge \quad g(n) = O(h(n))$$

vale:

$$f(n) = O(h(n))$$

Dimostrazione.

$$\begin{aligned} f(n) = O(g(n)) \wedge g(n) = O(h(n)) &\Rightarrow \\ f(n) \leq c_1 \cdot g(n) \wedge g(n) \leq c_2 \cdot h(n) &\Rightarrow \\ f(n) \leq c_2 \cdot h(n) &\Rightarrow \\ f(n) = O(h(n)) & \end{aligned}$$

□

2.2 Altre notazioni

Definizione 16 - Notazione o .

Sia $g(n)$ una funzione di costo. Si indica con $o(g(n))$ l'insieme delle funzioni di costo $f(n)$ tali per cui:

$$\forall c, \exists m : f(n) < c \cdot g(n) \quad \forall n \geq m$$

Le funzioni $f(n)$ che rispettano questa disuguaglianza sono dette essere *o-piccoli* di $g(n)$ e si scrive in simboli $f(n) = o(g(n))$.

Definizione 17 - Notazione ω .

Sia $g(n)$ una funzione di costo. Si indica con $\omega(g(n))$ l'insieme delle funzioni di costo $f(n)$ tali per cui:

$$\forall d, \exists m : f(n) > d \cdot g(n) \quad \forall n \geq m$$

Le funzioni $f(n)$ che rispettano questa disuguaglianza sono dette essere *ω -piccoli* di $g(n)$ e si scrive in simboli $f(n) = \omega(g(n))$.

Osservazioni Date due funzioni di costo $f(n)$ e $g(n)$ possiamo fare le seguenti affermazioni:

$$\begin{aligned} \lim_{n \rightarrow +\infty} \frac{f(n)}{g(n)} = 0 &\Rightarrow f(n) = o(g(n)) \\ \lim_{n \rightarrow +\infty} \frac{f(n)}{g(n)} = c \neq 0 &\Rightarrow f(n) = \Theta(g(n)) \\ \lim_{n \rightarrow +\infty} \frac{f(n)}{g(n)} = +\infty &\Rightarrow f(n) = \omega(g(n)) \end{aligned}$$

Si noti anche che:

$$\begin{aligned} f(n) = o(g(n)) &\Rightarrow f(n) = O(g(n)) \\ f(n) = \omega(g(n)) &\Rightarrow f(n) = \Omega(g(n)) \end{aligned}$$

2.3 Classificazione delle funzioni di costo

Giunti a questo punto possiamo definire un ordinamento per le principali *funzioni di costo*:

Definizione 18 - Ordinamento delle funzioni di costo.

Per ogni $r < s$, $h < k$ e $a < b$ vale:

$$\begin{aligned} O(1) \subset O(\log^r n) \subset O(\log^s n) \subset O(n^h) \subset O(n^h \log^r n) \\ \subset O(n^h \log^s n) \subset O(n^k) \subset O(a^n) \subset O(b^n) \end{aligned}$$

Capitolo Nr.3

Funzioni di ricorrenza

Quando si calcola la *complessità* di un algoritmo ricorsivo, questa viene espressa tramite un'*equazione* o una *funzione di ricorrenza*, ovvero una formula matematica definita in maniera ricorsiva.

Un esempio di *equazione di ricorrenza* è quella che descrive il costo dell'algoritmo *Merge sort*:

$$T(n)^1 = \begin{cases} T(\lfloor n/2 \rfloor) + T(\lceil n/2 \rceil) + \Theta(n) & n > 1 \\ \Theta(1) & n \leq 1 \end{cases}$$

Il nostro obiettivo è ottenere, quando possibile, una *formula*, o *forma chiusa*, che rappresenti la *classe di complessità* della funzione. Nell'esempio del *Merge sort* la *formula chiusa* è:

$$T(n) = \Theta(n \log n)$$

In realtà, le *equazioni di ricorrenza* possono essere usate anche per risolvere problemi. Vediamone un esempio:

Problema 1 - Applicazione delle funzioni di ricorrenza nella risoluzione di problemi.

Un bambino scende una scala composta da n scalini. Ad ogni passo, può decidere di fare 1, 2, 3, 4 scalini alla volta. Determinare in quanti modi diversi può scendere le scale.

Ad esempio, se $n = 7$, alcuni dei modi possibili sono i seguenti:

- 1,1,1,1,1,1,1;
- 1,2,4;
- 4,2,1;
- 2,2,2,1;
- 1,2,2,1,1;

¹Questa descrive il caso generale, mentre la precedente si basava sull'ipotesi di array di dimensione 2^i con $i \in \mathbb{N}$

Soluzione Sia $M(n)$ il numero di modi in cui è possibile scegliere n scalini. $M(n)$ può essere espresso come segue:

$$M(n) = \begin{cases} 0 & n < 0 \\ 1 & n = 0 \\ \sum_{k=1}^4 M(n-k) & n > 0 \end{cases}$$

Questa *ricorrenza* può poi essere trasformata in un algoritmo² tramite *ricorsione* o *programmazione dinamica*.

3.1 Studio delle equazioni di ricorrenza

Vediamo di seguito alcuni modi per analizzare le *funzioni di ricorrenza* e ricavarne la *forma chiusa*.

3.1.1 Metodo dell'albero di ricorsione

Il *metodo dell'albero di ricorsione* (o per *livelli*) prevede che la *ricorrenza* venga “srotolata” in un albero i cui nodi rappresentano il costo ai vari livelli della ricorsione.

Esempio 3 - Esempio semplice di analisi con albero di ricorsione.

Sia $T(n)$ la seguente funzione di ricorrenza:

$$T(n) = \begin{cases} T(n/2) + b & n > 1 \\ c & n \leq 1 \end{cases}$$

Ipotizziamo, per semplicità, che $n = 2^k$, quindi proviamo a risolvere $T(n) = T(2^k)$:

$$\begin{aligned} T(n) &= b + T(n/2) \\ &= b + b + T(n/4) \\ &= b + b + b + T(n/8) \\ &= \dots \\ &= \underbrace{b + \dots + b}_{\log n} + T(1) \\ &= b \cdot \log n + c \end{aligned}$$

Quindi, $T(n) = b \cdot \log k + c = \Theta(\log n)$.

Esempio 4 - Utilizzo delle serie matematiche nel processo di semplificazione.

Sia $T(n)$ la seguente funzione di ricorrenza:

$$T(n) = \begin{cases} 4T(n/2) + n & n > 1 \\ 1 & n \leq 1 \end{cases}$$

²Il risultato sono i *numeri di Tetranacci*: 1, 1, 2, 4, 8, 15, 29, 56, 108, 208, 401, 773, 1490, 2872, 5536, ...

Ipotizziamo sempre $n = 2^k$:

$$\begin{aligned}
T(n) &= n + 4T(n/2) \\
&= n + 4(n/2) + 16T(n/4) \\
&= n + 2n + 16(n/4) + 64T(n/8) \\
&= \dots \\
&= n + 2n + 4n + 8n + \dots + 2^{\log(n)-1}n + 4^{\log n}T(1) \\
&= \left(\sum_{i=0}^{\log(n)-1} 2^i\right) \cdot n + 4^{\log n}
\end{aligned}$$

A questo punto, possiamo sfruttare la *serie geometrica finita* e riscrivere la funzione come:

$$\begin{aligned}
T(n) &= \left(\sum_{i=0}^{\log(n)-1} 2^i\right) \cdot n + 4^{\log n} \\
&= \left(\frac{2^{\log n} - 1}{2 - 1}\right) \cdot n + 4^{\log n} && \text{Serie geometrica finita} \\
&= (2^{\log n} - 1) \cdot n + 4^{\log n} \\
&= (n^{\log 2} - 1) \cdot n + n^{\log 4} && \text{Teorema di scambio base-argomento} \\
&= (n^1 - 1) \cdot n + n^2 \\
&= (n - 1) \cdot n + n^2 \\
&= n^2 - n + n^2 \\
&= 2n^2 - n
\end{aligned}$$

Da qui è evidente che $T(n) = \Theta(n^2)$.

Esempio 5 - Utilizzo di una tabella dei costi.

Sia $T(n)$ la seguente funzione di ricorrenza:

$$t(n) = \begin{cases} 4T(n/2) + n^3 & n > 1 \\ 1 & n \leq 1 \end{cases}$$

Proviamo ad analizzare l'albero delle chiamate per i primi 3 livelli:

$$\begin{array}{c}
n^3 \\
\hline
\overbrace{\left(\frac{n}{2}\right)^3 \quad \left(\frac{n}{2}\right)^3 \quad \left(\frac{n}{2}\right)^3 \quad \left(\frac{n}{2}\right)^3} \\
\hline
\overbrace{\left(\frac{n}{4}\right)^3 \quad \left(\frac{n}{4}\right)^3 \quad \left(\frac{n}{4}\right)^3 \quad \left(\frac{n}{4}\right)^3} \quad \overbrace{\left(\frac{n}{4}\right)^3 \quad \left(\frac{n}{4}\right)^3 \quad \left(\frac{n}{4}\right)^3 \quad \left(\frac{n}{4}\right)^3} \quad \overbrace{\left(\frac{n}{4}\right)^3 \quad \left(\frac{n}{4}\right)^3 \quad \left(\frac{n}{4}\right)^3 \quad \left(\frac{n}{4}\right)^3} \quad \overbrace{\left(\frac{n}{4}\right)^3 \quad \left(\frac{n}{4}\right)^3 \quad \left(\frac{n}{4}\right)^3 \quad \left(\frac{n}{4}\right)^3}
\end{array}$$

È chiaro che con una funzione di questo tipo, non è possibile procedere in questo modo. Una strategia migliore è quella di usare una tabella come la seguente.

Livello	Dim. input	Costo per chiamata	N. chiamate	Costo livello
0	n	n^3	1	n^3
1	$n/2$	$(n/2)^3$	4	$4(n/2)^3$
2	$n/4$	$(n/4)^3$	16	$16(n/4)^3$
...
i	$n/2^i$	$(n/2^i)^3$	4^i	$4^i(n/2^i)^3$
...
$\log(n) - 1$	$n/2^{\log(n)-1}$	$(n/2^{\log(n)-1})^3$	$4^{\log(n)-1}$	$4^{\log(n)-1} (n/2^{\log(n)-1})^3$
$\log n$	1	$T(1)$	$4^{\log n}$	$4^{\log n}$

Ora, il costo totale è dato dalla somma del costo di ogni livello. Vale quindi:

$$\begin{aligned}
T(n) &= \sum_{i=0}^{\log(n)-1} \left(4^i (n/2^i)^3 \right) + 4^{\log n} \\
&= \left(\sum_{i=0}^{\log(n)-1} \frac{4^i}{2^{3i}} \right) \cdot n^3 + 4^{\log n} \\
&= \left(\sum_{i=0}^{\log(n)-1} \frac{2^{2i}}{2^{3i}} \right) \cdot n^3 + 4^{\log n} \\
&= \left(\sum_{i=0}^{\log(n)-1} \left(\frac{1}{2} \right)^i \right) \cdot n^3 + 4^{\log n} \\
&= \left(\sum_{i=0}^{\log(n)-1} \left(\frac{1}{2} \right)^i \right) \cdot n^3 + n^{\log 4} && \text{Teorema di scambio base-argomento} \\
&= \left(\sum_{i=0}^{\log(n)-1} \left(\frac{1}{2} \right)^i \right) \cdot n^3 + n^2 \\
&\leq \left(\sum_{i=0}^{+\infty} \left(\frac{1}{2} \right)^i \right) \cdot n^3 + n^2 && \text{Estensione ad infinito della sommatoria}
\end{aligned}$$

Giunti a questo punto abbiamo:

$$T(n) \leq \left(\sum_{i=0}^{+\infty} \left(\frac{1}{2} \right)^i \right) \cdot n^3 + n^2$$

Tuttavia, possiamo riconoscere nella sommatoria la *serie geometrica infinita decrescente* e riscrivere $T(n)$ come:

$$\begin{aligned}
T(n) &\leq \frac{1}{1-\frac{1}{2}} n^3 + n^2 \\
&\leq 2n^3 + n^2
\end{aligned}$$

Abbiamo quindi dimostrato che $T(n) = O(n^3)$, tuttavia, poiché $T(n) \geq n^3$, possiamo anche affermare che $T(n) = \Omega(n^3)$, quindi $T(n) = \Theta(n^3)$.

NB. Abbiamo indicato n^3 come costo della prima chiamata perché di sicuro lo si dovrà pagare, ma in realtà il vero costo avrebbe dovuto essere $c \cdot n^3$ per qualche $c > 0$. Abbiamo rimosso c per semplicità, ma grazie alla *Proprietà di eliminazione delle costanti* sappiamo che considerarlo non sarebbe comunque servito.

Esempio 6 - Ulteriore esempio di utilizzo di una tabella dei costi.

Sia $T(n)$ la seguente funzione di ricorrenza:

$$T(n) = \begin{cases} 4T(n/2) + n^2 & n > 1 \\ 1 & n \leq 1 \end{cases}$$

Procediamo utilizzando una tabella come quella di prima.

Livello	Dim. input	Costo per chiamata	N. chiamate	Costo livello
0	n	n^2	1	n^2
1	$n/2$	$(n/2)^2$	4	$4(n/2)^2$
2	$n/4$	$(n/4)^2$	16	$16(n/4)^2$
...
i	$n/2^i$	$(n/2^i)^2$	4^i	$4^i(n/2^i)^2$
...
$\log(n) - 1$	$n/2^{\log(n)-1}$	$(n/2^{\log(n)-1})^2$	$4^{\log(n)-1}$	$4^{\log(n)-1}(n/2^{\log(n)-1})^2$
$\log n$	1	$T(1)$	$4^{\log n}$	$4^{\log n}$

Scriviamo come prima la funzione di ricorrenza sotto forma di sommatoria:

$$\begin{aligned}
T(n) &= \sum_{i=0}^{\log(n)-1} (4^i \cdot (n/2^i)^2) + 4^{\log n} \\
&= \left(\sum_{i=0}^{\log(n)-1} \frac{4^i}{2^{2i}} \right) \cdot n^2 + 4^{\log n} \\
&= \left(\sum_{i=0}^{\log(n)-1} \frac{2^{2i}}{2^{2i}} \right) \cdot n^2 + 4^{\log n} \\
&= \left(\sum_{i=0}^{\log(n)-1} 1 \right) \cdot n^2 + 4^{\log n} \\
&= \left(\sum_{i=0}^{\log(n)-1} 1 \right) \cdot n^2 + n^{\log 4} \\
&= \left(\sum_{i=0}^{\log(n)-1} 1 \right) \cdot n^2 + n^2 \\
&= n^2 \log n + n^2
\end{aligned}$$

Quindi $T(n) = \Theta(n^2 \log n)$.

3.1.2 Metodo della sostituzione

Con questo metodo si cerca di "indovinare" la *forma chiusa* di una *funzione di ricorrenza* e di dimostrarne la validità per induzione.

Esempio 7 - Esempio di intuizione corretta.

Sia $T(n)$ la seguente funzione di ricorrenza:

$$T(n) = \begin{cases} T(\lfloor n/2 \rfloor) + n & n > 1 \\ 1 & n \leq 1 \end{cases}$$

Proviamo ad indovinarne la forma chiusa:

$$\begin{aligned}
T(n) &= n + T(\lfloor n/2 \rfloor) \\
&= n + \frac{n}{2} + T(\lfloor n/4 \rfloor) \\
&= n + \frac{n}{2} + \frac{n}{4} + T(\lfloor n/8 \rfloor) \\
&= \dots \\
&= n + \frac{n}{2} + \frac{n}{4} + \frac{n}{8} + \dots + T(1) \\
&\leq 2n
\end{aligned}$$

Vedendo questa catena di uguaglianze possiamo supporre $T(n) = O(n)$. Proviamo quindi a dimostrarlo procedendo per induzione.

Caso base Dimostriamo la disequazione per $T(1)$:

$$T(1) = 1 \leq 1 \cdot c \quad \forall c \geq 1$$

Il caso base è verificato.

Passo induttivo Ipotizziamo che $T(k) \leq c \cdot k \quad \forall k < n$ e dimostriamo la disequazione per $T(n)$:

$$\begin{aligned} T(n) &= T(\lfloor n/2 \rfloor) + n \\ &\leq c \cdot \lfloor n/2 \rfloor + n && \text{Sostituzione} \\ &\leq c \cdot (n/2) + n && \text{Rimozione dell'intero inferiore} \\ &= (c/2 + 1) \cdot n \\ &\leq c \cdot n \end{aligned}$$

L'ultima disequazione è vera per:

$$\frac{c}{2} + 1 \leq c \Leftrightarrow c \geq 2$$

Anche il passo induttivo è verificato, dunque vale:

$$T(n) \leq c \cdot n \text{ per } \begin{cases} c \geq 1 & \text{Caso base} \\ c \geq 2 & \text{Passo induttivo} \end{cases}$$

Quelle condizioni sono verificate $\forall c \geq 2$ e $\forall n \geq 1 = m$. Abbiamo quindi dimostrato la correttezza della nostra ipotesi, cioè $T(n) = O(n)$.

Riusciamo a fare lo stesso sul limite inferiore? Al primo livello la funzione costa, di sicuro, almeno n . È quindi plausibile aspettarsi che $T(n) = \Omega(n)$. Procediamo di nuovo per induzione.

Caso base Dimostriamo la disequazione per $T(1)$:

$$T(1) = 1 \geq 1 \cdot d \quad \forall d \leq 1$$

Il caso base è verificato.

Passo induttivo Ipotizziamo che $T(k) \geq d \cdot k \quad \forall k < n$ e dimostriamo la disequazione per $T(n)$:

$$\begin{aligned} T(n) &= T(\lfloor n/2 \rfloor) + n \\ &\geq d \cdot \lfloor n/2 \rfloor + n && \text{Sostituzione} \\ &\geq d \cdot (n/2) - 1 + n && \text{Rimozione dell'intero inferiore} \\ &= (d/2 - 1/n + 1) \cdot n \\ &\geq d \cdot n \end{aligned}$$

L'ultima disequazione è vera per:

$$\frac{d}{2} - \frac{1}{n} + 1 \geq d \Leftrightarrow d \leq 2 - \frac{2}{n}$$

Anche il passo induttivo è verificato, dunque vale:

$$T(n) \geq d \cdot n \text{ per } \begin{cases} d \leq 1 & \text{Caso base} \\ d \leq 2 - \frac{2}{n} & \text{Passo induttivo} \end{cases}$$

Quelle condizioni sono verificate per $d = 1$ e $\forall n \geq 2 = m$. Abbiamo quindi verificato la correttezza della nostra ipotesi, cioè $T(n) = \Omega(n)$.

Avendo dimostrato che $T(n)$ è sia un $O(n)$ che un $\Omega(n)$, ne consegue che $T(n)$ è anche un $\Theta(n)$.

NB. Nel precedente esempio abbiamo provato per induzione che

$$T(n) = \begin{cases} T(\lfloor n/2 \rfloor) + n & n > 1 \\ 1 & n \leq 1 \end{cases} = \Omega(n)$$

Tuttavia, è possibile giungere allo stesso risultato senza ricorrere alla ricorsione. Ricordiamo la definizione di Ω :

$$\exists d > 0, \exists m \geq 0 : f(n) \geq d \cdot g(n) \quad \forall n \geq m$$

Vale la seguente catena di disequazioni:

$$T(n) = T(\lfloor n/2 \rfloor) + n \geq n \geq d \cdot n \quad \forall d \leq 1$$

Questa condizione è la stessa del caso base della dimostrazione per induzione, dunque $T(n) = \Omega(n)$.

Esempio 8 - Esempio di intuizione errata.

Sia $T(n)$ la seguente funzione di ricorrenza:

$$T(n) = \begin{cases} T(n-1) + n & n > 1 \\ 1 & n \leq 1 \end{cases}$$

Risolvendo questa funzione di ricorrenza per livelli vale:

$$\begin{aligned} T(n) &= n + T(n-1) \\ &= (n-1) + n + T(n-2) \\ &= (n-2) + (n-1) + n + T(n-3) \\ &= \dots \\ &= 1 + \dots + (n-2) + (n-1) + n \\ &= \sum_{i=1}^n i = \frac{n(n-1)}{2} = \Theta(n^2) \end{aligned}$$

Supponiamo però di voler provare a dimostrare che $T(n) = O(n)$. Procediamo quindi per induzione.

Caso base Dimostriamo la disequazione per $T(1)$:

$$T(1) = 1 \leq 1 \cdot c \quad \forall c \geq 1$$

Il caso base è verificato.

Passo induttivo Ipotizziamo che $T(k) \leq c \cdot k \quad \forall k < n$ e dimostriamo la disequazione per $T(n)$:

$$\begin{aligned} T(n) &= T(n-1) + n \\ &\leq c \cdot (n-1) + n && \text{Sostituzione} \\ &= c \cdot n - c + n \\ &= (c+1)n - c \\ &\leq (c+1)n && \text{Rimozione elemento negativo} \\ &\leq c \cdot n \end{aligned}$$

L'ultima disequazione è impossibile poiché per essere vera dovrebbe vale:

$$c+1 \leq c$$

Dunque, $T(n) \neq O(n)$.

Esempio 9 - Limiti inferiori e superiori.

Sia $T(n)$ la seguente funzione di ricorrenza:

$$T(n) = \begin{cases} T(\lfloor n/2 \rfloor) + T(\lceil n/2 \rceil) + 1 & n > 1 \\ 1 & n \leq 1 \end{cases}$$

Qual è il costo di questa funzione?

Iniziamo di nuovo procedendo per livelli.

Livello	Dim. input	Costo per chiamata	N. chiamate	Costo livello
0	n	1	1	1
1	$n/2$	2	2	2
2	$n/4$	4	4	4
...
i	$n/2^i$	2^i	2^i	2^i
...
$\log(n) - 1$	$n/2^{\log(n)-1}$	1	$2^{\log(n)-1}$	$2^{\log(n)-1}$
$\log n$	1	$T(1)$	$2^{\log n}$	$2^{\log n}$

Scrivendo ora la funzione di ricorrenza come sommatoria otteniamo:

$$\begin{aligned} T(n) &= \sum_{i=0}^{\log n} 2^i \\ &= 2^0 + 2^1 + \dots + 2^{\log(n)-1} + 2^{\log n} \\ &= 2^0 + 2^1 + \dots + \frac{2^{\log n}}{2} + 2^{\log n} && \text{Teorema del prodotto} \\ &= 2^0 + 2^1 + \dots + \frac{n^{\log 2}}{2} + n^{\log 2} && \text{Teorema di scambio base-argomento} \\ &= 2^0 + 2^1 + \dots + \frac{n}{2} + n = \Theta(n) \end{aligned}$$

Quindi, $T(n) = \Theta(n)$. Adesso però proviamo a procedere per sostituzione. Dimostriamo quindi che $T(n) = O(n)$.

Caso base Dimostriamo la disequazione per $T(1)$:

$$T(1) = 1 \leq 1 \cdot c \quad \forall c \geq 1$$

Il caso base è verificato.

Passo induttivo Ipotizziamo che $T(k) \leq c \cdot k \quad \forall k < n$ e dimostriamo la disequazione per $T(n)$:

$$\begin{aligned} T(n) &= T(\lfloor n/2 \rfloor) + T(\lceil n/2 \rceil) + 1 \\ &\leq c \cdot (\lfloor n/2 \rfloor) + c \cdot (\lceil n/2 \rceil) + 1 && \text{Sostituzione} \\ &= c \cdot n + 1 \\ &\leq c \cdot n \end{aligned}$$

L'ultima disequazione è impossibile poiché si riduce a $1 \leq 0$ che è un'affermazione falsa.

Non siamo riusciti a dimostrare che $T(n) = O(n)$, tuttavia noi sappiamo che in realtà lo è. Cosa c'è di sbagliato? Abbiamo ottenuto a sinistra un termine troppo grande, quindi proviamo a partire da un'ipotesi più ristretta. Proviamo con $O(n - b)$.

Caso base Dimostriamo la disequazione per $T(1)$:

$$T(1) = 1 \leq 1 \cdot c - b \quad \forall c \geq b + 1$$

Il caso base è verificato.

Passo induttivo Ipotizziamo che $\exists b > 0 : T(k) \leq c \cdot k - b \quad \forall k < n$ e dimostriamo la disequazione per $T(n)$:

$$\begin{aligned} T(n) &= T(\lfloor n/2 \rfloor) + T(\lceil n/2 \rceil) + 1 \\ &\leq c \cdot (\lfloor n/2 \rfloor) - b + c \cdot (\lceil n/2 \rceil) - b + 1 && \text{Sostituzione} \\ &= c \cdot n - 2b + 1 \\ &\leq c \cdot n - b \end{aligned}$$

L'ultima disequazione è vera per:

$$-2b + 1 \leq -b \Leftrightarrow b \geq 1$$

Quindi, per verificare sia il caso base che il passo induttivo, è sufficiente porre $b = 1$ e $c = 2$. Inoltre, questo vale per ogni $n \geq 1$, quindi poniamo anche $m = 1$. Abbiamo dunque provato che $T(n) = O(n - b) = O(n)$.

A questo punto resta da dimostrare soltanto il limite inferiore, cioè che $T(n) = \Omega(n)$.

Caso base Dimostriamo la disequazione per $T(1)$:

$$T(1) = 1 \geq 1 \cdot d \quad \forall d \leq 1$$

Il caso base è verificato.

Passo induttivo Ipotizziamo che $T(k) \geq d \cdot k \quad \forall d \leq k$ e dimostriamo la disequazione per $T(n)$:

$$\begin{aligned} T(n) &= T(\lfloor n/2 \rfloor) + T(\lceil n/2 \rceil) + 1 \\ &\geq d \cdot (\lfloor n/2 \rfloor) + d \cdot (\lceil n/2 \rceil) + 1 && \text{Sostituzione} \\ &= d \cdot n + 1 \\ &\geq d \cdot n \end{aligned}$$

Chiaramente, l'ultima disequazione è sempre vera, dunque è dimostrato che $T(n) = \Omega(n)$ e, di conseguenza, anche che $T(n) = \Theta(n)$.

Esempio 10 - Problemi con i casi base.

Sia $T(n)$ la seguente funzione di ricorrenza:

$$T(n) = \begin{cases} 2T(\lfloor n/2 \rfloor) + n & n > 1 \\ 1 & n \leq 1 \end{cases}$$

Calcoliamo la forma chiusa procedendo per livelli.

$$\begin{aligned} T(n) &= n + 2T(\lfloor n/2 \rfloor) \\ &= n + 2(n/2) + 4T(\lfloor n/4 \rfloor) \\ &= n + 2(n/2) + 4(n/4) + 8T(\lfloor n/8 \rfloor) \\ &= \dots \\ &= n + 2(n/2) + 4(n/4) + \dots + 2^{\log(n)-1}(n/2^{\log(n)-1}) + 2^{\log(n)} \cdot T(1) \\ &= n \log n = \Theta(n \log n) \end{aligned}$$

Passiamo al procedimento per sostituzione e proviamo a dimostrare che $T(n) = O(n \log n)$.

Caso base Dimostriamo la disequazione per $T(1)$:

$$T(1) = 1 \leq c \cdot 1 \log 1 = 0 \Rightarrow 1 \leq 0$$

Come nell'esempio precedente, pur avendo fatto un tentativo corretto non siamo riusciti a dimostrarlo. Sta volta il problema è nel caso base, tuttavia, non dimentichiamo che stiamo lavorando con notazioni asintotiche, che quindi valgono a partire da un certo valore di n .

Proviamo quindi a partire da un valore di $n > 1$:

$$\begin{aligned} T(2) &= 2T(\lfloor 2/2 \rfloor) + 2 = 2T(1) + 2 = 4 \leq c \cdot 2 \log 2 \Leftrightarrow c \geq 2 \\ T(3) &= 2T(\lfloor 3/2 \rfloor) + 3 = 2T(1) + 3 = 5 \leq c \cdot 3 \log 3 \Leftrightarrow c \geq \frac{5}{3 \log 3} \\ T(4) &= 2T(\lfloor 4/2 \rfloor) + 2 = 2T(2) + 4 \end{aligned}$$

Non serve dimostrare l'ultima disequazione perché è basata su $T(2)$ che abbiamo già verificato.

Passo induttivo Ipotizziamo che $T(k) \leq c \cdot k \log k \quad \forall k < n$ e dimostriamo la disequazione per $T(n)$:

$$\begin{aligned} T(n) &= 2T(\lfloor n/2 \rfloor) + n \\ &\leq 2 \cdot c \cdot (\lfloor n/2 \rfloor) \log(\lfloor n/2 \rfloor) + n && \text{Sostituzione} \\ &\leq 2 \cdot c \cdot (n/2) \log(n/2) + n && \text{Rimozione intero inferiore} \\ &= c \cdot n(\log n - 1) + n && \text{Teorema del rapporto} \\ &= c \cdot n \log n - c \cdot n + n \\ &\leq c \cdot n \log n \end{aligned}$$

L'ultima disequazione è vera per:

$$-c \cdot n + n \leq 0 \Leftrightarrow -c \cdot n \leq -n \Leftrightarrow c \geq 1$$

A questo punto vale che $T(n) \leq c \cdot n \log n$ per:

- Caso base: $\forall c \geq 2, \forall c \geq \frac{5}{3 \log 3}$;
- Passo induttivo: $\forall c \geq 1$

Siccome in tutti e tre i casi c è in una relazione \geq con il secondo termine, è sufficiente porre $c = \max \left\{ 1, 2, \frac{5}{3 \log 3} \right\}$. Inoltre, poiché abbiamo dimostrato il caso base per $n = 2$ e $n = 3$, prendiamo $m = 2$. Ecco provato che $T(n) = O(n)$.

La dimostrazione del limite inferiore non è necessaria poiché alla prima invocazione di T pagheremo certamente n , dunque $T(n) = \Omega(n)$.

Ricapitolando, questo metodo risolutivo si basa sul tentare di intuire la forma chiusa delle funzioni di ricorrenza e sul dimostrare per induzione, andando a sostituire il tentativo all'interno della funzione, la correttezza dell'intuizione.

3.1.3 Metodo delle ricorrenze comuni

Molte funzioni di ricorrenza sono risolubili velocemente mediante l'applicazione di un qualche teorema. Esistono diversi teoremi che sono specifici per particolari classi di funzioni di ricorrenza.

Ricorrenze lineari con partizione bilanciata Per questa classe di *funzioni di ricorrenza* esistono due versioni di uno stesso teorema. Vediamo prima la versione *ridotta*.

Definizione 19 - Teorema delle ricorrenze lineari con partizione bilanciata - Rid.

Siano a e b costanti intere tali che $a \geq 1$ e $b \geq 2$. Siano poi c e β costanti reali tali che $c > 0$ e $\beta \geq 0$. Sia $T(n)$ una funzione di ricorrenza della seguente forma:

$$T(n) = \begin{cases} aT(n/b) + cn^\beta & n > 1 \\ d & n \leq 1 \end{cases}$$

Allora, posto $\alpha = \frac{\log a}{\log b} = \log_b a$ vale:

$$T(n) = \begin{cases} \Theta(n^\alpha) & \alpha > \beta \\ \Theta(n^\alpha \log n) & \alpha = \beta \\ \Theta(n^\beta) & \alpha < \beta \end{cases}$$

NB. Di seguito, nella dimostrazione, ipotizziamo che $n = b^k \Rightarrow k = \log n$ così da semplificare i calcoli.

Dimostrazione. Sia $T(n)$ la seguente funzione di ricorrenza:

$$T(n) = \begin{cases} aT(n/b) + cn^\beta & n > 1 \\ d & n \leq 1 \end{cases}$$

Calcoliamone la *forma chiusa* procedendo per *livelli*:

Livello	Dim. input	Costo per chiamata	N. chiamate	Costo livello
0	b^k	$cb^{k\beta}$	1	$cb^{k\beta}$
1	b^{k-1}	$cb^{(k-1)\beta}$	a	$acb^{k\beta}$
2	b^{k-2}	$cb^{(k-2)\beta}$	a^2	$a^2cb^{k\beta}$
...
i	b^{k-i}	$cb^{k\beta}$	a^i	$a^i cb^{k\beta}$
...
$k-1$	b	cb^β	a^{k-1}	$a^{k-1}cb^{k\beta}$
k	1	d	a^k	$a^k d$

Sommando i costi di tutti i livelli si ottiene la seguente espressione:

$$T(n) = da^k + cb^{k\beta} \sum_{i=0}^{k-1} \frac{a^i}{b^{i\beta}} = da^k + cb^{k\beta} \sum_{i=0}^{k-1} \left(\frac{a}{b^\beta} \right)^i$$

Possiamo osservare che:

$$a^k = a^{\log_b n} = a^{\frac{\log n}{\log b}} = 2^{\log a \frac{\log n}{\log b}} = n^{\frac{\log a}{\log b}} = n^\alpha$$

$$\alpha = \frac{\log a}{\log b} \Rightarrow \alpha \log b = \log a \Rightarrow \log b^\alpha = \log a \Rightarrow a = b^\alpha$$

A questo punto, se poniamo $q = \frac{\alpha}{b^\beta} = \frac{b^\alpha}{b^\beta} = b^{\alpha-\beta}$ possiamo riscrivere $T(n)$ come:

$$T(n) = da^k + cb^{k\beta} \sum_{i=0}^{k-1} \left(\frac{\alpha}{b^\beta} \right)^i = dn^\alpha + cb^{k\beta} \sum_{i=0}^{k-1} q^i$$

Passiamo dunque alla dimostrazione, caso per caso, del teorema.

Caso $\alpha > \beta$ Se $\alpha > \beta$ segue che $q = b^{\alpha-\beta} > 1$, quindi:

$$\begin{aligned}
T(n) &= dn^\alpha + cb^{k\beta} \sum_{i=0}^{k-1} q^i \\
&= dn^\alpha + cb^{k\beta} \frac{q^k - 1}{q - 1} && \text{Serie geometrica finita} \\
&\leq dn^\alpha + cb^{k\beta} \frac{q^k}{q - 1} \\
&= dn^\alpha + \frac{cb^{k\beta} \alpha^k}{b^{k\beta}(q-1)} && \text{Sostituzione di } q \\
&= dn^\alpha + \frac{c\alpha^k}{q-1} \\
&= dn^\alpha + \frac{cn^\alpha}{q-1} && a^k = n^\alpha \\
&= n^\alpha \left(d + \frac{c}{q-1} \right)
\end{aligned}$$

Quindi, $T(n) = O(n^\alpha)$ e, per via della componente dn^α , $T(n) = \Omega(n^\alpha)$, dunque $T(n) = \Theta(n^\alpha)$.

Caso $\alpha = \beta$ Se $\alpha = \beta \Rightarrow q = b^{\alpha-\beta} = 1$, quindi:

$$\begin{aligned}
T(n) &= dn^\alpha + cb^{k\beta} \sum_{i=0}^{k-1} q^i \\
&= dn^\alpha + cb^{k\alpha} k && \alpha = \beta \wedge q^i = 1^i = 1 \\
&= dn^\alpha + cn^\alpha k && b^\alpha = a \wedge a^k = n^\alpha \\
&= n^\alpha (d + ck) \\
&= n^\alpha \left(d + c \frac{\log n}{\log b} \right) && k = \log_b n = \frac{\log n}{\log b}
\end{aligned}$$

Quindi, $T(n) = \Theta(n^\alpha \log n)$.

Caso $\alpha < \beta$ Se $\alpha < \beta \Rightarrow q = b^{\alpha-\beta} < 1$, quindi:

$$\begin{aligned}
T(n) &= dn^\alpha + cb^{k\beta} \sum_{i=0}^{k-1} q^i \\
&= dn^\alpha + cn^{k\beta} \frac{q^k - 1}{q - 1} && \text{Serie geometrica finita} \\
&= dn^\alpha + cb^{k\beta} \frac{1 - q^k}{1 - q} && \text{Cambio di segno} \\
&\leq dn^\alpha + cb^{k\beta} \frac{1}{1 - q} \\
&= dn^\alpha + \frac{cn^\beta}{1 - q} && b^k = n
\end{aligned}$$

Quindi, $T(n) = O(n^\beta)$ e, per il termine non ricorsivo $cb^{k\beta} = cn^\beta$, $T(n) = \Omega(n^\beta)$, dunque $T(n) = \Theta(n^\beta)$. \square

Definizione 20 - Teorema delle ricorrenze lineari con partizione bilanciata - Est.

Siano $a \geq 1$, $b > 1$ e $f(n)$ una funzione asintoticamente positiva. Sia poi $T(n)$ una funzione di ricorrenza della seguente forma:

$$T(n) = \begin{cases} aT(n/b) + f(n) & n > 1 \\ d & n \leq 1 \end{cases}$$

Valgono le seguenti casistiche:

- 1) Se $\exists \epsilon > 0 : f(n) = O(n^{\log_b(a) - \epsilon}) \Rightarrow T(n) = \Theta(n^{\log_b a})$
- 2) Se $f(n) = \Theta(n^{\log_b a}) \Rightarrow T(n) = \Theta(f(n) \log n)$
- 3) Se $\exists \epsilon > 0 : f(n) = \Omega(n^{\log_b(a) + \epsilon}) \wedge \exists c : 0 < c < 1, \exists m \geq 0 : af(n/b) \leq cf(n) \quad \forall n \geq m \Rightarrow T(n) = \Theta(f(n))$

NB. Nonostante non siano presenti nelle definizioni, i teoremi appena visti valgono anche per funzioni di ricorrenza espresse usando gli operatori di *intero-inferiore* e *intero-superiore*.

Esempio 11 - Applicazione della forma estesa del teorema.

Sia $T(n)$ la seguente funzione di ricorrenza:

$$T(n) = \begin{cases} 9T(n/3) + n \log n & n > 1 \\ 1 & n \leq 1 \end{cases}$$

Trovarne la forma chiusa.

Proviamo ad applicare il teorema appena visto. Poiché il fattore non ricorsivo, cioè $n \log n$, non è una semplice potenza di n e, contemporaneamente, è una funzione crescente, quindi asintoticamente positiva, applichiamo la versione estesa del teorema.

Ricorrenza	a	b	$\log_b a$	Caso	Funzione
$9T(n/3) + n \log n$	9	3	2	(1)	$\Theta(n^2)$

In questo esempio si applica la prima casistica del teorema perché $f(n) = n \log n$ è certamente un $O(n^2)$, ma cresce comunque meno rispetto a n^2 e più di n , quindi se prendiamo $\epsilon < 1$, vale:

$$f(n) = n \log n = O(n^{\log_b(a)-\epsilon}) = O(n^{2-\epsilon})$$

E quindi, per il primo caso del Teorema delle ricorrenze lineari con partizione bilanciata esteso, $T(n) = \Theta(n^2)$.

Esempio 12 - Applicazione della forma ridotta del teorema.

Sia $T(n)$ la seguente funzione di ricorrenza:

$$T(n) = \begin{cases} T(2n/3) + 1 & n > 1 \\ 1 & n \leq 1 \end{cases}$$

Trovare la forma chiusa.

Questa volta possiamo applicare la versione ridotta del teorema.

Ricorrenza	a	b	$\log_b a$	β	Caso	Funzione
$T(2n/3) + 1$	1	3/2	0	0	(2)	$\Theta(\log n)$

In questo caso $\alpha = \log_b a = \log_{3/2} 1 = 0$ e poiché anche $\beta = 0$ si applica la seconda casistica, cioè $T(n) = \Theta(\log n)$.

In tutti i casi in cui è applicabile la versione ridotta del teorema, si può usare anche quella estesa. Con questa funzione, ad esempio, avremmo potuto notare che:

$$f(n) = 1 = \Theta(1) = \Theta(n^0) = \Theta(n^{\log_b a})$$

E quindi, per il secondo caso, avremmo ottenuto:

$$T(n) = \Theta(f(n) \log n) = \Theta(1 \log n) = \Theta(\log n)$$

Esempio 13 - Applicazione della forma estesa del teorema.

Sia $T(n)$ la seguente funzione di ricorrenza:

$$T(n) = \begin{cases} 3T(n/4) + n \log n & n > 1 \\ 1 & n \leq 1 \end{cases}$$

Trovare la forma chiusa.

$f(n) = n \log n$, quindi applico il teorema delle ricorrenze lineari con partizione bilanciata nella sua forma estesa.

Ricorrenza	a	b	$\log_b a$	Caso	Funzione
$3T(n/4) + n \log n$	3	4	≈ 0.79	(3)	$\Theta(f(n))$

Si tratta del terzo caso perché $f(n) = n \log n = \Omega(n) = \Omega(n^{\log_b(a)-\epsilon})$ con $\epsilon < 1 - \log_b a = 1 - \log_4 3 \approx 0.208$.

Tuttavia, dobbiamo anche dimostrare che $\exists c : 0 < c < 1$ e $\exists m \geq 0$ tali che:

$$af(n/b) \leq cf(n) \quad \forall n \geq m$$

Vale quanto segue:

$$\begin{aligned} af(n/b) &= a(n/b) \log(n/b) && \text{Sostituzione} \\ &= 3(n/4 \log(n/4)) \\ &= \frac{3}{4}n(\log(n) - \log(4)) && \text{Teorema del rapporto} \\ &\leq \frac{3}{4}n \log n \\ &\leq cn \log n \end{aligned}$$

L'ultima disequazione è vera per $c = \frac{3}{4}$ e per qualsiasi m , quindi è dimostrato che $T(n) = \Theta(f(n)) = \Theta(n \log n)$.

Esempio 14 - Esempio di inapplicabilità del teorema.

Sia $T(n)$ la seguente funzione di ricorrenza:

$$T(n) = \begin{cases} 2T(n/2) + n \log n & n > 1 \\ 1 & n \leq 1 \end{cases}$$

Trovare la forma chiusa.

In questo caso vale $\log_b a = \log_2 2 = 1$ e $f(n) = n \log n = \Omega(n) = \Theta(n \log n) = O(n^2)$, ma poiché:

- $n \log n \neq O(n^{1-\epsilon}) \quad \forall \epsilon > 0$
- $n \log n \neq \Theta(n^1)$
- $n \log n \neq \Omega(n^{1+\epsilon}) \quad \forall \epsilon > 0$

non è possibile applicare nessuna casistica del teorema, dunque è necessario usare altre tecniche.

Ricorrenze lineari di ordine costante

Definizione 21 - Teorema delle ricorrenze lineari di ordine costante.

Siano a_1, a_2, \dots, a_h costanti intere non negative con h costante e positiva. Siano poi c e β costanti reali tali che $c > 0$ e $\beta \geq 0$. Sia infine $T(n)$ definita dalla seguente funzione di ricorrenza:

$$T(n) = \begin{cases} \sum_{i=1}^h (a_i T(n-i)) + cn^\beta & n > m \\ \Theta(1) & n \leq m \leq h \end{cases}$$

Allora, posto $a = \sum_{i=1}^h a_i$ vale:

- 1) Se $a = 1 \Rightarrow T(n) = \Theta(n^{\beta+1})$
- 2) Se $a \geq 2 \Rightarrow T(n) = \Theta(a^n n^\beta)$

Esempio 15 - Applicazione del teorema delle ricorrenze lineari di ordine costante.

Sia $T(n)$ la seguente funzione di ricorrenza:

$$T(n) = \begin{cases} T(n-10) + n^2 & n > 1 \\ \Theta(1) & n \leq 1 \end{cases}$$

Trovare la forma chiusa.

Proviamo ad applicare il teorema delle ricorrenze lineari di ordine costante.

Ricorrenza	a	β	Caso	Funzione
$T(n-10) + n^2$	1	2	(1)	$\Theta(n^{\beta+1})$

In questo esempio vale il caso 1 perché il coefficiente di $T(n-10)$ è 1, quindi, per il teorema, risulta $T(n) = \Theta(n^{\beta+1}) = \Theta(n^3)$.

Esempio 16 - Applicazione del teorema delle ricorrenze lineari di ordine costante.

Sia $T(n)$ la seguente funzione di ricorrenza:

$$T(n) = \begin{cases} T(n-2) + T(n-1) + 1 & n > 1 \\ \Theta(1) & n \leq 1 \end{cases}$$

Trovare la forma chiusa.

Applichiamo di nuovo il teorema appena enunciato.

Ricorrenza	a	β	Caso	Funzione
$T(n-2) + T(n-1) + 1$	2	0	(2)	$\Theta(a^n n^\beta)$

Qui si applica il caso 2 del teorema e poiché $\beta = 0$ vale $T(n) = \Theta(a^n n^\beta) = \Theta(2^n)$, cioè questa funzione ha costo esponenziale.

Problema 2 - Funzioni di ricorrenza parametriche.

Siano $T(n)$ e $T'(n)$ le seguenti funzioni di ricorrenza:

$$T(n) = \begin{cases} 7T(n/2) + n^2 & n > 1 \\ 1 & n \leq 1 \end{cases} \quad \text{Algoritmo } A^3$$

$$T'(n) = \begin{cases} aT'(n/4) + n^2 & n > 1 \\ 1 & n \leq 1 \end{cases} \quad \text{Algoritmo } A'$$

Trovare il massimo valore intero di a che renda A' asintoticamente più veloce di A .

Iniziamo calcolando la complessità di A e, in particolare, applichiamo il **Teorema delle ricorrenze lineari con partizione bilanciata - Rid.**

Ricorrenza	a	b	$\log_b a$	β	Caso	Funzione
$7T(n/2) + n^2$	7	2	≈ 2.81	2	(1)	$\Theta(n^{\log_2 7})$

Ora, poiché in $T'(n)$ compare $T'(n/4)$, trasformo $\log_2 7$ in un qualche \log_4 :

$$\begin{aligned} \log_2 7 &= \frac{\log_4 7}{\log_4 2} && \text{Cambio di base} \\ &= \frac{\log_4 7}{1/2} \\ &= 2 \log_4 7 \\ &= \log_4 7^2 && \text{Teorema della potenza} \\ &= \log_4 49 \end{aligned}$$

A questo punto studio $T'(n)$ in base al variare di a :

$$\begin{aligned} a < 16 &\Rightarrow \alpha = \log_4 a < 2 \Rightarrow \alpha < \beta \Rightarrow T'(n) = \Theta(n^2) = O(T(n)) \\ a = 16 &\Rightarrow \alpha = \log_4 a = 2 \Rightarrow \alpha = \beta \Rightarrow T'(n) = \Theta(n^2 \log n) = O(T(n)) \\ 16 < a \leq 49 &\Rightarrow \alpha = \log_4 a > 2 \Rightarrow \alpha > \beta \Rightarrow T'(n) = \Theta(n^\alpha) = O(T(n)) \\ a > 49 &\Rightarrow \alpha = \log_4 a > 2 \Rightarrow \alpha > \beta \Rightarrow T'(n) = \Theta(n^\alpha) = \Omega(T(n)) \end{aligned}$$

Nel terzo caso $2 < a \leq \log_4 49$ quindi $T'(n)$ cresce al più come $T(n)$, mentre per $a > 49$, $\alpha > \log_4 49$ e quindi $T'(n)$ cresce più di $T(n)$.

Quindi, il massimo valore intero di a che rende A' asintoticamente più veloce di A è 49.

³Funzione di ricorrenza dell'algoritmo di Strassen

Capitolo Nr.4

Strutture dati

4.1 Introduzione

Iniziamo con delle definizioni:

Definizione 22 - Dato.

In un linguaggio di programmazione, un dato è un valore che una variabile può assumere.

Definizione 23 - Tipo di dato astratto.

Un tipo di dato astratto è un modello matematica dato da una collezione di valori e un insieme di operazioni ammesse su questi valori.

Definizione 24 - Tipo di dato primitivo.

Un tipo di dato fornito direttamente da un linguaggio è detto essere primitivo.

Definizione 25 - Specifica e implementazione di un tipo di dato.

Per ogni tipo di dato astratto si definiscono due livelli:

- *Specifica: costituisce l'interfaccia di utilizzo del tipo di dato e ne nasconde i dettagli implementativi;*
- *Implementazione: è la realizzazione vera e propria del tipo di dato;*

I seguenti sono esempi di *specifica* e *implementazione* di un tipo di dato:

Specifica	Implementazione
Numeri reali	IEEE754
Pile	Pile basate su vettori Pile basate su puntatori
Code	Code basate su vettori circolari Code basate su puntatori

Definizione 26 - Struttura di dati.

Una struttura di dati è una collezione di dati caratterizzata dalla struttura della collezione, piuttosto che dal tipo dei dati contenuti.

Una *struttura di dati* è caratterizzata da un insieme di operatori che consentono di manipolarne la struttura e da un modo sistematico di organizzare i dati in essa contenuti.

Le *strutture di dati* possono essere categorizzate sulla base di tre parametri:

- *Lineari/Non lineari*: se è presente o meno una sequenza;
- *Dinamiche/Statiche*: se è possibile modificare o meno la dimensione della struttura dopo averla creata;
- *Omogenee/Disomogenee*: se i dati contenuti sono tutti dello stesso tipo o di tipi diversi;

4.2 Strutture di dati astratte

4.2.1 Sequenze

Definizione 27 - Sequenza.

Una sequenza è una struttura dati dinamica e lineare rappresentante una sequenza ordinata di valori che possono comparire anche più di una volta.

NB. L'ordine dei valori all'interno di una *sequenza* è importante!

Operazioni ammesse Su una *sequenza* sono ammesse le seguenti operazioni:

- Data una posizione è possibile aggiungere o togliere elementi, cioè se $s = s_1, s_2, \dots, s_n$ è la *sequenza*, l'elemento s_i è in posizione pos_i , inoltre, esistono posizioni fittizie quali pos_0 e pos_{n+1} che rappresentano la posizione del primo elemento e dell'elemento successivo a quello in pos_n ;
- Accesso diretto alla testa o alla coda;
- Accesso sequenziale alle altre posizioni;

Specifica

Frammento 8 - Sequenza.

```
% Restituisce true se la sequenza è vuota
boolean isEmpty()

% Restituisce true se  $p = pos_0$  o se  $p = pos_{n+1}$ 
boolean finished(POS p)

% Restituisce la posizione del primo elemento
POS head()

% Restituisce la posizione dell'ultimo elemento
POS tail()
```

```

% Restituisce la posizione dell'elemento che segue p
POS next(POS p)
% Restituisce la posizione dell'elemento che precede p
POS prev(POS p)
% Inserisce l'elemento v di tipo ITEM nella posizione p e restituisce
% la posizione del nuovo elemento, che diviene il predecessore di p
POS insert(POS p, ITEM v)
% Rimuove l'elemento contenuto nella posizione p e restituisce la posizione
% del successore di p, che diviene il successore del predecessore di p
POS remove(POS p)
% Legge l'elemento di tipo ITEM contenuto nella posizione p
ITEM read(POS p)
% Scrive l'elemento v di tipo ITEM nella posizione p
write(POS p, ITEM v)

```

4.2.2 Insiemi

Definizione 28 - Insieme.

Un insieme è una struttura dati dinamica e non lineare che memorizza una collezione non ordinata di valori non ripetuti.

NB. L'ordinamento fra elementi è dato dall'eventuale relazione d'ordine definita sul tipo degli elementi stessi.

Operazioni ammesse Su un *insieme* sono ammessi diversi tipi di operazioni:

- *Operazioni di base:*
 - Inserimento;
 - Rimozione;
 - Test di appartenenza;
- *Operazioni di ordinamento:* estrazione valore massimo/minimo;
- *Operazioni insiemistiche:*
 - Unione;
 - Intersezione;
 - Differenza;
- *Iterazione sugli elementi:* esecuzione di operazioni su tutti gli elementi dell'*insieme* (e.g. `foreach (x ∈ S) do ...`);

Specifica

Frammento 9 - Insieme.

```
% Restituisce la cardinalità dell'insieme
int size()

% Restituisce true se  $x$  è contenuto nell'insieme
boolean contains(ITEM x)

% Inserisce  $x$  nell'insieme, se non è già presente
insert(ITEM x)

% Rimuove  $x$  dall'insieme, se è presente
remove(ITEM x)

% Restituisce un nuovo insieme che è l'unione di  $A$  e  $B$ 
Set union(Set A, Set B)

% Restituisce un nuovo insieme che è l'intersezione di  $A$  e  $B$ 
Set intersection(Set A, Set B)

% Restituisce un nuovo insieme che è la differenza di  $A$  e  $B$ 
Set difference(Set A, Set B)
```

4.2.3 Dizionari

Definizione 29 - Dizionario.

Un dizionario è una struttura dati dinamica che rappresenta il concetto matematico di relazione univoca, o associazione chiave-valore, $R : D \rightarrow C$, dove D è un insieme dominio i cui elementi sono detti chiavi e C è un insieme codominio di elementi detti valori.

Operazioni ammesse Su un *dizionario* sono ammesse le seguenti operazioni:

- Data una chiave è possibile accedere al valore associato o a `nil` se la chiave non è associata a nulla;
- Creazione di una nuova associazione chiave-valore, eventualmente sostituendo il precedente valore associato a quella chiave;
- Eliminazione di un'associazione chiave-valore;

Specifica

Frammento 10 - Dizionario.

```
% Restituisce il valore associato alla chiave  $k$  se presente, nil altrimenti
ITEM lookup(ITEM k)

% Associa il valore  $v$  alla chiave  $k$ 
insert(ITEM k, ITEM v)

% Rimuove l'associazione della chiave  $k$ 
remove(ITEM k)
```

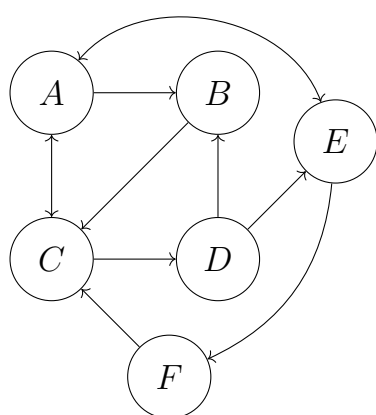
4.2.4 Grafi e alberi

Definizione 30 - Grafo.

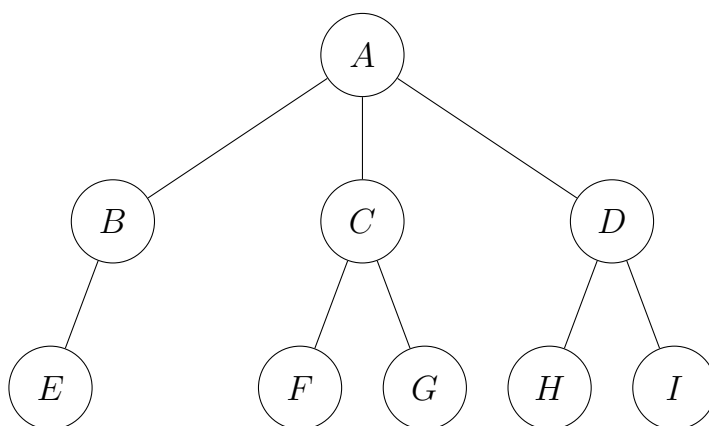
Un grafo è una struttura composta da un insieme di elementi detti nodi o vertici e un insieme di coppie di nodi, ordinate o meno, dette archi.

Definizione 31 - Albero ordinato.

Un albero ordinato è una struttura dati dinamica composta da un insieme finito di elementi detti nodi. Uno di questi nodi è detto radice, tutti gli altri sono partizionati in insiemi ordinati e disgiunti che sono anch'essi alberi ordinati.



(a) Grafo con archi orientati



(b) Albero ordinato

Fig. 4.1: Grafo VS Albero ordinato

Operazioni ammesse Oltre a inserimento e rimozione, le operazioni ammesse su *grafi* e *alberi* ruotano attorno alla possibilità di accedere a tutti gli elementi delle strutture secondo diverse *visite*¹.

4.2.5 Criticità nell'implementazione di strutture dati astratte

Quando si implementa una *struttura di dati astratta* può essere naturale prediligere una particolare *struttura di dati elementare* piuttosto che un'altra. Ad esempio, viene naturale implementare una *sequenza* usando una *lista*, o un *albero astratto* come *albero di puntatori*. Tuttavia, esistono possibilità meno scontate, come l'utilizzo di un vettore di booleani per l'implementazione di un *insieme*, o un *albero* implementato come *vettore dei padri*.

La *struttura di dati elementare* scelta per l'implementazione di una *struttura dati astratta* si ripercuote sull'efficienza delle singole operazioni. Per esempio, un *dizionario* implementato come *tabella hash* permette di avere *complessità* $O(1)$ nella funzione `lookup` e $O(n)$ nella ricerca del minimo. Invece, lo stesso *dizionario* implementato come *albero* porta la *complessità* della `lookup` a $O(\log n)$, ma riduce a $O(1)$ quella per la ricerca del minimo.

¹Tutte le *visite* saranno discusse nel dettaglio nel prossimo capitolo

4.3 Strutture di dati elementari

Vediamo ora le possibili implementazioni delle più comuni *strutture dati elementari*: *liste*, *pile* e *code*.

4.3.1 Liste

Definizione 32 - Lista.

Una lista è una sequenza di nodi contenenti dati arbitrari e uno o due puntatori all'elemento successivo e/o precedente.

NB. È importante notare che nodi contigui nella *lista* non lo sono necessariamente anche in memoria.

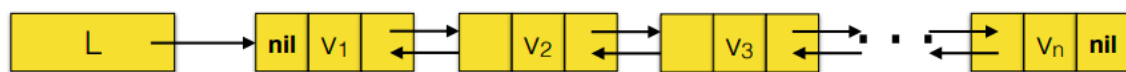
NB. In una *lista* tutte le operazioni hanno costo $O(1)$.

Diverse implementazioni di una *lista* possono essere categorizzate sulla base di tre parametri:

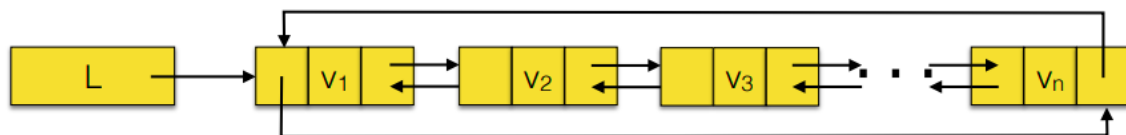
- *Monodirezionali/Bidirezionali*: sono *bidirezionali* le implementazioni in cui ogni nodo contiene due puntatori: uno all'elemento precedente, l'altro al successivo;
- *Con sentinelle/Senza sentinella*: sono *con sentinella* tutte le implementazioni in cui una *lista* vuota ha un elemento;
- *Circolare/Non circolare*: sono *circolari* le implementazioni in cui l'ultimo elemento ha come successivo il primo, o viceversa;



Monodirezionale



Bidirezionale



Bidirezionale circolare



Monodirezionale con sentinella

Fig. 4.2: Diverse implementazioni di una *lista*

Implementazione del tipo di dato POS

Frammento 11 - POS.

```
POS succ                                % Elemento successivo
POS pred                                % Elemento precedente
ITEM value                              % Valore

POS Pos(ITEM v)
    POS p = new POS
    p.succ = nil
    p.pred = nil
    p.value = v
    return p
```

Implementazione di una lista bidirezionale con sentinella

Frammento 12 - Lista bidirezionale con sentinella.

```
LIST pred                                % Predecessore    % Restituisce true se p è la
LIST succ                                % Successore      % posizione sentinella
ITEM value                              % Valore          boolean finished(POS p)
                                                    return (p == this)

LIST List()
    LIST t = new LIST
    t.pred = t
    t.succ = t
    return t

boolean isEmpty()
    return (pred == succ == this)

POS head()
    return succ

POS tail()
    return pred

POS next(POS p)
    return p.succ

POS prev(POS p)
    return p.pred

ITEM read(POS p)
    return p.value

write(POS p, ITEM v)
    p.value = v

POS insert(POS p, ITEM v)
    POS t = Pos(v)
    t.pred = p.pred
    p.pred.succ = t
    t.succ = p
    p.pred = t
    return t

POS remove(POS p)
    p.pred.succ = p.succ
    p.succ.pred = p.pred
    POS t = p.succ
    delete p
    return t
```

NB. I tipi POS e LIST sono equivalenti.

NB. Alla posizione *sentinella* non è stato assegnato alcun valore in quanto serve soltanto a semplificare le operazioni di inserimento e rimozione.

4.3.2 Pila

Definizione 33 - Pila.

Una *pila* è una struttura dati dinamica e lineare, nella quale l'accesso agli elementi è definito in base all'ordine in cui sono stati inseriti. In particolare, è possibile accedere direttamente soltanto all'ultimo elemento inserito.

NB. Le *pila* sono basate sull'approccio *LIFO* (*Last In, First Out*).

Specifica

Frammento 13 - Pila.

```
% Restituisce true se la pila è vuota
boolean isEmpty()
% Inserisce v in cima alla pila
push(ITEM v)
% Rimuove l'elemento in cima alla pila e lo restituisce
ITEM pop()
% Legge l'elemento in cima alla pila
ITEM top()
```

NB. Nel gergo delle *pila*, l'elemento “in cima” è l'ultimo elemento inserito, mentre quello “in fondo”, il primo.

Una *pila* può essere implementata come *vettore*, quindi di dimensione limitata, o come *lista bidirezionale* con puntatore all'elemento in testa.

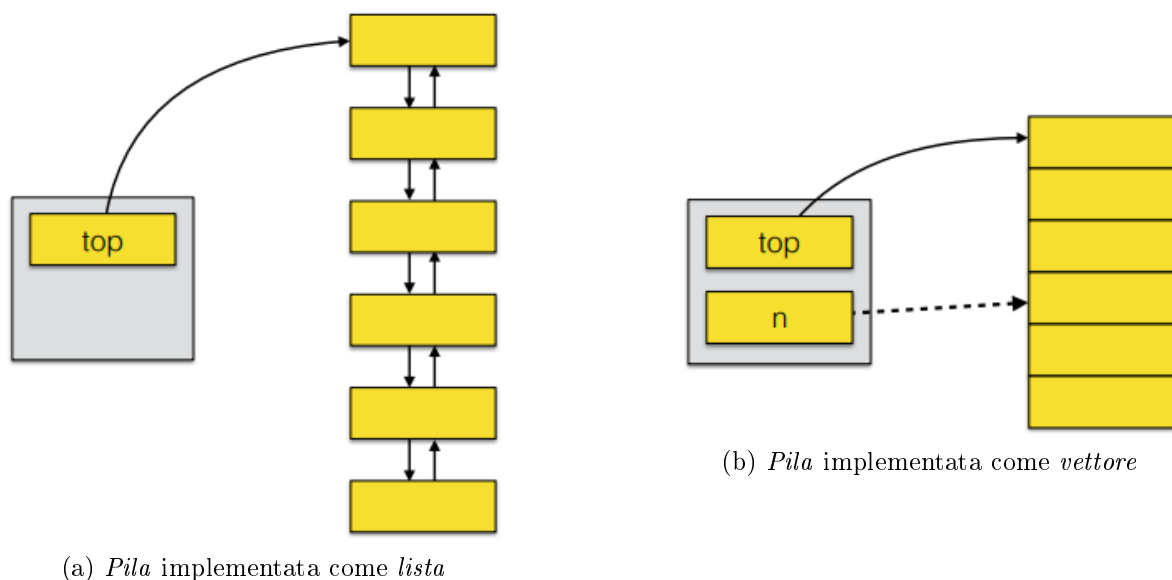


Fig. 4.3: Diverse implementazioni di una *pila*

Implementazione di una pila basata su vettore

Frammento 14 - Pila basata su vettore.

```
ITEM[] V           % Elementi
int cur            % Posizione cursore
int max_dim        % Dimensione massima

STACK Stack(int dim)
    STACK t = new STACK
    t.V = new int[1...dim]
    t.V = new int[1...dim]
    t.cur = 0
    t.max_dim = dim
    return t

ITEM top()
    precondition:(cur > 0)
    return V[cur]

boolean isEmpty()
    return (cur == 0)

ITEM pop()
    precondition:(cur > 0)
    ITEM t = V[cur]
    cur = cur - 1
    return t

push(ITEM v)
    precondition:(cur < max_dim)
    cur = cur + 1
    V[cur] = v
```

NB. In una *pila* implementata come *lista* non è necessario specificare una dimensione massima.

4.3.3 Code

Definizione 34 - Coda.

Una *coda* è una struttura dati dinamica e lineare, nella quale l'accesso agli elementi è definito in base all'ordine in cui sono stati inseriti. In particolare, è possibile accedere direttamente soltanto al primo elemento inserito.

NB. Le *code* sono basate sull'approccio *FIFO* (*First In, First Out*).

Specifica

Frammento 15 - Coda.

```
% Restituisce true se la coda è vuota
boolean isEmpty()
% Inserisce v in fondo alla coda
enqueue(ITEM v)
% Estrae l'elemento in testa alla coda e lo restituisce al chiamante
ITEM dequeue()
% Legge l'elemento in testa alla coda
ITEM top()
```

NB. Nel gergo delle *code* l'elemento “in testa” è il primo inserito, mentre quello “in coda”, l'ultimo.

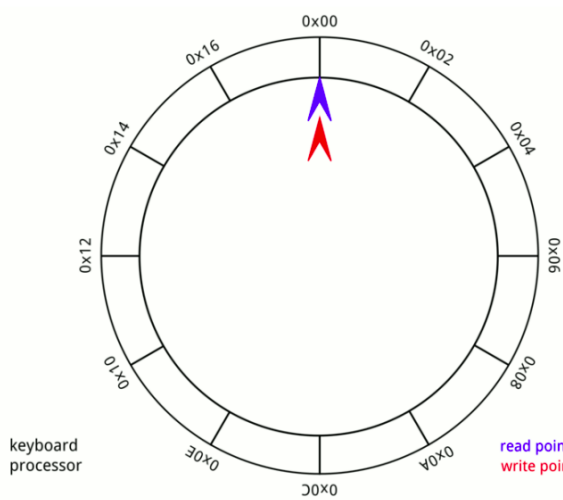
Una *coda* può essere implementata come *vettore circolare*, quindi di dimensione limitata, o come *lista monodirezionale* con puntatore *head* per l'estrazione e *tail* per l'inserimento.

Implementazione di una coda basata su vettore circolare In un *vettore circolare* gli indici possono essere gestiti efficientemente facendo uso dell'operazione di *modulo*. Bisogna tuttavia fare attenzione alla gestione dei problemi di *overflow*, provocati, non dall'accesso ad aree di memoria esterne alla struttura (usando correttamente il *modulo* ciò è impossibile), ma dalla scrittura su valori non ancora letti. Questo problema si propone quando il *vettore* è pieno.

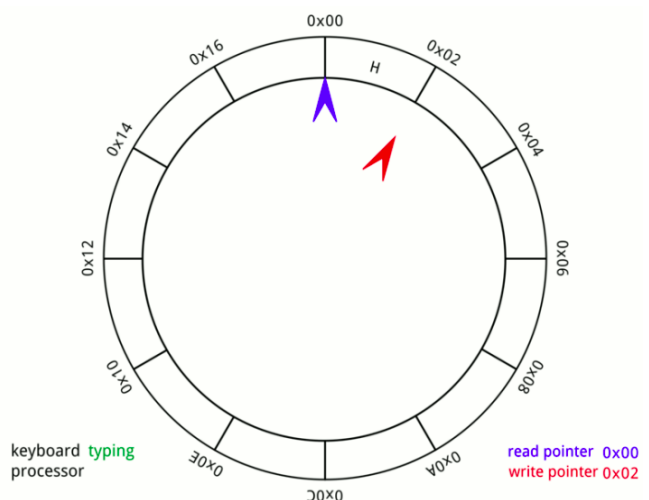
Vediamo come questo problema può essere evitato.

Utilizzando due puntatori, *read pointer* e *write pointer*, che sono, rispettivamente, il puntatore alla *testa* e alla *coda* della struttura, si può fare in modo che non sia possibile scrivere nuovi valori quando il *write pointer* punta alla cella immediatamente precedente a quella puntata dal *read pointer*.

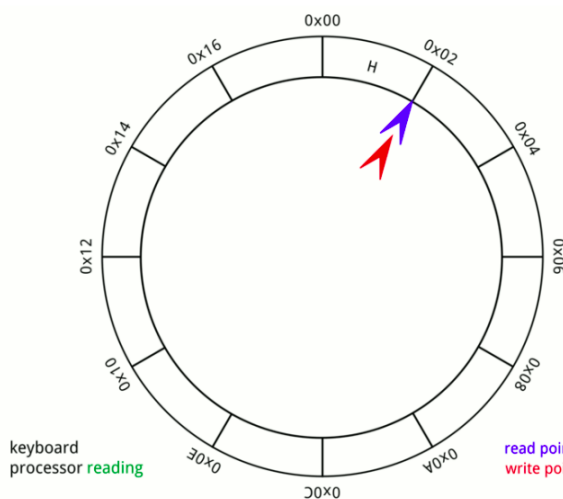
Vediamo nel dettaglio graficamente:



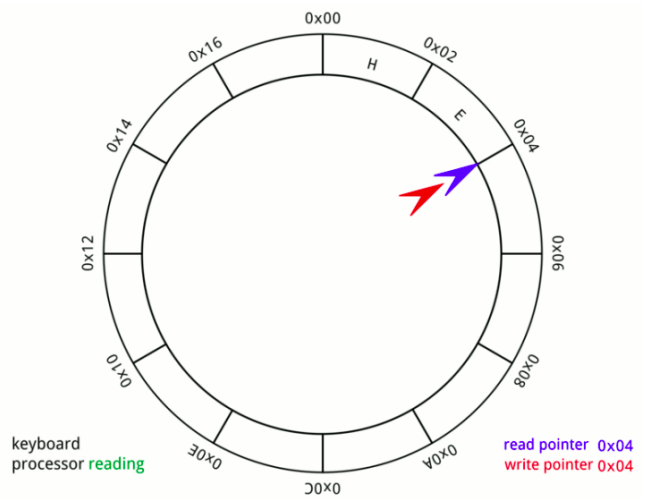
(a) Prima scrittura



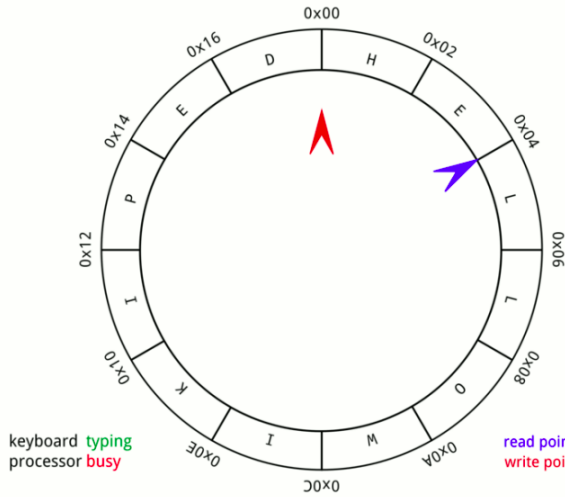
(b) Seconda scrittura



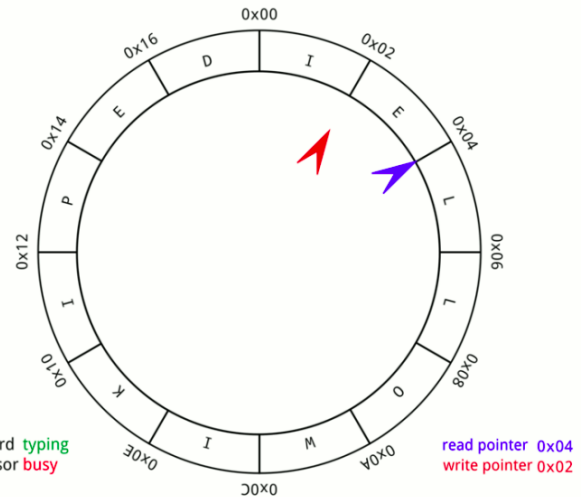
(c) Prima lettura



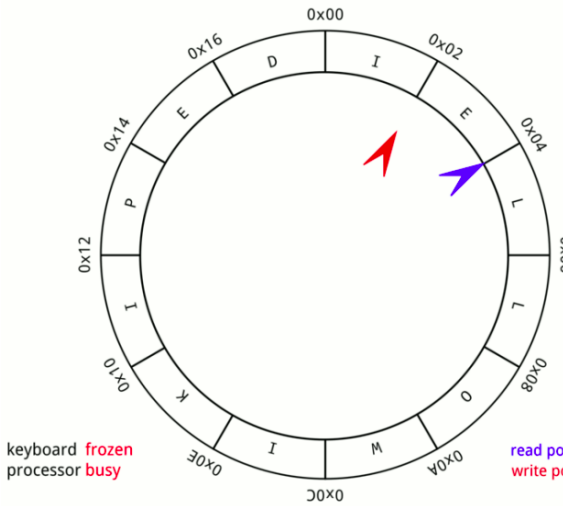
(d) Seconda lettura



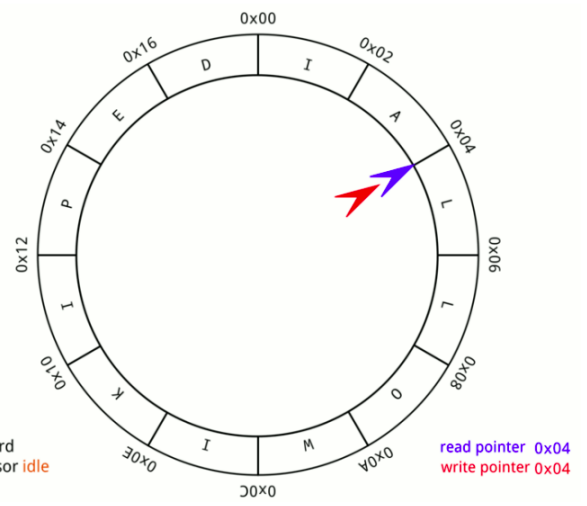
(e) Ultima scrittura su cella libera



(f) Prima scrittura su cella già letta



(g) Blocco delle scritture



(h) Lettura di tutte le celle

Fig. 4.3: Esempio di esecuzione di una *coda* implementata come *vettore circolare*

Nelle immagini di cui sopra si vede molto bene come, dopo ogni operazione di lettura e scrittura, i rispettivi puntatori avanzino fino alla cella successiva. Fino a quando ci sono celle libere i due puntatori avanzano in modo indipendente, ma quando queste finiscono, le successive scritture vanno a sovrascrivere i valori precedentemente inseriti e che sono stati già letti.

Una volta che il *write pointer* si trova sulla cella che precede quella indicata dal *read pointer*, la *coda* è piena e quindi non è possibile inserire nuovi elementi fino a quando non ne vengono letti alcuni. Quando invece il *read pointer* va a sovrapporsi al *write pointer* significa che la *coda* è stata svuotata e di conseguenza vengono bloccate le letture.

Frammento 16 - Coda basata su vettore circolare.

```
ITEM[] V                % Elementi
int cur_dim              % Dimensione attuale
int head                 % Testa della coda
int max_dim              % Dimensione massima

QUEUE Queue(int dim)
    QUEUE t = new QUEUE
    t.V = new int[dim]
    t.max_dim = dim
    t.head = 0
    t.cur_dim = 0
    return t

ITEM top()
    precondition:(cur_dim > 0)
    return V[head]

boolean isEmpty()
    return (cur_dim == 0)

ITEM dequeue()
    precondition:(cur_dim > 0)
    ITEM t = V[head]
    head = (head + 1) mod max_dim
    cur_dim = cur_dim - 1
    return t

enqueue(ITEM v)
    precondition:(cur_dim < max_dim)
    V[(head + cur_dim) mod max_dim] =
    v
    cur_dim = cur_dim + 1
```

NB. In una *coda* implementata come *lista* non è necessario specificare una dimensione massima.

Capitolo Nr.5

Alberi binari e alberi generici

5.1 Introduzione

Iniziamo con una definizione:

Definizione 35 - Albero radicato.

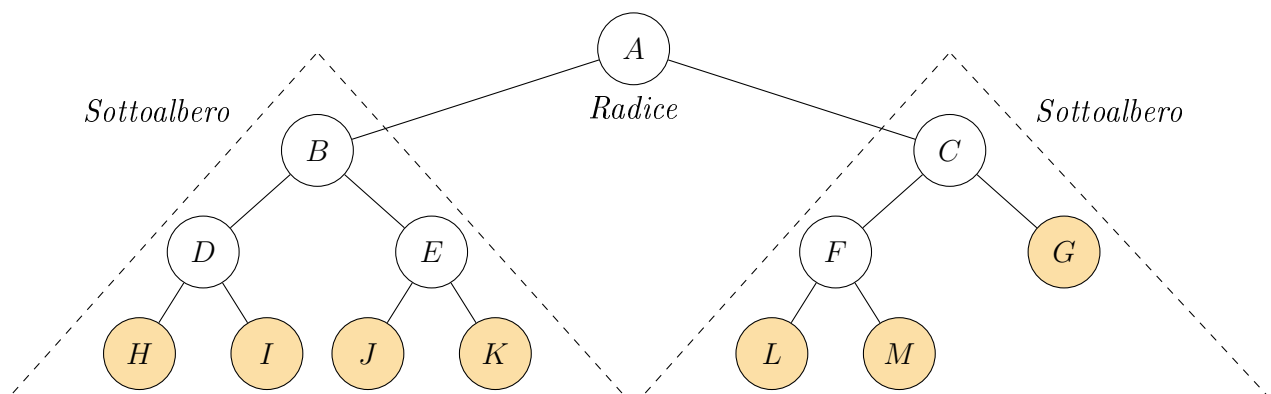
Un albero radicato consiste di un insieme di nodi e di archi orientati che connettono coppie di nodi con le seguenti proprietà:

- *Un nodo dell'albero è designato come nodo radice;*
- *Ogni nodo n , a parte la radice, ha esattamente un arco entrante;*
- *Per ogni nodo esiste un unico cammino che parte dalla radice e raggiunge quel nodo;*
- *L'albero è connesso;*

Definizione 36 - Albero radicato (definizione ricorsiva).

Un albero radicato è definito come un insieme vuoto, oppure un nodo radice e zero o più sottoalberi, ognuno dei quali è un albero radice. La radice è connessa alla radice di ogni sottoalbero con un arco orientato.

5.1.1 Terminologia



Partendo dallo schema di cui sopra possiamo definire la seguente terminologia:

- A è la *radice*;
- B e C sono *radici* dei *sottoalberi*;
- D ed E sono *fratelli*;
- D ed E sono *figli* di B ;
- B è il *padre* di D ed E ;
- I *nodi* gialli sono *foglie*;
- Gli altri *nodi* sono *nodi interni*;

Per ogni *albero* possiamo poi definire i seguenti parametri:

Definizione 37 - Profondità di un nodo (depth).

È definita *profondità di un nodo* la *lunghezza del cammino semplice dalla radice al nodo*. La *lunghezza* è misurata in *numero di archi attraversati*.

Definizione 38 - Livello (level).

È definito *livello* l'insieme dei *nodi* alla *stessa profondità*

Definizione 39 - Altezza di un albero (height).

È definita *altezza di un albero* la *profondità massima della sue foglie*.

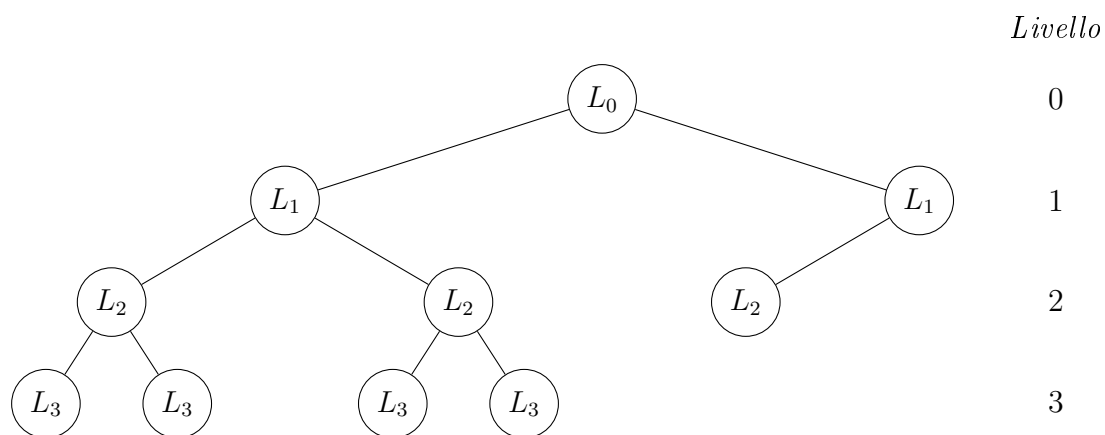


Fig. 5.1: Albero di altezza 3

5.2 Alberi binari

Definizione 40 - Albero binario.

Un *albero binario* è un *albero radicato* in cui ogni *nodo* ha al più due *figli*, identificati come *figlio sinistro* e *figlio destro*.

NB. Due *alberi* T e U che hanno gli stessi *nodi*, gli stessi *figli* per ogni *nodo* e la stessa *radice*, sono distinti qualora un *nodo* u sia designato come *figlio sinistro* di v in T e come *figlio destro* di v in U .

5.2.1 Specifica

Frammento 17 - Albero binario.

```
% Costituisce un nuovo nodo, contenente v, senza figli o genitori
Tree(ITEM v)
% Legge il valore memorizzato nel nodo
ITEM read()
% Modifica il valore memorizzato nel nodo
write(ITEM v)
% Restituisce il padre, oppure nil se questo è il nodo radice
TREE parent()
% Restituisce il figlio sinistro di questo nodo, oppure nil se è assente
TREE left()
% Restituisce il figlio destro di questo nodo, oppure nil se è assente
TREE right()
% Inserisce il sottoalbero radicato t come figlio sinistro di questo nodo
insertLeft(TREE t)
% Inserisce il sottoalbero radicato t come figlio destro di questo nodo
insertRight(TREE t)
% Distrugge ricorsivamente il figlio sinistro di questo nodo
deleteLeft()
% Distrugge ricorsivamente il figlio destro di questo nodo
deleteRight()
```

5.2.2 Memorizzazione di un albero binario

In memoria, per ogni *nodo*, memorizziamo un puntatore al *nodo padre* (P), che sarà *nil* nel caso della *radice*, e altri due puntatori, contenenti, rispettivamente, il riferimento al *figlio sinistro* (L) e *destro* (R) di quel *nodo*. Nel caso di *nodi foglia*, quei puntatori saranno entrambi *nil*.

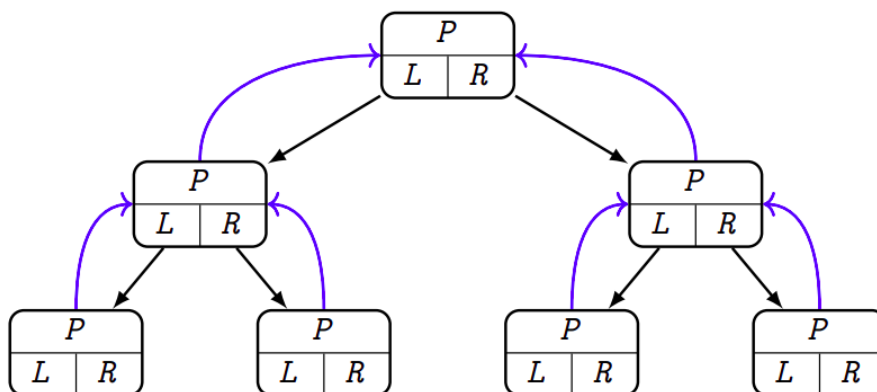


Fig. 5.2: Schema di memorizzazione di un *albero binario*

5.2.3 Implementazione di un albero binario

Frammento 18 - Implementazione di un albero binario.

```
Tree(ITEM v)
    TREE t = new TREE
    t.parent = nil
    t.left = nil
    t.right = nil
    t.value = v
    return t

ITEM read()
    return value

write(ITEM v)
    value = v

TREE parent()
    return parent

TREE left()
    return left

TREE right()
    return right

% Ipotizziamo che sia possibile
% inserire un figlio solo se non
% ne esiste già uno
insertLeft(TREE t)
    if (left == nil) then
        t.parent = this
        left = t

insertRight(TREE t)
    if (right == nil) then
        t.parent = this
        right = t

deleteLeft()
    if (left ≠ nil) then
        left.deleteLeft()
        left.deleteRight()
        delete left
        left = nil

deleteRight()
    if (right ≠ nil) then
        right.deleteLeft()
        right.deleteRight()
        delete right
        right = nil
```

5.2.4 Visite di un albero binario

Definizione 41 - Visita di un albero.

Una visita è una strategia per visitare tutti i nodi dell'albero.

Definizione 42 - Visita in profondità (depth first search).

Una visita in profondità è un tipo di visita nella quale vengono visitati ricorsivamente tutti i sottoalberi.

NB. Questo tipo di *visita* richiede l'utilizzo di una *pila*.

Per la *visita in profondità* esistono tre varianti che si differenziano per il momento in cui viene utilizzato il valore di un *nodo*:

- *Visita in Pre-order*: il valore del *nodo* viene usato prima di visitare i *sottoalberi*;

- *Visita in In-order*: il valore del *nodo* viene usato dopo aver visitato il *sottoalbero di sinistra* e prima di visitare quello di *destra*;
- *Visita in Post-order*: il valore del *nodo* viene usato dopo aver visitato i *sottoalberi*;

Definizione 43 - Visita in ampiezza (breadth first search).

Una visita in ampiezza è un tipo di visita nella quale l'albero viene visitato un livello alla volta partendo dalla radice.

NB. Questo tipo di *visita* richiede l'utilizzo di una *coda*.

Implementazione delle visite

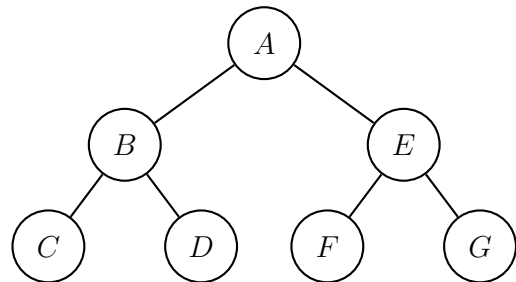
Frammento 19 - Implementazione delle visite.

<pre>% Depth-First-Search in Pre-order dfs_pre_order(TREE t) if (t ≠ nil) then print t.value dfs_pre_order(t.left) dfs_pre_order(t.right) % Depth-First-Search in In-order dfs_in_order(TREE t) if (t ≠ nil) then dfs_in_order(t.left) print t.value dfs_in_order(t.right) % Depth-First-Search in Post-order dfs_post_order(TREE t) if (t ≠ nil) then dfs_post_order(t.left) dfs_post_order(t.right) print t.value</pre>	<pre>% Breadth-First-Search bfs(TREE t) QUEUE q = Queue() q.enqueue(t) while (not q.isEmpty()) do TREE u = q.dequeue() print u.value % Accoda entrambi i figli % del nodo corrente u = u.left() if (u ≠ nil) then q.enqueue(u) u = u.parent().right() if (u ≠ nil) then q.enqueue(u)</pre>
---	--

Esempi di visite

Vediamo un esempio di utilizzo delle *visite*:

- *DFS Pre-order*: A B C D E F G;
- *DFS In-order*: C B D A F E G;
- *DFS Post-order*: C D B F G E A;
- *BFS*: A B E C D F G;



NB. Il *costo computazionale* di una *visita* di un *albero* contenente n *nodi* è $\Theta(n)$ poiché ogni *nodo* viene visitato una sola volta.

5.3 Alberi generici

Negli *alberi generici*, ovviamente, non possiamo parlare di *figlio sinistro* e *figlio destro*, ma possiamo comunque distinguere i *figli* di un *nodo* in base alla loro posizione. Parleremo infatti, di *figlio più a sinistra*, o *primo figlio*, e *fratello destro*, o *prossimo fratello*.

5.3.1 Specifica

Frammento 20 - Albero generico.

```
% Costituisce un nuovo nodo, contenente v, senza figli o genitori
Tree(ITEM v)
% Legge il valore memorizzato nel nodo
ITEM read()
% Modifica il valore memorizzato nel nodo
write(ITEM v)
% Restituisce il padre, oppure nil se questo è il nodo radice
TREE parent()
% Restituisce il primo figlio da sinistra, oppure nil se questo nodo
% è una foglia
TREE leftmostChild()
% Restituisce il primo fratello sulla destra, oppure nil se è assente
TREE rightSibling()
% Inserisce il sottoalbero t come primo figlio di questo nodo
insertChild(TREE t)
% Inserisce il sottoalbero t come prossimo fratello di questo nodo
insertSibling(TREE t)
% Distrugge l'albero radicato identificato dal primo figlio
deleteChild()
% Distrugge l'albero radicato identificato dal prossimo fratello
deleteSibling()
% Distrugge l'albero radicato identificato dal nodo
delete(TREE t)
```

5.3.2 Memorizzazione di un albero generico

A differenza degli *alberi binari*, per gli *alberi generici* esistono diversi modi per rappresentarli in memoria. I modi sono tre:

- *Vettore dei figli*: ogni *nodo* contiene un riferimento al *padre* e un vettore contenente i puntatori ai *figli*;
- *Primo figlio, prossimo fratello*: ogni *nodo* contiene un riferimento al *padre* e un riferimento al prossimo *fratello*;
- *Vettore dei padri*: l'*albero* è rappresentato da un vettore di coppie nelle quali il primo valore è il valore associato al *nodo* e il secondo è l'indice della posizione del *padre* nel vettore;

NB. L'approccio con *vettore dei figli* può causare uno spreco di memoria se molte celle di quei vettori puntano a nil.

Tecniche di memorizzazione a confronto

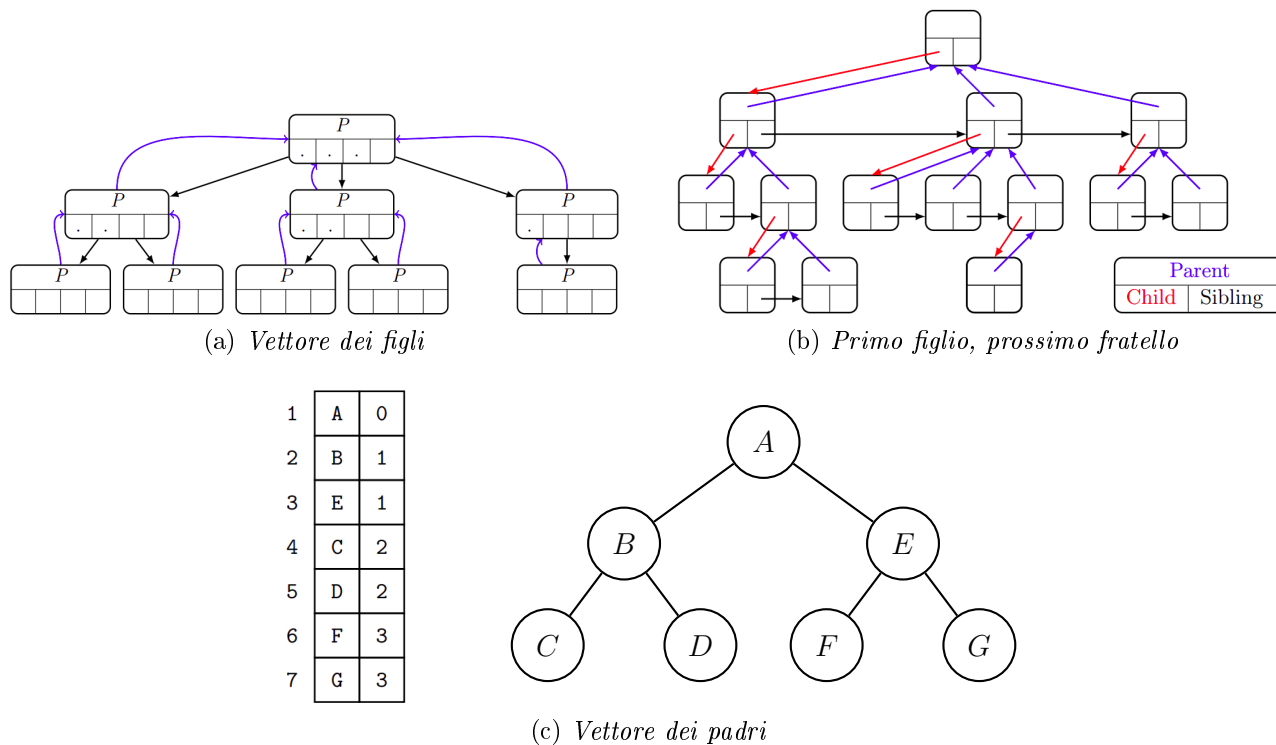


Fig. 5.3: Memorizzazione di *alberi generici*

5.3.3 Implementazione di un albero generico

Di seguito, è riportata l'implementazione di un *albero generico* realizzato con la tecnica di memorizzazione *primo figlio, prossimo fratello*.

Frammento 21 - Implementazione di un albero generico.

```

TREE parent    % Riferimento al padre          % Inserisce t prima dell'attuale
TREE child      % Riferimento al                 % figlio
                  % primo figlio                  insertChild(TREE t)
TREE sibling     % Riferimento al                 t.parent = this
                  % prossimo fratello             t.sibling = child
ITEM value      % Valore                         child = t

TREE Tree(ITEM v)                                % Inserisce t prima dell'attuale
    TREE t = new TREE                             % prossimo fratello
    t.value = v                                       insertSibling(TREE t)
    t.parent = nil                                   t.parent = parent
    t.child = nil                                    t.sibling = sibling
    t.sibling = nil                                  sibling = t
    return t

```

```

ITEM read()
    return value

write(ITEM v)
    value = v

TREE parent()
    return parent

TREE leftmostChild()
    return child

TREE rightSibling()
    return sibling

deleteChild()
    TREE c = child.rightSibling()
    delete(child)
    child = c

deleteSibling()
    TREE s = sibling.rightSibling()
    delete(sibling)
    sibling = s

delete(TREE t)
    TREE u = t.leftmostChild()
    while (u ≠ nil) do
        TREE next = u.rightSibling()
        delete(u)
        u = next
    delete t

```

5.3.4 Visite di un albero generico

Anche per gli *alberi generici* valgono le stesse *visite* viste per gli *alberi binari* ad eccezione della *visita In-order* perché in questo caso non è ben definibile una situazione intermedia.

Implementazione delle visite

Frammento 22 - Implementazione delle visite.

```

% Depth-First-Search in Pre-order
dfs_pre_order(TREE t)
    if (t ≠ nil) then
        print t.value
        TREE u = t.leftmostChild()
        while (u ≠ nil) do
            dfs_pre_order(u)
            u = u.rightSibling()

% Depth-First-Search in Post-order
dfs_post_order(TREE t)
    if (t ≠ nil) then
        TREE u = t.leftmostChild()
        while (u ≠ nil) do
            dfs_post_order(u)
            u = u.rightSibling()
        print t.value

% Breadth-First-Search
bfs(TREE t)
    QUEUE q = Queue()
    q.enqueue(t)
    while (not q.isEmpty()) do
        TREE u = q.dequeue()
        print u.value
        % Accoda tutti i figli
        % del nodo corrente
        u = u.leftmostChild()
        while (u ≠ nil) do
            q.enqueue(u)
            u = u.rightSibling()

```

Capitolo Nr.6

Alberi binari di ricerca e alberi bilanciati

6.1 Alberi binari di ricerca

Precedentemente abbiamo accennato al concetto di *dizionario*, ovvero una *struttura dati* costituita da coppie chiave-valore. Un *dizionario* può essere implementato in molteplici modi e come abbiamo visto, implementazioni diverse possono comportare *complessità* diverse per le singole operazioni che agiscono su quella *struttura*.

Ad, esempio, la seguente tabella riporta le complessità delle operazioni per tre possibili implementazioni:

Struttura dati	lookup	insert	remove
Vettore ordinato	$O(\log n)$	$O(1)$	$O(n)$
Vettore non ordinato	$O(n)$	$O(1)$	$O(1)$
Lista non ordinata	$O(n)$	$O(1)$	$O(1)$

NB. La *complessità* delle operazioni di **insert** e **remove** per le implementazioni con *vettore* e *lista non ordinati* è $O(1)$ solo se si ipotizza che, nella **insert** non si debba verificare se l'elemento è già presente o meno, e nella **remove**, che si conosca già la posizione dell'elemento da rimuovere. Se non si fanno queste ipotesi, la *complessità* diventa $O(n)$.

A questo punto ci chiediamo se non sia possibile implementare un *dizionario* sfruttando gli *alberi binari*. Supponiamo che ogni *nodo* contenga una coppia formata da una chiave e un valore. E ipotizziamo anche che le chiavi appartengano ad un insieme totalmente ordinato, cioè che sia sempre possibile stabilire se una chiave precede o meno un'altra. L'implementazione di un *nodo* dell'*albero* potrebbe quindi essere:

Frammento 23 - Nodo albero.

TREE parent

TREE left

TREE right

ITEM key, value

Poiché le chiavi sono ordinabili, possiamo ipotizzare che per ogni *nodo* u valgano le seguenti due proprietà:

1. Le chiavi contenute nel *sottoalbero sinistro* di u sono tutte minori di $u.key$;
2. Le chiavi contenute nel *sottoalbero destro* di u sono tutte maggiori di $u.key$;

NB. Le proprietà appena definite permettono di realizzare un algoritmo di ricerca dicotomica. Definiamo una regola generale:

Definizione 44 - Albero binario di ricerca (ABR).

Un albero binario di ricerca è un albero binario nel quale per ogni nodo u vale:

- 1. Ogni nodo del sottoalbero sinistro ha un valore inferiore di quello del nodo u ;*
- 2. Ogni nodo del sottoalbero destro ha un valore maggiore di quello del nodo u ;*

6.1.1 Specifica

Per un *dizionario* implementato come *albero binario di ricerca* possiamo definire la seguente *specifica*.

Frammento 24 - Dizionario come albero binario di ricerca.

```
% Getter
TREE parent()
TREE left()
TREE right()
ITEM key()
ITEM value()

% Operazioni per il dizionario
ITEM lookup(ITEM k)
insert(ITEM k, ITEM v)
remove(ITEM k)

% Operazioni definibili su un albero di ricerca binaria

% Restituisce il nodo successore di  $t$ , cioè il più piccolo
% tra i nodi più grandi di  $t$ 
TREE successorNode(TREE t)

% Restituisce il nodo predecessore di  $t$ , cioè il più grande
% tra i nodi più piccoli di  $t$ 
TREE predecessorNode(TREE t)

% Restituisce il nodo con valore minore
TREE min(TREE t)

% Restituisce il nodo con valore maggiore
TREE max(TREE t)

% Funzioni interne ausiliarie

% Restituisce il nodo dell'albero  $t$  contenente la chiave  $k$ ,
% se è presente, altrimenti nil
TREE lookupNode(TREE t, ITEM k)

% Inserisce l'associazione chiave-valore  $(k,v)$  nell'albero  $t$ .
% Se la chiave è già presente sostituisce il valore associato, altrimenti
```

```

% viene inserita una nuova associazione.
% Se t è nil restituisce il nodo creato, altrimenti t
TREE insertNode(TREE t, ITEM k, ITEM v)
% Rimuove il nodo associato alla chiave k e restituisce la radice
% dell'albero (potrebbe essere stata cambiata)
TREE removeNode(TREE t, ITEM k)

```

Implementazione dizionario Operazioni a parte, l'implementazione di *dizionario* mediante *albero binario di ricerca* non è altro che:

Frammento 25 - Dizionario come ABR.

```
TREE tree
```

```
Dictionary()
    tree = nil
```

6.1.2 Ricerca

L'implementazione della funzione `lookup` è banale:

Frammento 26 - Dizionario come ABR - lookup.

```
ITEM lookup(ITEM k)
    TREE t = lookupNode(tree, k)
    if (t ≠ nil) then
        return t.value()
    else
        return nil
```

La `lookupNode` può essere implementata sia in modo ricorsivo che iterativo:

Frammento 27 - Implementazione ricorsiva e iterativa di lookupNode.

<pre> % Versione ricorsiva TREE lookupNode(TREE t, ITEM k) if (t == nil or t.key == k) then return t else if (k < t.key) then return lookupNode(t.left, k) else return lookupNode(t.right, k) </pre>	<pre> % Versione iterativa TREE lookupNode(TREE t, ITEM k) TREE u = t while (t ≠ nil and t.key ≠ k) do if (k < u.key) then % Sottoalbero di sinistra u = u.left else % Sottoalbero di destra u = u.right return u </pre>
---	---

6.1.3 Massimo e minimo

Grazie alle proprietà degli *alberi binari di ricerca* la ricerca del minimo e del massimo si riduce alla ricerca del *nodo* più a sinistra e più a destra dell'*albero*.

Frammento 28 - Implementazione minimo e massimo.

% Ricerca del minimo

```
TREE min(TREE t)
    TREE u = t
    if (u == nil) do
        return u
    while (u.left ≠ nil) do
        u = u.left
    return u
```

% Ricerca del massimo

```
TREE max(TREE t)
    TREE u = t
    if (u == nil) do
        return u
    while (u.right ≠ nil) do
        u = u.right
    return u
```

6.1.4 Successore e predecessore

Definizione 45 - Successore.

Il successore di un nodo u è il più piccolo nodo maggiore di u .

Definizione 46 - Predecessore.

Il predecessore di un nodo u è il più grande nodo minore di u .

Ricerca del successore Poiché il *successore* di un *nodo* è il più piccolo tra i *nodi* maggiori di esso, la ricerca del *successore* di un *nodo* u non può che essere fatta sui *nodi* maggiori di u . A questo punto quindi, si configurano due casi:

1. u ha un figlio destro: il *successore* di u è il *minimo* tra i *nodi* del suo *sottoalbero destro*;
2. u non ha un figlio destro: il *successore*, se esiste, è uno degli *avi* di u , ovvero uno dei *nodi* superiori. Quindi, risalendo l'*albero*, il *successore* è il primo *nodo* maggiore di u , cioè il primo *nodo* per il quale u sta nel suo *sottoalbero sinistro*;

Frammento 29 - Implementazione successore.

```
TREE successorNode(TREE t)
```

```
    if (t == nil) then return t
```

```
    if (t.right ≠ nil)
```

```
        return min(t.right)
```

```
    else
```

```
        TREE p = t.parent
```

```
        while (p ≠ nil and t == p.right) do
```

```
            t = p
```

```
            p = p.parent
```

```
        return p
```

% Caso 1: t ha un figlio destro

% Caso 2: t non ha un figlio destro

Ricerca del predecessore Per il *predecessore* valgono le stesse considerazioni fatte per la ricerca del *successore*, ad eccezione del fatto che, se esiste, il *predecessore* di un *nodo u* sarà il massimo del suo *sottoalbero sinistro* oppure il primo *avo* che ha *u* come elemento del *sottoalbero destro*.

Frammento 30 - Implementazione predecessore.

```
TREE successorNode(TREE t)
    if (t == nil) then return t
    if (t.left ≠ nil)                                % Caso 1: t ha un figlio sinistro
        return max(t.left)
    else                                              % Caso 2: t non ha un figlio sinistro
        TREE p = t.parent
        while (p ≠ nil and t == p.left) do
            t = p
            p = p.parent
        return p
```

6.1.5 Inserimento

Per eseguire l'inserimento di una nuova associazione nel *dizionario* è sufficiente fare uso della funzione `insertNode`, che va ad inserire un *nodo*, o a modificarne uno esistente, in un *albero binario di ricerca*.

Frammento 31 - Dizionario come ABR - insert.

```
insert(ITEM k, ITEM v)
    tree = insertNode(tree, k, v)
```

Nemmeno l'implementazione della `insertNode` crea molti problemi, in quanto è sufficiente applicare ricorsivamente le *proprietà degli alberi binari di ricerca*. Andremo quindi ad invocare ricorsivamente l'`insertNode`, sul *sottoalbero sinistro* o *destro* a seconda della relazione d'ordine tra il valore del nodo (i.e il valore della chiave associata al *nodo*) e il valore della chiave da inserire, fino al raggiungimento di una *foglia*, alla quale aggiungeremo un nuovo *figlio destro* o *sinistro*, o di un *nodo*, associato ad una chiave con lo stesso valore di quella da inserire, nel qual caso sostituiremo il valore salvato nel *nodo* con quello nuovo.

Prima però di passare all'implementazione vera e propria, definiamo una funzione ausiliaria `link` che presi due *nodi* *p*, *u* e una chiave *k*, inserisce il *nodo u* come *figlio sinistro* di *p* se $p.key > k$, altrimenti lo inserisce come *figlio destro*.

Frammento 32 - Implementazione link.

```
link(TREE p, TREE u, ITEM k)
    if (u ≠ nil) then
        u.parent = p                                % Inserisce il riferimento al padre
    if (p ≠ nil) then
        if (p.key > k) then
            p.left = u                              % Inserisce u come figlio sinistro
        else
            p.right = u                             % Inserisce u come figlio destro
```

Frammento 33 - Implementazione ricorsiva insertNode.

```
TREE insertNode(TREE t, ITEM k, ITEM v)
    % L'albero è vuoto, quindi creo un nuovo nodo e lo restituisco
    if (t == nil) then
        return Tree(k, v)
    if (t.key == k) then
        t.value = v      % L'associazione esiste già, per cui aggiorno il valore
        return t
    if (t.key > k) then
        if (t.left == nil) then      % Inserisce il nodo come figlio sinistro
            link(t, Tree(k, v), k)
            return t
        else
            return insertNode(t.left, k, v)    % Continua a discendere l'albero
    else
        if (t.right == nil) then      % Inserisce il nodo come figlio destro
            link(t, Tree(k, v), k)
            return t
        else
            return insertNode(t.right, k, v)    % Continua a discendere l'albero
```

È anche possibile realizzare una versione iterativa della insertNode

Frammento 34 - Implementazione iterativa insertNode.

```
TREE insertNode(TREE t, ITEM k, ITEM v)      % Riferimento al padre
    TREE p = nil
    TREE u = t
    while (u ≠ nil and u.key ≠ k) do          % Cerca la posizione di inserimento
        p = u
        u = iif(k < u.key, u.left, u.right)
    if (u ≠ nil and u.key == k) then          % La chiave è già presente
        u.value = v
    else
        TREE newTree = Tree(k, v)            % Crea un nuovo nodo
        link(p, newTree, k)                  % Inserisce il nuovo nodo
        if (p == nil) then                   % Il nodo creato è il primo dell'albero
            t = newTree
    return t                                  % Restituisce il nuovo albero o quello modificato
```

6.1.6 Cancellazione

Come per la ricerca e l'inserimento, anche l'implementazione della funzione **remove** è basata su un'altra funzione, la **removeNode**.

Frammento 35 - Dizionario come ABR - remove.

```
remove(ITEM k)
    tree = removeNode(tree, k)
```

L'implementazione della `removeNode` non è per nulla banale, in quanto, se u è il *nodo* da eliminare, vanno considerate diverse casistiche:

1. u non ha figli: si elimina il *nodo*;
2. u ha un figlio f : si elimina u e si rende f il figlio dell'*ex-padre* di u in sostituzione di u (*short-cut*);
3. u ha due figli: si individua il *successore* s di u . Poiché il *successore*, per definizione, non ha figlio sinistro, si stacca s dal proprio padre e si attacca il suo figlio destro, se esiste, come figlio sinistro del padre di s . Infine, si copia il valore di s su u e si elimina s ;

Frammento 36 - Implementazione `removeNode`.

```

TREE removeNode(TREE t, ITEM k)
    TREE u = lookupNode(t, k)
    if (u  $\neq$  nil) then
        if (u.left == nil and u.right == nil) then                                % Caso 1
            link(u.parent, nil, k)                                                % Rimuove il figlio
            delete u
        else if (u.left  $\neq$  nil and u.right  $\neq$  nil) then                        % Caso 3
            TREE s = successorNode(u)
            link(s.parent, s.right, s.key) % Attacca il figlio di s al padre di s
            u.key = s.key                                                         % Copia su u la chiave di s
            u.value = s.value                                                     % Copia su u il valore di s
            delete s
        else if (u.left  $\neq$  nil and u.right == nil) % Caso 2 con figlio sinistro
            link(u.parent, u.left, k)
            if (u.parent == nil) then
                t = u.left
        else                                                                    % Caso 2 con figlio destro
            link(u.parent, u.right, k)
            if (u.parent == nil) then
                t = u.right
    return t

```

Vediamo la dimostrazione della correttezza di questa implementazione:

Dimostrazione. Procediamo caso per caso. Sia u il *nodo* che si intende rimuovere dall'*albero*:

1. *Nessun figlio*: eliminare foglie non inficia sull'ordine degli altri nodi;
2. *Un solo figlio*:
 - (a) *Figlio destro*: se u è il figlio destro di p tutti i valori nell'*albero* che ha come radice u sono maggiori di p ;
 - (b) *Figlio sinistro*: se u è il figlio sinistro di p tutti i valori nell'*albero* che ha come radice u sono minori di p ;
3. *Due figli*: il *successore* s del *nodo* è maggiore di tutti i nodi del *sottoalbero di sinistra* di u , ma allo stesso tempo, è anche minore di tutti i nodi nel *sottoalbero destro* di u , di conseguenza può essere sostituito a u ;

□

6.1.7 Costo computazionale delle operazioni

Tutte le operazioni viste per gli *alberi binari di ricerca* sono confinate ai *nodi* posizionati lungo un cammino semplice dalla *radice* a una *foglia*. Di conseguenza, se h è l'altezza di un *albero*, tutte le operazioni sono $O(h)$.

A questo punto vale la pena chiedersi quali siano il caso peggiore e quello migliore.

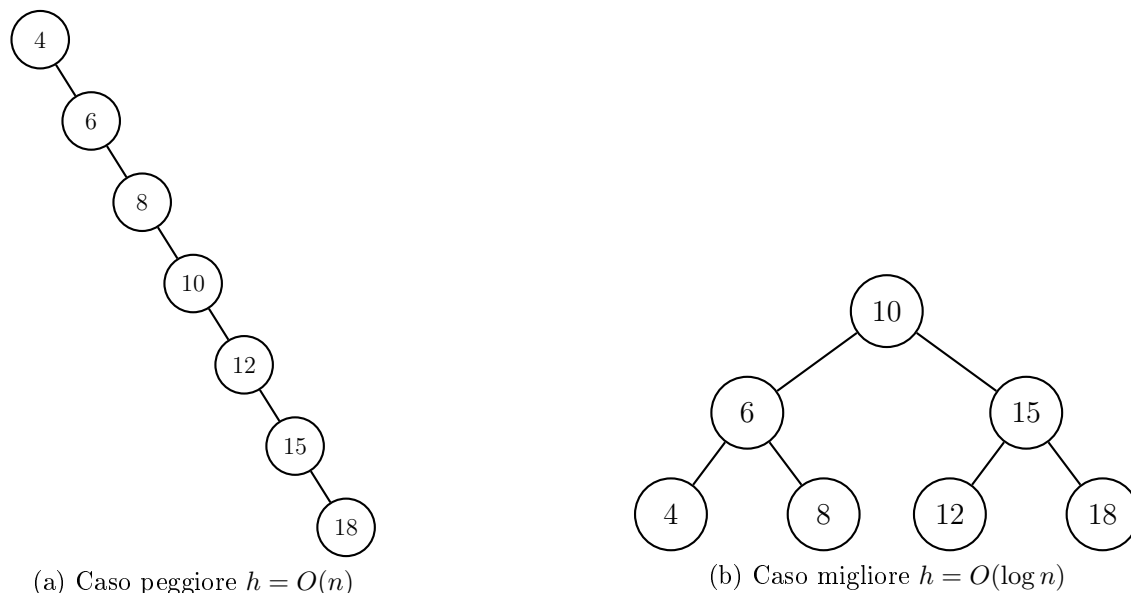


Fig. 6.1: *Caso peggiore e caso migliore*

Da queste immagini risulta evidente che sia meglio evitare di lavorare con *alberi* troppo “*sbilanciati*”.

6.2 Alberi binari di ricerca bilanciati

Abbiamo terminato l'ultima sezione discutendo brevemente il modo in cui l'altezza di un *albero* determini il costo delle operazioni. In particolare, abbiamo accennato al fatto che nel caso pessimo tutti i *nodi* sono posizionati come in una *lista*, mentre nel caso migliore l'*albero* gode una struttura più simmetrica rispetto alla *radice*.

Ovviamente, nella maggior parte delle situazioni non si avrà a che fare né con l'uno che con l'altro ed è quindi ragionevole attendersi, generalmente, un'altezza inferiore a quella del caso pessimo. Addirittura, è dimostrato che se gli elementi vengono inseriti in ordine casuale, l'altezza *media* è $O(\log n)$.

Nella realtà però, difficilmente ci si affida alla casualità e quindi, per lavorare con *alberi* di altezza ottimale, si deve ricorrere a tecniche che garantiscano un buon livello di *bilanciamento*.

Definizione 47 - Fattore di bilanciamento.

È definito *fattore di bilanciamento* di un nodo v la massima differenza di l'altezza tra i suoi sottoalberi ed è indicato in simboli come $\beta(v)$.

Esistono diversi algoritmi di *bilanciamento*, quali ad esempio:

- *Alberi AVL* (Adelson-Veskii e Landis): bilanciano l'*albero* facendo uso di rotazioni e garantiscono che per ogni nodo v $\beta(v) \leq 1$;

- *B-alberi* (Bayer e McCreight): ogni *nodo* v ha $\beta(v) = 0$ e sono specializzati per *strutture* salvate in memoria secondaria;
- *Alberi 2-3* (Hopcroft): ogni *nodo* v ha $\beta(v) = 0$ e sono basati su una *struttura* dinamica che attraverso operazioni di merge e split, distribuisce i *nodi* su *alberi* con due o tre *figli*;

In questa trattazione analizzeremo un quarto tipo di *alberi*, gli *Alberi Red-Black*, ideati da Guibas e Sedgewick nel 1978.

6.3 Alberi Red-Black

Definizione 48 - Alberi Red-Black.

Gli *Alberi Red-Black* sono *alberi* binari di ricerca nei quali ogni *nodo* è colorato di rosso oppure di nero. Le chiavi vengono mantenute soltanto nei *nodi* interni e le foglie sono costituite da *nodi* speciali con valore *Nil*. Inoltre, vengono rispettati i seguenti vincoli:

1. La radice è nera;
2. Tutte le foglie sono nere;
3. Entrambi i figli di un *nodo* rosso sono neri;
4. Ogni cammino semplice da un *nodo* ad una delle sue foglie ha sempre lo stesso numero di *nodi* neri;

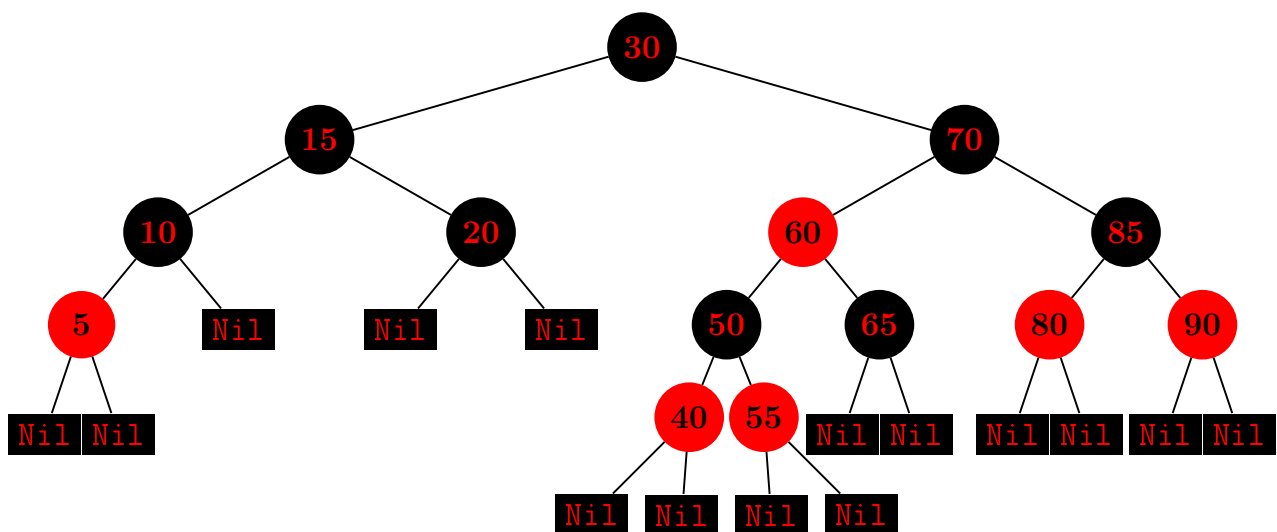


Fig. 6.2: *Albero Red-Black*

Diamo inoltre le seguenti definizioni:

Definizione 49 - Altezza nera di un nodo.

Dato un *nodo* v , l'altezza nera di quel *nodo*, indicata come $bh(v)$, è il numero di *nodi* neri lungo ogni cammino da v (escluso) ad ogni sua foglia (inclusa).

Definizione 50 - Altezza nera di un Albero Red-Black.

L'altezza nera di un *Albero Red-Black* corrisponde all'altezza nera della sua radice.

L'albero della figura 6.2 ha un'altezza nera pari a 3.

NB. Per ogni *Albero Red-Black* sono possibili più colorazioni che possono anche portare ad altezze nere diverse.

Foglie come nodi Nil Come detto, le *foglie* sono rappresentate da speciali *nodì* con valore Nil. Questo significa, che invece di un puntatore nil, c'è un riferimento ad un *nodo* Nil di colore nero. E, poiché tutti i *nodì* Nil sono uguali, ne esiste uno solo al quale fanno riferimento tutti i *nodì* che hanno delle *foglie*.

In realtà, i *nodì* che hanno come *figli* questo tipo di *nodì*, sono essi stessi le vere foglie dell'albero, in quanto, i *nodì* Nil non sono altro che *sentinelle* create per permettere di accedere al colore dei *figli* senza dover gestire il caso in cui questi siano nil.

Definizione di un nodo in un Albero Red-Black In generale, la definizione dei *nodì* per *Alberi Red-Black* non è diversa da quella dei *nodì* per *alberi binari di ricerca* classici. L'unica differenza sta nella necessità di tenere traccia del colore assegnato ad ogni *nodo*.

Frammento 37 - Implementazione di un nodo in un Albero Red-Black.

```

TREE parent
TREE left
TREE right
int color                                % Per efficienza si sarebbe potuto usare un boolean
ITEM key, value

```

6.3.1 Rotazioni

Ogni volta che si va ad inserire un elemento in un *Albero Red-Black* è possibile che le *condizioni di bilanciamento* vengano violate.

Quando ciò accade è possibile agire applicando delle operazioni di *rotazione*, oppure andando direttamente a modificare i colori dei *nodì* presenti nella zona in cui i vincoli risultano violati.

Una *rotazione* prevede che la *radice* di un *albero* venga sostituita con uno dei suoi *figli*. Un'operazione di questo tipo, se opportunamente sfruttata, permette di bilanciare un "albero sbilanciato".

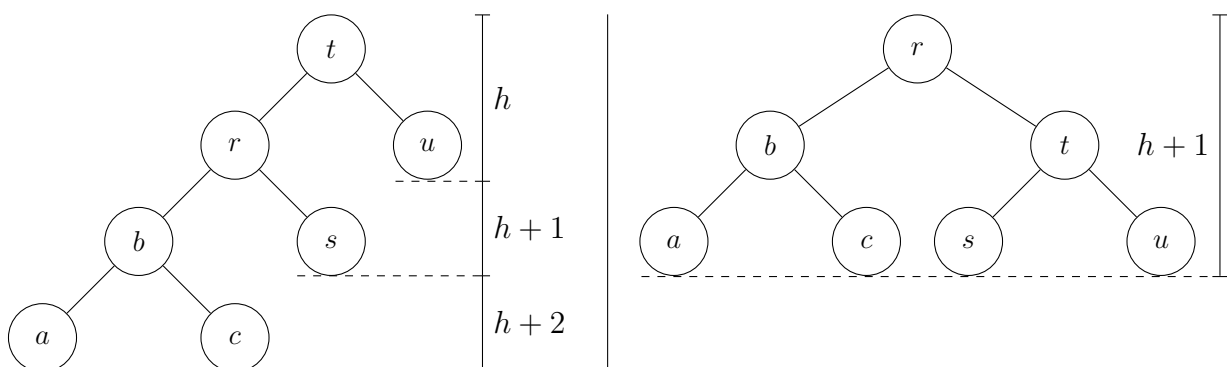


Fig. 6.3: Esempio di *rotazione verso destra* del *nodo t*

Frammento 38 - Implementazione funzioni di rotazione.

```
% Rotazione verso sinistra
TREE rotateLeft(TREE r)
    TREE s = r.right
    TREE t = r.parent
```

```
% Il sottoalbero sinistro di s
% diventa il figlio destro di r
r.right = s.left
if (s.left  $\neq$  nil) then
    s.left.parent = r
```

```
% r diventa il figlio sinistro
% di s
s.left = r
r.parent = s
```

```
% s diventa il figlio di t
s.parent = t
if (t  $\neq$  nil) then
    if (t.left == r) then
        t.left = s
    else
        t.right = s
return s
```

```
% Rotazione verso destra
TREE rotateRight(TREE r)
    TREE b = r.left
    TREE t = r.parent
```

```
% Il sottoalbero destro di b
% diventa il figlio sinistro di r
r.left = b.right
if (b.right  $\neq$  nil) then
    b.right.parent = r
```

```
% r diventa il figlio destro
% di b
b.right = r
r.parent = b
```

```
% b diventa il figlio di t
b.parent = t
if (t  $\neq$  nil) then
    if (t.left == r) then
        t.left = b
    else
        t.right = b
return b
```

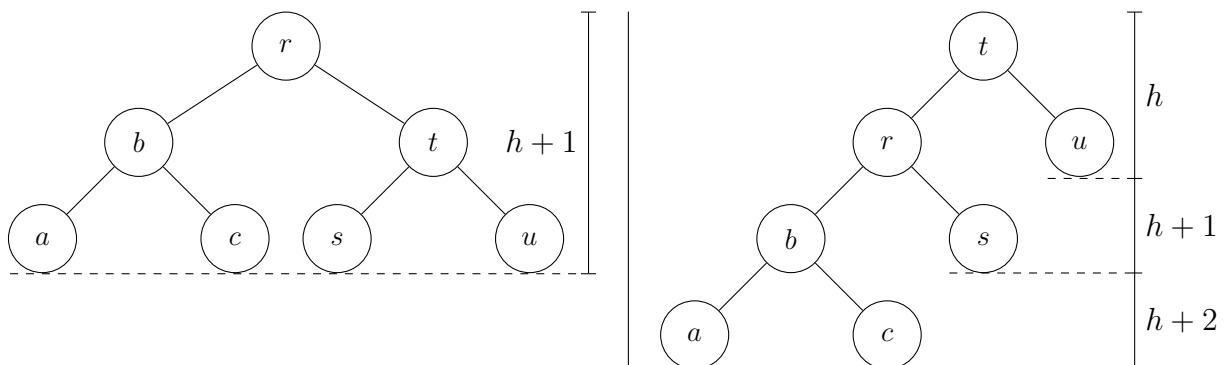


Fig. 6.4: Esempio di *rotazione verso sinistra* del *nodo t*

Da questo esempio si vede anche che applicando una rotazione su un *nodo* e applicando poi la rotazione inversa sullo stesso *nodo*, si riottiene l'*albero* di partenza della prima rotazione.

6.3.2 Inserimento

A questo punto, possiamo affrontare l'argomento *inserimento*, che nei principi generali non si discosta molto da quanto visto per gli *alberi binari di ricerca*. In particolare, la posizione di inserimento viene ricercata con le stesse regole, ma in più il nuovo *nodo* viene "colorato" di rosso.

Ovviamente, come già accennato, l'inserimento di un *nodo* può comportare la violazione dei *vincoli di bilanciamento*. È quindi opportuno che ad ogni inserimento, se necessario, l'*albero* venga ribilanciato.

Frammento 39 - Implementazione insertNode per Alberi Red-Black.

```

TREE insertNode(TREE t, ITEM k, ITEM v)
    TREE p = nil                                % Riferimento al padre
    TREE u = t
    while (u  $\neq$  nil and u.key  $\neq$  k) do        % Cerca la posizione di inserimento
        p = u
        u = iif(k < u.key, u.left, u.right)
    if (u  $\neq$  nil and u.key == k) then            % Chiave già presente
        u.value = v
    else
        TREE newTree = Tree(k, v)                % Crea un nuovo nodo
        link(p, newTree, k)                        % Inserisce il nuovo nodo
        balanceInsert(newTree)                    % Garantisce il bilanciamento dell'albero
        if (p == nil) then                        % Il nodo creato è il primo dell'albero
            t = newTree
    return t

```

La funzione `balanceInsert` si occupa di garantire il rispetto di tutti i *vincoli di bilanciamento* modificando la colorazione dei *nodi* o applicando delle rotazioni.

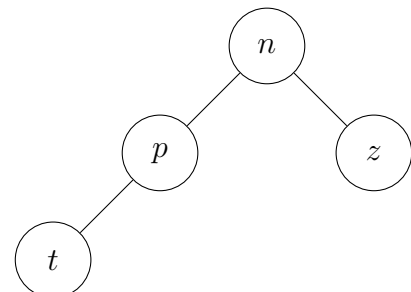
Questa funzione procede con un approccio bottom-up a partire dal *nodo* inserito. A mano a mano che risale, i *figli* di ogni *nodo* rosso vengono colorati di nero, garantendo così il rispetto del *vincolo 3*. Tutte le violazioni ancora irrisolte vengono in questo modo spostate verso l'alto. Quando infine viene raggiunta la *radice*, questa viene colorata di nero come richiesto dal *vincolo 1*, ma ciò si rende necessario solo se il *nodo* inserito era proprio la *radice* o se risalendo l'*albero* era stata colorata di rosso.

NB. Le operazioni di ripristino sono necessarie solo quando due *nodi* consecutivi sono rossi.

NB. Ogni *nodo* inserito viene in automatico colorato di rosso perché in questo modo c'è la possibilità che non sia necessario ribilanciare l'*albero*. Al contrario, l'inserimento di un *nodo* nero comporta sempre un ribilanciamento perché viene modificata l'*altezza nera* dell'*albero*.

Quando si inserisce un *nodo*, per capire se i *vincoli di bilanciamento* sono stati violati, è sufficiente analizzare quattro *nodi*:

- Il *nodo* *t* appena inserito;
- Il *padre* *p* di *t*;
- Il *padre* *n* di *p* o, equivalentemente, il *nonno* di *t*;
- Il secondo *figlio* *z* di *n*, ovvero il *fratello* di *p* o lo *zio* di *t*;



A seconda del colore di questi quattro *nodi* si configurano 7 possibili casi.

Caso 1 - Il nuovo nodo non ha padre Questo caso si configura quando viene inserita la *radice* dell'*albero* o quando risalendo si è arrivati alla *radice*. A questo punto l'unico vincolo che potrebbe risultare violato è il numero 1, quindi è sufficiente colorare il *nodo* di nero.

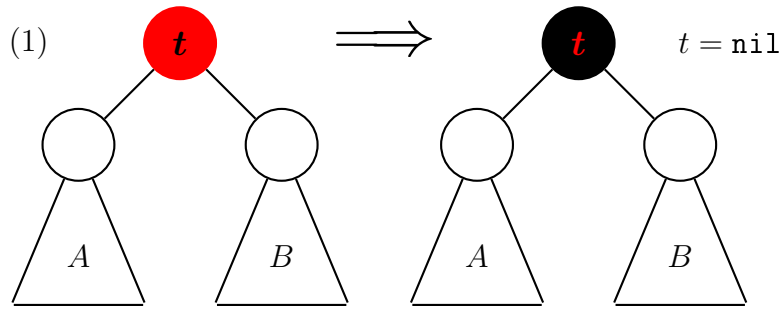


Fig. 6.5: Risoluzione del caso 1

Caso 2 - Il padre di t è nero In questo caso non serve fare nulla, poiché nessun vincolo può risultare violato.

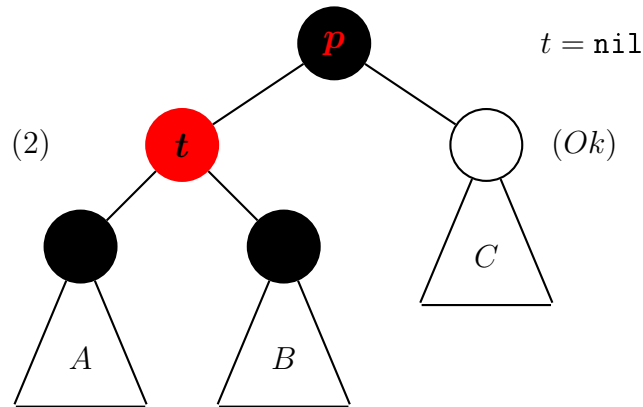


Fig. 6.6: Risoluzione del caso 2

Caso 3 - Sia t che il padre e lo zio sono rossi Poiché z è rosso, è possibile colorare di nero p e z e poi colorare di rosso n . A questo punto, poiché tutti i cammini che passano per p e z passano anche per n , l'altezza nera certamente non avrà subito variazioni.

A questo punto, è ancora possibile che n violi i vincoli 1 e 3. Di conseguenza, risaliamo l'albero ponendo $t = n$ e ripetiamo il ciclo.

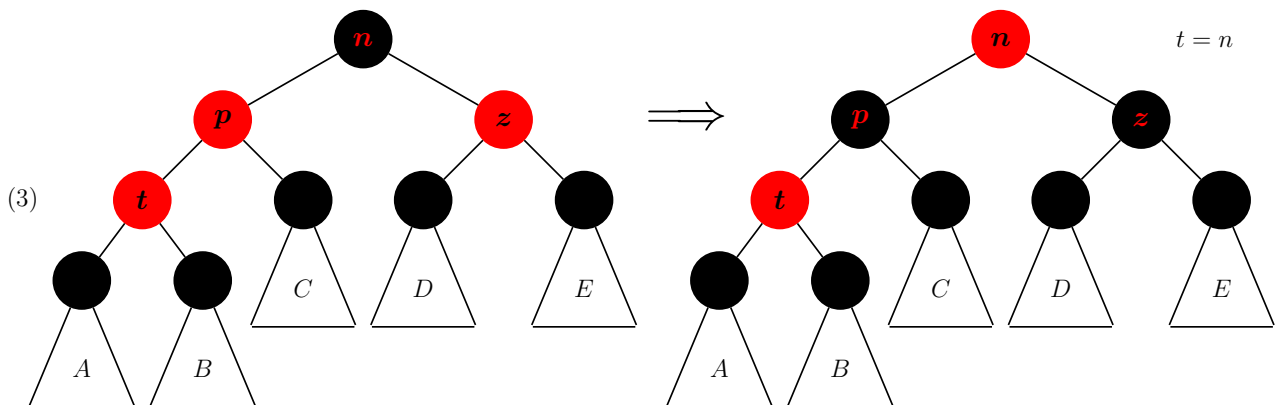


Fig. 6.7: Risoluzione del caso 3

Caso 4 - Sia t che il padre sono rossi e lo zio è nero

- t è il figlio destro di p e p è il figlio sinistro di n : viene eseguita una rotazione verso sinistra partendo dal nodo p in modo da rendere p il figlio sinistro di t e potersi così ricondurre al caso 5a. In questo caso vengono coinvolti soltanto i nodi p e t che essendo entrambi rossi, non influenzano l'altezza nera;
- t è il figlio sinistro di p e p è il figlio destro di n : viene eseguita una rotazione verso destra partendo dal nodo p in modo da rendere p il figlio destro di t e potersi così ricondurre al caso 5b. In questo caso vengono coinvolti soltanto i nodi p e t che essendo entrambi rossi, non influenzano l'altezza nera;

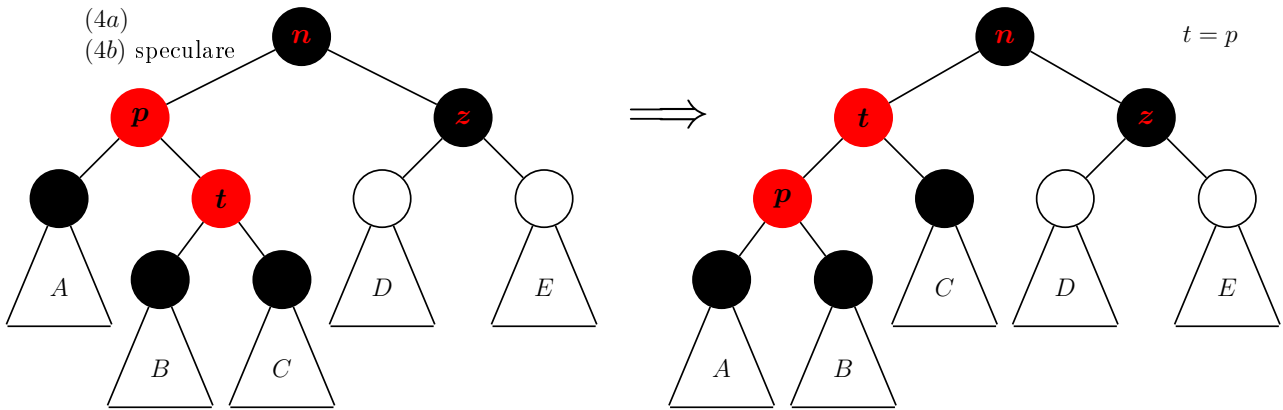


Fig. 6.8: Risoluzione del caso 4

Caso 5 - Sia t che il padre sono rossi e lo zio è nero

- t è il figlio sinistro di p e p è il figlio sinistro di n : viene eseguita una rotazione verso destra partendo dal nodo n in modo da rendere t ed n figli di p . Quindi, colorando di rosso n e di nero p ci si porta in una situazione nella quale tutti i vincoli di bilanciamento sono rispettati e l'altezza nera è uguale alla situazione iniziale;
- t è il figlio destro di p e p è il figlio destro di n : viene eseguita una rotazione verso sinistra partendo dal nodo n in modo da rendere t ed n figli di p . Quindi, colorando di rosso n e di nero p ci si porta in una situazione nella quale tutti i vincoli di bilanciamento sono rispettati e l'altezza nera è uguale alla situazione iniziale;

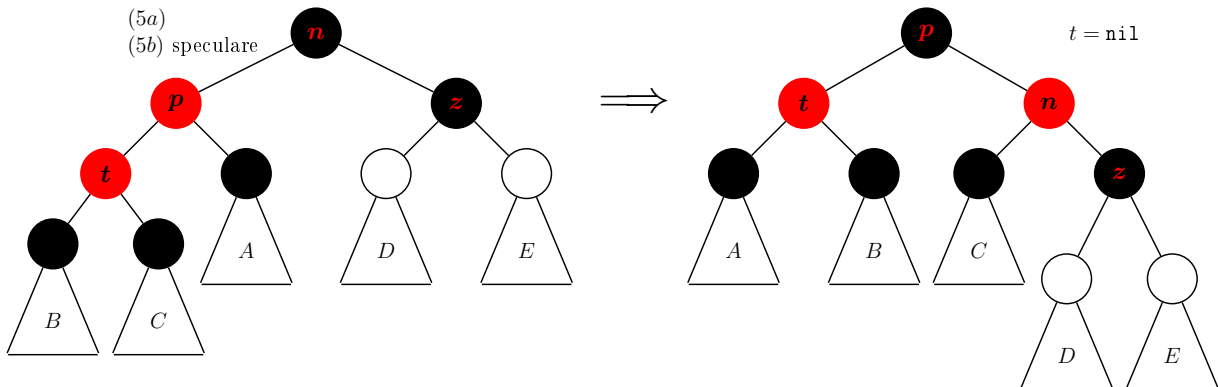


Fig. 6.9: Risoluzione del caso 5

Frammento 40 - Implementazione balanceInsert.

```
balanceInsert(TREE t)
  t.color = RED           % Il nodo inserito viene inizialmente colorato di rosso
  while (t ≠ nil) do
    TREE p = t.parent      % Padre
    TREE n = iif(p ≠ nil, p.parent, nil) % Nonno
    TREE z = iif(n ≠ nil, iif(n.left == p, n.right, n.left), nil) % Zio
    if (p == nil) then      % Caso 1
      t.color = BLACK
      t = nil
    else if (p.color == BLACK) then % Caso 2
      t = nil
    else if (z.color == RED) then % Caso 3
      p.color = BLACK
      z.color = BLACK
      n.color = RED
      t = n
    else
      if (t == p.right and p == n.left) then % Caso 4a
        rotateLeft(p)
        t = p
      else if (t == p.left and p == n.right) then % Caso 4b
        rotateRight(p)
        t = p
      else
        if (t == p.left and p == n.left) then % Caso 5a
          rotateRight(n)
        else if (t == p.right and p == n.right) then % Caso 5b
          rotateLeft(n)

        p.color = BLACK
        n.color = RED
        t = nil
```

Complessità della balanceInsert Non è difficile rendersi conto che tutte le istruzioni interne al ciclo `while` hanno complessità $O(1)$. Di conseguenza, la complessità totale della funzione dipende dal numero di volte che viene eseguito il ciclo. Poiché t quando viene inserito è una *foglia*,¹ il caso peggiore è quello in cui il ribilanciamento coinvolge tutti i *livelli* e quindi termina solo dopo aver raggiunto la *radice*.

In quel caso la complessità dipende dall'*altezza* dell'*albero*. Vedremo che negli *Alberi Red-Black* esiste una relazione ben precisa tra la loro *altezza* il numero di noi *nodi* che contengono.

¹I *nodi* di un *albero* vengono sempre inseriti al *livello* delle *foglie*

6.3.3 Altezza degli Alberi Red-Black

Definizione 51 - Numero massimo di nodi interni.

In un Albero Red-Black, un sottoalbero di radice u contiene un numero $n \geq 2^{bh(u)} - 1$ di nodi interni.

Dimostrazione. Procediamo per induzione sull'altezza del nodo u .

Caso base Dimostro la disequazione per $h = 0$:

Se $h = 0$, u è una foglia Nil e quindi il sottoalbero con radice u non può contenere nodi interni, quindi n deve per forza valere 0. Poiché:

$$n \geq 2^{bh(u)} - 1 = 2^0 - 1 = 1 - 1 = 0$$

ciò è confermato e quindi il caso base risulta verificato.

Passo induttivo Ipotizzo che $n \geq 2^{bh(u)} - 1$ per ogni albero con radice u e altezza $h \leq 1$. Dimostro la disequazione per $h > 1$:

Se $h > 1$, u è certamente un nodo interno e quindi ha due figli. Ogni figlio di u ha un'altezza nera che è pari a $bh(u)$ se è rosso e $bh(u) - 1$ se è nero. Ogni sottoalbero con radice uno dei figli di u ha altezza $h - 1$ e quindi, per ipotesi induttiva, so che ognuno di quei sottoalberi ha almeno $2^{bh(u)-1} - 1$ nodi interni. Allora, per il sottoalbero con radice u vale:

$$n \geq 2 \cdot (2^{bh(u)-1} - 1) + 1$$

cioè, il numero di nodi interni del sottoalbero con radice u è pari, almeno, alla somma dei limiti inferiori al numero di nodi interni dei sottoalberi che hanno come radice un figlio di u , più 1, perché bisogna contare anche il nodo u stesso. Poiché vale seguente catena di uguaglianze:

$$n \geq 2 \cdot (2^{bh(u)-1} - 1) + 1 = 2 \cdot \left(\frac{2^{bh(u)}}{2} - 1 \right) + 1 = 2^{bh(u)} - 2 + 1 = 2^{bh(u)} - 1$$

il passo induttivo risulta verificato. □

Definizione 52 - Quota nera minima.

In un Albero Red-Black, almeno la metà dei nodi dalla radice a una foglia deve essere nera.

Dimostrazione. Per il secondo vincolo di bilanciamento, i figli di un nodo rosso devono essere neri e, poiché la situazione in cui sono presenti il minor numero di nodi neri è quando nodi rossi e neri sono alternati, almeno la metà dei nodi devono essere neri. □

Definizione 53 - Lunghezza massima dei cammini semplici.

In un Albero Red-Black, dati due cammini semplici, dalla radice a due foglie, non è possibile che uno sia lungo più del doppio dell'altro.

Dimostrazione. Per il quarto vincolo di bilanciamento, tutti i cammini da un nodo ad una foglia hanno lo stesso numero di nodi neri e, per la definizione precedente, almeno la metà di quei nodi deve essere nera. Di conseguenza, il caso più sbilanciato è quello in cui un cammino è costituito da soli nodi neri e l'altro da nodi neri e rossi alternati e quindi si ha che il secondo cammino è lungo il doppio del primo. □

Definizione 54 - Altezza massima di un Albero Red-Black.

L'altezza massima di un Albero Red-Black con n nodi interni è al più $2\log(n+1)$.

Dimostrazione. Vale quanto segue:

$$\begin{aligned} n \geq 2^{bh(u)} - 1 &\Leftrightarrow n \geq 2^{h/2} - 1 && \text{Quota nera minima} \\ &\Leftrightarrow n + 1 \geq 2^{h/2} \\ &\Leftrightarrow \log(n+1) \geq \frac{h}{2} && \text{Applicazione logaritmo ai termini} \\ &\Leftrightarrow h \leq 2\log(n+1) \end{aligned}$$

□

Complessità funzione insertNode Fatte queste considerazioni sull'altezza massima degli Alberi Red-Black, possiamo affermare che la complessità della funzione `insertNode` è $O(\log n)$. Questo è vero in quanto costa $O(\log n)$ discendere l'albero dalla radice al punto di inserimento, costa $O(1)$ inserire il nodo e infine, costa $O(\log n)$ ribilanciare l'albero nel caso peggiore, il caso 3.

6.3.4 Cancellazione

Similmente a quanto visto per l'inserimento, la funzione `removeNode` per gli Alberi Red-Black è costruita a partire dalla controparte per alberi binari di ricerca ed è costituita da un insieme di casistiche che possono comportare o meno, la necessità di ribilanciare l'albero.

Le operazioni di ribilanciamento si rendono necessarie soltanto quando viene rimosso un nodo nero. Ciò è vero in quanto l'eliminazione di un nodo nero comporta sicuramente una variazione dell'altezza nera e potrebbe anche portare ad avere due nodi rossi in posizioni consecutive.

Più in generale, quando viene rimosso un nodo nero possono essere violati i vincoli 1, 3 o 4. In particolare:

- *Violato il vincolo 1:* viene eliminata la radice e la nuova radice è rossa;
- *Violato il vincolo 3:* viene eliminato un nodo che aveva il padre e almeno uno dei figli rossi;
- *Violato il vincolo 4:* ogni volta che viene eliminato un nodo nero, l'altezza nera cambia;

Frammento 41 - Implementazione removeNode per Alberi Red-Black.

```
TREE removeNode(TREE t, ITEM k)
    TREE u = lookupNode(t, k)
    if (u ≠ nil) then
        if (u.left == nil and u.right == nil) then                % Caso 1
            link(u.parent, nil, k)                                % Rimuove il figlio
            delete u
        else if (u.left ≠ nil and u.right ≠ nil) then              % Caso 3
            TREE s = successorNode(u)
            link(s.parent, s.right, s.key) % Attacca il figlio di s al padre di s
            u.key = s.key                                           % Copia su u la chiave di s
            u.value = s.value                                       % Copia su u il valore di s
            delete s
```

```

else                                     % Caso 2 con figlio destro
    link(u.parent, u.right, k)
    if (u.parent == nil) then
        t = u.right

    if (u.color == BLACK) then
        balanceRemove(t, u)           % Garantisce il bilanciamento dell'albero
return t

```

Frammento 42 - Implementazione balanceRemove.

```

balanceRemove(TREE r, TREE t)
    while (t ≠ r and t.color == BLACK) do    % Il ciclo termina quando t diventa
                                                % la radice o un nodo rosso
        TREE p = t.parent                    % Padre
        if (t == p.left) then                % È sicuramente diverso da nil
            TREE f = p.right                 % Fratello
            TREE ns = f.left                 % Nipote sinistro
            TREE nd = f.right                % Nipote destro
            (f.color == RED) then            % Caso 1
                p.color = RED
                f.color = BLACK
                rotateLeft(p)
            else
                if (ns.color == BLACK and nd.color == BLACK) then    % Caso 2
                    f.color = RED
                    t = p
                else if (ns.color == RED and nd.color == BLACK) then    % Caso 3
                    ns.color = BLACK
                    f.color = RED
                    rotateRight(f)
                else if (nd.color == RED) then    % Caso 4
                    f.color = p.color
                    p.color = BLACK
                    nd.color = BLACK
                    rotateLeft(p)
                    t = r                                % Fa terminare il ciclo
            else                                         % I casi si ripetono simmetrici ai precedenti
                ...

```

Complessità della funzione removeNode Come per la insertNode anche in questo caso la complessità totale è $O(\log n)$ in quanto dipende dall'altezza dell'albero.

Capitolo Nr.7

Grafi

Moltissimi problemi possono essere visti come problemi sui *grafi* e sebbene questi abbiano forma astratta, le loro soluzioni trovano applicazioni negli ambiti più disparati.

Abbiamo già dato una prima definizione generica di *grafo*, ma adesso andremo ad estenderla introducendo una serie di concetti che saranno necessari per affrontare proficuamente questo capitolo.

7.1 Introduzione

Diamo le seguenti definizioni.

Definizione 55 - Grafo orientato.

Un grafo orientato è una coppia $G = (V, E)$ dove V è l'insieme dei vertici, o dei nodi, del grafo ed E è un insieme di coppie ordinate di nodi (u, v) per $u, v \in V$ dette archi, o lati.

$$\begin{aligned} V &= \{a, b, c, d, e, f\} \\ E &= \{(a, b), (a, d), (d, a), \\ &\quad (b, c), (d, c), (d, e), \\ &\quad (e, c)\} \end{aligned}$$

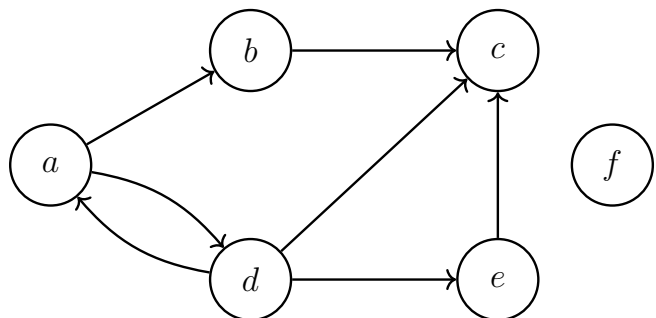


Fig. 7.1: Grafo orientato

Definizione 56 - Grafo non orientato.

Un grafo non orientato è una coppia $G = (V, E)$ dove V è l'insieme dei vertici, o dei nodi, del grafo ed E è un insieme di coppie non ordinate di nodi (u, v) per $u, v \in V$ dette archi, o lati.

$$V = \{a, b, c, d, e, f\}$$

$$E = \{(a, b), (a, d), (b, c), (d, c), (d, e), (e, c)\}$$

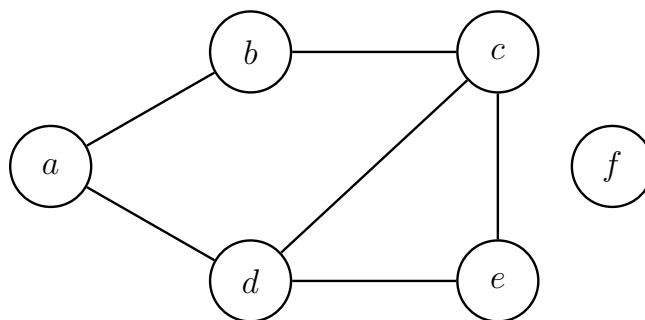
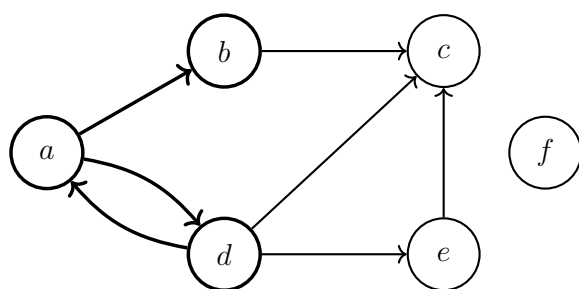


Fig. 7.2: Grafo non orientato

NB. I grafi orientati e non orientati sono anche detti, rispettivamente, *directed* e *undirected*. In un grafo orientato un vertice v è detto *adiacente* a u se esiste un arco (u, v) e, quell'arco, è detto *incidente* da u in v . Ovviamente, in un grafo non orientato la nozione di *adiacenza* è simmetrica.

Esempio 17 - Adiacenza e incidenza.

Nel seguente grafo orientato, per la parte evidenziata, valgono le seguenti relazioni di adiacenza e incidenza.



- (a, b) è incidente da a a b ;
- (a, d) è incidente da a a d ;
- (d, a) è incidente da d ad a ;
- b è adiacente ad a ;
- d è adiacente ad a ;
- a è adiacente a d ;

7.1.1 Dimensioni dei grafi

Definizione 57 - Numero di nodi e di archi.

Dato un grafo $G = (V, E)$ il numero di nodi e il numero di archi corrispondono, rispettivamente, alla cardinalità degli insiemi V ed E .

NB. Di seguito, salvo diverse indicazioni, indicheremo con n il numero di *nodi* e con m il numero di *archi*, cioè $n = |V|$ ed $m = |E|$.

Definizione 58 - Numero massimo di archi in un grafo orientato.

Dato un grafo orientato vale la seguente relazione sul numero di archi:

$$m \leq n(n - 1)$$

Definizione 59 - Numero massimo di archi in un grafo non orientato.

Dato un grafo non orientato vale la seguente relazione sul numero di archi:

$$m \leq \frac{n(n - 1)}{2}$$

NB. In entrambi i casi possiamo affermare che il numero di *archi* è $O(n^2)$.

NB. Per gli algoritmi che operano sui *grafi* la *complessità* è espressa sia in termine di n che m e quindi considereremo *lineare* un algoritmo con *complessità* $O(n + m)$.

7.1.2 Casi speciali

Definizione 60 - Grafo completo.

Un grafo con un arco tra tutte le coppie di nodi è detto completo.

Definizione 61 - Grafi sparsi e densi.

Un grafo con “pochi archi” è detto essere sparso, mentre uno con “tanti archi”, denso.

NB. La definizione di *grafi sparsi* e *densi* non è formale, ma tendenzialmente un *grafo* con $m = O(n)$ o $O(n \log n)$ è considerato *sparso*. Sono invece considerati *densi* i *grafi* con $m = \Omega(n^2)$.

Possiamo definire gli *alberi* in funzione dei *grafi*.

Definizione 62 - Albero libero.

Un grafo con $m = n - 1$ è detto albero libero.

Definizione 63 - Albero radicato (visto come grafo).

Un grafo connesso con $m = n - 1$ archi nel quale uno dei nodi è stato designato come radice è detto albero radicato.

Definizione 64 - Foresta.

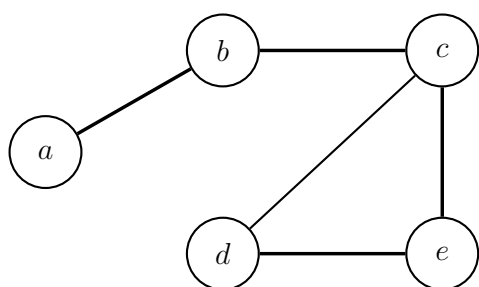
Un insieme di alberi è detto foresta.

7.1.3 Cammini e gradi dei nodi

Definizione 65 - Cammino.

In un grafo $G = (V, E)$, una sequenza di $k + 1$ nodi u_0, u_1, \dots, u_k è definita essere un cammino di lunghezza k , se è tale per cui l'arco $(u_i, u_{i+1}) \in E \forall 0 \leq i \leq k - 1$.

Esempio 18 - Cammino.



La sequenza di nodi a, b, c, e, d è un cammino di lunghezza 4.

NB *In questo caso il cammino è anche detto semplice perché tutti i nodi che lo compongono sono distinti.*

Definizione 66 - Grado di un nodo in un grafo orientato.

In un grafo orientato si definisce grado entrante di un nodo il numero di archi incidenti su di esso, mentre il grado uscente è definito come il numero di archi incidenti da esso.

Definizione 67 - Grado di un nodo in un grafo non orientato.

In un grafo non orientato il grado di un nodo è definito come il numero di archi incidenti su di esso.

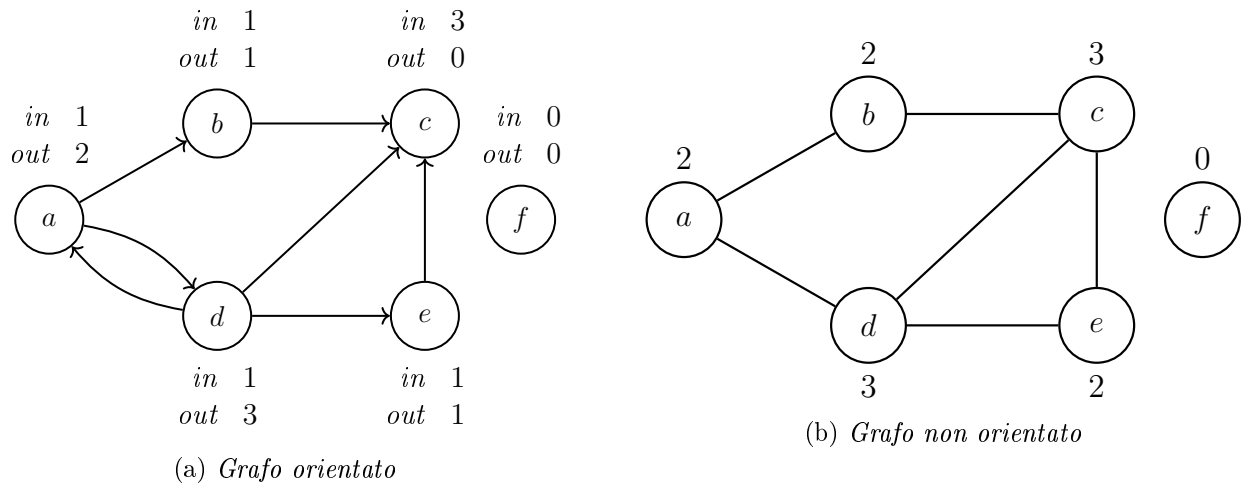


Fig. 7.3: Grado dei nodi nei grafi orientati e non orientati

7.1.4 Specifica

Frammento 43 - Grafo.

```
% Crea un nuovo grafo
GRAPH Graph()
% Restituisce l'insieme di tutti i vertici
SET V()
% Restituisce il numero di nodi
int size()
% Restituisce l'insieme dei nodi adiacenti ad u
SET adj(NODE u)
% Aggiunge un nodo u al grafo
insertNode(NODE u)
% Aggiunge l'arco (u,v) al grafo
insertEdge(NODE u, NODE v)
% Rimuove il nodo u dal grafo
removeNode(NODE u)
% Rimuove l'arco (u,v) dal grafo
removeEdge(NODE u, NODE v)
```

7.1.5 Memorizzare un grafo

Per la memorizzazione dei *grafi* esistono due approcci: *matrici di adiacenza* e *liste di adiacenza*.

Memorizzazione di un grafo orientato L'approccio con *matrice di adiacenza* prevede che venga realizzata una matrice $n \times n$ tale che:

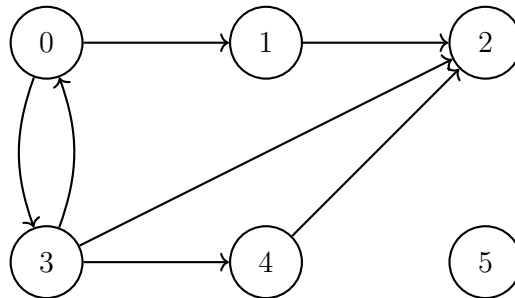
$$m_{u,v} = \begin{cases} 1 & \text{se } (u,v) \in E \\ 0 & \text{se } (u,v) \notin E \end{cases}$$

Quindi, il *grafo* viene rappresentato tramite una matrice di valori binari nella quale la cella che si trova nell' u -esima riga e v -esima colonna è 1 se e solo se esiste un *arco* da u a v .

D'altra parte, le *liste di adiacenza* sfruttano il concetto di *adiacenza* espresso nelle specifiche dalla funzione `adj`. Vale infatti, la seguente definizione matematica di `adj` e quindi di *adiacenza*:

$$G.adj(u) = \{v \mid (u,v) \in E\}$$

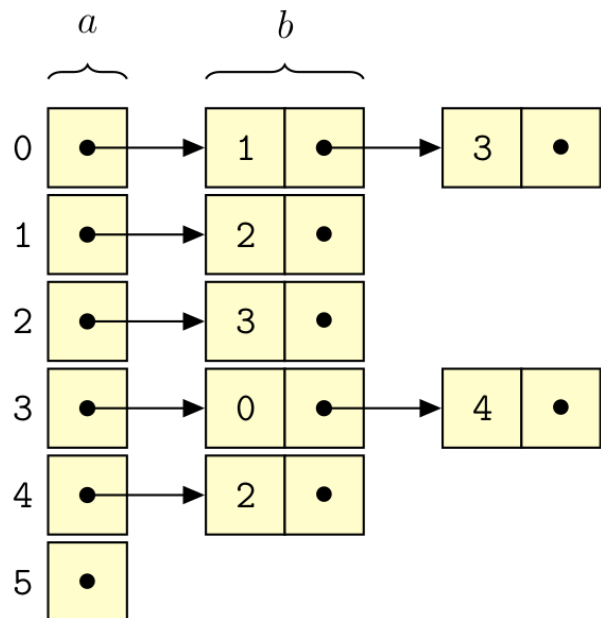
Quindi, nell'approccio basato su *liste*, viene realizzata una *lista* contenente tutti i *vertici* del *grafo*. Ogni *nodo* della *lista* poi, a cascata, contiene un riferimento ad un uno dei *vertici* che nel *grafo* sono *adiacenti* al *vertice* associato a quel *nodo* della *lista*.



(a) Grafo orientato

	0	1	2	3	4	5
0	0	1	0	1	0	0
1	0	0	1	0	0	0
2	0	0	0	1	0	0
3	1	0	0	0	1	0
4	0	0	1	0	0	0
5	0	0	0	0	0	0

(b) Matrice di adiacenza



(c) Lista di adiacenza

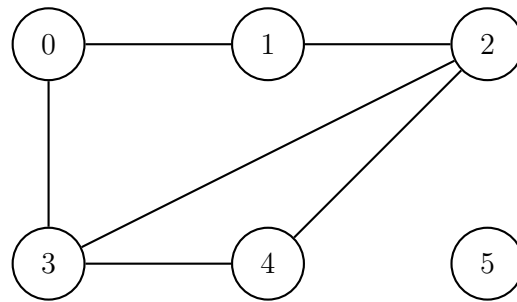
Fig. 7.4: Memorizzazione di un *grafo orientato*

Dall'immagine si può derivare facilmente il *costo di memorizzazione* nei due casi. Per le *matrici di adiacenza* lo spazio necessario dipende unicamente dal quadrato del numero di *nodi* e quindi

è $O(n^2)$, mentre per memorizzare una *lista di adiacenza*, lo spazio usato dipende sia dal numero di *nod*i che di *lati* e quindi è $O(a \cdot n + b \cdot m) = O(n + m)$.

Memorizzazione di un grafo non orientato Per i *grafi non orientati* non cambia molto, ma nel caso delle *matrici di adiacenza* possiamo dimezzare lo spazio grazie alla simmetria dell'*adiacenza*. Di conseguenza lo spazio necessario diventa $O(\frac{n(n-1)}{2})$, che però è comunque un $O(n^2)$.

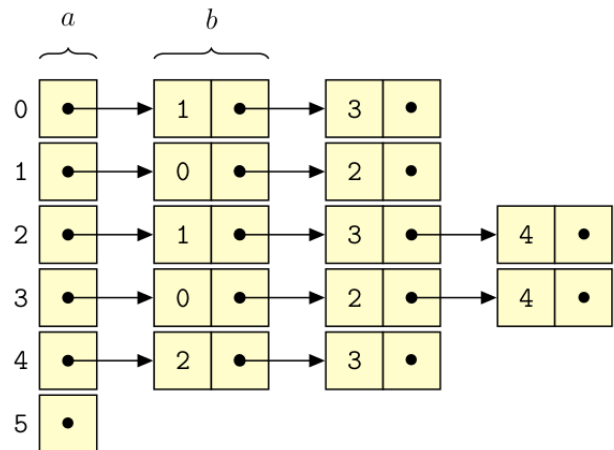
Per le *liste di adiacenza* invece, un *arco* (u, v) deve essere memorizzato sia tra i *nod*i *adiacenti* a u che a v . Di conseguenza, il numero di *archi* da memorizzare raddoppia e, conseguentemente, lo spazio necessario diventa $O(a \cdot n + 2b \cdot m)$ che anche sta volte è comunque $O(n + m)$.



(a) *Grafo non orientato*

	0	1	2	3	4	5
0		1	0	1	0	0
1			1	0	0	0
2				1	1	0
3					1	0
4						0
5						

(b) *Matrice di adiacenza*



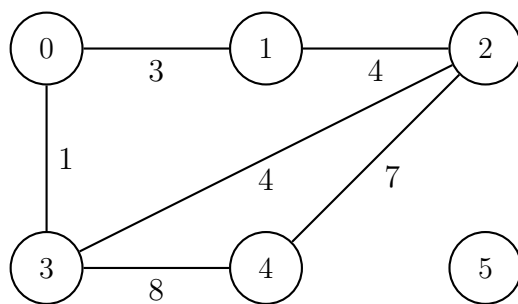
(c) *Lista di adiacenza*

Fig. 7.5: Memorizzazione di un *grafo non orientato*

Giunti a questo punto possiamo dire che, a parità di metodo di memorizzazione scelto, lo spazio necessario per memorizzare un *grafo orientato* non è diverso da quello che servirebbe se il *grafo* fosse *non orientato*.

Memorizzazione di un grafo pesato Un caso interessante è quello in cui si debba memorizzare un *grafo pesato*, cioè un *grafo* nel quale ad ogni *arco* è associato un costo (o un profitto). Matematicamente possiamo pensare il peso di un *arco* come l'applicazione di una funzione di peso del tipo: $w : V \times V \rightarrow \mathbb{R}$. L'unico caso particolare è quello in cui non esista un *arco* tra due *vertici* e quindi, a seconda del problema che si sta affrontando, si può decidere di assegnare un peso nullo o infinito.

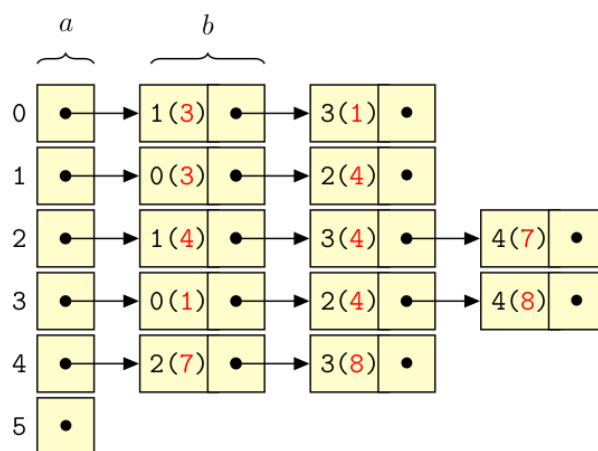
Ciò che cambia nella memorizzazione è che nella *matrice di adiacenza*, invece di indicare un valore binario, si indica il peso dell'*arco*, mentre nella *lista di adiacenza* si aggiunge il valore del peso all'interno di ogni *nodo*.



(a) Grafo pesato non orientato

	0	1	2	3	4	5
0	0	3	0	1	0	0
1		0	4	0	0	0
2			0	4	7	0
3				0	8	0
4					0	0
5						0

(b) Matrice di adiacenza



(c) Lista di adiacenza

Fig. 7.6: Memorizzazione di un grafo pesato non orientato

Dettagli sull'implementazione Se non diversamente specificato, di seguito assumeremo che i *grafi* siano memorizzati come *vettori di adiacenza*, ovvero *liste* implementate tramite *vettori* (*dinamici* o *statici*). Inoltre, consideriamo $O(1)$ il costo per l'accesso alle informazioni di un oggetto di classe **NODE** e per l'inserimento e la rimozione di *nodi* e *archi* dai *grafi*. Ipotizzeremo inoltre, che dopo l'inizializzazione i *grafi* siano *statici* e che quindi non sia più possibile modificarli.

Fatte queste premesse, possiamo concludere questa sezione discutendo del *costo* complessivo di alcune operazioni basilari sui *grafi*.

Frammento 44 - Iterazione su *nodi* e *archi*.

```
% Iterazione su tutti i nodi
foreach (u ∈ G.V()) do
    { Operazioni sul nodo u }
```

```
% Iterazione su tutti i nodi e tutti
gli archi
foreach (u ∈ G.V()) do
    { Operazioni sul nodo u }
    foreach (v ∈ G.adj(u)) do
        { Operazioni sull'arco (u,v) }
```

Matrici di adiacenza

- Lo spazio richiesto per la memorizzazione è $O(n^2)$;
- Il tempo richiesto per verificare se un nodo u è *adiacente* a un altro nodo v è $O(1)$;
- Il tempo necessario per iterare su tutti gli *archi* è $O(n^2)$

Liste di adiacenza

- Lo spazio richiesto per la memorizzazione è $O(n + m)$;
- Il tempo richiesto per verificare se un nodo u è *adiacente* a un altro nodo v è $O(n)$;
- Il tempo necessario per iterare su tutti gli *archi* è $O(n + m)$

In generale, le *matrici di adiacenza* sono adatte per memorizzare *grafi densi* e, al contrario, per *grafi sparsi* risulta più convenienti l'utilizzo delle *liste*.

7.2 Visite di un grafo

Definizione 68 - Visita di un grafo.

Dati, un grafo $G = (V, E)$ e un vertice $r \in V$ detto radice o sorgente, visitare un grafo significa visitare una e una sola volta tutti i nodi del grafo che possono essere raggiunti da r .

Come per gli *alberi*, esistono *visite in profondità* e *in ampiezza*. Un primo approccio alla *visita* di un *grafo* potrebbe essere il doppio ciclo visto in precedenza:

Frammento 45 - Visita scriteriata di un grafo.

```
% Iterazione su tutti i nodi e tutti gli archi
foreach (u ∈ G.V()) do
    { Operazioni sul nodo u }
    foreach (v ∈ G.adj(u)) do
        { Operazioni sull'arco (u,v) }
```

Questo approccio tuttavia non va bene in quanto non viene tenuto conto della struttura del *grafo* e, cosa più importante, non segue alcun criterio nell'ordine di visita dei *nodi* e degli *archi*. A dirla tutta, questo approccio viola la definizione stessa di *visita* in quanto non offre alcuna garanzia sul fatto che *nodi* e *archi* verranno visitati una sola volta.

7.2.1 Visita in ampiezza

Nella *visita in ampiezza* andiamo a visitare i *nodi* in ordine crescente di distanza dalla *radice*. Ciò può essere utile per calcolare il *cammino* più breve da r a tutti gli altri *nodi* e quindi riuscire a generare un *albero breadth-first*.

Definizione 69 - Albero breadth-first.

Un albero breadth-first è un albero contenente tutti i nodi raggiungibili da r e tale per cui il cammino dalla radice r al nodo u nell'albero corrisponde al cammino più breve da r a u nel grafo.

Dunque, per l'implementazione della *visita in ampiezza* possiamo partire da quanto fatto per gli *alberi generici* ipotizzando di trattare i *nodi adiacenti* come *figli*. Con le opportune modifiche si ottiene una funzione di questo tipo:

Frammento 46 - Implementazione errata visita in ampiezza di un grafo.

```
bfsTraversal(GRAPH G, NODE r)
    QUEUE q = Queue()
    q.enqueue(r)
    while (not q.isEmpty()) do
        NODE u = q.dequeue()
        { Visita il nodo u }
        foreach (v ∈ G.adj(u)) do
            q.enqueue(v)                                % Accoda tutti i nodi adiacenti a u
```

Questa implementazione, come la precedente, è errata e il motivo è che non solo lo stesso *nodo* può essere visitato più volte, ma la funzione non termina mai l'esecuzione. Infatti, se *u* e *v* sono *adiacenti*, quando viene visitato *u* viene messo in coda *v* e quando viene visitato *v* viene messo in coda *u*. È quindi necessario utilizzare un sistema che permetta di tenere traccia dei *nodi* già visitati.

Frammento 47 - Implementazione generica visita in ampiezza di un grafo.

```
graphTraversal(GRAPH G, NODE r)
    SET s = Set()                                       % Insieme generico
    s.insert(r)                                         % Da specificare
    { Marca il nodo r }
    while (s.size() > 0) do
        NODE u = s.remove()                             % Da Specificare
        foreach (v ∈ G.adj(u)) do
            { Visita l'arco (u,v) }
            if (v non è stato ancora marcato) then
                { Marca il nodo v }
                s.insert(v)                             % Da specificare
```

Frammento 48 - Implementazione visita in ampiezza di un grafo.

```
bfs(GRAPH G, NODE r)
    QUEUE q = Queue()
    q.enqueue(r)
    boolean[] visited = new boolean[G.size()]          % Vettore marcatore
    foreach (u ∈ G.V() - {r}) do                       % Estrae tutti i nodi tranne r
        visited[u] = false
    visited[r] = true                                    % Marca r
    while (not q.isEmpty()) do
        NODE u = q.dequeue()
        { Visita il nodo u }
        foreach (v ∈ G.adj(u)) do
```

```

    { Visita l'arco (u,v) }
    if (not visited[v]) then % Controlla se il nodo è già stato visitato
        visited[v] = true % Marca v
        q.enqueue(v)

```

7.2.2 Visita in profondità

La *visita in profondità* viene usata per visitare tutti i *nodi* e non solo quelli raggiungibili da una singola sorgente come avviene con la *visita in ampiezza*. Proprio per questo, la ricerca restituisce una *foresta* di *alberi depth-first* e non un solo *albero*.

La funzione `dfs` per la *visita in profondità* può essere realizzata in modo ricorsivo o iterativo. In entrambi i casi si fa uso di una *pila*, ma nel caso della ricorsione la *pila* ha una dimensione massima definita dal sistema operativo, quindi per grafi di grandi dimensioni è bene considerare attentamente il tipo di implementazione più opportuno.

Frammento 49 - Implementazione ricorsiva visita in profondità di un grafo.

```

dfs(GRAPH G, NODE u, boolean[] visited)
    visited[u] = true
    { Visita il nodo u } % Pre-order
    foreach (v ∈ G.adj(u)) do
        if (not visited[v]) then
            { Visita l'arco (u,v) }
            dfs(G, v, visited)
    { Visita il nodo u } % Post-order

```

Frammento 50 - Implementazione iterativa visita in profondità di un grafo.

```

dfs(GRAPH G, NODE u)
    STACK s = Stack()
    s.push(r)
    boolean[] visited = new boolean[G.size()]
    foreach (u ∈ G.V()) do
        visited[u] = false
    while (not s.isEmpty()) do
        NODE u = s.pop()
        if (not visited[u]) then
            { Visita il nodo u } % Pre-order
            visited[u] = true
            foreach (v ∈ G.adj(u)) do
                { Visita l'arco (u,v) }
                s.push(v)

```

È bene esplicitare alcuni particolari dell'implementazione iterativa. In particolare, un *nodo* può essere inserito più volte nella *pila*, in quanto il controllo sul marcatore viene effettuato al momento dell'estrazione e non dell'inserimento.

La *visita in post-order* prevede che alla scoperta di un *nodo*, questo venga inserito nella *pila* impostando un tag `discovery`. Alla successiva estrazione di quel *nodo*, il tag `discovery` viene cambiato in `finish` e il *nodo* viene inserito nuovamente nella *pila*. Alla terza estrazione viene effettuata la *visita*. L'utilizzo dei tag `discovery` e `finish` garantisce che un *nodo* con tag `finish` venga estratto soltanto dopo che tutti i suoi vicini sono già stati inseriti.

7.2.3 Complessità delle visite

Tutte e tre le implementazioni viste per i due tipi di *visite* hanno *complessità* $O(n + m)$. Nello specifico, l'implementazione iterativa della *dfs* costa $O(n + m)$ perché effettua $O(m)$ *visite* degli *archi*, $O(m)$ inserimenti ed estrazioni e $O(n)$ *visite* dei *nodi*.

7.3 Problemi sui grafi risolubili con visite in ampiezza

Le *visite in ampiezza* e in *profondità* sono solitamente sfruttate all'interno di altri algoritmi per risolvere problemi più complessi.

L'utilizzo più importante delle *visite in ampiezza* è nella ricerca del *cammino più breve* tra due *nodi*. Tuttavia, prima di arrivare a discutere quel problema, facciamo una piccola digressione e andiamo a vedere come calcolare la distanza tra un *nodo* e tutti gli altri.

7.3.1 Calcolo della distanza tra nodi

Per distanza tra *nodi* si intende il numero di *archi* che collegano un *nodo* a tutti gli altri. Il *nodo* di partenza ha distanza 0 da se stesso, i suoi vicini, ovvero i *nodi* ad esso *adiacenti*, hanno distanza 1 e così via. In generale, se un *nodo* u è *adiacente* ad un *nodo* che è a distanza k e, contemporaneamente, non è *adiacente* a nessun *nodo* che ha distanza inferiore a k dal *nodo* di partenza, allora u è a distanza $k + 1$.

NB. Il concetto di distanze appena espresso deriva dai cosiddetti *Numeri di Erdős*. Erdős fu un matematico estremamente prolifico che scrisse più di 1500 articoli con più di 500 coautori. I *numeri di Erdős* descrivono la distanza di una persona da Erdős. Quindi, Erdős ha valore *erdos* = 0, i suoi coautori hanno *erdos* = 1 e, in generale, una persona coautore di qualcuno con *erdos* = k e che non è coautore con nessuno che abbia *erdos* < k , ha *erdos* = $k + 1$.

Quindi, l'algoritmo per il calcolo della distanza di un *nodo* dagli altri, non fa altro che calcolare il *numero di Erdős* di ogni *nodo*, e lo fa, sfruttando il meccanismo delle *visite in ampiezza*.

Frammento 51 - Implementazione distance per il calcolo dei numeri di Erdős.

```
distance(GRAPH G, NODE r, int[] distances)
    QUEUE q = Queue()
    q.enqueue(r)
    foreach (u ∈ G.V() - {r}) do
        distances[u] = ∞                % Tutti i nodi non visitati sono a ∞
    distances[r] = 0
    while (not q.isEmpty()) do
        NODE u = q.dequeue()
        foreach (v ∈ G.adj(u)) do
            if (distances[v] == ∞) then    % Un nodo mai visitato ha erdos = ∞
                distances[v] = distance[u] + 1    % Se u ha erdos = k, v ha k+1
                q.enqueue(v)
```

7.3.2 Ricerca del cammino più breve fra due nodi

Partendo dalla funzione per il calcolo dei *numeri di Erdős* è possibile implementare una funzione che, dato un *nodo* r , costruisca l'*albero di copertura* con *radice* r . Quell'*albero* è realizzato tramite un *vettore dei padri* e può essere utilizzato per determinare, non solo la distanza di r da uno dei *nodi*, ma anche il *cammino semplice* più breve.

Frammento 52 - Implementazione distance per la ricerca dei cammini brevi.

```
distance(GRAPH G, NODE r, int[] distances, NODE[] parents)
    QUEUE q = Queue()
    q.enqueue(r)
    foreach (u ∈ G.V() - {r}) do
        distances[u] = ∞                % Tutti i nodi non visitati sono a ∞
        parents[u] = nil                % Tutti i nodi non visitati non avranno un padre
    distances[r] = 0
    parents[r] = nil                    % r è la radice dell'albero di copertura
    while (not q.isEmpty()) do
        NODE u = q.dequeue()
        foreach (v ∈ G.adj(u)) do
            if (distances[v] == ∞) then    % Un nodo mai visitato ha erdos = ∞
                distances[v] = distance[u] + 1    % Se u ha erdos = k, v ha k + 1
                parents[v] = u                % Popola il vettore dei padri
                q.enqueue(v)
```

Il cammino semplice tra due nodi può poi essere stampato con la seguente funzione.

Frammento 53 - Implementazione printPath per la stampa del cammino tra due nodi.

```
printPath(NODE r, NODE s, NODE[] parents)
    if (r == s) then                    % Cammino tra un nodo e se stesso
        print s
    else if (parents[s] == nil) then % Se il padre è nil, non esiste un cammino
        print "error"
    else
        printPath(r, parents[s], parents)
        print s
```

Complessità Entrambe le funzioni hanno *complessità lineare* $O(n + m)$ perché si basano sull'algoritmo per la visita in ampiezza.

7.4 Problemi sui grafi risolubili con visite in profondità

7.4.1 Ricerca delle componenti connesse

Definizione 70 - Raggiungibilità.

Un nodo v è raggiungibile da un nodo u se esiste almeno un cammino da u a v .

NB. Nei grafi non orientati il concetto di raggiungibilità è simmetrico.

Definizione 71 - Grafo connesso.

Un grafo non orientato $G = (V, E)$ è connesso se e solo se ogni suo nodo è raggiungibile da ogni altro nodo.

Definizione 72 - Sottografo.

$G' = (V', E')$ è un sottografo di $G = (V, E)$, ovvero $G' \subseteq G$ se e solo se $V' \subseteq V$ e $E' \subseteq E$.

Definizione 73 - Grafo massimale.

G' è massimale se e solo se non esiste un altro sottografo G'' di G che sia connesso e più grande di G' , ovvero se non esiste un grafo G'' che soddisfi la seguente relazione:

$$G' \subseteq G'' \subseteq G$$

Definizione 74 - Componente connessa.

Un grafo $G' = (V', E')$ è una componente connessa di $G = (V, E)$ se e solo se è un sottografo connesso e massimale di G .

Fatte salve queste definizioni, possiamo dire che una *visita in profondità* può sia permetterci di capire se un grafo è *connesso* o meno, che di conoscere le sue *componenti connesse*. Per fare ciò, è sufficiente usare un *vettore* contenente gli identificatori delle *componenti connesse*. In particolare, il valore in posizione u del *vettore* è l'identificatore della *componente connessa* alla quale appartiene il *nodo* u .

Frammento 54 - Implementazione cc per la ricerca delle componenti connesse.

```
int[] cc(GRAPH G)
    int[] id = new int[G.size()]           % Vettore degli identificatori
    foreach (u ∈ G.V()) do
        id[u] = 0
    int counter = 0                         % Contatore Componenti Connesse
    foreach (u ∈ G.V()) do
        if (id[u] == 0) then
            counter = counter + 1
            ccdfs(G, counter, u, id)
    return id

% Funzione ausiliaria che marca tutti i nodi tra loro raggiungibili
ccdfs(GRAPH G, int counter, NODE u, int[] id)
    id[u] = counter
    foreach (v ∈ G.adj(u)) do
        if (id[v] == 0) then
            ccdfs(G, counter, v, id)
```

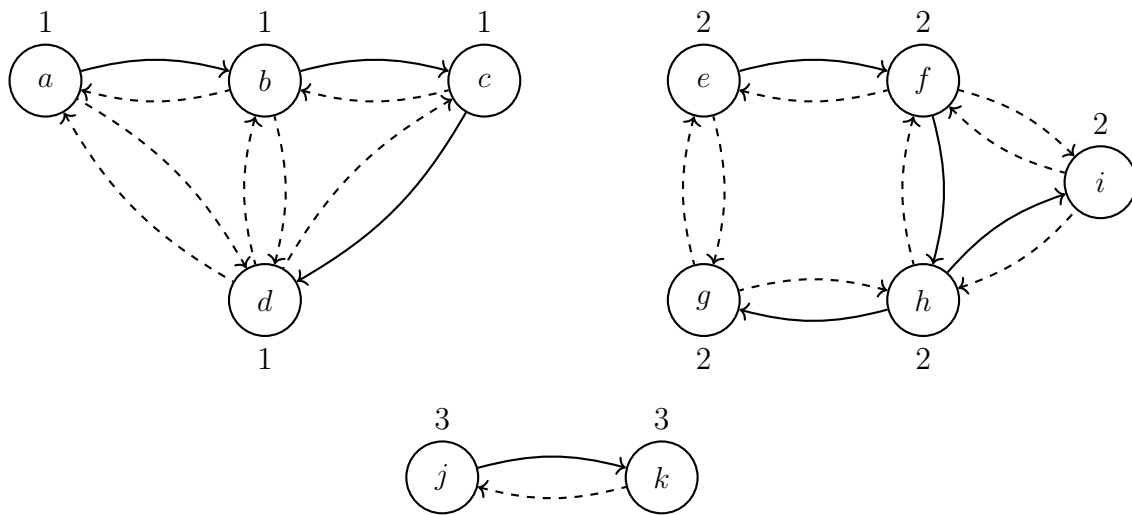


Fig. 7.7: Grafo con tre componenti connesse

NB. Nella figura di cui sopra è rappresentato un *grafo orientato*, ma poiché per ogni coppia di nodi esiste un *arco* in entrambe le direzioni, il *grafo* può anche essere visto come *non orientato*.

7.4.2 Verifica di esistenza di cicli nei grafi non orientati

Definizione 75 - Ciclo in un grafo non orientato.

In grafo non orientato $G = (V, E)$, un ciclo C di lunghezza $k > 2$ è una sequenza di nodi $u_0, u_1 \dots, u_k$ tale che $(u_i, u_{i+1}) \in E \ \forall 0 \leq i \leq k-1$ e $u_0 = u_k$.

NB. Il vincolo $k > 2$ serve per escludere i *cicli* banali composti da coppie di archi (u, v) e (v, u) che sono onnipresenti nei *grafi non orientati*.

Definizione 76 - Grafo aciclico.

Un grafo non orientato che non contiene cicli è detto *aciclico*.

Definizione 77 - Grafo ciclico.

Un grafo che contiene almeno un ciclo è detto *essere ciclico*.

Riuscire a capire se un *grafo non orientato* contiene cicli, non è difficile. È sufficiente modificare l'algoritmo per la *visita in profondità* aggiungendo un parametro che ad ogni invocazione contenga un riferimento all'ultimo *nodo* visitato. Quindi, si esegue la normale *visita in profondità*, evitando però di visitare il *nodo* indicato dal parametro, e, se si arriva a un altro *nodo* già visitato, allora si è in presenza di un *ciclo*.

Frammento 55 - Implementazione hasCycle per la verifica di esistenza di cicli in grafi non orientati.

```
boolean hasCycle(GRAPH G)
    boolean[] visited = new boolean[G.size()]
    foreach (u ∈ G.V()) do
```

```

    visited[u] = false
foreach (u ∈ G.V()) do
    if (not visited[u]) then
        if (hasCycleRec(G, u, nil, visited)) then
            return true
        return false

% Funzione ausiliaria
boolean hasCycleRec(GRAPH G, NODE u, NODE p, boolean[] visited)
    visited[u] = true
    foreach (v ∈ G.adj(u) - {p}) do      % Visita tutti i nodi adiacenti tranne p
        if (visited[v]) then
            return true
        else if (hasCycleRec(G, v, u, visited)) then
            return true
    return false

```

7.4.3 Verifica di esistenza di cicli nei grafi orientati

Definizione 78 - Ciclo in un grafo orientato.

In un grafo orientato $G = (V, E)$, un ciclo C di lunghezza $k \geq 2$ è una sequenza di nodi u_0, u_1, \dots, u_k tale che $(u_i, u_{i+1}) \in E \forall 0 \leq i \leq k-1$ e $u_0 = u_k$.

NB. Nei grafi orientati i cicli banali formati da coppie di archi (u, v) e (v, u) sono accettati.

NB. I cicli in cui tutti i nodi ad eccezione del primo e l'ultimo sono distinti sono detti *cicli semplici*.

Definizione 79 - Grafi orientati aciclici - DAG.

Un grafo orientato che non contiene cicli è detto essere un DAG (Directed Acyclic Graph).

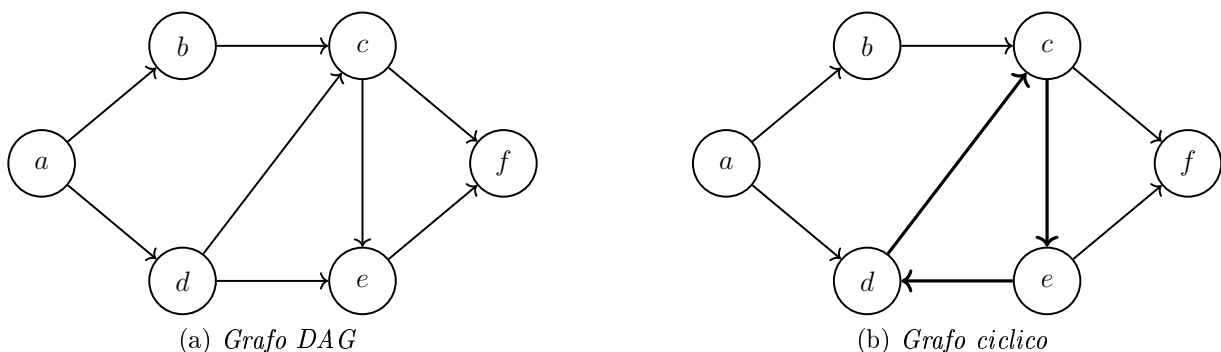


Fig. 7.8: Grafi DAG VS grafi ciclici

Verificare l'esistenza di *cicli* all'interno di *grafi orientati* non è semplice come nei *grafi non orientati*, infatti non è possibile usare l'algoritmo visto in precedenza, ma ne serve uno ad-hoc.

Definizione 80 - Archi degli alberi di copertura.

Ogni volta che si esamina un arco da un nodo marcato a uno non marcato, tale arco viene detto *arco dell'albero*.

Quando si effettua una *visita*, l'*albero di copertura* risultante non contiene tutti gli *archi* del grafo. Gli archi (u, v) non compresi nell'*albero di copertura* T possono essere classificati in tre modi:

- Se u è un antenato di v in T , (u, v) è detto *arco in avanti*;
- Se u è un discendente di v in T , (u, v) è detto *arco all'indietro*;
- Se (u, v) non è né un *arco in avanti*, né un *arco all'indietro*, è un *arco di attraversamento*;

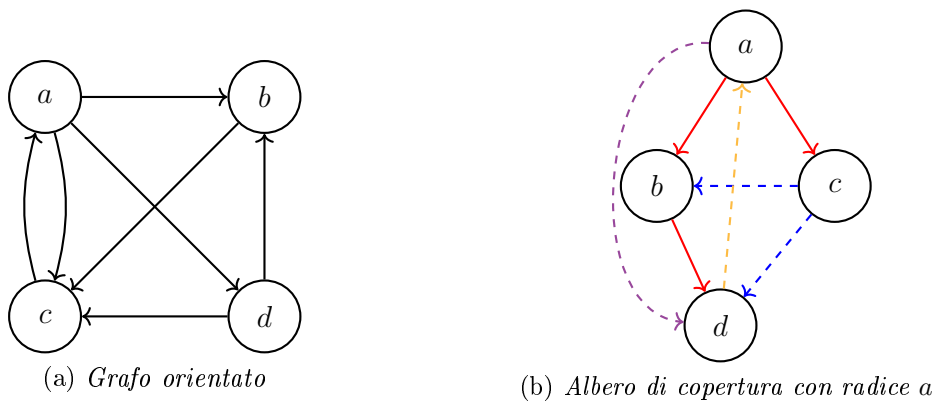


Fig. 7.9: *Classificazione degli archi*

Nella figura si vede l'*albero di copertura* ottenuto dopo aver effettuato una *visita in profondità* partendo dal *nodo* a . Gli archi rossi sono gli *archi dell'albero di copertura*, mentre gli archi (a, d) e (d, a) sono, rispettivamente, un *arco in avanti* e un *arco all'indietro*. Infine, gli archi uscenti da c sono *archi di attraversamento*.

Frammento 56 - Implementazione dfs_schema per la classificazione degli archi.

```
dfs_schema(GRAPH G, NODE u, int &time, int[] dt, int[] ft)
    time = time + 1                                % Aggiorna il contatore del tempo
    dt[u] = time                                    % Imposta il tempo di scoperta del nodo u
    foreach (v ∈ G.adj(u)) do
        if (dt[v] == 0) then                        % Arco dell'albero
            { Visita l'arco (u,v) }
            dfs_schema(G, v, time, dt, ft)
        else if (dt[u] > dt[v] and ft[v] == 0) then % Arco all'indietro
            { Visita l'arco (u,v) }
        else if (dt[u] < dt[v] and ft[v] ≠ 0) then % Arco in avanti
            { Visita l'arco (u,v) }
        else                                         % Arco di attraversamento
            { Visita l'arco (u,v) }
    time = time + 1                                % Aggiorna il contatore del tempo
    ft[u] = time                                    % Imposta il tempo di fine visita del nodo u
```

In questa funzione, il *tempo di scoperta* viene impostato quando per la prima volta si visita un *nodo*, mentre il *tempo di fine visita* quando sono già stati visitati tutti i *nodi adiacenti*.

Definizione 81 - Teorema di caratterizzazione delle coppie di nodi.

Data una visita in profondità di un grafo $G = (V, E)$, per ogni coppia di nodi $u, v \in V$, vale una delle seguenti condizioni:

- Gli intervalli $[dt[u], ft[u]]$ e $[dt[v], ft[v]]$ non si intersecano, né sovrappongono in alcun modo dunque, u e v , nella foresta *depth-first*, non sono discendenti l'uno dell'altro;
- L'intervallo $[dt[u], ft[u]]$ è contenuto in $[dt[v], ft[v]]$ dunque, in uno degli alberi *depth-first*, u è un discendente di v ;
- L'intervallo $[dt[u], ft[u]]$ contiene $[dt[v], ft[v]]$ dunque, in uno degli alberi *depth-first*, u è un antenato di v ;

Grazie a questa definizione possiamo dire che, dati due nodi u e v , se vale la seguente:

$$[dt[u], ft[u]] \subset [dt[v], ft[v]]$$

valgono le seguenti affermazioni:

- Il nodo v è un *antenato* di u ;
- Il nodo u è un *discendente* di v ;
- L'arco (v, u) , se non è un arco dell'albero, è un arco *in avanti*;
- L'arco (u, v) , se non è un arco dell'albero, è un arco *all'indietro*;

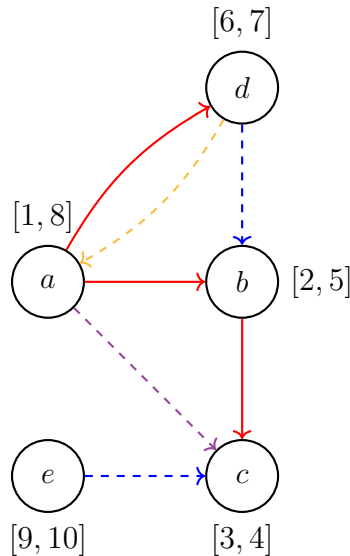


Fig. 7.10: Classificazione degli archi e intervalli di tempo

Nella figura di cui sopra, si è effettuata una *visita in profondità* a partire dal nodo a e si sono visitati, nell'ordine, i nodi a, b, c, d . Gli archi (a, b) , (b, c) e (a, d) sono *archi dell'albero* poiché quando b, c e d sono stati scoperti avevano certamente $dt = 0$.

L'intervallo associato al nodo d è contenuto in quello associato al nodo a , quindi l'arco (d, a) , poiché non è un *arco dell'albero*, è un *arco all'indietro*. In realtà, quando l'algoritmo

arriva a visitare il *nodo* d , $ft[a]$ non è ancora stato impostato, ma dato che la condizione $dt[u] > dt[v]$ and $ft[v] == 0$ con $u = d$ e $v = a$ risulta verificata, l'*arco* (d, a) è effettivamente un *arco all'indietro*.

L'*arco* (a, c) , invece, è un *arco in avanti* perché verifica la relazione sugli intervalli, ovvero $[dt[c], ft[c]] \subset [dt[a], ft[a]]$, ma anche perché durante l'esecuzione dell'algoritmo la condizione $dt[u] < dt[v]$ and $ft[v] \neq 0$ con $u = a$ e $v = c$ viene verificata.

Infine, il *nodo* e non appartiene alla *componente connessa* di a , quindi non fa nemmeno parte dello stesso *albero di copertura*. L'*arco* (e, c) non soddisfa nessuna condizione quindi è un *arco di attraversamento*.

Fatta questa digressione, possiamo ritornare al problema originale: verificare l'esistenza di *cicli* in un *grafo orientato*. Enunciamo il seguente teorema.

Definizione 82 - Teorema di esistenza dei cicli.

Un grafo orientato è aciclico se e solo se non esistono archi all'indietro.

Dimostrazione.

- (\Rightarrow) : si supponga esista un *ciclo*. Sia (v, u) un *arco* del *ciclo* e sia u il primo *nodo* ad essere visitato. Prima o poi verrà visitato il *cammino* che connette u a v e da v verrà scoperto l'*arco all'indietro* (v, u) ;
- (\Leftarrow) : se esiste un *arco all'indietro* (u, v) , dove v è un antenato di u , allora esistono un *cammino* da v a u e un *arco* da u a v , ovvero un *ciclo*;

□

Quindi, per verificare l'esistenza di un *ciclo* è sufficiente cercare un *arco all'indietro*. Se ne esiste almeno uno, esisterà sicuramente un anche un *ciclo*.

Frammento 57 - Implementazione hasCycle per la verifica di esistenza di cicli in un grafo orientato.

```
boolean hasCycle(GRAPH G, NODE u, int &time, int[] dt, int[] ft)
    time = time + 1
    dt[u] = time
    foreach (v ∈ G.adj(u)) do
        if (dt[v] == 0) then
            if (hasCycle(G, v, time, dt, ft)) then
                return true
        else if (dt[u] > dt[v] and ft[v] == 0) then
            return true
    time = time + 1
    ft[u] = time
    return false
```

7.4.4 Realizzare un ordinamento topologico di un grafo orientato

Definizione 83 - Ordinamento topologico.

Dato un grafo orientato aciclico (DAG) G , un ordinamento topologico di G è un ordinamento lineare dei suoi nodi tale che se $(u, v) \in E$, allora u appare prima di v nell'ordinamento.

NB. Se il *grafo* è *aciclico* esistono più ordinamenti possibili, ma se esiste anche solo un *ciclo* non ne esiste alcuno.

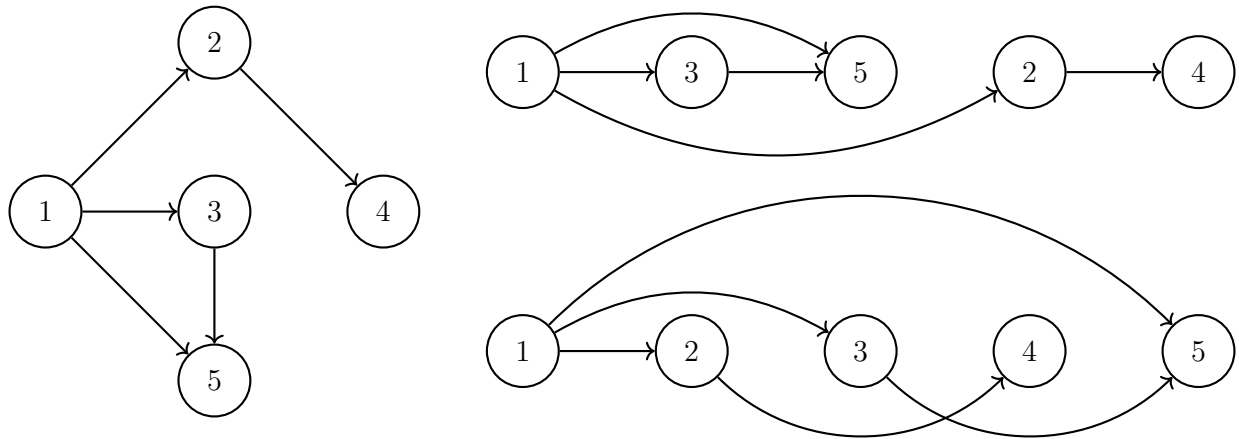


Fig. 7.11: Possibili *ordinamenti topologici* di un *grafo*

Approccio naive Un primo approccio a questo problema potrebbe essere quello in cui, scelto un *nodo* privo di *archi entranti*, lo si aggiunge all'ordinamento e si rimuovono tutti i suoi *archi*. Quindi, si ripete l'operazione per tutti gli altri *nodi*.

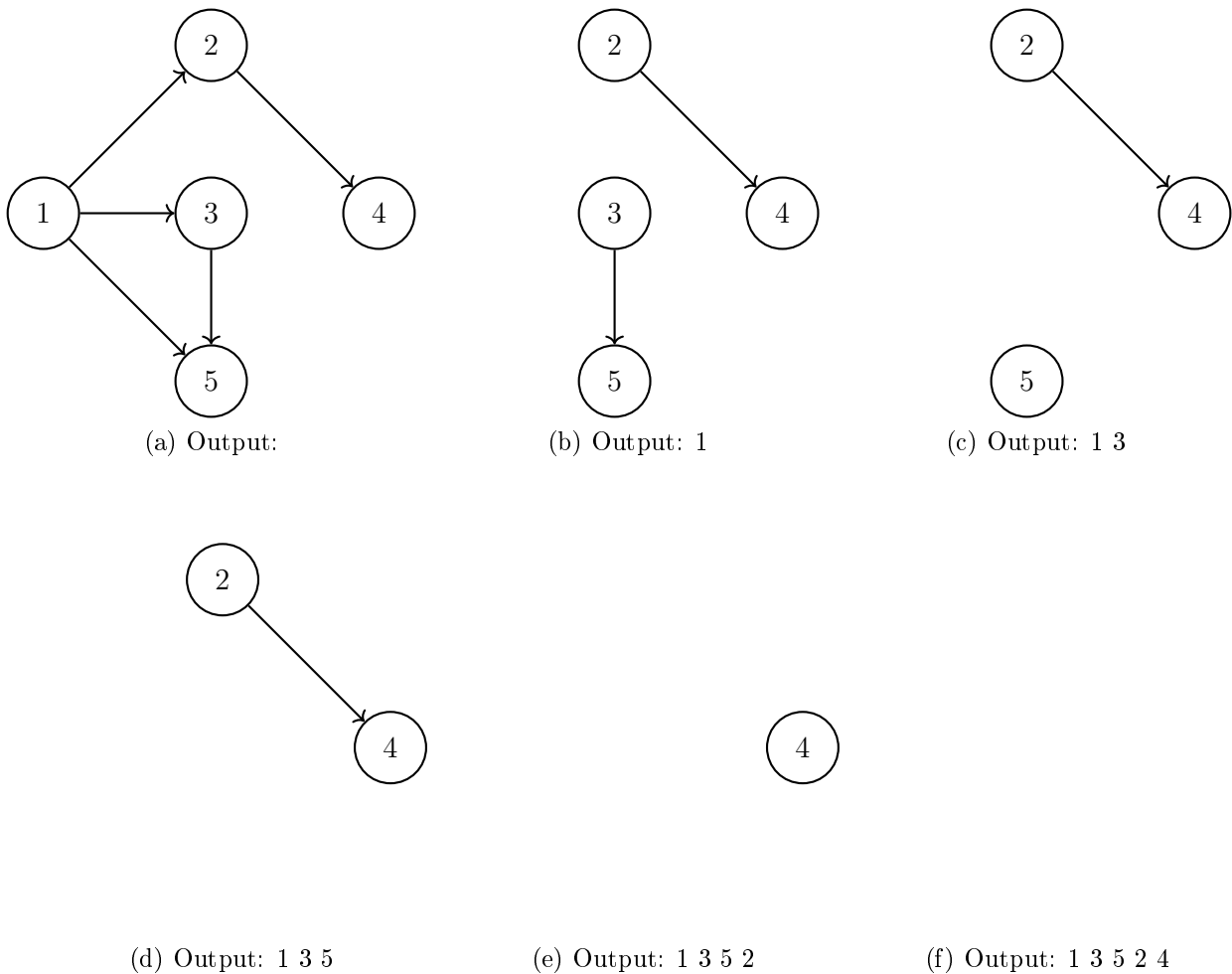


Fig. 7.11: Esecuzione dell'algoritmo "naive" per l'*ordinamento topologico*

Approccio efficiente Un approccio migliore prevede di effettuare una *visita in Post-order* inserendo tutti i *nodi* in una *pila* a mano a mano che li si visita. L'utilizzo della *visita in Post-order* fa sì che un *nodo* venga inserito solo quando tutti i suoi *discendenti* sono già stati scoperti ed aggiunti alla *pila*, e, l'utilizzo della *pila* permette di estrarre i *nodi* in ordine inverso a quello di *inserimento*. Ciò, permette di ottenere un corretto *ordinamento topologico*.

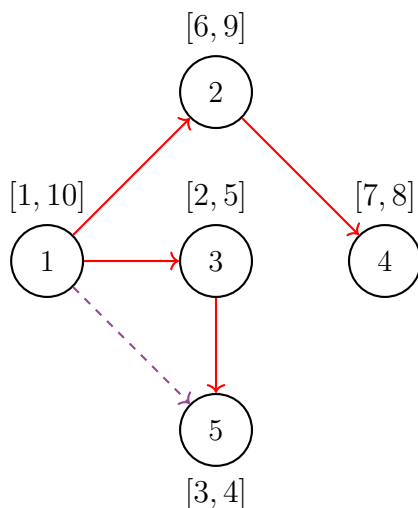
Frammento 58 - Implementazione topSort per l'ordinamento topologico di un grafo DAG.

```

STACK topSort(GRAPH G)
    STACK s = Stack()
    boolean[] visited = new boolean[G.size()]
    foreach (u ∈ G.V()) do
        visited[u] = false
    foreach (u ∈ G.V()) do
        if (not visited[u]) then
            ts-dfs(G, u, visited, s)           % Effettua la visita in Post-order
    return s

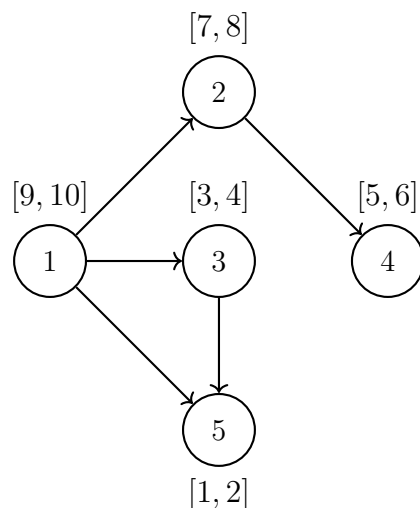
% Funzione ausiliaria
ts-dfs(GRAPH G, NODE u, boolean[] visited, STACK s)
    visited[u] = true
    foreach (v ∈ G.adj(u)) do
        if (not visited[v]) then
            ts-dfs(G, v, visited, s)
    s.push(u)

```



Stack = {1, 2, 4, 3, 5}

(a) *Ordinamento topologico* partendo da 1



Stack = {1, 2, 4, 3, 5}

(b) *Ordinamento topologico* partendo da 5

Fig. 7.12: Esempi di *ordinamenti topologici*

7.4.5 Ricerca delle componenti fortemente connesse

Definizione 84 - Grafo fortemente connesso.

Un grafo orientato $G = (V, E)$ è *fortemente connesso* se e solo se ogni suo nodo è raggiungibile da ogni altro nodo.

Definizione 85 - Componente fortemente connessa.

Un grafo $G' = (V', E')$ è una *componente fortemente connessa* di $G = (V, E)$ se e solo se G' è un sottografo fortemente connesso e massimale di G .

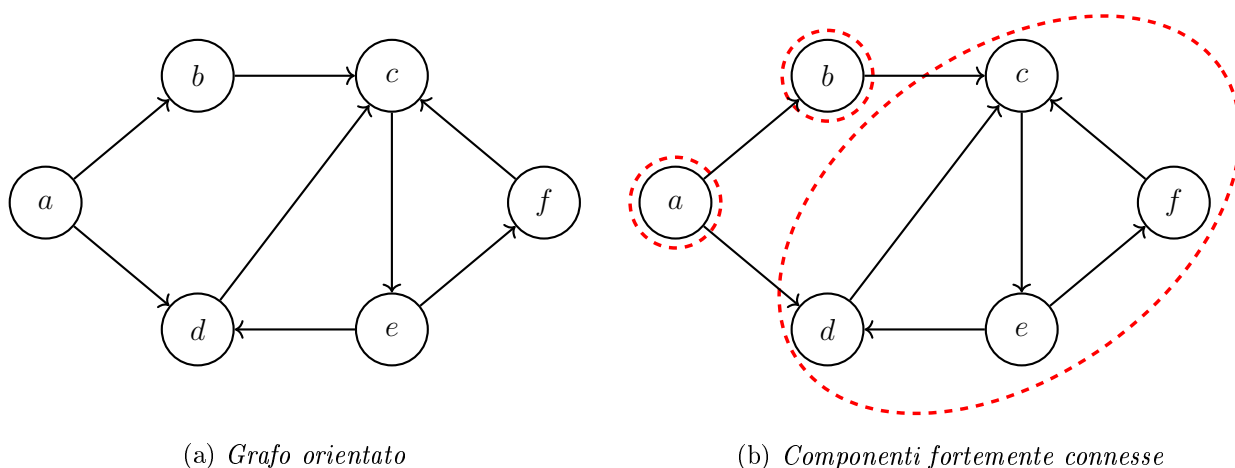


Fig. 7.13: Componenti fortemente connesse di un grafo

Una soluzione ingenua potrebbe essere quella di applicare `cc` al *grafo*, ma purtroppo il risultato dipenderebbe dal *nodo* di partenza. Ad esempio, nella figura di cui sopra, se si applicasse `cc` partendo da *a* si otterrebbe un'unica componente connessa, mentre applicandola da *b* se ne otterrebbero due. La soluzione corretta è quella fornita dall'*Algoritmo di Kosaraju*.

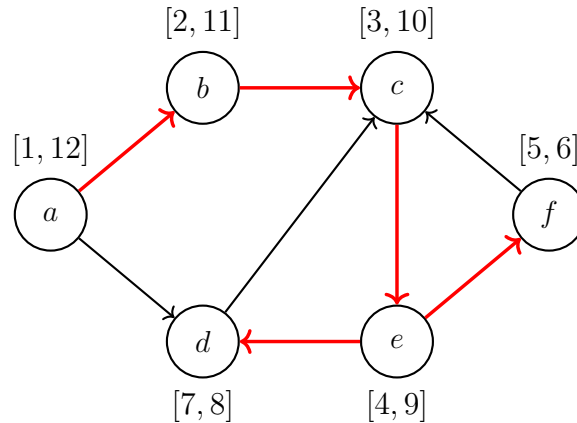
Frammento 59 - Implementazione Algoritmo di Kosaraju.

```
int[] scc(GRAPH G)
    STACK S = topsort(G)                                % Prima visita
    GRAPH GT = transpose(G)                            % Calcolo del grafo trasposto
    return cc(GT, S)                                    % Seconda visita
```

Utilizzo della `topsort` La prima cosa che dovrebbe farci storcere il naso è l'utilizzo della funzione `topsort` per *grafi* non *aciclici*. Quando abbiamo parlato di *ordinamento topologico* avevamo infatti posto come condizione l'*aciclicità* dei *grafi*, vincolo necessario per poter definire un ordine di visita dei *nodi*. In questo caso però, una *componente connessa* non banale, che quindi include più di un *nodo*, è sicuramente un *ciclo* e l'ordine di visita dei *nodi* in un *ciclo* non ha importanza.

Di conseguenza, applicando la **topsort** su un *grafo* generico siamo sicuri che:

- Se l'*arco* (u, v) non appartiene a un *ciclo*, il *nodo* u compare prima di v nell'ordinamento;
- Se l'*arco* (u, v) appartiene a un *ciclo*, i *nodi* compaiono in un qualche ordine ininfluenza;



Stack = {a, b, c, e, d, f}

Fig. 7.14: Ordinamento topologico di un *grafo* generico

NB. L'utilizzo della **topsort** permette di ottenere i *nodi* di un *grafo* in ordine decrescente di tempo di fine.

Utilizzo della traspose

Definizione 86 - Grafo trasposto.

Dato un *grafo* orientato $G = (V, E)$, il *grafo* trasposto $G_T = (V, E_T)$ ha gli stessi *nodi* e archi di G , ma gli archi sono orientati in senso opposto. Cioè:

$$E_T = \{(u, v) \mid (v, u) \in E\}$$

Frammento 60 - Implementazione traspose per la generazione di un *grafo* trasposto.

```
GRAPH traspose(GRAPH G)
    GRAPH GT = Graph()                                % Crea un grafo vuoto
    foreach (u ∈ G.V()) do
        GT.insertNode(u)                                % Aggiunge a GT tutti i nodi di G
    foreach (u ∈ G.V()) do
        foreach (v ∈ G.adj(u)) do
            GT.insertEdge(v, u)                        % Aggiunge a GT gli archi di G invertendoli
```

Il costo di questa funzione è *lineare*, cioè $O(n + m)$, perché vengono aggiunti tutti i *nodi* e tutti gli *archi*.

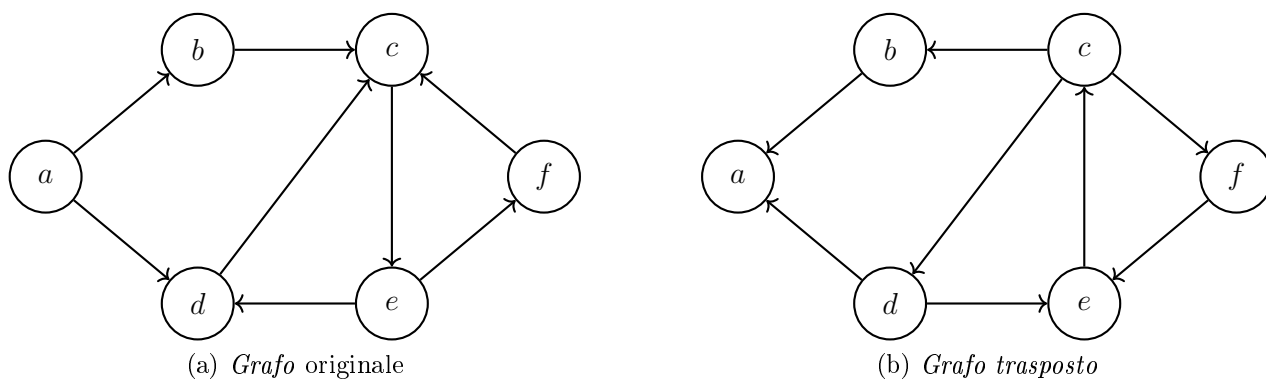


Fig. 7.15: Grafo orientato e relativo grafo trasposto

Utilizzo della cc Usiamo un'implementazione della cc diversa rispetto a quella vista precedentemente. In particolare, l'ordine di visita dei *nod*i viene stabilito in base all'ordine in cui questi sono stati inseriti nella *pila* ricevuta come parametro.

Frammento 61 - Implementazione alternativa di cc.

```
int[] cc(GRAPH G, STACK S)
    int[] id = new int[G.size()]
    foreach (u ∈ G.V()) do
        id[u] = 0
    int counter = 0
    while (not S.isEmpty()) do
        NODE u = S.pop()
        if (id[u] == 0) then
            counter = counter + 1
            ccdfs(G, counter, u, id)
    return id

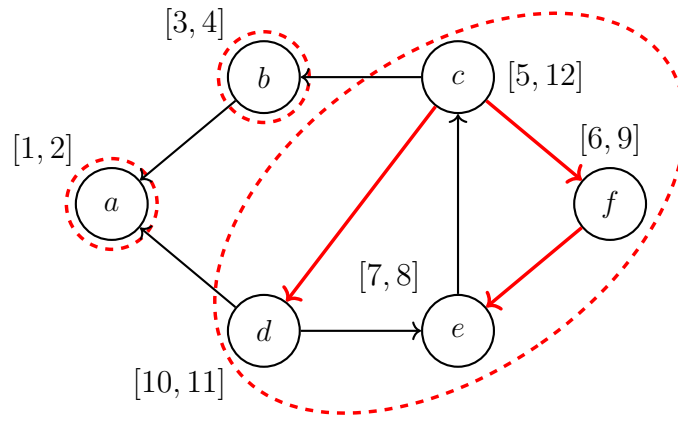
% L'implementazione della funzione ausiliaria non cambia
ccdfs(GRAPH G, int counter, NODE u, int[] id)
    id[u] = counter
    foreach (v ∈ G.adj(u)) do
        if (id[v] == 0) then
            ccdfs(G, counter, v, id)
```

Quindi, per ricercare le *componenti fortemente connesse*, applichiamo l'algoritmo di ricerca delle *componenti connesse* al grafo trasposto e procediamo a visitare i *nod*i nell'ordine stabilito dall'*ordinamento topologico* del grafo originale.

Questo significa che, nell'esempio che stiamo considerando, la cc inizia esaminando il *nodo* *a*. Poiché nel grafo trasposto *a* non ha alcun *arco uscente*, quel singolo *nodo* viene marcato come appartenente alla *componente connessa* 1. Alla seconda iterazione del ciclo **while** viene estratto il *nodo* *b*. Nel grafo trasposto *b* è adiacente soltanto ad *a* che è già stato marcato, quindi *b* viene assegnato alla *componente connessa* banale con identificativo 2.

Infine viene estratto *c* che fa parte di due *cicli*. Poiché da *c* possono essere raggiunti tutti i *nod*i che fanno parte di quei *cicli*, i *nod*i *c, e, f, d*, vengono assegnati a un'unica *componente connessa*.

Al termine dell'esecuzione di cc siamo riusciti ad identificare con successo tutte e tre le *componenti fortemente connesse* del grafo.



Stack = {a, b, c, f, e, d}

Fig. 7.16: Componenti fortemente connesse del grafo

NB. Per quanto riguarda i *nod*i della *componente fortemente connessa* contenente c , l'ordine di estrazione dei *nod*i di quella *componente* non ha importanza.

Costo computazionale Poiché tutte e tre le funzioni che vengono richiamate dalla `scc` hanno costo $O(n + m)$ anche la `scc` ha la stessa *complessità*.

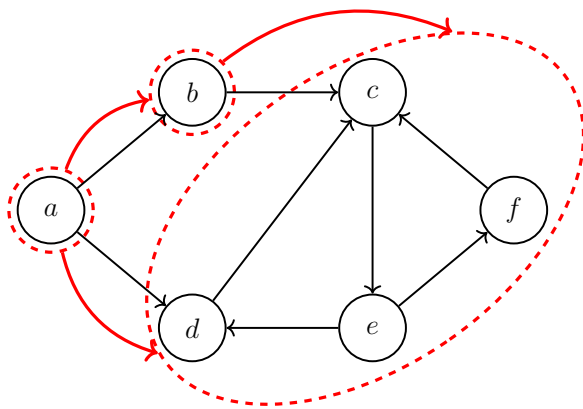
Dimostrazione di correttezza

Definizione 87 - Grafo delle componenti.

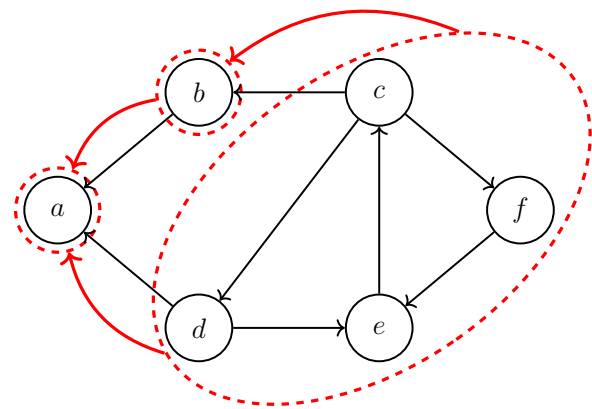
Dato un grafo $G = (V, E)$, il suo grafo delle componenti $C(G) = (V_C, E_C)$ è un grafo tale che:

- $V_C = \{C_1, C_2, \dots, C_k\}$, dove C_i è la i -esima componente fortemente connessa di G ;
- $E_C = \{(C_i, C_j) \mid \exists (u_i, u_j) \in E : u_i \in C_i \wedge u_j \in C_j\}$

NB. Dati un grafo G e il suo trasposto G^T , vale $C(G^T) = [C(G)]^T$.



(a) Grafo delle componenti del grafo originale



(b) Grafo delle componenti del grafo trasposto

Fig. 7.17: Grafi delle componenti del grafo d'esempio e del suo trasposto

NB. I grafi delle componenti sono sempre *aciclici*.

Definizione 88 - Relazione tra discovery e finish time nei grafi delle componenti.

Dato un nodo C di un grafo delle componenti, valgono le seguenti equivalenze:

$$dt(C) = \min\{dt(u) \mid u \in C\}$$

$$ft(C) = \max\{ft(u) \mid u \in C\}$$

NB. Il *discovery* e *finish time* di una componente fortemente connessa corrispondono al *discovery* e *finish time* del primo nodo visitato in quella componente.

Definizione 89 - Teorema di ordinamento delle componenti fortemente connesse.

Siano C e C' due distinte componenti fortemente connesse del grafo orientato $G = (V, E)$. Se esiste un arco $(C, C') \in E_C$, allora $ft(C) > ft(C')$.

NB. Le componenti fortemente connesse vengono visitate in ordine decrescente di tempo di fine.

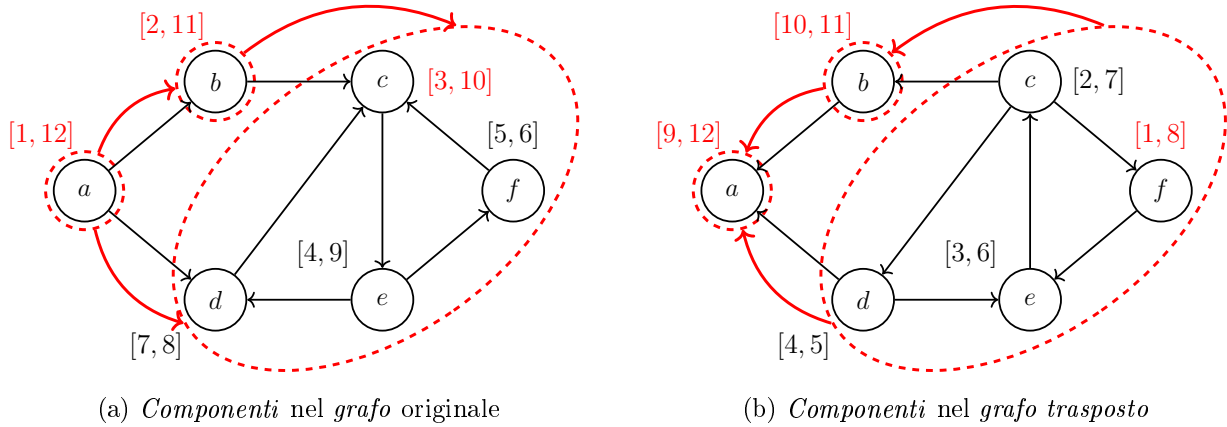


Fig. 7.18: *Discovery* e *finish time* delle componenti

NB. Quando si considera un arco tra due componenti fortemente connesse, in realtà, bisogna considerare l'arco che collega i due nodi appartenenti ad una e all'altra componente.

Nelle immagini sopra si vede bene come le componenti banali a e b siano collegate dall'arco (a, b) nel grafo originale e da (b, a) nel grafo trasposto. Quindi, se C_a è la componente che contiene a e C_b è la componente che contiene b , vale $ft(C_a) > ft(C_b)$. Equivalentemente, possiamo dire che nel grafo trasposto vale $ft(C_b) < ft(C_a)$.

A questo punto abbiamo tutti gli strumenti per dimostrare la correttezza dell'Algoritmo di Kosaraju.

Dimostrazione. Se le componenti C_x e C_y sono connesse mediante un arco $(x, y) \in E_T$, sicuramente $ft(C_x) < ft(C_y)$. Di conseguenza, la visita di C_y inizierà prima di quella di C_x . Inoltre, poiché in $C(G)$ esiste l'arco (C_x, C_y) non può esistere un cammino da C_y a C_x , perché altrimenti il grafo sarebbe ciclico e quindi la visita di C_y non raggiungerà mai nodi appartenenti a C_x .

In altre parole, la funzione cc assegnerà correttamente tutti gli identificatori delle componenti. \square

Capitolo Nr.8

Hashing

Nei capitoli precedenti abbiamo affrontato il concetto di **Dizionario** e abbiamo visto come la scelta della *struttura dati* si ripercuota sulla *complessità* delle operazioni.

In particolare, abbiamo osservato le seguenti *complessità*:

Operazione	Array non ordinato	Array ordinato	Lista	Alberi RB
insert()	$O(1), O(n)$	$O(n)$	$O(1), O(n)$	$O(\log n)$
lookup()	$O(n)$	$O(\log n)$	$O(n)$	$O(\log n)$
remove()	$O(n)$	$O(n)$	$O(n)$	$O(\log n)$
foreach	$O(n)$	$O(n)$	$O(n)$	$O(n)$

Esiste una *struttura dati* che ci permetta di fare meglio di così?

La risposta è sì e si chiama *hash table* (o *tabella hash*) e ci consente di ottenere *complessità* $O(1)$ in tutte le operazioni (tranne ovviamente nel **foreach**). Vediamo quindi che cos'è una *tabella hash*.

Definizione 90 - Tabella hash.

Una *tabella hash* è una *struttura dati dinamica* che permette di memorizzare associazioni chiave-valore. L'insieme delle possibili chiavi è rappresentato da un insieme universo \mathcal{U} di dimensione u e le chiavi sono memorizzati in un vettore $T[0..m-1]$ di dimensione m detto *tabella hash*.

Ogni elemento di \mathcal{U} è associato ad una sola posizione all'interno della tabella e questa associazione è definita da una *funzione hash*.

Definizione 91 - Funzione hash.

Una *funzione hash* è una *funzione definita* come $h : \mathcal{U} \rightarrow \{0, \dots, m-1\}$.

Le *funzioni hash* sono tali per cui la coppia chiave-valore $\langle k, v \rangle$ è memorizzata nella tabella alla posizione $h(k)$.

NB. L'insieme universo \mathcal{U} ha potenzialmente una dimensione infinità, mentre la *tabella hash* ha una dimensione limitata, quindi è sicuro che più elementi saranno associati alla stessa posizione.

Definizione 92 - Collisione.

Quando due o più chiavi nel dizionario hanno lo stesso valore hash diciamo che è avvenuta una *collisione*.

Idealmente, vorremmo realizzare *funzioni hash* che non generino collisioni.

Tabelle ad accesso diretto In alcuni casi in cui l'insieme delle chiavi è noto a priori ed è un sottoinsieme “piccolo” di \mathbb{Z}^+ (e.g insieme dei giorni dell'anno) è possibile definire la *funzione hash* come la funzione identità, ovvero:

$$h(k) = k \quad \forall k \in \mathcal{U} \subset \mathbb{Z}^+$$

e definire una *tabella hash* di dimensione $m = u$.

In situazioni di questo tipo, ovviamente, non si genereranno mai *collisioni*, ma purtroppo non sono quasi mai praticabili perché se u è molto grande è richiesto uno spazio eccessivo se vengono usate poche delle possibili chiavi si va a sprecare memoria.

8.1 Caratteristiche e implementazioni di funzioni hash

Andiamo a vedere come possono essere definite le *funzioni hash* e quali sono le loro caratteristiche.

Definizione 93 - Funzione hash perfetta.

Una funzione hash h si dice perfetta se è iniettiva, ovvero se vale:

$$\forall k_1, k_2 \in \mathcal{U} : k_1 \neq k_2 \Rightarrow h(k_1) \neq h(k_2)$$

NB. Le funzioni iniettive non danno mai origine a *collisioni*.

Per le stesse ragioni espresse per le *tabelle ad accesso diretto*, è molto difficile poter ottenere una *funzione hash perfetta* e inoltre, lo spazio delle chiavi è spesso grande, sconosciuto e sparso.

Quindi, se non possiamo evitare che si generino *collisioni* cerchiamo almeno di minimizzarne il numero. Per farlo cerchiamo *funzioni hash* che distribuiscano in modo uniforme le chiavi all'interno della *tabella hash*. Ma cosa significa “in modo uniforme”?

Definizione 94 - Uniformità semplice.

Siano $P(k)$ la probabilità che una chiave k sia inserita nella tabella e $Q(i)$ la probabilità che una chiave venga inserita nella posizione i , cioè:

$$Q(i) = \sum_{k \in \mathcal{U} : h(k)=i} P(k)$$

Una funzione hash gode di uniformità semplice se vale:

$$Q(i) = \frac{1}{m} \quad \forall i \in [0, \dots, m-1]$$

Esempio 19 - Funzione hash con uniformità semplice.

Se \mathcal{U} è l'insieme dei numeri reali $[0, 1[$ e ogni chiave ha la stessa probabilità di essere scelta, la funzione hash $h(k) = \lfloor km \rfloor$ gode di uniformità semplice.

Notiamo che poter ottenere una *funzione hash* con *uniformità semplice* dobbiamo conoscere la distribuzione delle probabilità P e quindi si ripropongono gli stessi problemi di prima: insieme delle chiavi sconosciuto in principio. Per questo motivo, nella realtà si usano tecniche “euristiche”.

8.1.1 Realizzare una funzione hash

Assumiamo che ogni chiave sia traducibile in valori numerici interpretando la propria rappresentazione in memoria come un valore intero.

Ad esempio, ipotizziamo di avere a disposizione le seguenti funzioni per trasformare una stringa in un intero:

Frammento 62 - Funzioni per la manipolazioni di chiavi stringhe.

```
% Restituisce il valore ordinale binario di c in una qualche codifica1
bin bin(char c)
% Restituisce la rappresentazione binaria di una chiave k, concatenando
% i valori binari dei caratteri che la compongono
bin bin(string k)
% Restituisce il valore numerico associato al numero binario b
int int(bin b)
% Restituisce la traduzione in valore intero di una chiave k
int int(string k)
    return int(bin(k))
```

Esempio 20 - Tradurre una chiave stringa in un intero.

Calcolare il valore intero corrispondente alla chiave $k = \text{"DOG"}$ di tipo **string**.

Applichiamo la funzione $\text{bin}(\text{"DOG"})$ e calcoliamo i valori binari dei caratteri della stringa "DOG":

$$\text{ord}('D') = 01000100 \quad \text{ord}('O') = 01001111 \quad \text{ord}('G') = 01000111$$

Quindi, restituiamo la concatenazione dei tre valori binari:

$$01000100 \ 01001111 \ 01000111$$

Ora, usando la funzione $\text{int}(\text{"DOG"})$ traduciamo in intero la stringa:

$$\text{int}(01000100) = 68 \cdot 256^2 \quad \text{int}(01001111) = 79 \cdot 256^1 \quad \text{int}(01000111) = 71 \cdot 256^0$$

Terminiamo restituendo la somma dei tre valori:

$$68 \cdot 256^2 + 79 \cdot 256^1 + 71 \cdot 256^0 = 4476743$$

Come facciamo a trasformare quel numero in un valore compreso tra 0 e $m - 1$?

8.1.2 Metodo dell'estrazione

Il primo metodo sarebbe quello di scegliere $m = 2^p$ e definire la *funzione hash* come la traduzione in intero di p bit scelti tra i bit di $\text{bin}(k)$, ovvero:

$$h(k) = \text{bin}(b) \quad \text{dove} \quad b \subset \text{bin}(k) : \#b = p$$

Il problema di questa tecnica è che ha un alta probabilità di generare *collisioni*.

¹D'ora in avanti useremo la codifica binaria ASCII a 8 bit

Esempio 21 - Estrazione dei bit meno significativi.

Sia $m = 2^p = 2^{16}$ e si supponga che $\text{bin}(k)$ restituisca i 16 bit meno significativi della rappresentazione.

$$\begin{aligned}\text{bin}(\text{"Alberto"}) &= 01110100\ 01101111 \Rightarrow h(\text{"Alberto"}) = 29\ 807 \\ \text{bin}(\text{"Roberto"}) &= 01110100\ 01101111 \Rightarrow h(\text{"Roberto"}) = 29\ 807\end{aligned}$$

Esempio 22 - Estrazione di un gruppo arbitrario di bit.

Sia $m = 2^p = 2^{16}$ e si supponga che $\text{bin}(k)$ restituisca i 16 bit che vanno dal quinto al ventesimo (estremi inclusi).

$$\begin{aligned}\text{bin}(\text{"Alberto"}) &= 0001\ 01101100\ 0110 \Rightarrow h(\text{"Alberto"}) = 5\ 830 \\ \text{bin}(\text{"Alessio"}) &= 0001\ 01101100\ 0110 \Rightarrow h(\text{"Alessio"}) = 5\ 830\end{aligned}$$

In entrambi i casi abbiamo rilevato delle *collisioni*.

8.1.3 Metodo dello XOR

Prendiamo sempre $m = 2^p$, ma questa volta la *funzione hash* restituisce la traduzione in intero di un valore b ottenuto effettuando la somma modulo 2, bit-a-bit, di sottoinsiemi di dimensione p di $\text{bin}(k)$. Questo tipo di somma altro non è se non la funzione XOR.

Tuttavia, neanche questo metodo va bene perché permutazioni della stessa stringa possono generare lo stesso valore.

Esempio 23 - Utilizzo della funzione XOR.

$$\begin{aligned}\text{Sia } m &= 2^p = 2^{16}. \\ \text{bin}(\text{"Montresor"}) &= \begin{array}{llll} 01101101 & 01101111 & \oplus & \\ 01101110 & 01110100 & \oplus & \\ 01110010 & 01100101 & \oplus & \\ 01110011 & 01101111 & \oplus & \\ 01110010 & 00000000 & \oplus & \\ = 01110000 & 00010001 & & \end{array} \\ \text{bin}(\text{"Sontremor"}) &= \begin{array}{llll} 01110011 & 01101111 & \oplus & \\ 01101110 & 01110100 & \oplus & \\ 01110010 & 01100101 & \oplus & \\ 01101101 & 01101111 & \oplus & \\ 01110010 & 00000000 & \oplus & \\ = 01110000 & 00010001 & & \end{array}\end{aligned}$$

Le due codifiche sono uguali quindi $h(\text{"Montresor"}) = h(\text{"Sontremor"}) = 28\ 689$.

8.1.4 Metodo della divisione

In questo caso scegliamo per m un valore dispari e possibilmente primo. La *funzione hash* è definita come $h(k) = \text{int}(k) \bmod m$.

Esempio 24 - Utilizzo del metodo della divisione.

Sia $m = 383$.

$$\begin{aligned}h(\text{"Alberto"}) &= 18\ 415\ 043\ 350\ 787\ 183 \bmod 383 = 221 \\ h(\text{"Alessio"}) &= 18\ 415\ 056\ 470\ 632\ 815 \bmod 383 = 77 \\ h(\text{"Cristin"}) &= 4\ 860\ 062\ 892\ 481\ 405\ 294 \bmod 383 = 130\end{aligned}$$

Siamo finalmente riusciti a non ottenere alcuna *collisione*.

Abbiamo deciso di assegnare ad m un numero primo perché se avessimo scelto una potenza di 2, l'operatore di modulo avrebbe finito per farci considerare solo i p bit meno significativi facendoci così ritornare alla situazione del *metodo dell'estrazione*. Inoltre, anche un valore del tipo $2^p - 1$ avrebbe creato problemi in quanto si può dimostrare che permutazioni di stringhe con un set di caratteri di dimensione 2^p hanno sempre lo stesso *hash*.

In definitiva, questo metodo funziona se ad m viene assegnato un valore primo che sia “distante” da potenze di 2 o di 10.

8.1.5 Metodo della moltiplicazione

Per questo metodo scegliamo un m qualsiasi, ma è meglio se è una potenza di due. Definiamo poi una costante reale C compresa tra 0 e 1 (estremi esclusi). A questo punto, se $i = \text{int}(k)$, la *funzione hash* è definita come $h(k) = \lfloor m(C \cdot i - \lfloor C \cdot i \rfloor) \rfloor$.

Notiamo che $C \cdot i - \lfloor C \cdot i \rfloor$ estrae la componente decimale di $C \cdot i$.

Esempio 25 - Utilizzo del metodo della moltiplicazione.

Siano $m = 2^{16}$ e $C = \frac{\sqrt{5}-1}{2}$ ².

$$\begin{aligned} h(\text{“Alessio”}) &= 65\,536 \cdot 0.51516739168 = 33\,762 \\ h(\text{“Alberto”}) &= 65\,536 \cdot 0.78732161432 = 51\,598 \\ h(\text{“Cristian”}) &= 65\,536 \cdot 0.72143641000 = 47\,280 \end{aligned}$$

Per questo metodo è consigliato scegliere un valore di m che sia una potenza di 2 perché consente di rendere più efficiente l'implementazione dell'algoritmo.

Infatti, se $m = 2^p$ e w è la dimensione in bit di una parola in memoria, sia per $i = \text{bit}(k)$ che per m , vale $i, m \leq 2^w$. Se poi $s = \lfloor C \cdot 2^w \rfloor$, $i \cdot s$ può essere espresso come $r_1 \cdot 2^w + r_0$ dove r_1 e r_0 contengono rispettivamente la parte intera e frazionaria di $i \cdot C$. Infine, il valore $h(k)$ di ogni chiave corrisponde ai p bit più significativi di r_0 .

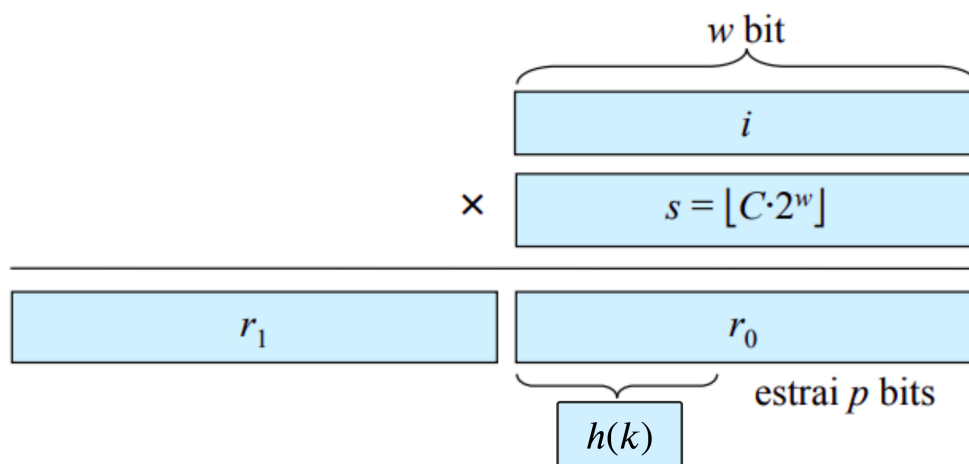


Fig. 8.1: Implementazione efficiente del *metodo della moltiplicazione*

In definitiva, il *metodo della moltiplicazione* sembra offrire una soluzione sufficientemente efficace anche se in realtà, la *funzione hash* ottenuta non gode di *uniformità semplice* e non viene mai usata in applicazioni reali.

² $C = \frac{\sqrt{5}-1}{2}$ è il valore suggerito da Knuth, l'ideatore dell'algoritmo

8.2 Gestione delle collisioni

Se abbiamo stabilito che non possiamo garantire l'assenza di *collisioni* dobbiamo definire una metodologia per la loro gestione. In particolare, se, in fase di inserimento, la posizione in cui dovremmo inserire una chiave è già occupata, dovremo trovarne una alternativa e, allo stesso modo, se in fase di ricerca la chiave cercata non è nella posizione attesa, dovremo cercarla altrove. Questa ricerca dovrebbe costare $O(1)$ nel caso medio e $O(n)$ nel caso pessimo.

Le tecniche possibili sono due: *liste di trabocco* e *indirizzamento aperto*. Queste due sono anche chiamate rispettivamente tecniche a *memorizzazione esterna* e *interna*.

8.2.1 Liste di trabocco

Questa metodologia prevede che tutte le chiavi con lo stesso *hash* vengano memorizzate in una *lista monodirezionale* o in un *vettore dinamico*. Ogni cella della *tabella hash* poi conterrà un puntatore alla testa della *lista* o al primo elemento del *vettore* associato a quella posizione.

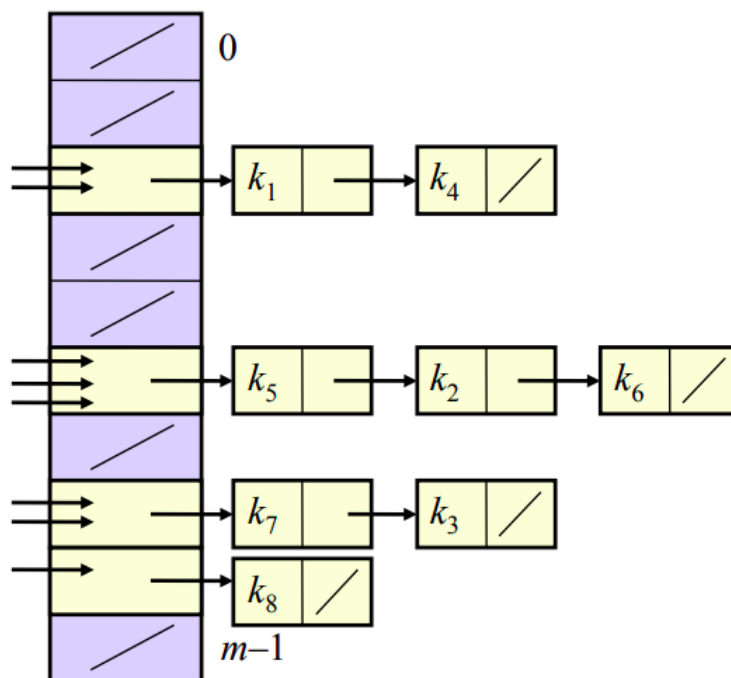


Fig. 8.2: Struttura di una *tabella hash* con *liste di trabocco*

La funzione `insert()` va ad aggiungere in testa la chiave, mentre la `lookup()` e la `remove()` scansionano la *lista* fino a trovare la chiave cercata.

Complessità Per studiare la *complessità* di questa soluzione abbiamo bisogno di definire alcuni parametri:

- n : numero di chiavi memorizzate nella *tabella hash*;
- m : dimensione della *tabella hash*;
- $\alpha = \frac{n}{m}$: *fattore di carico* della *tabella hash*;
- $I(\alpha)$: numero medio di accessi per la ricerca di una chiave che non è presente nella *tabella hash* (ricerca con insuccesso);

- $S(\alpha)$: numero medio di accessi per la ricerca di una chiave che è presente nella *tabella hash* (ricerca con successo);

Ovviamente il caso pessimo è quello in cui $m = 1$ e quindi tutte le chiavi sono inserite in un'unica lista e quindi la *complessità* è pari a quella che si avrebbe in un *dizionario* implementato come *vettore non ordinato* o come *lista*.

Studiare il caso medio invece, ci richiede di fare delle assunzioni. In particolare, ipotizziamo che la *funzione hash* costi $\Theta(1)$ e che consenta un *hashing uniforme*. Fatte queste ipotesi, ci aspettiamo che ogni *lista/vettore* abbia una lunghezza pari ad $\alpha = \frac{n}{m}$.

In una ricerca con insuccesso devono ovviamente essere toccati tutti i valori di una lista, mentre nelle ricerche con successo, in media, ne vengono toccati la metà. Questo, unito al fatto che calcolare la posizione di ricerca costa $\Theta(1)$, ci porta ad avere, rispettivamente, un costo $\Theta(1) + \alpha$ e $\Theta(1) + \frac{\alpha}{2}$ per le ricerche con insuccesso e con successo.

Ma qual è il significato del *fattore di carico*?

Il parametro influenza il costo delle operazioni sulla *tabella hash*. Per esempio, se $n = O(m)$, $\alpha = O(1)$ e quindi tutte le operazioni costano $O(1)$.

8.2.2 Indirizzamento aperto

Il problema delle *liste di trabocco* è che ci costringono a realizzare una *struttura dati* complessa. L'idea alla base delle implementazioni a *indirizzamento aperto* invece, è di memorizzare tutte le chiavi nella *tabella stessa* in modo tale che ogni posizione contenga una chiave, oppure **nil**.

Quindi, se in fase di inserimento la posizione calcolata è già occupata, se ne sceglie un'altra. Per la ricerca invece, si parte dalla posizione prescelta e si visitano tutte le posizioni alternative fino a quando non si trova la chiave cercata. Per affrontare nel dettaglio questa tecnica implementativa diamo alcune definizioni:

Definizione 95 - Ispezione.

Un'ispezione è l'esame di una posizione durante la ricerca.

Ovviamente, il numero massimo di ispezioni coincide con la dimensione m della *tabella hash*. Modifichiamo la definizione di *funzione hash* in modo da includere il concetto di *ispezione*.

Definizione 96 - Funzione hash per tabelle ad indirizzamento aperto.

In una tabella hash implementata con la tecnica dell'indirizzamento aperto, la funzione hash H è definita come

$$H : \mathcal{U} \times \overbrace{\{0, \dots, m-1\}}^{\text{Numero di ispezione}} \rightarrow \overbrace{\{0, \dots, m-1\}}^{\text{Posizione nella tabella}}$$

Definizione 97 - Sequenza di ispezione.

Una sequenza di ispezione $[H(k, 0), H(k, 1), \dots, H(k, m-1)]$ è una permutazione degli indici $[0, \dots, m-1]$ corrispondente all'ordine in cui vengono visitate le posizioni della tabella.

Generalmente non dovrebbe servire visitare tutte le posizioni, ma in ogni caso, vogliamo evitare di visitarne più volte la stessa.

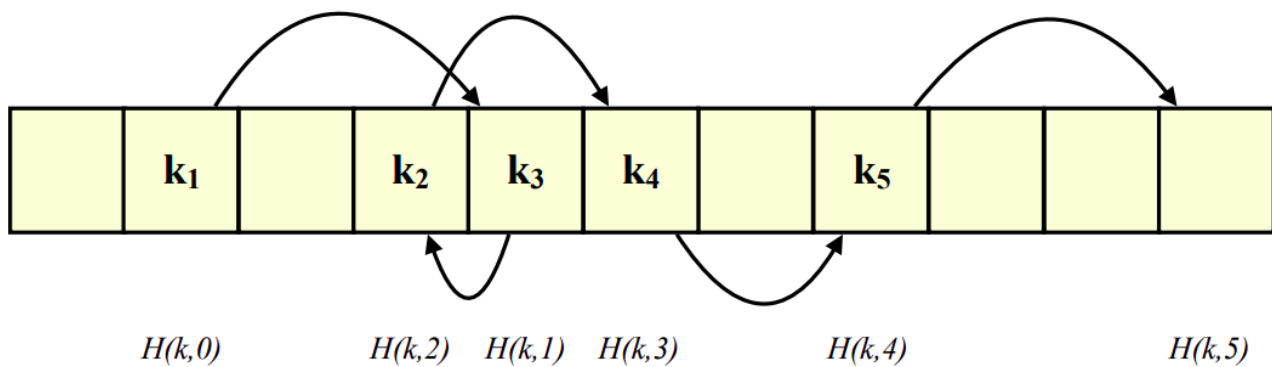


Fig. 8.3: Esempio di *sequenza di ispezione*

Ma cosa succede al *fattore di carico*?

Siccome $0 \leq n \leq m$, anche α è compreso tra 0 e 1 (estremi inclusi). Questo comporta che la *tabella hash* potrebbe andare in overflow, ovvero occupare tutte le posizioni.

Generalizzando il concetto di *uniformità semplice*, diamo la definizione di *hashing uniforme*.

Definizione 98 - Hashing uniforme.

Si parla di hashing uniforme quando una chiave ha la stessa probabilità di avere come sequenza di ispezione una qualsiasi delle $m!$ permutazioni di $[0, \dots, m-1]$.

Nella realtà è difficile raggiungere l'*hashing uniforme*, per cui ci si accontenta di ottenere alcune delle possibili permutazioni.

Le *sequenze di ispezione* dipendono dalla tecnica di ispezione che si usa e le più diffuse sono: *ispezione lineare*, *quadratica* e *doppio hashing*.

Ispezione lineare Questa tecnica definisce la *funzione hash* come:

$$H(k, i) = (H_1(k) + h \cdot i) \bmod m$$

In questa definizione, h è il passo della sequenza e di conseguenza la *sequenza di ispezione* per una chiave k diventa: $[H_1(k), H_1(k) + h, H_1(k) + 2h, \dots]$.

Questa tecnica ci permette di avere al massimo m possibili *sequenze* che sono tutte determinabili dalla posizione della prima *ispezione*. Inoltre, c'è il problema della cosiddetta *aggregazione primaria* (o *primary clustering*) che comporta la formazione di sotto sequenze sempre più lunghe di posizioni occupate, col risultato che una posizione libera preceduta da i posizioni occupate viene riempita con probabilità $\frac{i+1}{m}$ e i tempi di inserimento e cancellazione crescono sempre di più.

Ispezione quadratica L'*ispezione quadratica* segue lo stesso principio di quella *lineare*, ma le *ispezioni* hanno un offset che dipende da una funzione quadratica del numero di *ispezione*. La *funzione hash* è infatti definita come:

$$H(k, i) = (H_1(k) + h \cdot i^2) \bmod m$$

Anche in questo caso sono possibili m *sequenze di ispezione*, ma nessuna di esse è una permutazione e si propone il problema dell'*aggregazione secondaria* (o *secondary clustering*) ovvero, chiavi con la stessa posizione iniziale hanno anche *sequenze di ispezione* identiche.

Doppio hashing In questo caso H è:

$$H(k, i) = (H_1(k) + i \cdot H_2(k)) \bmod m$$

dove $H_1(k)$ fornisce la posizione iniziale e $H_2(k)$ l'offset per le successive *ispezioni*. Sono possibili m^2 sequenze, ma per garantire una permutazione completa $H_2(k)$ deve essere coprimo con $H_1(k)$. Cioè, se $m = 2^p$, $H_2(k)$ deve essere un numero dispari, mentre se m è primo, $H_2(k)$ deve essere un valore minore di m .

Frammento 63 - Implementazione tabella hash con hashing doppio.

```
ITEM[] K                                % Tabella delle chiavi
ITEM[] V                                % Tabella dei valori
int m                                    % Dimensione della tabella hash

HASH Hash(int dim)
    HASH t = new HASH
    t.m = dim
    t.K = new ITEM[dim]
    t.V = new ITEM[dim]
    for (i = 0 to dim - 1) do
        t.K[i] = nil
    return t

% Cerca la posizione associata ad una chiave
int scan(ITEM k, boolean insert)
    int delpos = m                        % Prima posizione deleted
    i = 0                                % Numero di ispezione
    j = H(k)                             % Posizione di partenza
    while (K[i] ≠ k and K[j] ≠ nil and i < m) do
        if (K[j] == deleted and delpos == m) then
            delpos = j
        j = (j + H'(k)) mod m
        i = i + 1
    if (insert and K[j] ≠ k and delpos < m) then
        return delpos
    return j

% Cerca e restituisce il valore associato a una chiave oppure nil
ITEM lookup(ITEM k)
    int i = scan(k, false)
    if (K[i] == k) then
        return V[i]
    else
        return nil
```

```

% Inserisce un'associazione chiave-valore o ne modifica il valore associato
insert(ITEM k, ITEM v)
    int i = scan(k, true)
    if (K[i] == nil or K[i] == deleted or K[i] == k) then
        k[i] = k                                % È inutile se K[i] == k
        V[i] = v
    else
        { Errore: la tabella è piena }
% Rimuove un'associazione chiave-valore
remove(ITEM k)
    int i = scan(k, false)
    if (K[i] == k) then
        K[i] = deleted
        V[i] = nil

```

Esaminiamo più nel dettaglio la funzione `scan()`. Può infatti risultare strano il fatto che abbiamo introdotto il valore `deleted` che usiamo al posto di `nil` per marcare una posizione come libera dopo una cancellazione. Questo serve a evitare situazioni come quella in figura, in cui la ricerca della chiave k_5 si interrompe all'ispezione della posizione `nil` facendo erroneamente credere che k_5 non sia presente nella *tabella hash*.

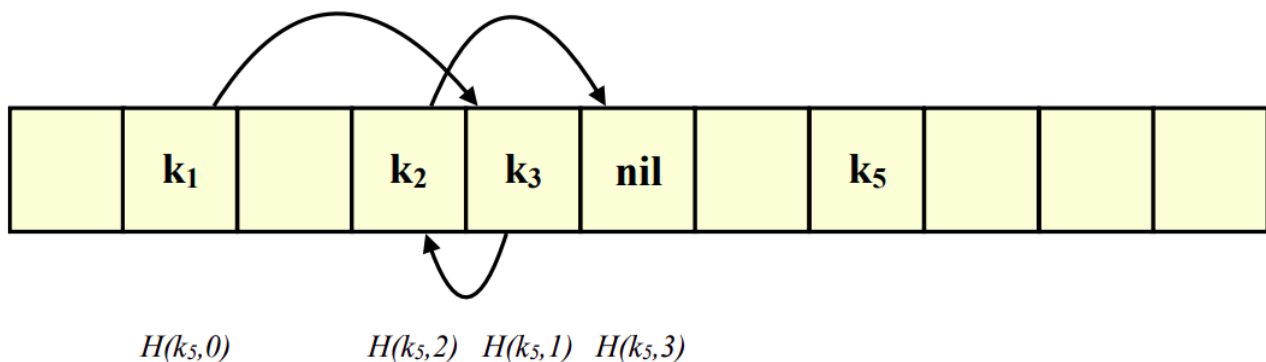


Fig. 8.4: *Sequenza di ispezione errata*

L'utilizzo del valore `deleted` al posto di `nil` ci consente di marcare le posizioni nelle quali in precedenza c'era un valore che poi è stato eliminato. In particolare, le posizioni `deleted` sono considerate come piene in fase di ricerca e vuote in fase di inserimento.

Questa soluzione risolve il problema della ricerca, ma rende il tempo di ricerca non più dipendente dal *fattore di carico* α e fa sì che il concatenamento sia più comune nelle implementazioni che ammettono la rimozione.

8.2.3 Complessità delle diverse implementazioni

Mettiamo a confronto le *complessità* di alcune delle implementazioni viste:

Tecnica	α	$I(\alpha)$	$S(\alpha)$
Lineare	$0 \leq \alpha < 1$	$\frac{(1-\alpha)^2+1}{2(1-\alpha)^2}$	$\frac{1-\frac{\alpha}{2}}{1-\alpha}$
Hashing doppio	$0 \leq \alpha < 1$	$\frac{1}{1-\alpha}$	$-\frac{1}{\alpha} \ln(1-\alpha)$
Liste di trabocco	$\alpha \geq 0$	$1 + \alpha$	$1 + \frac{\alpha}{2}$

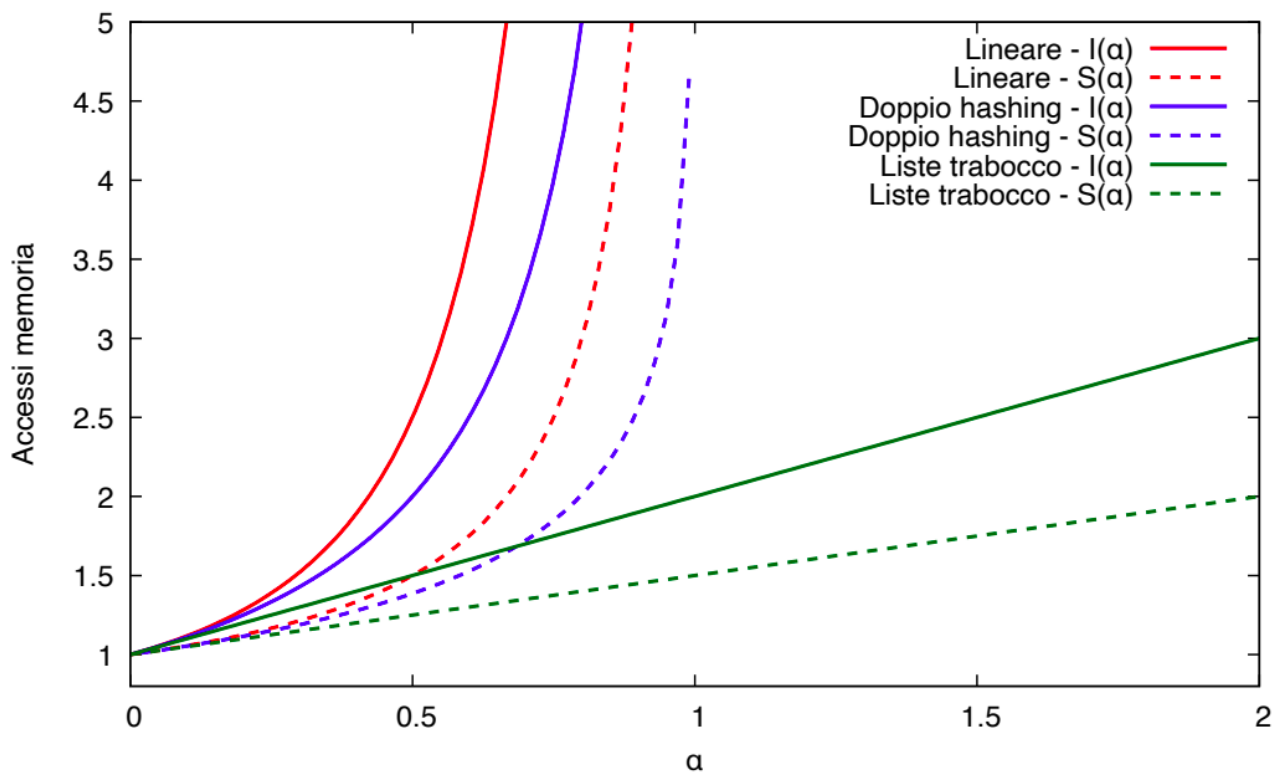


Fig. 8.5: *Complessità delle tabelle hash a confronto*

Conclusioni Per concludere possiamo dire che le *tabelle hash* permettono di implementare in modo molto efficiente dei *dizionari*, ma hanno una scarsa “locality of reference” che genera molte *cache miss* e inoltre, non permettono di ordinare le chiavi.

Capitolo Nr.9

Analisi ammortizzata

Dopo aver parlato a lungo di *strutture dati*, vediamo ora una tecnica che ci permette di effettuare una stima dall'alto del costo che è necessario pagare per eseguire delle operazioni su quelle strutture.

Definizione 99 - Analisi ammortizzata.

L'analisi ammortizzata è una tecnica per l'analisi di complessità che valuta il tempo richiesto per eseguire, nel caso pessimo, una sequenza di operazioni su una struttura dati.

In generale, esistono operazioni più o meno costose, ma se le operazioni più costose sono poco frequenti, il loro costo può essere ammortizzato da quelle meno costose.

È importante esplicitare la profonda differenza tra l'analisi del caso medio e l'*analisi ammortizzata*, effettuata sul caso pessimo. Nella prima, l'analisi è di tipo probabilistico ed è effettuata sulle singole operazioni, mentre, la seconda, è un'analisi deterministica effettuata su una sequenza di operazioni, e in particolare, sulla sequenza più costosa possibile.

9.1 Metodi per l'analisi ammortizzata

Esistono diverse tecniche per realizzare un'*analisi ammortizzata*.

Definizione 100 - Metodo dell'aggregazione.

Col metodo dell'aggregazione si calcola la complessità $T(n)$ necessaria per eseguire n operazioni in sequenza nel caso pessimo, quindi, si calcola il costo ammortizzato come $T(n)/n$ cioè come media su n operazioni.

Definiamo meglio i termini presenti nella definizione:

- *Sequenza*: è una sequenza di operazioni che permettono alla *struttura dati* di evolvere;
- *Caso pessimo*: è la *sequenza* con *complessità* più alta tra tutte quelle possibili;
- *Aggregazione*: la *complessità* $T(n)$ è ottenuta mediante una sommatoria dei costi delle singole operazioni;

Definizione 101 - Metodo degli accantonamenti.

Col metodo degli accantonamenti si assegna ad ogni operazione un costo ammortizzato che può anche essere diverso dal costo effettivo. In particolare, le operazioni meno costose vengono caricate di un costo aggiuntivo detto credito, che viene poi usato per pagare le operazioni più costose.

NB. Potenzialmente, ad ogni operazione potrebbe essere assegnato un costo diverso.

In generale, utilizzando il *metodo degli accantonamenti*, il *costo ammortizzato* assegnato alle operazioni meno costose è definito come:

$$\text{Costo ammortizzato} = \text{Costo effettivo} + \text{Credito prodotto}$$

Viceversa, il costo per le operazioni più costose è:

$$\text{Costo ammortizzato} = \text{Costo effettivo} - \text{Credito consumato}$$

Quindi, l'obiettivo, quando si utilizza questo metodo, è dimostrare che la somma dei *costi ammortizzati* a_i è un limite superiore ai costi effettivi c_i , ovvero che vale:

$$\sum_{i=1}^n c_i \leq \sum_{i=1}^n a_i$$

Ovviamente, la dimostrazione deve valere per tutte le sequenze, ma se vale per il caso pessimo vale anche per tutti gli altri. Il *credito* dopo la t -esima operazione è sempre positivo ed è espresso dalla seguente formula:

$$\sum_{i=1}^t a_i - \sum_{i=1}^t c_i \geq 0$$

Definizione 102 - Metodo del potenziale.

Col metodo del potenziale si associa una funzione di potenziale Φ ad uno stato S della struttura. La funzione Φ definisce la "quantità di lavoro" $\Phi(S)$ che è stato contabilizzata nell'analisi ammortizzata, ma non ancora spesa.

NB. $\Phi(S)$ rappresenta la quantità di energia potenziale che è stata "immagazzinata" in quello stato e che può essere spesa per eseguire le operazioni più costose.

In pratica, $\Phi(S)$ può essere vista come la cumulazione dei *crediti prodotti* dalle operazioni fino al raggiungimento dello stato S .

In generale, il *costo ammortizzato* di un'operazione è pari a:

$$\begin{aligned} \text{Costo ammortizzato} &= \text{Costo effettivo} + \text{Variazione di potenziale} \\ a_i &= c_i + (\Phi(S_i) - \Phi(S_{i-1})) \end{aligned}$$

dove S_i è lo stato associato all' i -esima operazione. Se proviamo a calcolare il *costo ammortizzato* di una sequenza di n operazioni, vale quanto segue:

$$\begin{aligned} A &= \sum_{i=1}^n a_i \\ &= \sum_{i=1}^n (c_i + \Phi(S_i) - \Phi(S_{i-1})) \\ &= \sum_{i=1}^n c_i + \sum_{i=1}^n (\Phi(S_i) - \Phi(S_{i-1})) \\ &= C + (\Phi(S_1) - \Phi(S_0)) + (\Phi(S_2) - \Phi(S_1)) + \cdots + (\Phi(S_n) - \Phi(S_{n-1})) \\ &= C + \Phi(S_n) - \Phi(S_0) \end{aligned}$$

Allora, se $\Phi(S_n) - \Phi(S_0) \geq 0$, A è un limite superiore al costo effettivo.

9.2 Esempio di analisi ammortizzata

Esempio 26 - Contatore binario.

Si consideri un contatore binario a k bit implementato come un vettore A di booleani, nel quale $A[0]$ è il bit meno significativo e $A[k-1]$ il più significativo. Il valore del contatore corrisponde al risultato della seguente sommatoria:

$$x = \sum_{i=0}^{k-1} A[i] \cdot 2^i$$

Alla struttura del contatore è associata soltanto la funzione *increment* per l'incremento del valore.

Frammento 64 - Implementazione funzione increment per contatori binari.

```
increment(int[] A, int k)
    int i = 0
    while (i < k and A[i] == 1) do    % Imposta a 0 la prima sequenza di bit a 1
        A[i] = 0
        i = i + 1
    if (i < k) then                    % Se i < k non c'è overflow
        A[i] = 1
```

Se $k = 8$, si ottiene un contatore binario a 8 bit e la seguente tabella ne mostra lo stato dopo il sedicesimo incremento.

x	$A[7]$	$A[6]$	$A[5]$	$A[4]$	$A[3]$	$A[2]$	$A[1]$	$A[0]$
0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	1
2	0	0	0	0	0	0	1	0
3	0	0	0	0	0	0	1	1
4	0	0	0	0	0	1	0	0
5	0	0	0	0	0	1	0	1
6	0	0	0	0	0	1	1	0
7	0	0	0	0	0	1	1	1
8	0	0	0	0	1	0	0	0
9	0	0	0	0	1	0	0	1
10	0	0	0	0	1	0	1	0
11	0	0	0	0	1	0	1	1
12	0	0	0	0	1	1	0	0
13	0	0	0	0	1	1	0	1
14	0	0	0	0	1	1	1	0
15	0	0	0	0	1	1	1	1
16	0	0	0	1	0	0	0	0

Proviamo ora a realizzare una stima dall'alto del costo della funzione *increment* usando i metodi per l'analisi ammortizzata.

Metodo dell'aggregazione Utilizzando questo metodo ci chiediamo quale sia il costo $T(n)$ da pagare per eseguire n operazioni in sequenza.

Inizialmente potremmo notare che per rappresentare n in binario servono $k = \lceil \log(n+1) \rceil$ bit. Un'invocazione di **increment** nel caso pessimo richiede un tempo $O(k)$, quindi n invocazioni costano $T(n) = O(nk)$. A questo punto, il costo ammortizzato di ogni operazione è $T(n)/n = O(nk)/n = O(k) = O(\log n)$.

Se però notiamo che il costo per eseguire l'intera sequenza è proporzionale al numero di bit che vengono modificati, possiamo provare a realizzare una stima più ristretta. Proviamo quindi a contare per ogni riga e per ogni colonna il numero di bit che sono stati modificati.

x	A[7]	A[6]	A[5]	A[4]	A[3]	A[2]	A[1]	A[0]	#bit
0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	1	1
2	0	0	0	0	0	0	1	0	2
3	0	0	0	0	0	0	1	1	1
4	0	0	0	0	0	1	0	0	3
5	0	0	0	0	0	1	0	1	1
6	0	0	0	0	0	1	1	0	2
7	0	0	0	0	0	1	1	1	1
8	0	0	0	0	1	0	0	0	4
9	0	0	0	0	1	0	0	1	1
10	0	0	0	0	1	0	1	0	2
11	0	0	0	0	1	0	1	1	1
12	0	0	0	0	1	1	0	0	3
13	0	0	0	0	1	1	0	1	1
14	0	0	0	0	1	1	1	0	2
15	0	0	0	0	1	1	1	1	1
16	0	0	0	1	0	0	0	0	5
#bit	0	0	0	1	2	4	8	16	

Guardando l'ultima riga possiamo notare che il bit in posizione A[0] viene modificato ad ogni incremento, quello in A[1] ogni due, quello in A[2] ogni 4 e così via. Di conseguenza, il costo $T(n)$ è una funzione di questo tipo:

$$T(n) = \sum_{i=0}^{k-1} \left\lfloor \frac{n}{2^i} \right\rfloor \leq n \sum_{i=0}^{k-1} \frac{1}{2^i} \leq n \sum_{i=0}^{+\infty} \left(\frac{1}{2}\right)^i = 2n$$

e quindi il costo ammortizzato vale:

$$\frac{T(n)}{n} \leq \frac{2n}{n} = 2 = O(1)$$

Metodo degli accantonamenti Supponiamo che il costo effettivo della **increment** sia d , dove d è il numero di bit che vengono modificati ad ogni incremento. Proviamo però ad assegnare alla **increment** un costo di 2 che include 1 per il costo effettivo che si paga per cambiare un bit da 0 a 1 e 1 per il futuro cambio dello stesso bit da 1 a 0.

Di conseguenza, in ogni istante, il credito è pari al numero di bit a 1 presenti e quindi il costo totale vale $O(n)$ e il costo ammortizzato $O(1)$.

Metodo del potenziale Scegliamo come funzione di potenziale Φ il numero di bit a 1 presenti nel contatore. Ad ogni utilizzo della funzione **increment**, se t è il numero di bit a 1 incontrati prima del primo 0, il costo effettivo è $1 + t$ perché cambiamo lo stato di tutti i primi t bit a 1 e del primo bit a 0.

La variazione di potenziale, invece, vale $1 - t$ in quanto tutti i bit a 1 incontrati diventano 0 e il primo bit a 0 diventa 1. Quindi, se allo stato S_i $\Phi(S_i) = t$, allo stato S_{i+1} $\Phi(S_{i+1}) = 1$ e, conseguentemente, la variazione di potenziale tra lo stato S_{i+1} e lo stato S_i vale:

$$\Phi(S_{i+1}) - \Phi(S_i) = 1 - t$$

Il costo ammortizzato di un'invocazione della **increment**, allora, vale:

$$\begin{aligned} \text{Costo ammortizzato} &= \text{Costo effettivo} + \text{Variazione di potenziale} \\ &= (1 + t) + (1 - t) \\ &= 2 = O(1) \end{aligned}$$

Inoltre, poiché allo stato S_0 , $\Phi(S_0) = 0$ perché non ci sono bit impostati a 1, e allo stato S_n , $\Phi(S_n) \geq 0$, risulta verificata anche la condizione $\Phi(S_n) - \Phi(S_0) \geq 0$.

9.3 Vettori dinamici

Possiamo utilizzare le tecniche di *analisi ammortizzata* per stimare la *complessità* delle operazioni di inserimento e cancellazione nei vettori con ridimensionamento dinamico della dimensione.

Prima però vediamo brevemente cosa significa “ridimensionamento dinamico della dimensione”. Quando si implementa una *sequenza* utilizzando un vettore, se ne specifica una dimensione iniziale detta *capacità*. Ovviamente, non è sempre nota a priori la quantità di elementi che dovranno essere inseriti nella struttura, e quindi la *capacità* iniziale può rivelarsi insufficiente. I *vettori dinamici* risolvono questo problema allocando un nuovo vettore con una capacità maggiore e ricopiando in esso tutti i valori presenti nel vettore originale, che poi può essere eliminato.

9.3.1 Inserimento

Quando in fase di inserimento si rende necessario aumentare la dimensione del vettore, esistono due principali metodologie di approccio: *incremento fisso* e *incremento variabile*.

Il primo approccio prevede che la *capacità* venga incrementata di un fattore fisso (e.g. +1, +2, +10, ...), mentre il secondo è dipendente dalla dimensione attuale della struttura (e.g. ·2, ·1.5, ...).

Ma qual è il metodo migliore? Proviamo a utilizzare l'*analisi ammortizzata* per realizzare una stima.

Incremento fisso Il costo effettivo di un inserimento con ridimensionamento fisso è descritto dalla seguente espressione:

$$c_i = \begin{cases} i & \text{se } (i \bmod d) = 1^1 \\ 1 & \text{altrimenti} \end{cases}$$

dove d è sia la dimensione iniziale che il valore di cui viene incrementata.

¹L'inserimento ha un costo i quando $(i \bmod d) = 1$ perché se è stata raggiunta la dimensione massima quell'operazione di modulo vale 1

Supponiamo $d = 4$ e vediamo il costo delle prime 17 operazioni di inserimento:

n	Costo
1	1
2	1
3	1
4	1
5	$1 + d = 5$
6	1
7	1
8	1
9	$1 + 2d = 9$
10	1
11	1
12	1
13	$1 + 3d = 13$
14	1
15	1
16	1
17	$1 + 4d = 17$

A questo punto, calcoliamo il costo effettivo di n operazioni:

$$\begin{aligned}
T(n) &= \sum_{i=1}^n c_i \\
&= n + \sum_{j=1}^{\lfloor n/d \rfloor} d \cdot j \\
&= n + d \sum_{j=1}^{\lfloor n/d \rfloor} j \\
&= n + d \frac{(\lfloor n/d \rfloor + 1) \lfloor n/d \rfloor}{2} \\
&\leq n + \frac{(n/d + 1)n}{2} && \text{Rimozione dell'intero inferiore} \\
&= n + \frac{n^2/d + n}{2} = O(n^2)
\end{aligned}$$

Il *costo ammortizzato* vale:

$$\frac{T(n)}{n} = \frac{O(n^2)}{n} = O(n)$$

Incremento dinamico Supponendo che ad ogni incremento la dimensione del vettore venga raddoppiata, il costo effettivo vale:

$$c_i = \begin{cases} i & \text{se } \exists k \in \mathbb{Z}_0^+ : i = 2^k + 1 \\ 1 & \text{altrimenti} \end{cases}$$

Se la dimensione iniziale è 1, il costo delle prime 17 operazioni è il seguente:

n	Costo
1	1
2	$1 + 2^0 = 2$
3	$1 + 2^1 = 3$
4	1
5	$1 + 2^2 = 5$
6	1
7	1
8	1
9	$1 + 2^3 = 9$
10	1
11	1
12	1
13	1
14	1
15	1
16	1
17	$1 + 2^4 = 17$

Il costo effettivo di n operazioni vale:

$$\begin{aligned}
T(n) &= \sum_{i=1}^n c_i \\
&= n + \sum_{j=0}^{\lfloor \log n \rfloor} 2^j \\
&= n + 2^{\lfloor \log n \rfloor + 1} - 1 \\
&\leq 2^{\log(n)+1} - 1 && \text{Rimozione dell'intero inferiore} \\
&= n + 2n - 1 = O(n)
\end{aligned}$$

e di conseguenza quello *ammortizzato* è:

$$\frac{T(n)}{n} = \frac{O(n)}{n} = O(1)$$

Giunti a questo punto, siamo riusciti a dimostrare che l'incremento dinamico ha un costo inferiore all'incremento fisso.

9.3.2 Cancellazione

Se il vettore non è ordinato, rimuovere un elemento significa spostare l'ultimo elemento della *sequenza* nella posizione dell'elemento da rimuovere.

Per ridurre lo spreco di memoria, è opportuno ridurre la dimensione del vettore quando il *fattore di carico* $\alpha = \frac{Dim}{Capacità}$ ² scende sotto una certa soglia.

L'operazione di riduzione della *capacità* è detta *contrazione* e, similmente a quanto avviene con l'*espansione*, prevede che venga allocato un nuovo vettore di dimensione inferiore a quella

²Il valore *Dim* rappresenta il numero di elementi che in dato momento sono presenti nella struttura

attuale, che vi vengano copiati i valori del vettore originale e che quindi quest'ultimo venga eliminato.

Ma qual è la soglia di *contrazione* ottimale?

Una prima strategia potrebbe essere quella di dimezzare la *capacità* quando *Dim* raggiunge la metà della *capacità*, cioè quando $\alpha = \frac{1}{2}$. Il problema di questa scelta è che dopo la *contrazione* nella struttura non rimangono posizioni libere e quindi un successivo inserimento richiederebbe di fare subito un'*espansione*.

Una strategia migliore è quella invece di scegliere $\alpha = \frac{1}{4}$. In questo modo, se quando viene raggiunta la *soglia di contrazione*, si va a dimezzare la *capacità*, il *fattore di carico* α , invece di aumentare fino a 1 si fermerà a $\frac{1}{2}$ concedendoci di aggiungere al vettore tanti elementi quanti sono quelli già presenti prima di dover effettuare un'*espansione*.

Analisi del costo Proviamo a stimare il costo di questa seconda soluzione utilizzando il *metodo del potenziale*. Scegliamo una funzione di potenziale Φ che vale 0 all'inizio e immediatamente dopo un'*espansione* o una *contrazione* e il cui valore cresca fino a raggiungere il numero di elementi presenti nella struttura quando $\alpha = 1$ e diminuisca fino a un quarto quando α si riduce alla *soglia di contrazione*.

Quindi, la definizione di Φ è la seguente:

$$\Phi = \begin{cases} 2 \cdot Dim - Capacità & \text{se } \alpha \geq \frac{1}{2} \\ \frac{Capacità}{2} - Dim & \text{se } \alpha < \frac{1}{2} \end{cases}$$

Vediamone il valore in alcuni casi particolari:

- $\alpha = \frac{1}{2}$: è stata appena effettuata un'*espansione* o una *contrazione* e quindi $\Phi = 0$;
- $\alpha = 1$ la struttura è piena, ovvero $Dim = Capacità$ e quindi $\Phi = Dim = Capacità$;
- $\alpha = \frac{1}{4}$: il *fattore di carico* ha raggiunto la *soglia di contrazione*, ovvero $Dim = \frac{Capacità}{4}$ e quindi $\Phi = Dim$;

Una funzione di potenziale di questo tipo ci garantisce che il potenziale presente nel momento in cui si effettua un'*espansione* o una *contrazione* sia sufficiente per “pagare” quelle stesse operazioni.

Proviamo quindi a calcolare il *costo ammortizzato* in base ai diversi valori di α :

- $\alpha \geq \frac{1}{2}$:

$$\begin{aligned} a_i &= c_i + \Phi_i - \Phi_{i-1} \\ &= 1 + (2Dim_i - Capacità_i) - (2Dim_{i-1} - Capacità_{i-1}) \\ &= 1 + 2(Dim_{i-1} + 1) - Capacità_i - 2Dim_{i-1} + Capacità_{i-1} \\ &= 1 + 2(Dim_{i-1} + 1) - Capacità_{i-1} - 2Dim_{i-1} + Capacità_{i-1} \\ &= 1 + 2Dim_{i-1} + 2 - Capacità_{i-1} - 2Dim_{i-1} + Capacità_{i-1} \\ &= 3 \end{aligned}$$

- $\alpha = 1$:

$$\begin{aligned} a_i &= c_i + \Phi_i - \Phi_{i-1} \\ &= 1 + Dim_{i-1} + (2Dim_i + Capacità_i) - (2Dim_{i-1} - Capacità_{i-1}) \\ &= 1 + Dim_{i-1} + 2(Dim_{i-1} + 1) - 2Dim_{i-1} - 2Dim_{i-1} + Dim_{i-1} \\ &= 3 \end{aligned}$$

In entrambi i casi il costo è $O(1)$.

NB. Poiché non è conveniente che il *fattore di carico* α cresca troppo, regole di ridimensionamento simili vengono usate anche nelle *tabelle hash*. Solitamente, quando α raggiunge una soglia di 0.5 o 0.75 la *tabella hash* viene estesa raddoppiandone la *capacità*. Quest'operazione, oltre che dimezzare il *fattore di carico*, rimuove anche tutti gli elementi *deleted*. Il costo nel caso peggiore è $O(m)$, ma quello *ammortizzato* è $O(1)$.

9.4 Discussione sugli insiemi

Giunti a questo punto della trattazione, abbiamo introdotto una vasta gamma di *strutture dati* più o meno complesse e caratterizzata da costi più o meno convenienti. Abbiamo anche visto che *strutture dati astratte* possono essere implementate utilizzando diverse *strutture dati concrete* e che questo influenza il costo delle operazioni. Per chiudere il cerchio, mettiamo ora a confronto diverse implementazioni di un *insieme*.

9.4.1 Insieme come vettore di booleani

Quando si sa che fare con m valori ordinabili e consecutivi è possibile implementare un *insieme* per quei valori come un vettore di booleani di dimensione m . Ogni posizione del vettore vale **true** se l'elemento associato appartiene all'insieme, altrimenti vale **false**.

Vediamone l'implementazione:

Frammento 65 - Insieme come vettore di booleani.

```
boolean[] V      % Vettore di booleani    int size()
int size % Numero di elementi presenti    return size
int capacity     % Dimensione massima

SET Set(int m)
    SET t = new SET
    t.size = 0
    t.capacity = m
    t.V = new int[1...m] = false
    return t

boolean contains(int x)
    if (1 ≤ x ≤ capacity) then
        return V[x]
    else
        return false

SET union(SET A, SET B)
    int newSize = max(A.capacity, B.capacity)
    SET C = Set(newSize)
    for (i = 1 to A.capacity) do
        if (A.contains(i)) then
            C.insert(i)
    for (i = 1 to B.capacity) do
        if (B.contains(i)) then
            C.insert(i)
    return C

insert(int x)
    if (1 ≤ x ≤ capacity) then
        if (not V[x]) then
            size = size + 1
            V[x] = true

remove(int x)
    if (1 ≤ x ≤ capacity) then
        if (V[x]) then
            size = size - 1
            V[x] = false
```

```

SET difference(SET A, SET B)
    SET C = Set(A.capacity)
    for (i = 1 to A.capacity) do
        if (A.contains(i) and not B.contains(i)) then
            C.insert(i)
    return C

```

```

SET intersection(SET A, SET B)
    int newSize = min(A.capacity, B.capacity)
    SET C = Set(newSize)
    for (i = 1 to newSize) do
        if (A.contains(i) and B.contains(i)) do
            C.insert(i)
    return C

```

I vantaggi di questa implementazione sono certamente la semplicità e l'efficienza delle operazioni di inserimento, rimozione e verifica di appartenenza. Tuttavia, la memoria occupata è indipendente dal numero di elementi effettivamente contenuti, così come lo sono le operazioni che visitano tutti gli elementi. Ad esempio, le operazioni di **union**, **difference** e **intersection** hanno costo $O(m)$.

9.4.2 Insieme come lista non ordinata

In un *insieme* implementato come *lista* non ordinata le operazioni di inserimento, ricerca e rimozione hanno un costo $O(n)$ dove n è il numero di elementi presenti. Per le operazioni di unione, intersezione e differenza non va meglio, infatti dati due *insiemi* A e B di dimensione n_A e n_B , la *complessità* è $O(n_A n_B)$.

NB. Se supponiamo di sapere che un elemento non appartiene all'*insieme*, l'inserimento può avere costo $O(1)$.

Frammento 66 - Insieme come lista non ordinata.

```

SET union(SET A, SET B)
    SET C = Set()
    foreach (s ∈ A) do                                     % Costa  $O(n_A)$ 
        C.insert(s)                                       % Costa  $O(n_C)$ 3
    foreach (s ∈ B) do                                     % Costa  $O(n_B)$ 
        C.insert(s)                                       % Costa  $O(n_C)$ 
    return C                                              % Abbiamo pagato  $O(n_A n_C + n_B n_C) = O(n_C m)$ 

```

```

SET difference(SET A, SET B)
    SET C = Set()
    foreach (s ∈ A) do                                     % Costa  $O(n_A)$ 
        if (not B.contains(s))                             % Costa  $O(n_B)$ 
            C.insert(s)                                     % Possiamo supporre ci costi  $O(1)$ 
    return C                                              % Abbiamo pagato  $O(n_A n_B 1) = O(n_A n_B)$ 

```

³Avremmo anche potuto supporre di pagare $O(1)$ in quanto sicuramente nessuno degli elementi di A apparteneva a C

```

SET intersection(SET A, SET B)
  SET C = Set()
  foreach (s ∈ A) do                                     % Costa  $O(n_A)$ 
    if (B.contains(s))                                   % Costa  $O(n_B)$ 
      C.insert(s)                                         % Possiamo supporre ci costi  $O(1)$ 
  return C                                                % Abbiamo pagato  $O(n_A n_B) = O(n_A n_B)$ 

```

NB. Il costo dell'implementazione come vettore non ordinato è in tutto e per tutto equivalente all'implementazione come *lista* non ordinata.

9.4.3 Insieme come lista ordinata

Con le *liste* ordinate i costi rimangono identici ad eccezione di unione, intersezione e differenza il cui costo si riduce a $O(n)$.

Frammento 67 - Insieme come lista ordinata.

```

SET union(SET A, SET B)
  SET C = Set()
  POS posA = A.head()
  POS posB = B.head()
  while (not A.finished(posA) and not B.finished(posB)) do %  $O(\max(n_A, n_B))$ 
    C.insert(C.tail(), A.read(posA)) % L'inserimento in coda costa  $O(1)$ 
    if (A.read(posA) == B.read(posB)) then
      posA = A.next(posA)
      posB = B.next(posB)
    else if (A.read(posA) < B.read(posB)) then
      posA = A.next(posA)
    else
      posB = B.next(posB)
  return C

```

```

SET difference(SET A, SET B)
  SET C = Set()
  POS posA = A.head();
  POS posB = B.head();
  while (not A.finished(posA) and not B.finished(posB)) do %  $O(\max(n_A, n_B))$ 
    if (A.read(posA) ≠ B.read(posB)) then
      C.insert(C.tail(), A.read(posA)) % L'inserimento in coda costa  $O(1)$ 
      if (A.read(posA) > B.read(posB)) then
        posB = B.next(posB)
      posA = A.next(posA)
  return C

```

```

SET intersection(SET A, SET B)
  SET C = Set()
  POS posA = A.head();
  POS posB = B.head();
  while (not A.finished(posA) and not B.finished(posB)) do    %  $O(\max(n_A, n_B))$ 
    if (A.read(posA) == B.read(posB)) then
      C.insert(C.tail(), A.read(posA))    % L'inserimento in coda costa  $O(1)$ 
      posA = A.next(posA)
      posB = B.next(posB)
    else if (A.read(posA) < B.read(posB)) then
      posA = A.next(posA)
    else
      posB = B.next(posB)
  return C

```

NB. Il costo dell'implementazione come vettore ordinato è equivalente all'implementazione come *lista* ordinata, tranne per la funzione **contains** che nel vettore ha costa $O(\log n)$ in quanto è possibile usare un algoritmo di ricerca dicotomica.

9.4.4 Confronto generale

In generale, se n è il numero di elementi presenti in un *insieme* ed m è la capacità di quell'*insieme*, la *complessità* delle operazioni a seconda dell'implementazione usata è riassunta dalla tabella sottostante.

Operazione Struttura	contains lookup	insert	remove	min	foreach	Ordinabile
Vettore booleano	$O(1)$	$O(1)$	$O(1)$	$O(m)$	$O(m)$	Sì
Lista non ordinata	$O(n)$	$O(n)$	$O(n)$	$O(n)$	$O(n)$	No
Lista ordinata	$O(n)$	$O(n)$	$O(n)$	$O(1)$	$O(n)$	Sì
Vettore ordinato	$O(\log n)$	$O(n)$	$O(n)$	$O(1)$	$O(n)$	Sì
Albero bilanciato	$O(\log n)$	$O(\log n)$	$O(\log n)$	$O(\log n)$	$O(n)$	Sì
Tabella Hash Mem. interna	$O(1)$	$O(1)$	$O(1)$	$O(m)$	$O(m)$	No
Tabella Hash Mem. esterna	$O(1)$	$O(1)$	$O(1)$	$O(m + n)$	$O(m + n)$	No

Capitolo Nr.10

Divide-et-impera

10.1 Introduzione

Giunti a questo punto della trattazione, è arrivato il momento di parlare delle tecniche per la risoluzione di problemi, ovvero di quelle tecniche che permettono di arrivare alla definizione di un algoritmo per la risoluzione di un particolare problema.

Esiste un'ampia gamma di categorie di problemi, tra le quali troviamo:

- *Problemi decisionali*: l'obiettivo è riuscire a stabilire se il dato in ingresso soddisfa o meno una proprietà;
- *Problemi di ricerca*: l'obiettivo è ricercare all'interno dell'insieme dei dati di input un sottoinsieme di dati che soddisfano una certa proprietà;
- *Problemi di ottimizzazione*: in un insieme di soluzioni alle quali è associata una funzione di costo, l'obiettivo è trovare la soluzione di costo minimo;

Le tecniche per la ricerca di soluzioni sono molteplici e nel corso della trattazione ne approfondiremo diverse. In questo capitolo, iniziamo a vedere la tecnica del *Divide-et-impera*.

Definizione 103 - Tecnica del Divide-et-impera.

La tecnica del Divide-et-impera prevede che il problema da risolvere venga suddiviso in sotto-problemi indipendenti e che le soluzioni ai sotto-problemi vengano combinate per ottenere la soluzione al problema di partenza.

La definizione lascia intendere molto chiaramente la simbiosi che esiste tra algoritmi *Divide-et-impera* e ricorsione. Infatti, la suddivisione in sotto-problemi viene realizzata applicando ricorsivamente l'algoritmo ad un sottoinsieme dei dati di input.

L'approccio *Divide-et-impera* si compone di tre fasi:

- *Divide*: il problema viene suddiviso in sotto-problemi indipendenti;
- *Impera*: vengono risolti i sotto-problemi;
- *Combina*: le soluzioni dei sotto-problemi vengono combinate per ottenere la soluzione al problema di partenza;

NB. La tecnica del *Divide-et-impera* trova applicazione soprattutto negli ambiti dei *problemi decisionali* e di *ricerca*.

Vediamo un esempio tipico di applicazione della tecnica *Divide-et-impera*.

Problema 3 - Problema della Torre di Hanoi.

Quello della Torre di Hanoi è un problema matematico molto famoso. Esistono tre pioli e n dischi di dimensione diversa. Inizialmente i dischi sono impilati in ordine decrescente sul piolo di sinistra. Lo scopo del gioco è riuscire a impilare quegli stessi dischi in ordine decrescente sul piolo di destra.

Ad ogni mossa è possibile spostare un disco ed è possibile usare il piolo centrale come appoggio. L'importante è non posizionare mai un disco sopra uno più piccolo.

Frammento 68 - Implementazione algoritmo per risoluzione della Torre di Hanoi.

```
hanoi(int n, int src, int dest, int aux)
    if (n == 1) then
        print src → dest
    else
        hanoi(n-1, src, aux, dest)      % Sposta n-1 dischi sul piolo ausiliario
        print src → dest                 % Sposta il disco più grande
        hanoi(n-1, aux, src, dest)       % Sposta n-1 dischi sul piolo finale
```

In questo codice è evidente la suddivisione in sotto-problemi in quanto l'invocazione `hanoi(n-1, src, aux, dest)` sposta tutti i dischi tranne l'ultimo sul piolo ausiliario usando il piolo che sarebbe di destinazione come piolo d'appoggio. Quindi, sposta il disco rimasto, il più grande, sul piolo di destinazione. L'invocazione `hanoi(n-1, aux, dest, src)` sposta di nuovo gli $n-1$ dischi di prima sul piolo di destinazione.

È interessante provare a studiare la complessità di questa soluzione. La funzione di ricorrenza è la seguente:

$$T(n) = 2T(n-1) + 1$$

*e per il **Teorema delle ricorrenze lineari di ordine costante** il costo di questo algoritmo è $\Theta(2^n)$, che nonostante sia un costo esponenziale è comunque il migliore possibile in quanto è dimostrabile l'ottimalità della soluzione proposta.*

Convenienza degli algoritmi Divide-et-impera L'utilizzo di una qualsiasi tecnica per la risoluzione di problemi non sempre permette di arrivare ad una soluzione ottima o anche solo conveniente. Ad esempio, la seguente implementazione ricorsiva della funzione di ricerca del minimo non è migliore della sua controparte iterativa:

Frammento 69 - Implementazione minrec per la ricerca ricorsiva del minimo.

```
int minrec(int[] A, int i, int j)
    if (i == j) then
        return A[i]
    else
        m = [(i + j) / 2]
        return min(minrec(A, i, m), minrec(A, m + 1, j))
```

La *funzione di ricorrenza* di questa soluzione è:

$$T(n) = \begin{cases} 2T(n/2) + 1 & n > 1 \\ 1 & n \leq 1 \end{cases}$$

e per il *Teorema delle ricorrenze lineari con partizione bilanciata - Rid* vale $T(n) = \Theta(n)$ che è la stessa *complessità* della versione iterativa dell'algoritmo. Di conseguenza, in questo caso non conviene usare la tecnica del *Divide-et-impera* perché l'algoritmo ottenuto è più complicato di quello che già conoscevamo.

10.2 Algoritmo Quicksort

All'inizio della trattazione abbiamo già esaminato un algoritmo di ordinamento basato sul *Divide-et-impera*: l'algoritmo *merge sort*.

10.2.1 Principi di funzionamento

Il *Quicksort* è un altro algoritmo di ordinamento, basato anch'esso su *Divide-et-impera*, che riceve in input un vettore $A[1 \dots n]$ e due indici lo e hi tali che $1 \leq lo \leq hi \leq n$.

Nella fase di *divide* viene scelto un valore $p \in A[lo \dots hi]$ detto *perno* o *pivot*. Quindi, tutti gli elementi del vettore vengono permutati in modo da portare i valori più piccoli del *pivot* alla sua sinistra e gli altri alla sua destra.

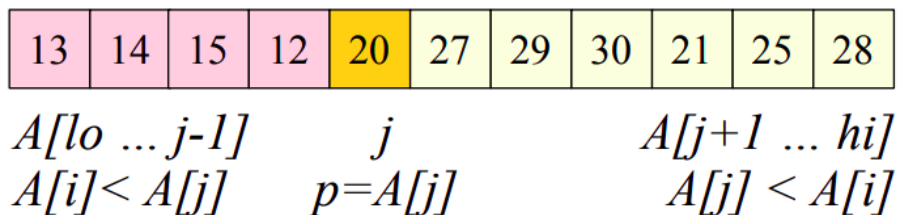


Fig. 10.1: Posizionamento dei valori dopo la scelta del *pivot*

Nella fase di *Impera* vengono ordinati i due sottoarray $A[lo \dots j-1]$ e $A[j+1 \dots hi]$ e infine nella fase di *Combina* non serve fare nulla in quanto il vettore risulta già ordinato.

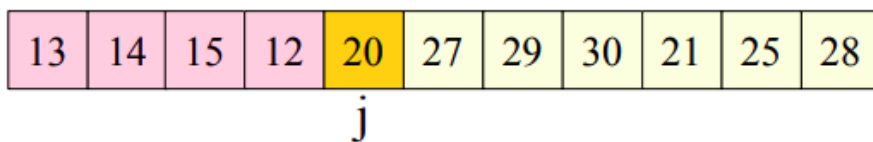


Fig. 10.2: Vettore ordinato al termine della fase di *Impera*

10.2.2 Implementazione

Frammento 70 - Implementazione dell'algoritmo Quicksort.

```
quicksort(ITEM[] A, int lo, int hi)
    if (lo < hi) then
        int j = pivot(A, lo, hi)                % Calcola il pivot
        quicksort(A, lo, j - 1)                 % Ordina il sottoarray sinistro
        quicksort(A, j + 1, hi)                 % Ordina il sottoarray destro

% Funzione per la ricerca del pivot
int pivot(ITEM[] A, int lo, int hi)
    ITEM pivot = A[lo]                          % Scelta del valore del pivot
    int j = lo                                  % Indice del pivot
    for (i = lo + 1 to hi) do
        if (A[i] < pivot) then
            j = j + 1                            % Aggiorna l'indice del pivot
            swap(A, i, j)                        % Scambia di posizione gli elementi agli indici i e j
    A[lo] = A[j]
    A[j] = pivot                                % Mette in posizione j il pivot
    return j

% Funzione ausiliaria per lo scambio di posizione di due elementi
swap(ITEM[] A, int i, int j)
    int tmp = A[i]
    A[i] = A[j]
    A[j] = tmp
```

10.2.3 Costo computazionale

Per studiare il costo della funzione principale `quicksort` analizziamo le funzioni ausiliarie. La `swap` è banale e ha un costo $\Theta(1)$, mentre la `pivot` costa $\Theta(n)$ perché va a confrontare il *pivot* con ogni elemento dell'array. Per il costo della `quicksort` consideriamo separatamente il caso pessimo e il caso migliore.

Il caso pessimo è quello in cui la scelta del *pivot* porti sempre ad avere due sottoarray di dimensione 0 e $n - 1$ e ciò induce una *funzione di ricorrenza* di questo tipo:

$$T(n) = T(n - 1) + T(0) + \Theta(n) = T(n - 1) + \Theta(n) = \Theta(n^2)$$

Nel caso migliore invece, il vettore viene sempre diviso in due sottoarray di dimensione $n/2$ e quindi il costo è descritto dalla seguente funzione:

$$T(n) = 2T(n/2) + \Theta(n) = \Theta(n \log n)$$

E il caso medio? Fortunatamente, i partizionamenti nel caso medio sono molto più simili al caso migliore che al peggiore. Ad esempio, con un partizionamento $9 - a - 1$ vale:

$$T(n) = T(n/10) + T(9n/10) + cn = \Theta(n \log n)$$

Lo stesso accade anche un partizionamento $99 - a - 1$:

$$T(n) = T(n/100) + T(99n/100) + cn = \Theta(n \log n)$$

NB. In questi esempi non stiamo considerando i fattori moltiplicativi, ma in certi contesti potrebbero essere importanti.

Riassumendo, il costo dell'algoritmo *Quicksort* è $\Theta(n \log n)$ nei casi ottimo e medio, mentre è $\Theta(n^2)$ nel caso pessimo. Il *Merge sort* aveva un costo $\Theta(n \log n)$ in tutti i casi, quindi a prima vista sembrerebbe essere più conveniente. In verità, poiché il *Quicksort* non usa memoria addizionale, gode di fattori moltiplicativi inferiori rispetto al *Merge sort*, è parallelizzabile ed esistono tecniche euristiche che consentono di evitare il caso pessimo, è spesso preferito al *Merge sort*.

Funzione euristica per la selezione del pivot Vediamo di seguito un esempio di una funzione per la selezione del *pivot* in maniera euristica.

Frammento 71 - Implementazione euristica della funzione pivot.

```
int pivot(ITEM[] A, int lo, int hi)
    int m = [(lo + hi) / 2]
    if (A[lo] > A[hi]) then
        swap(A, lo, hi)                % Sposta il massimo in ultima posizione
    if (A[m] > A[hi]) then
        swap(A, m, hi)                % Sposta il massimo in ultima posizione
    if (A[m] > A[lo]) then
        swap(A, m, lo)                % Sposta il mediano in prima posizione
    ITEM pivot = A[lo]
    int j = lo                         % Indice del pivot
    for (i = lo + 1 to hi) do
        if (A[i] < pivot) then
            j = j + 1                  % Aggiorna l'indice del pivot
            swap(A, i, j)              % Scambia di posizione gli elementi agli indici i e j
    A[lo] = A[j]
    A[j] = pivot                      % Mette in posizione j il pivot
    return j
```

10.3 Esercizio di applicazione del Divide-et-impera

Problema 4 - Ricerca di un gap in un vettore.

Dato un vettore V contenente $n \geq 2$ interi, un gap è un indice i con $1 < i \leq n$ tale che $V[i-1] < V[i]$.

- Dimostrare che se $n \geq 2$ e $V[1] < V[n]$, allora V contiene sicuramente almeno un gap;
- Progettare un algoritmo che, dato un vettore V contenente $n \geq 2$ valori e tale che $V[1] < V[n]$, restituisce la posizione di un gap nel vettore;

Iniziamo dimostrando il primo punto.

Dimostrazione. Supponiamo per assurdo che non esista alcun gap all'interno di V . Di conseguenza, deve per forza valere la seguente catena di disuguaglianze:

$$V[1] \geq V[2] \geq \dots \geq V[n]$$

Questo però è impossibile in quanto per ipotesi $V[1] < V[n]$ e quindi deve per forza esistere almeno un gap all'interno di V . \square

Per il secondo punto proviamo a fare un ragionamento per induzione. Siano i e j due indici tali che $1 \leq i < j \leq n$ e supponiamo che $V[i] < V[j]$. In base a questa definizione di i e j , nel sottoarray $V[i \dots j]$ ci sono almeno due elementi e il primo, $V[i]$, è minore dell'ultimo, $V[j]$.

Proviamo a dimostrare per induzione sulla dimensione del sottoarray $V[i \dots j]$ che esiste almeno un gap.

Caso base $n = 2$ e quindi $j - i + 1 = 2$, ovvero $i = j - 1$. Da questo segue che $V[i] < V[j] \Leftrightarrow V[j - 1] < V[j]$ e quindi alla posizione j esiste un gap.

Passo induttivo *Ipotizziamo che dato un qualunque sottoarray $V[h \dots k]$, di dimensione $n' < n$ e tale che $V[h] < V[k]$, contenga un gap. A questo punto, se consideriamo un qualunque indice $m \in]i, j[$, si verificherà sicuramente almeno uno dei seguenti casi:*

- $V[m] < V[j]$: per ipotesi induttiva, esiste sicuramente un gap all'interno di $V[m \dots j]$;
- $V[i] < V[m]$: per ipotesi induttiva, esiste sicuramente un gap all'interno di $V[i \dots m]$;

Fatte queste considerazioni, possiamo usare la tecnica del Divide-et-impera per definire un algoritmo per la ricerca di gap:

Frammento 72 - Implementazione algoritmo Divide-et-impera per la ricerca di gap.

```
int gap(int[] V, int n)
    return gaprec(V, 1, n)

int gaprec(int[] V, int i, int j)
    if (j == i + 1) then                                     % Caso base
        return j
    int m = [(i + j) / 2]                                     % Scelta dell'indice  $m \in ]i, j[$ 
    if (V[m] < V[j]) then
        return gaprec(V, m, j)
    else
        return gaprec(V, i, m)
```

Capitolo Nr.11

Strutture dati specializzate

Finora abbiamo esaminato una serie di *strutture dati* e per ciascuna ne abbiamo anche analizzato il costo delle operazioni. È possibile però definire *strutture dati speciali*, o per meglio dire “*specializzate*”, nelle quali vengono implementate soltanto alcune delle operazioni e per questo motivo quelle implementazioni possono essere realizzare in modo più efficiente.

In questo capitolo vedremo due esempi di *strutture specializzate*: le *code a priorità* e gli *insiemi disgiunti*.

11.1 Code a priorità

Definizione 104 - Coda a priorità.

Una *coda a priorità* è una *struttura dati astratta*, simile a una *coda*, in cui ogni elemento possiede un valore che ne indica la “*priorità*” e che viene usato per stabilire l’ordine di estrazione degli elementi dalla struttura.

Esistono due tipi di *code a priorità* (*priority queue*):

- *Min-priority queue*: l’estrazione avviene per valori crescenti di priorità;
- *Max-priority queue*: l’estrazione avviene per valori decrescenti di priorità;

Specifica

Frammento 73 - MINPRIORITYQUEUE.

```
% Crea una coda a priorità con capacità n
PRIORITYQUEUE PriorityQueue(int n)
% Restituisce true se la coda a priorità è vuota
boolean isEmpty()
% Restituisce l’elemento minimo di una coda a priorità non vuota
ITEM min()
% Rimuove e restituisce l’elemento minimo di una coda a priorità non vuota
deleteMin()
% Inserisce l’elemento x con priorità p nella coda a priorità e restituisce
% un oggetto PRIORITYITEM che identifica x all’interno della coda
PRIORITYITEM insert(ITEM x, int p)
```

% Diminuisce la priorità dell'oggetto identificato da y portandola a p
 decrease(PRIORITYITEM y , int p)

NB. La specifica di una *max-priority queue* è uguale, ma invece delle operazioni `min`, `deleteMin` e `decrease` ha `max`, `deleteMax` e `increase`.

Utilizzando le *strutture dati* viste finora possiamo calcolare i seguenti costi:

Operazione	Lista Vettore non ordinato	Lista ordinata	Vettore ordinato	Albero Red-Black
<code>min</code>	$O(n)$	$O(1)$	$O(1)$	$O(\log n)$
<code>deleteMin</code>	$O(n)$	$O(1)$	$O(n)$	$O(\log n)$
<code>insert</code>	$O(n)$	$O(n)$	$O(n)$	$O(\log n)$
<code>decrease</code>	$O(n)$	$O(n)$	$O(\log n)$	$O(\log n)$

È possibile fare meglio di così?

La risposta è sì, utilizzando uno *heap*, una *struttura dati speciale* che associa i vantaggi di un *albero*, cioè la *complessità* $O(\log n)$, e la memorizzazione efficiente ottenibile con i normali vettori.

11.1.1 Heap

La *struttura dati* dello *heap* fu inventata da J. Williams nel 1964 con l'obiettivo di realizzare l'algoritmo di ordinamento *HeapSort*. Vediamo come si è arrivati all'ideazione dello *heap*.

Consideriamo un *albero binario perfetto* come il seguente:

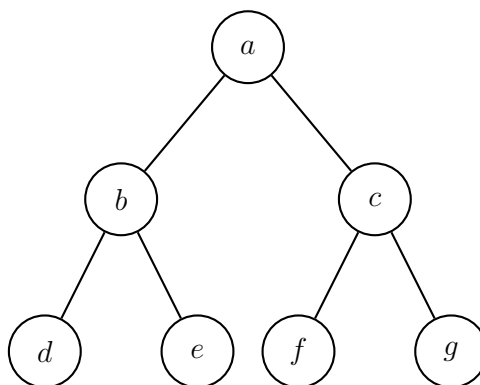


Fig. 11.1: *Albero binario perfetto*

Tutte le *foglie* sono alla stessa *profondità* h e tutti i *nodi* interni hanno *grado uscente* pari a 2. Se n è il numero di *nodi*, l'*altezza* vale $h = \lfloor \log n \rfloor$ e, data l'*altezza* h , il numero di *nodi* è $n = 2^{h+1} - 1$.

Cosa accade se si aggiunge un *nodo*?

Ovviamente l'*albero* non può più essere definito *perfetto*. Supponiamo però di “accatastare” tutti i nuovi *nodi* a partire da sinistra. In questo modo otteniamo un *albero binario completo* nel quale tutte le *foglie* hanno *profondità* h o $h - 1$, i *nodi* al *livello* h sono “accatastati” a sinistra e tutti *nodi* interni hanno *grado uscente* pari a 2 eccetto al più uno. Come prima poi, per l'*altezza* h vale $h = \lfloor \log n \rfloor$.

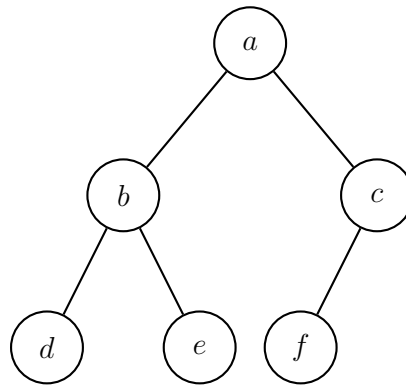


Fig. 11.2: *Albero completo*

Definizione 105 - Albero min-heap.

Un albero min-heap è un albero binario completo tale per cui il valore memorizzato in ogni nodo è minore dei valori memorizzati nei suoi figli.

Definizione 106 - Albero max-heap.

Un albero max-heap è un albero binario completo tale per cui il valore memorizzato in ogni nodo è maggiore dei valori memorizzati nei suoi figli.

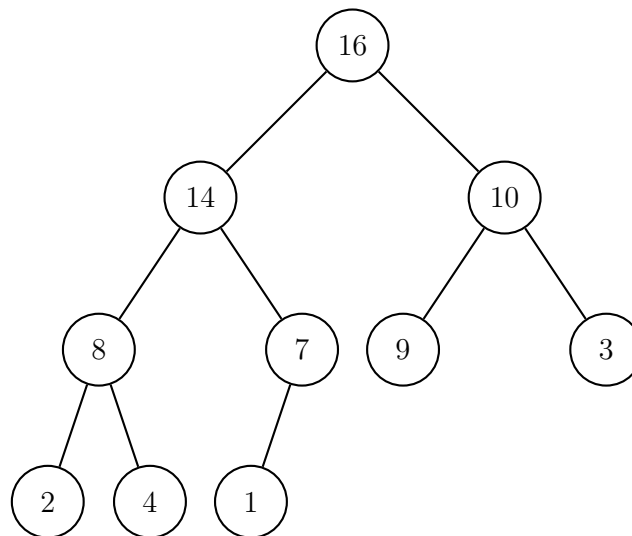


Fig. 11.3: *Albero max-heap*

Un *albero heap* non impone un ordinamento totale tra i *figli* di un *nodo*, bensì definisce un *ordinamento parziale* e quindi soddisfa le seguenti tre proprietà:

- *Riflessività*: ogni *nodo* è maggiore o uguale a se stesso;
- *Antisimmetria*: se $n \geq m$ e $m \geq n$, allora $m = n$;
- *Transitività*: se $n \geq m$ e $m \geq r$, allora $n \geq r$;

NB. Un ordinamento parziale è una nozione più debole, ma più facile da realizzare.

Memorizzazione Un *albero heap* può essere rappresentato mediante un *vettore heap*, cioè un vettore nel quale, se la *radice* è in posizione 1, il *padre* e i *figli sinistro* e *destro* di un *nodo* in posizione i si trovano rispettivamente alle posizioni $p(i) = \lfloor i/2 \rfloor$, $l(i) = 2i$ e $r(i) = 2i + 1$.

Se invece la *radice* è in posizione 0, gli indici del *padre* e dei *figli* diventano rispettivamente $p(i) = \lfloor (i - 1)/2 \rfloor$, $l(i) = 2i + 1$ e $r(i) = 2i + 2$.

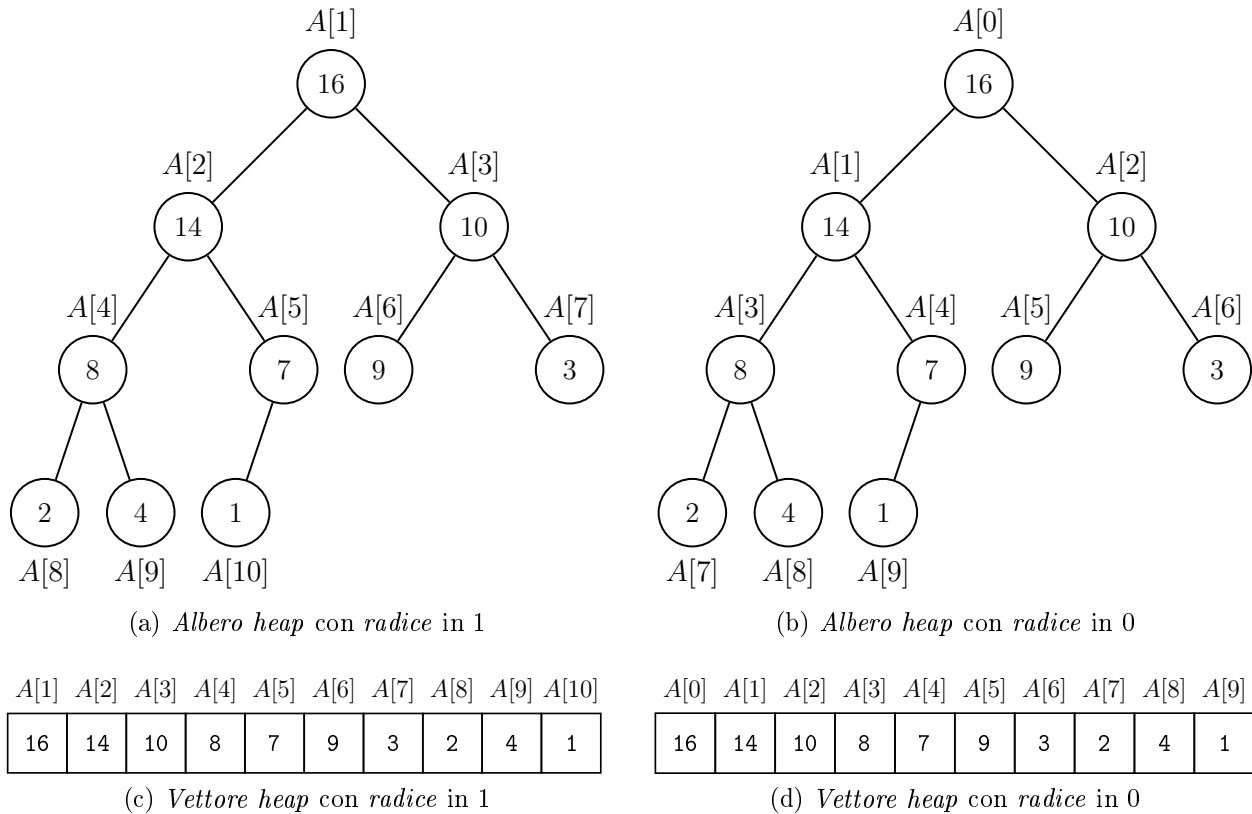


Fig. 11.4: Memorizzazione *alberi heap*

Dalle definizioni di *alberi min-heap* e *max-heap* possiamo dedurre delle proprietà sui relativi *vettori*:

- *Vettori min-heap*: $A[i] \leq A[l(i)]$ e $A[i] \leq A[r(i)]$;
- *Vettori max-heap*: $A[i] \geq A[l(i)]$ e $A[i] \geq A[r(i)]$;

11.1.2 Algoritmo HeapSort

Vediamo quindi l'algoritmo *HeapSort* per l'ordinamento in senso crescente di un vettore.

L'algoritmo ordina un vettore "in-place", costruendo prima su di esso un *albero max-heap* e poi spostando progressivamente in fondo gli elementi massimi. Ad ogni spostamento vengono ripristinate le proprietà degli *alberi max-heap*.

Le funzioni utilizzate sono due:

- **heapBuild**: costruisce un *albero max-heap* a partire da un vettore non ordinato;
- **maxHeapRestore**: ripristina le proprietà degli *alberi max-heap*;

NB. Tutte le operazioni vengono effettuate "in-place", cioè agiscono sul vettore stesso senza usare *strutture ausiliarie*. Quindi, l'*albero max-heap* non viene realmente creato, ma grazie alle sue proprietà di memorizzazione, il vettore di input viene semplicemente interpretato come fosse un *albero max-heap*.

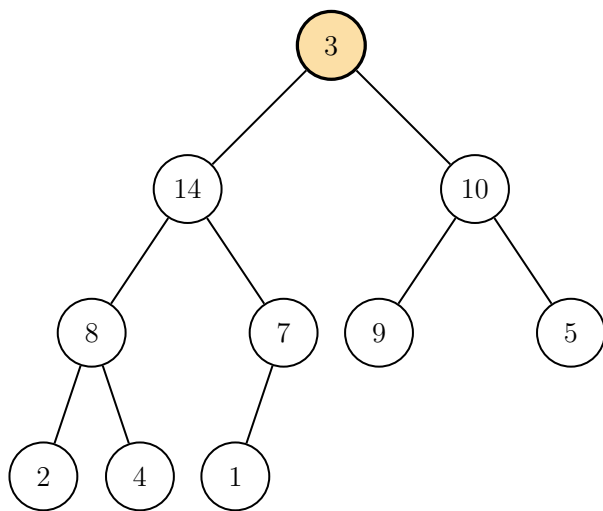
Per questo motivo, la funzione *maxHeapRestore* riceve in input un vettore *A* e un indice *i* e si occupa di garantire che al termine dell'esecuzione, l'albero binario con radice *i* sia un albero *max-heap*.

Esempio 27 - Esempio di esecuzione della *maxHeapRestore*.

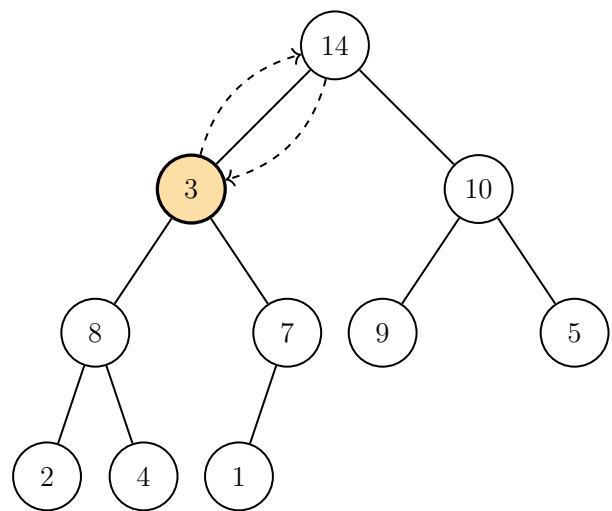
Consideriamo il seguente vettore:

(3, 14, 10, 8, 7, 9, 5, 2, 4, 1)

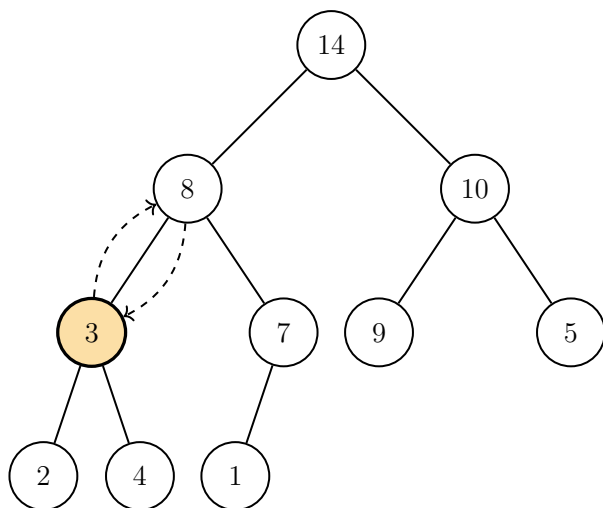
e ipotizziamo di invocare *maxHeapRestore* su di esso usando come radice la posizione 1.



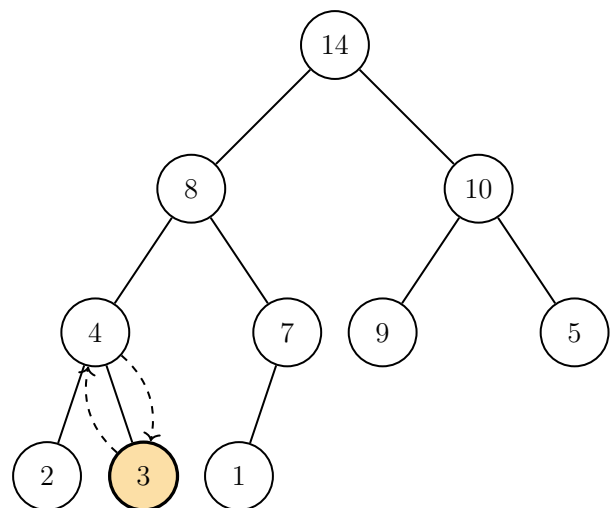
(a) Situazione iniziale



(b) Il 3 viene scambiato con il 14



(c) Il 3 viene scambiato con l'8



(d) Il 3 viene scambiato con il 4

Al termine dell'esecuzione abbiamo ottenuto un albero *max-heap* corretto.

Implementazione

Frammento 74 - Implementazione maxHeapRestore.

```
maxHeapRestore(ITEM[] A, int i, int dim)
    int max = i                                % Sceglie la radice
    if (l(i) ≤ and dim A[l(i)] > A[max]) then
        max = l(i)
    if (r(i) ≤ and dim A[r(i)] > A[max]) then
        max = r(i)
    if (i ≠ max) then                          % Se i == max l'albero è apposto
        swap(A, i, max)                       % Scambia la radice e il maggiore tra i suoi figli
        maxHeapRestore(A, max, dim)           % Controlla il sottoalbero con radice max
```

Complessità Ad ogni chiamata vengono eseguiti $O(1)$ confronti e se il *nodo* i non è massimo, si invoca ricorsivamente **maxHeapRestore** su uno dei *figli*. L'esecuzione nel caso peggiore termina al raggiungimento di una *foglia*, quindi il costo è limitato dall'*altezza* dell'*albero*, ovvero $T(n) = O(\log n)$.

Dimostrazione di correttezza

Dimostrazione. Dobbiamo dimostrare che al termine dell'esecuzione l'*albero* con *radice* in $A[i]$ rispetta le proprietà degli *alberi max-heap*. Procediamo per induzione sull'*altezza* h dell'*albero*.

Caso base $h = 0$. L'*albero* è composto da un solo *nodo* e quindi rispetta per forza le proprietà.

Passo induttivo Ipotizziamo che l'algoritmo funzioni correttamente su tutti gli *alberi* di *altezza* minore di h . A questo punto si configurano tre casi:

- *Caso 1:* $A[i] \geq A[l(i)]$ e $A[i] \geq A[r(i)]$, ovvero il *nodo* $A[i]$ è più grande dei propri *figli* e quindi l'*albero* radicato in $A[i]$ è un *albero max-heap* e l'algoritmo termina;
- *Caso 2:* $A[i] < A[l(i)]$ e $A[r(i)] < A[l(i)]$, ovvero il *figlio sinistro* è più grande sia del *padre* che del *fratello*. Quindi, vengono scambiati di posizione il *figlio sinistro* e il *padre*. A questo punto valgono $A[i] \geq A[l(i)]$ e $A[i] \geq A[r(i)]$, ma il *sottoalbero* con *radice* in $A[l(i)]$ potrebbe non rispettare più le proprietà per gli *alberi max-heap*. Di conseguenza, l'algoritmo continua applicando ricorsivamente **maxHeapRestore** sull'*albero* con *radice* in $A[l(i)]$, ma siccome quell'*albero* ha *altezza* minore di h , per ipotesi induttiva, l'algoritmo riesce correttamente a garantire il rispetto di tutte le proprietà;
- *Caso 3:* $A[i] < A[r(i)]$ e $A[l(i)] < A[r(i)]$, ovvero il *figlio destro* è più grande sia del *padre* che del *fratello*. Quindi, vengono scambiati di posizione il *figlio destro* e il *padre*. A questo punto valgono $A[i] \geq A[l(i)]$ e $A[i] \geq A[r(i)]$, ma il *sottoalbero* con *radice* in $A[r(i)]$ potrebbe non rispettare più le proprietà per gli *alberi max-heap*. Di conseguenza, l'algoritmo continua applicando ricorsivamente **maxHeapRestore** sull'*albero* con *radice* in $A[r(i)]$, ma siccome quell'*albero* ha *altezza* minore di h , per ipotesi induttiva, l'algoritmo riesce correttamente a garantire il rispetto di tutte le proprietà;

□

Consideriamo ora la funzione `heapBuild`. La funzione riceve in input un vettore da ordinare e, per le proprietà di memorizzazione degli *alberi heap*, sappiamo che tutti i *nod*i con indici compresi tra $\lfloor n/2 \rfloor + 1$ e n sono *foglie*, ovvero *alberi heap* contenenti un solo elemento.

La funzione `heapBuild` quindi, non fa altro che applicare `maxHeapRestore` a tutti i *nod*i, partendo dal primo che non è una *foglia*, cioè da $A[\lfloor n/2 \rfloor]$, e risalendo fino alla *radice*.

Implementazione

Frammento 75 - Implementazione `heapBuild`.

```
heapBuild(ITEM[] A, int n)
  for (i =  $\lfloor n/2 \rfloor$  downto 1) do
    maxHeapRestore(A, i, n)
```

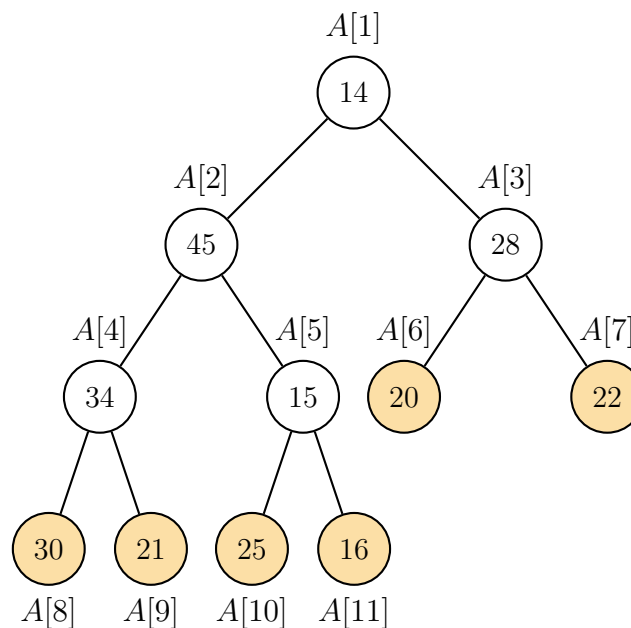


Fig. 11.5: `maxHeapRestore` viene applicata solo ai *nod*i interni

Dimostrazione di correttezza

Dimostrazione. Dimostriamo la seguente *invariante di ciclo*:

All'inizio di ogni iterazione del ciclo `for`, i nodi $[i + 1 \dots n]$ sono radici di alberi heap.

Inizializzazione All'inizio $i = \lfloor n/2 \rfloor$. Supponiamo che $\lfloor n/2 \rfloor + 1$ non sia una *foglia* e che quindi abbia almeno il *figlio sinistro*. Se così fosse, il *figlio* dovrebbe trovarsi alla posizione $2(\lfloor n/2 \rfloor + 1) = 2\lfloor n/2 \rfloor + 2$, ma ciò sarebbe assurdo perché le posizioni $n + 1$ e $n + 2$ eccedono la dimensione massima n . La stessa dimostrazione vale anche per tutti gli indici successivi.

Conservazione È possibile applicare `maxHeapRestore` a ogni *nodo* i perché entrambi i *nod*i alle posizioni $2i < 2i + 1 \leq n$ sono *radici* di *alberi heap*. Al termine di ogni iterazioni, i *nod*i $[i \dots n]$ sono *radici* di *alberi heap*.

Conclusione Al termine, $i = 0$ e quindi il *nodo* 1 è *radice* di un *albero heap*. □

Complessità Sicuramente possiamo dire che `heapBuild` è un $O(n \log n)$ in quanto il costo di `maxHeapRestore` è $O(\log n)$ e viene invocata $\lfloor n/2 \rfloor = O(n)$ volte. Ma è veramente questo il costo?

In realtà, le operazioni `maxHeapRestore` vengono eseguite un numero decrescente di volte in un *albero* di *altezza* crescente. E in particolare, vale la seguente tabella:

Altezza	# esecuzioni
0	$\lfloor n/2^1 \rfloor$
1	$\lfloor n/2^2 \rfloor$
2	$\lfloor n/2^3 \rfloor$
...	...
h	$\lfloor n/2^{h+1} \rfloor$

Se questo è vero, possiamo scrivere la *funzione di ricorrenza* come:

$$\begin{aligned}
 T(n) &\leq \sum_{h=1}^{\lfloor n/2 \rfloor} \frac{n}{2^{h+1}} h \\
 &= n \sum_{h=1}^{\lfloor n/2 \rfloor} \left(\frac{1}{2}\right)^{h+1} h \\
 &= \frac{n}{2} \sum_{h=1}^{\lfloor n/2 \rfloor} \left(\frac{1}{2}\right)^h h \\
 &\leq \frac{n}{2} \sum_{h=1}^{+\infty} \left(\frac{1}{2}\right)^h h && \text{Estensione ad infinito della somatoria} \\
 &= \frac{n}{2} \frac{\frac{1}{2}}{\left(1 - \frac{1}{2}\right)^2} && \text{Serie geometrica infinita decrescente} \\
 &= \frac{n}{2} \frac{\frac{1}{2}}{\left(\frac{1}{2}\right)^2} \\
 &= \frac{n}{2} 2 \\
 &= n = O(n)
 \end{aligned}$$

Implementazione

Frammento 76 - Implementazione dell'algoritmo HeapSort.

```

heapsort(ITEM[] A, int n)
    heapBuild(A, n)
    for (i = n downto 2) do
        swap(A, 1, i)           % L'elemento massimo viene spostato in fondo
        maxHeapRestore(A, 1, i - 1) % Ripristina le proprietà

```

In pratica, l'*HeapSort* inizia costruendo un *albero max-heap* sul vettore da ordinare. Ad ogni ciclo il primo elemento sarà sicuramente il massimo e quindi viene scambiato di posto con l'elemento in posizione i . Dopo ogni scambio viene invocata `maxHeapRestore` per ripristinare le proprietà degli *alberi max-heap*. Il tutto viene ripetuto fino a quando i non diventa 1.

Complessità La funzione `heapBuild` costa $\Theta(n)$, mentre le `maxHeapRestore` ha un costo $\Theta(\log i)$ per i che varia da n a 2. La *funzione di ricorrenza* è la seguente:

$$T(n) = \sum_{i=n}^2 \log i + \Theta(n) = \Theta(n \log n)$$

Dimostrazione di correttezza

Dimostrazione. Dimostriamo la seguente *invariante di ciclo*:

All'inizio di ogni passo i il sottovettore $A[i+1 \dots n]$ è ordinato e ogni elemento in $A[1 \dots i]$ è minore o al più uguale a ogni elemento in $A[i+1 \dots n]$. Inoltre, l'elemento $A[1]$ è la radice di un albero heap di dimensione i .

Inizializzazione Dopo la `heapBuild` il primo elemento è la *radice* di un *albero max-heap* e quindi è maggiore dei propri *figli* e, per la proprietà transitiva, anche dei *figli dei figli* fino alle *foglie*. Alla prima iterazione quindi, $A[1]$ è il massimo e viene scambiato di posto con $A[i] = A[n]$. A quel punto, $A[n]$ è l'elemento massimo e il sottovettore $A[1 \dots n-1]$ viene riorganizzato secondo le regole della `maxHeapRestore` portandone il massimo in $A[1]$.

Conservazione Per ogni $i \in [2 \dots n]$, il sottovettore $A[i+1 \dots n]$ è ordinato in senso crescente e in posizione $A[1]$ si trova l'elemento massimo del sottovettore $A[1 \dots i]$. Al termine dell'iterazione, $A[i \dots n]$ è ordinato e $A[1]$ è il massimo in $A[1 \dots i-1]$.

Conclusione Al termine, $i = 1$ e quindi il sottovettore $A[2 \dots n]$ è ordinato. Poiché vale sicuramente $A[1] \leq A[2]$, possiamo dire che tutto il vettore $A[1 \dots n]$ è ordinato in senso crescente. \square

11.1.3 Implementazione di code a priorità

Prima di vedere l'implementazione di una *coda a priorità*, vediamo la definizione di un **PRIORITYITEM** e l'implementazione della funzione `swap`.

Frammento 77 - PRIORITYITEM.

```
int priority                                % Valore di priorità
PRIORITYITEM value                         % Valore
int pos                                    % Posizione nel vettore
```

Frammento 78 - Implementazione funzione swap.

```
swap(PRIORITYITEM H, int i, int j)
    PRIORITYITEM temp = H[i]
    H[i] = H[j]
    H[j] = temp
    H[i].pos = i
    H[j].pos = j
```

Di seguito quindi, vediamo l'implementazione di una *min-priority queue* come vettore di coppie $\langle \text{valore}, \text{priorità} \rangle$. Ovviamente, tutto ciò che vedremo vale simmetrico per le *max-priority queue*.

Frammento 79 - Min-priority queue.

```
int capacity                                % Dimensione massima della coda
int dim                                    % Numero di elementi attualmente presenti nella coda
PRIORITYQUEUE[] H                         % Vettore heap
```

```

PRIORITYQUEUE PriorityQueue(int n)
    PRIORITYQUEUE t = new PRIORITYQUEUE
    t.capacity = n
    t.dim = 0
    t.H = new PRIORITYITEM[1...n]
    return t

PRIORITYITEM insert(ITEM x, int p)
    precondition: dim < capacity
    dim = dim + 1
    H[dim] = new PRIORITYITEM                                % Aggiunge l'elemento in fondo
    H[dim].value = x
    H[dim].priority = p
    H[dim].pos = dim
    int i = dim
    while (i > 1 and H[i].priority < H[p(i)].priority) do
        swap(H, i, p(i))    % Sposta il nuovo elemento nella posizione giusta
        i = p(i)
    return H[i]

minHeapRestore(PRIORITYITEM[] A, int i, int n)
    int min = i
    if (l(i) ≤ dim and A[l(i)].priority < A[min].priority) then
        min = l(i)
    if (r(i) ≤ dim and A[r(i)].priority < A[min].priority) then
        min = r(i)
    if (i ≠ min) then
        swap(A, i, min)
        minHeapRestore(A, min, dim)

ITEM deleteMin()
    precondition: dim > 0
    swap(H, 1, dim)                                            % Scambia il primo e l'ultimo elemento
    dim = dim - 1
    minHeapRestore(H, 1, dim)                                  % Ripristina le proprietà
    return H[dim + 1].value    % Restituisce il valore dell'elemento rimosso

ITEM min()
    precondition: dim > 0
    return H[1].value

decrease(PRIORITYITEM x, int p)
    precondition: p < x.priority
    x.priority = p
    int i = x.pos
    while (i > 1 and H[i].priority < H[p(i)].priority) do
        swap(H, i, p(i))    % Sposta l'elemento modificato nella posizione giusta
        i = p(i)

```

Complessità Tutte le operazioni che modificano gli *heap* ne ripristinano anche le proprietà. In particolare, questo viene fatto nel *cammino radice-foglia* dalla `deleteMin` e in quello *nodo-radice* dalla `insert` e dalla `decrease`. Poiché l'altezza è $\lfloor \log n \rfloor$, il costo di tali operazioni è $O(\log n)$.

La *complessità* di tutte le operazioni è riassunta dalla seguente tabella:

Operazione	Costo
<code>insert</code>	$O(\log n)$
<code>deleteMin</code>	$O(\log n)$
<code>min</code>	$\Theta(1)$
<code>decrease</code>	$O(\log n)$

11.2 Insiemi disgiunti

In alcune applicazioni (e.g. ricerca delle *componenti connesse* di un *grafo*) siamo interessati a gestire una collezione $S = \{S_1, \dots, S_n\}$ di *insiemi dinamici disgiunti*, ovvero tali per cui:

$$S_i \cap S_j = \emptyset \quad \forall i, j : i \neq j$$

$$\bigcup_{i=1}^n S_i = S \quad \text{con } n = |S|$$

Definiamo quindi un *merge-find set* come una struttura dati con le seguenti primitive:

- Creazione di n *insiemi disgiunti* composti da un elemento ciascuno;
- `find`: identificazione dell'*insieme* a cui appartiene un elemento;
- `merge`: unione di due *insiemi*;

Ogni insieme è identificato da un proprio elemento che viene designato come *rappresentante*. Di conseguenza, la primitiva `find` deve restituire il *rappresentante* dell'*insieme* in cui si trova elemento cercato. Definito l'elemento *rappresentante*, questo può essere cambiato solo in caso di unione con un altro *insieme*.

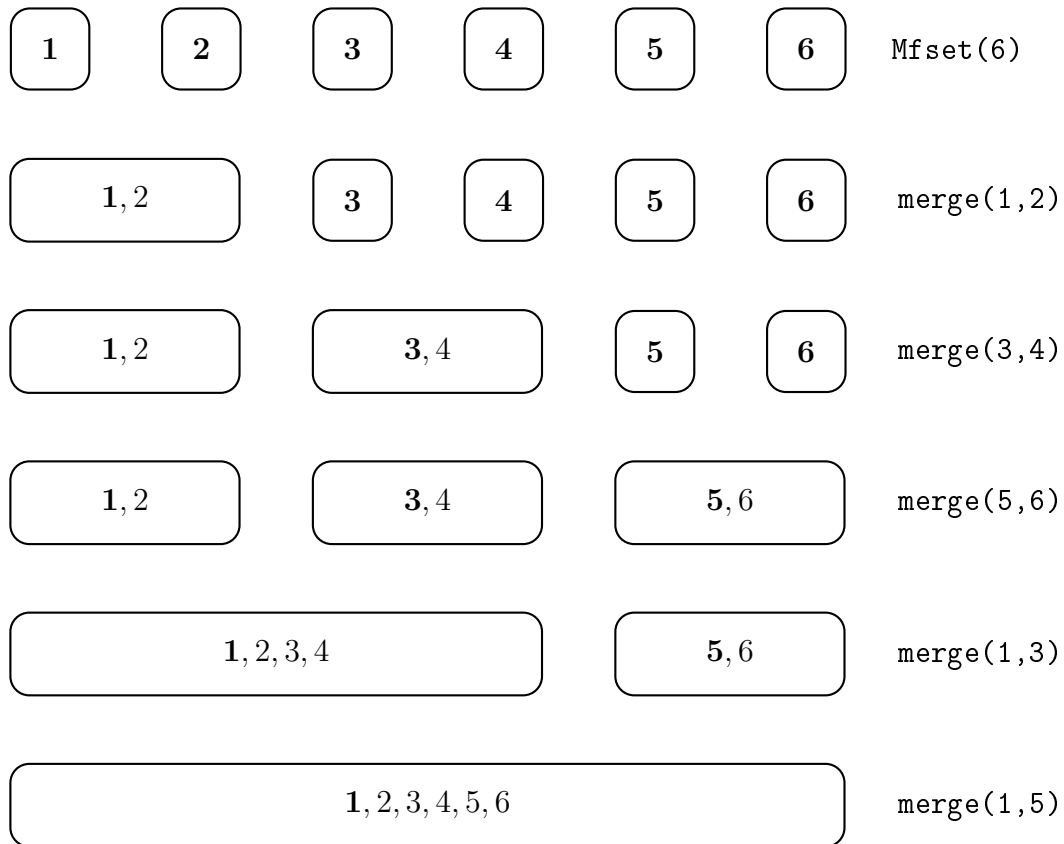
NB. Per semplicità di esposizione, ipotizziamo che gli elementi siano identificati da un valore intero $1 \dots n$ e che l'associazione con il valore vero e proprio sia memorizzata esternamente.

Specifica

Frammento 80 - Merge-find set.

```
% Crea n insiemi {1}, ..., {n}
MFSET Mfset(int n)
% Restituisce il rappresentante dell'insieme contenente x
int find(int x)
% Unisce gli insiemi che contengono x e y
merge(int x, int y)
```

Esempio 28 - Esempio di utilizzo di un merge-find set.



Utilizzando la struttura dati così definita, possiamo semplificare molto l'algoritmo per la ricerca delle *componenti connesse* di un *grafo* $G = (V, E)$. In particolare, è sufficiente inizializzare la struttura vedendo ogni *nodo* come un *insieme*, quindi procedere ad unire ogni *arco* $(u, v) \in E$ con una `merge(u, v)`. Al termine, ogni *insieme disgiunto* rappresenterà una *componente connessa*.

Frammento 81 - Ricerca delle componenti connesse con merge-find set.

```
MFSET cc(GRAPH G)
  MFSET M = Mfset(G.size())
  foreach (u ∈ G.V()) do
    foreach (v ∈ G.adj(u)) do
      M.merge(u, v)
  return M
```

La *complessità* di questa versione è $T(n) = O(n + m)$ dove m è il numero di invocazioni di `merge`.

NB. Questa versione dell'algoritmo è particolarmente indicata nel caso di *grafi dinamici*, nei quali vengono spesso aggiunti nuovi *archi*.

Ma quanto costa eseguire una **merge** e come è implementata?

Ci sono diverse implementazioni possibili, tra le quali, implementazioni basate su *insiemi di liste* e su *insiemi di alberi* (i.e. *foreste*).

11.2.1 Implementazione basata su insiemi di liste

Ogni *insieme* viene rappresentato da una *lista concatenata* in cui ciascun elemento, oltre al proprio valore, contiene un puntatore all'elemento successivo e un puntatore al *rappresentante*. Il *rappresentante* è il primo elemento della lista.

La primitiva **find** è banale, perché dato un *elemento*, è sufficiente restituire il valore del puntatore al *rappresentante*, e quindi richiede un tempo $O(1)$. Per la **merge**, si prende il *rappresentante* del secondo parametro, lo si “appende” alla *lista* del primo e si aggiornano tutti i *rappresentanti* della *lista* “appesa”. Nel caso pessimo, eseguire n operazioni **merge** costa $O(n^2)$, quindi il *costo ammortizzato* è $O(n)$.

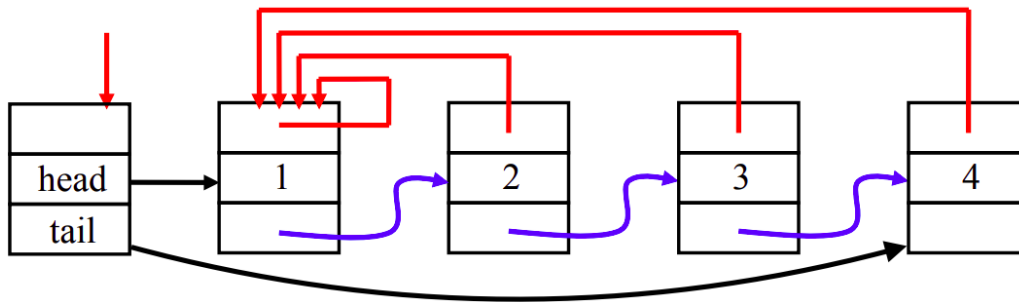


Fig. 11.6: Merge-find set implementati come *insiemi di liste*

11.2.2 Implementazione basata su insiemi di alberi

Ogni *insieme* viene rappresentato da un *albero* in cui ogni *nodo* contiene il proprio valore e un puntatore al *padre*. La *radice* è il *rappresentante* dell'*insieme* e il puntatore al *padre* punta al *nodo* stesso.

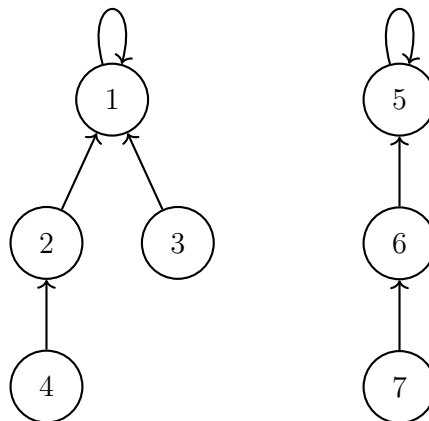


Fig. 11.7: Merge-find set implementati come *insiemi di alberi*

In questo caso la **find** risale *albero* fino alla *radice* e nel caso pessimo, questo costa $O(n)$. Per fare la **merge** invece, è sufficiente cambiare il puntatore al *padre* della *radice* in uno dei due *alberi* in modo che punti all'altra *radice*. Il costo, se non consideriamo quello per la ricerca dei *rappresentanti*, è $O(1)$.

11.2.3 Implementazioni con tecniche euristiche

Possiamo migliorare le implementazioni viste sfruttando qualche piccolo accorgimento?

Nel caso dei *merge-find set* basati su *liste* potremmo provare ad eseguire la **merge** modificando i puntatori della *lista* più corta. Per quanto riguarda gli *alberi* invece, potremmo ridurre il costo della **find** minimizzando l'*altezza* degli *alberi*. Tecniche di questo tipo sono chiamate *euristiche*, e gli algoritmi che le usano sono detti *euristici*.

Definizione 107 - Algoritmo euristico.

È detto *euristico* un algoritmo progettato per risolvere un problema più velocemente, qualora i metodi tradizionali non siano sufficienti, oppure per ricavare una soluzione approssimata, qualora non sia possibile ricavarne una esatta.

Euristica sul peso Approfondendo l'intuizione sulla lunghezza delle *liste*, possiamo memorizzare in ogni *lista* l'informazione sulla propria lunghezza e implementare la **merge** in modo che modifichi i puntatori dei *nodi* che stanno nella *lista* più corta. La lunghezza può essere mantenuta in $O(1)$ ed è possibile dimostrare che in questo tipo di implementazione il *costo ammortizzato* della **merge** si riduce ad $O(\log n)$.

Euristica sul rango Abbiamo detto che per gli *alberi* conviene cercare di minimizzare le *altezze*, quindi associamo ad ogni *nodo* k il proprio *rango*, $rank[k]$, definito come segue:

Definizione 108 - Rango di un nodo.

È definito *rango* di un *nodo* il numero di archi del più lungo cammino tra quel *nodo* e una delle proprie foglie.

NB. Possiamo anche definire il *rango* di un *nodo* come l'*altezza* del *sottoalbero* in esso radicato. A questo punto, per ridurre il costo della **find**, è sufficiente modificare la **merge** in modo che, nel caso di *alberi* con *ranghi* diversi, sia l'*albero* con *rango* minore ad essere "agganciato" all'altro. In questa situazione, l'*altezza* dell'*albero* con *rango* maggiore non cambia, ma se invece i due *alberi* avessero pari *rango*, l'*altezza* finale andrebbe incrementata di 1.

Definizione 109 - Legame tra il rango di un albero e il proprio numero di nodi.

Un albero MFSET con radice r e ottenuto tramite euristica sul rango ha almeno $2^{rank[r]}$ nodi.

Dimostrazione. Procediamo per induzione sul *rango* di r .

Caso base: $rank[r] = 0$ Dopo l'inizializzazione della struttura ogni *albero* ha $2^{rank[r]} = 2^0 = 1$ nodi;

Passo induttivo: $rank[r] > 0$ Facendo la **merge** di due *alberi* x e y di rango $rank[x]$ e $rank[y]$ distinguiamo due casi:

1. $rank[x] > rank[y]$: il *rango* dell'*albero* r ottenuto è $rank[r] = rank[x]$. Per induzione, il numero di *nodi* di r è almeno pari a $2^{rank[x]} + 2^{rank[y]}$ e quindi è maggiore di $2^{rank[x]}$;
2. $rank[x] = rank[y]$: il *rango* dell'*albero* r ottenuto è $rank[r] = rank[x] + 1$. Per induzione, il numero di *nodi* di r è almeno pari a $2^{rank[x]} + 2^{rank[y]} = 2^{rank[x]} + 2^{rank[x]} = 2^{rank[x]+1}$;

□

Dal teorema appena dimostrato possiamo ricavare il seguente corollario:

Definizione 110 - Corollario.

Un albero MFSET con radice r e n nodi ha un'altezza minore o uguale di $\log n$.

Dimostrazione. Vale la seguente relazione:

$$n \geq 2^{\text{rank}[r]} \Leftrightarrow \text{rank}[r] \leq \log n$$

□

Detto questo, la *complessità* della primitiva `find` è $O(\log n)$.

Euristica di compressione dei cammini Poiché il costo della `find` è legato all'*altezza* degli *alberi*, se tutti avessero *altezza* pari a 1, cioè se il *padre* di ogni *nodo* fosse la *radice*, riusciremmo a ridurre il costo di ogni invocazione a $O(1)$. Per raggiungere questo obiettivo possiamo modificare l'implementazione della `find` in modo che, quando risaliamo l'*albero*, ogni *nodo* modifichi il proprio puntatore al *padre* indicando il *padre* del proprio *padre*. In questo modo, la prima invocazione di `find` costa $O(\log n)$, mentre ricerche successive su *nod*i già visitati costeranno $O(1)$.

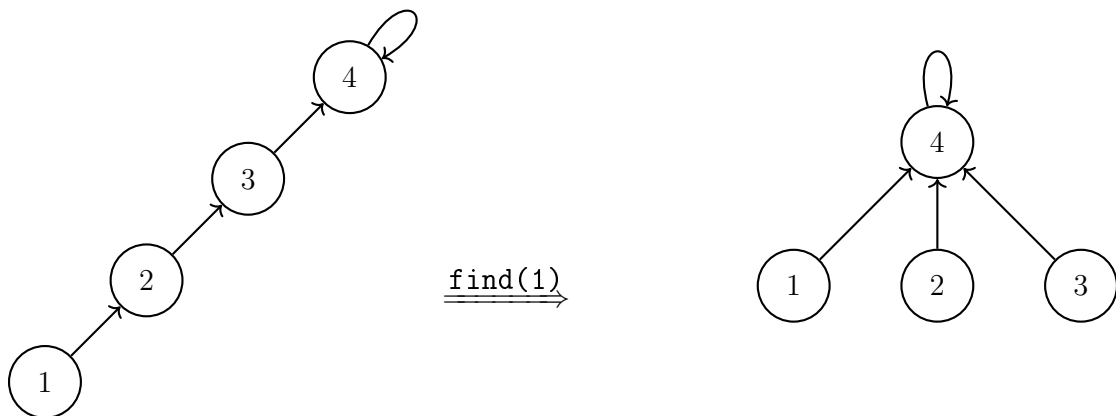


Fig. 11.8: Esempio di *compressione dei cammini*

NB. Applicando entrambe le *euristiche*, il *rango* non indica più l'*altezza* del *nodo*, bensì un limite superiore. Questo perché mantenere il valore corretto diventa troppo difficile (i.e. complesso) e comunque non è necessario.

Complessità Utilizzando entrambe le *tecniche euristiche* viste, il *costo ammortizzato* di m operazioni `merge-find` in un *insieme* di n elementi è $O(m \cdot \alpha(n))$, dove $\alpha(n)$ è la *funzione inversa di Ackermann* che ha una crescita estremamente lenta¹. Di conseguenza, il *costo ammortizzato* di una singola operazione è $O(1)$.

¹Ad esempio, per $n \leq 2^{65536}$, $\alpha(n) \leq 5$

Implementazione

Frammento 82 - Merge-find set basati su insiemi di alberi.

```
int[] parent      % Vettore dei padri      merge(int x, int y)
int[] rank        % Ranghi dei nodi        int rx = find(x)
                                                int ry = find(y)
                                                if (rx ≠ ry) then
MFSET Mfset(int n)                            if (rank[rx] > rank[ry]) then
    MFSET t = new MFSET                        parent[ry] = rx
    t.parent = new int[1...n]                  else if (rank[ry] > rank[rx]) then
    t.rank = new int[1...n]                    parent[rx] = ry
    for (i = 1 to n) do                        else
        t.parent[i] = i                        parent[rx] = ry
        t.rank[i] = 1                          rank[ry] = rank[ry] + 1
    return t
int find(int x)
    if (parent[x] ≠ x) then
        parent[x] = find(parent[x])
    return parent[x]
```

11.2.4 Complessità

Riassumendo, le *complessità* delle possibili implementazioni dei *merge-find set* sono le seguenti²:

Algoritmo	find	merge
Liste	$O(1)$	$O(n)$
Alberi	$O(n)$	$O(1)^+$
Liste con euristica sul peso	$O(1)$	$O(\log n)^*$
Albero con euristica sul rango	$O(\log n)$	$O(1)^+$
Albero con euristica sul rango e compressione dei cammini	$O(1)^*$	$O(1)$

² *: *complessità ammortizzata*;

+ : Consideriamo solo il costo della **merge**, senza la ricerca dei *rappresentanti* con la **find**;

12.1 Introduzione

Finora, l'unica tecnica di risoluzione di problemi, o di ricerca di algoritmi, che abbiamo visto è il **Divide-et-impera**. La *programmazione dinamica* è un'altra tecnica che però si basa su un approccio molto simile.

Entrambe prevedono di spezzare il problema di partenza in sotto-problemi più semplici e di ricostruire la soluzione del problema originale a partire dalle soluzioni dei sotto-problemi. La differenza fondamentale sta nel fatto che la *programmazione dinamica* risolve ogni sotto-problema una sola volta, mentre il *Divide-et-impera* non pone questo vincolo.

Per realizzare ciò, le soluzioni di tutti i sotto-problemi vengono salvate in una tabella che viene consultata ogni volta che è necessario risolvere uno dei sotto-problemi. In particolare, se la tabella non contiene la soluzione al sotto-problema considerato, questa viene calcolata e aggiunta alla tabella. In caso contrario, viene sfruttata la soluzione già nota.

12.1.1 Approccio generale

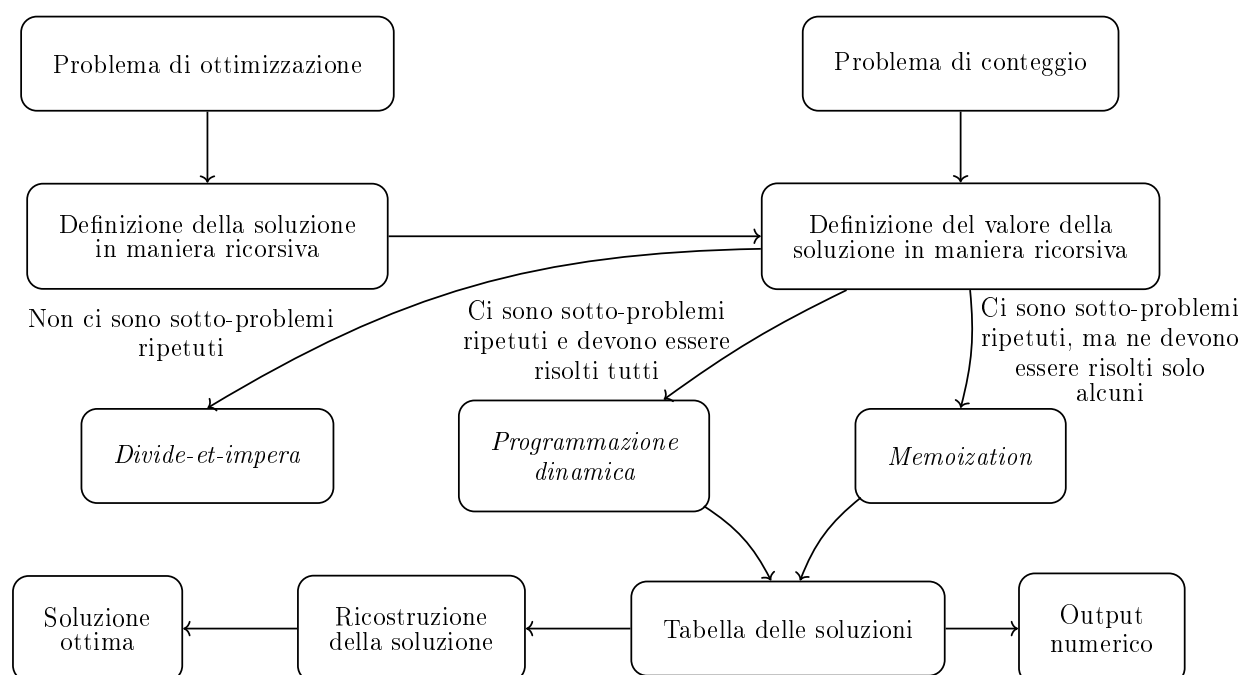


Fig. 12.1: Schema generale di approccio

Lo schema di cui sopra mostra un approccio generale alla scelta della tecnica risolutiva più adatta al tipo di problema da risolvere. In particolare, distinguiamo la soluzione del problema dal proprio valore, ad esempio, nella ricerca del *cammino breve* tra due *nodi*, la soluzione è il *cammino*, mentre il valore potrebbe esserne la lunghezza.

Una volta definiti questi parametri, se la suddivisione del problema non porta a dover risolvere più volte uno stesso sotto-problema, possiamo procedere ad implementare la soluzione sfruttando l'approccio *Divide-et-impera* classico. Caso contrario, ricadiamo nei casi d'uso della *programmazione dinamica*, ma dobbiamo ancora fare una distinzione: la risoluzione di alcuni dei sotto-problemi potrebbe non essere necessaria per arrivare alla soluzione del problema di partenza. Se il caso è questo, procediamo con la tecnica della *memoization* che non è altro se non un approccio top-down alla *programmazione dinamica* “classica”.

In ogni caso, utilizzando la *programmazione dinamica* si arriva alla definizione di una *tabella delle soluzioni* e, se il problema è un *problema di conteggio*, da cui possiamo ricavare direttamente la soluzione. Se invece stiamo considerando un *problema di ottimizzazione*, la soluzione dovrà essere ricavata attraverso un processo di “ricostruzione”.

12.2 Gioco del domino

Vediamo adesso un primo esempio di utilizzo della *programmazione dinamica* e di come la soluzione si differenzi, soprattutto in termini di efficienza, da una classica soluzione *divide-et-impera*.

Problema 5 - Domino lineare.

Il gioco del domino è basato su tessere di dimensione 2×1 . Scrivere un algoritmo che prenda in input un valore intero positivo n e restituisca il numero di modi in cui è possibile disporre le tessere del gioco in modo che creino un rettangolo $2 \times n$.

Ad esempio, per n pari a 0, 1, 2, 3, 4 l'algoritmo dovrà restituire 1, 1, 2, 3, 5.

NB. Per $n = 0$, restituiamo 1 perché per ottenere un rettangolo di dimensione 2×0 c'è un solo modo: non disporre nessuna tessera.

Basandoci sullo schema di cui sopra possiamo identificare questo problema con un *problema di conteggio* e quindi la prima cosa da fare è la “definizione del valore della soluzione in maniera ricorsiva”.

12.2.1 Approccio basato su divide-et-impera

Chiamiamo $DP[n]$ la funzione che per ogni valore $n \in \mathbb{N}$ restituisce il numero di disposizioni possibili delle tessere. La funzione DP è ricorsiva, quindi, prima di tutto identifichiamo il caso base: poiché per $n = 0$ e $n = 1$ $DP[n] = 1$, possiamo decidere che $DP[n] = 1$ per $n \leq 1$.

A questo punto, consideriamo il caso più complesso in cui dobbiamo esaminare un rettangolo $2 \times n$ arbitrariamente grande. Poiché le tessere del domino hanno dimensione 2×1 e non posso essere posizionate in modo sfalsato, i possibili posizionamenti di ogni tessera sono due: in verticale o in orizzontale. Se ipotizzassimo di mettere una tessera verticale a destra o a sinistra del rettangolo, il problema si ridurrebbe all'analisi di un rettangolo $2 \times (n - 1)$. Se invece posizionassimo la tessera in orizzontale otterremmo una figura composta da un rettangolo $2 \times (n - 2)$ e da un altro rettangolo più piccolo di dimensione 1×2 . Chiaramente, il rettangolo più piccolo corrisponde ad una tessera orizzontale posta sopra o sotto quella che abbiamo già messo e, di conseguenza, arriviamo alla conclusione che qualora si disponga una tessera in orizzontale se ne debba sempre aggiungere anche una seconda sopra o sotto la prima.

Fatta questa riflessione possiamo terminare dicendo che, dato un generico rettangolo $2 \times n$, il numero di possibili disposizioni di tessere è pari a $DP[n-1] + DP[n-2]$ e che quindi la definizione della funzione ricorsiva è:

$$DP[n] = \begin{cases} 1 & n \leq 1 \\ DP[n-1] + DP[n-2] & n > 1 \end{cases}$$

NB. Le soluzioni per $n-1$ e $n-2$ si sommano perché per ogni rettangolo $2 \times n$ ho la possibilità di mettere una tessera verticale o due tessere orizzontali.

Provando a valutare la funzione per valori crescenti di n otteniamo la seguente sequenza:

$$1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, \dots$$

che corrisponde esattamente alla *sequenza di Fibonacci*. L'implementazione dell'algoritmo corrisponde perciò all'implementazione della funzione per il calcolo di tale sequenza.

Frammento 83 - Prima implementazione della soluzione.

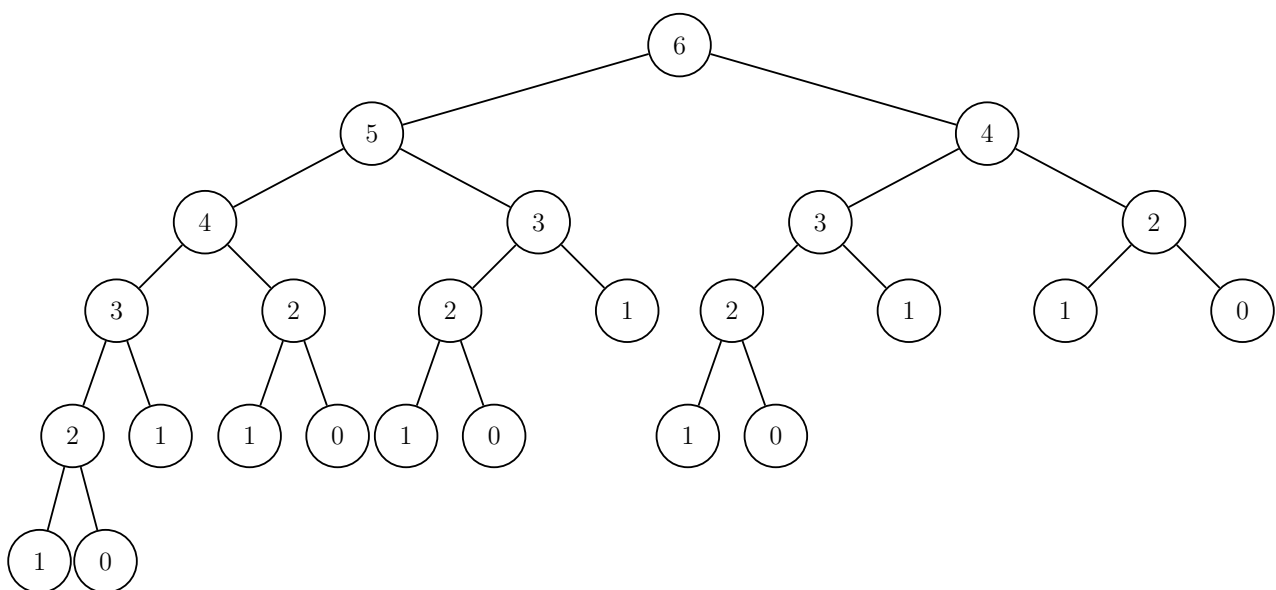
```
int domino1(int n)
    if (n ≤ 1) then
        return 1
    else
        return domino1(n-1) + domino1(n-2)
```

La *funzione di ricorrenza* associata a questa funzione è la seguente:

$$T(n) = \begin{cases} 1 & n \leq 1 \\ T(n-1) + T(n-2) & n > 1 \end{cases}$$

Per il *Teorema delle ricorrenze lineari di ordine costante* la *forma chiusa* di tale funzione è $T(n) = \Theta(2^n)$. Un costo esponenziale è tutt'altro che efficiente, non possiamo fare meglio di così?

Proviamo ad esaminare l'*albero delle invocazioni* per $n = 6$:



12.2.2 Approccio basato su programmazione dinamica

La presenza di sotto-problemi ripetuti ci costringe quindi ad evitare una soluzione realizzata con un classico *divide-et-impera*, e ad utilizzare invece la *programmazione dinamica*, con conseguente costruzione di una *tabella delle soluzioni*. In particolare, tale tabella conterrà un elemento per ogni sotto-problema da risolvere.

Poiché per ogni $n \in \mathbb{N}$ il calcolo della soluzione si basa su i due valori precedenti di n , ci troviamo in una situazione in cui tutti i sotto-problemi devono essere risolti e quindi procediamo con la *programmazione dinamica* classica o bottom-up. Ciò significa che per costruire la *tabella delle soluzioni* partiamo dai casi base e risaliamo fino ad n .

Frammento 84 - Seconda implementazione della soluzione.

```
int domino2(int n)
    int[] DP = new int[0...n]                % Tabella delle soluzioni
    DP[0] = 1
    DP[1] = 1
    for (i = 2 to n) do
        DP[i] = DP[i - 1] + DP[i - 2]
    return DP[n]
```

NB. Sebbene finora abbiamo parlato soltanto di soluzioni ricorsive, nulla ci vieta di realizzarne di iterative, se queste sono estensionalmente equivalenti.

L'aumento in efficienza è evidente in quanto la *complessità* si è ridotta a $\Theta(n)$, tuttavia, l'utilizzo di una *tabella delle soluzioni* ci porta a considerare anche la *complessità* dal punto di vista spaziale, cioè quanto spazio di memoria è necessario per la sua memorizzazione. In questo caso memorizziamo i valori in un vettore di dimensione $n + 1$, quindi la *complessità* è $S(n) = \Theta(n)$.

Possiamo fare meglio di così?

Da un punto di vista temporale no, ma da quello spaziale sì. Se consideriamo bene la funzione ci rendiamo facilmente conto del fatto che ad ogni iterazione utilizziamo soltanto i valori delle due posizioni precedenti. Di conseguenza, potremmo eliminare il vettore e utilizzare delle semplici variabili.

Frammento 85 - Terza implementazione della soluzione.

```
int domino3(int n)
    int DP0 = 1
    int DP1 = 1
    int DP2 = 1
    for (i = 2 to n) do
        DP0 = DP1
        DP1 = DP2
        DP2 = DP1 + DP0
    return DP2
```

Ora, la *complessità spaziale* si è ridotta a $\Theta(1)$.

Criterio di costo logaritmico Le *complessità* calcolate finora sono tali fino a quando ci basiamo su un *criterio di costo uniforme* nel quale consideriamo soltanto il numero di elementi. Se utilizzassimo un *criterio di costo logaritmico*, dovremmo ragionare sul modo in cui crescono i numeri della serie di Fibonacci.

Se $F(n)$ è l' n -esimo numero della serie, vale la *formula di Binet*:

$$F(n) = \frac{\phi^n}{\sqrt{5}} - \frac{(1-\phi)^n}{\sqrt{5}} \quad \text{con} \quad \phi = \frac{1+\sqrt{5}}{2} = 1.6180339887\dots^1$$

NB. Vale la pena far notare che $F(n) = DP[n-1]$ perché DP vale per $n \geq 0$, mentre $F(n)$ per $n \geq 1$.

Il primo termine della formula è un'esponenziale con base un numero maggiore di 1, quindi cresce esponenzialmente, mentre il secondo tende a 0. Questo ci dice che i bit necessari a memorizzare l' n -esimo valore della serie crescono linearmente. Ovvero, per memorizzare l' n -esimo termine servono $\Theta(n)$ bit e sommare due valori consecutivi costa $\Theta(n)$.

Detto questo, possiamo ricalcolare tutte le *complessità* delle tre soluzioni sotto il *criterio di costo logaritmico*.

Soluzione	Complessità temporale	Complessità spaziale
domino1	$O(n2^n)$	$O(n^2)$
domino2	$O(n^2)$	$O(n^2)$
domino3	$O(n^2)$	$O(n)$

NB. In pratica, abbiamo semplicemente moltiplicato per n tutti i valori.

12.3 Problema di Hateville

Problema 6 - Hateville.

Hateville è un villaggio particolare composto da n case numerate e disposte linearmente lungo una singola strada. Ad Hateville ognuno odia i propri vicini: l'abitante i odia i suoi vicini $i-1$ e $i+1$ (se esistono).

Si vuole realizzare una sagra per la quale è necessario raccogliere dei fondi. Ogni abitante i è disposto a donare una quantità $D[i]$, ma solo se nessuno dei suoi vicini partecipa.

Scrivere un algoritmo che restituisca la quantità massima di fondi che può essere raccolta e un secondo algoritmo che restituisca il sottoinsieme di indici $S \subseteq \{1, \dots, n\}$ tale per cui la donazione totale $T = \sum_{i \in S} D[i]$ sia massimale.

Ad esempio, se il vettore delle donazioni è $D = [10, 5, 5, 10]$ la massima quantità di fondi che può essere raccolta è 20 e l'insieme S è $\{1, 4\}$.

12.3.1 Approccio basato su divide-et-impera

Iniziamo provando a ridefinire il problema. Se $HV(i)$ è uno dei possibili insiemi di indici che consente di ottenere una raccolta massimale dalle prime i case, la soluzione al problema è $HV(n)$.

Consideriamo ora il vicino i . Possiamo accettare o meno la sua donazione. Se non la accettiamo possiamo accettare quella del vicino $i-1$ e quindi $HV(i) = HV(i-1)$. Se invece accettiamo la donazione di i , dobbiamo rifiutare quella di $i-1$, perciò $HV(i) = HV(i-2) \cup \{i\}$.

¹ ϕ rappresenta il valore della *sezione aurea* ed è anche definito come $\lim_{n \rightarrow +\infty} \frac{F(n+1)}{F(n)}$, ovvero come il rapporto tra numeri consecutivi della serie di Fibonacci

Per decidere se accettare o meno la donazione di i è sufficiente vedere quale scelta permette di raccogliere più fondi. Se *highest* è una funzione che dati due insiemi seleziona quello che garantisce una donazione più alta, $HV(i)$ diventa:

$$HV(i) = \text{highest}(HV(i-1), HV(i-2) \cup \{i\})$$

Ciò che abbiamo detto finora è vero secondo quanto affermato dal seguente teorema:

Definizione 111 - Sottostruttura ottima.

Siano P_i il problema dato dalle prime i case e S_i una sua soluzione ottima. Valgono le seguenti:

1. Se $i \notin S_i$, allora $S_i = S_{i-1}$;
2. Se $i \in S_i$, allora $S_i = S_{i-2} \cup \{i\}$;

Dimostrazione. Procediamo dimostrando separatamente i due punti.

Caso 1: $i \notin S_i$ Secondo il teorema, $S_i = S_{i-1}$, cioè S_i è la soluzione ottima sia di P_i che di P_{i-1} . Se supponessimo, per assurdo, che non fosse così, dovrebbe esistere una soluzione S'_{i-1} per il problema P_{i-1} tale che $S'_{i-1} = \text{highest}(S'_{i-1}, S_i)$. Cioè, la soluzione S'_{i-1} dovrebbe garantire una donazione maggiore di quella di S_i . Ma, se fosse così, S'_{i-1} dovrebbe essere anche la soluzione di P_i , ma ciò sarebbe assurdo dato che abbiamo supposto che S_i fosse una soluzione ottima per P_i .

Caso 2: $i \in S_i$ Se $i \in S_i$, $i-1 \notin S_i$, altrimenti non sarebbe una soluzione ammissibile. Quindi, $S_i - \{i\}$ deve essere una soluzione ottima per P_{i-2} . Se supponessimo, per assurdo, che non fosse così, dovrebbe esistere una soluzione S'_{i-2} per il problema P_{i-2} tale che $S'_{i-2} = \text{highest}(S'_{i-2}, S_i - \{i\})$. Ma, se fosse così, $S'_{i-2} \cup \{i\}$ dovrebbe essere una soluzione di P_i migliore di S_i , generando un assurdo. \square

Dimostrate le nostre ipotesi, possiamo procedere con la definizione ricorsiva della soluzione. I casi base sono due:

- Se $i = 0$, $HV(0) = \emptyset$;
- Se $i = 1$, $HV(1) = \{1\}$;

La funzione che otteniamo è la seguente:

$$HV(i) = \begin{cases} \emptyset & i = 0 \\ \{1\} & i = 1 \\ \text{highest}(HV(i-1), HV(i-2) \cup \{i\}) & i \geq 2 \end{cases}$$

Osservando la definizione possiamo notare la similitudine con la funzione per il calcolo della *sequenza di Fibonacci*. Essendoci sotto-problemi ripetuti, possiamo scartare subito qualunque soluzione basata su *divide-et-impera*.

12.3.2 Approccio basato su programmazione dinamica

La *programmazione dinamica* ci richiede di costruire una *tabella delle soluzioni*. Proviamo a costruire tale tabella partendo da un esempio. Sia il vettore delle donazioni D definito come segue:

$$D = [10, 5, 5, 8, 4, 7, 12]$$

La *tabella delle soluzioni* è la seguente:

i	0	1	2	3	4	5	6	7
HV	\emptyset	$\{1\}$	$\{1\}$	$\{1, 3\}$	$\{1, 4\}$	$\{1, 3, 5\}$	$\{1, 4, 6\}$	$\{1, 3, 5, 7\}$

Ci sono un paio di problemi. Prima di tutto dobbiamo definire la funzione *highest*, ma il problema più grosso è la difficoltà di memorizzazione degli insiemi in una tabella.

Per risolvere entrambi i problemi in una volta potremmo definire il valore delle soluzioni come la quantità di fondi che vengono raccolti. Definiamo quindi una funzione DP tale per cui $DP[i]$ sia la massima quantità di fondi che si possono ottenere dalle prime i case di Hateville. Partendo da HV possiamo definire DP come segue:

$$DP[i] = \begin{cases} 0 & i = 0 \\ D[1] & i = 1 \\ \max(DP[i-1], DP[i-2] + D[i]) & i \geq 2 \end{cases}$$

Adesso implementare la soluzione è una banalità.

Frammento 86 - Implementazione della soluzione al primo problema.

```
int hateville(int[] D, int n)
    int[] DP = new int[0...n]                % Tabella delle soluzioni
    DP[0] = 0
    DP[1] = D[1]
    for (i = 2 to n) do
        DP[i] = max(DP[i - 1], DP[i - 2] + D[i])
    return DP[n]
```

La *complessità temporale* e *spaziale* della soluzione sono entrambe $\Theta(n)$. Possiamo notare che, come nel problema del domino, per ogni i la soluzione è ottenuta a partire dalle soluzioni dei due valori precedenti di i , quindi, potremmo eliminare l'array e introdurre delle variabili semplici riducendo a $\Theta(1)$ la *complessità spaziale*.

Frammento 87 - Implementazione migliorata della soluzione al primo problema.

```
int hateville(int[] D, int n)
    int DP0 = 0
    int DP1 = 1
    int DP2 = 1
    for (i = 2 to n) do
        DP2 = max(DP1, DP0 + D[i])
    return DP2
```

Ricostruire la soluzione A questo punto abbiamo risolto la prima richiesta del problema. Per quanto riguarda la seconda, dobbiamo riuscire a ricostruire l'insieme di indici a partire dalla *tabella delle soluzioni*. Grazie al teorema sulla *Sottostruttura ottima* possiamo fare le seguenti assunzioni:

- Se la casa i non è stata selezionata, $i \notin S_i$ e quindi $S_i = S_{i-1}$. Questo, riflesso sulla *tabella delle soluzioni*, garantisce che se $DP[i] = DP[i-1]$, i non appartiene all'insieme S di indici;

- Se la casa i è stata selezionata, $i \in S_i$ e quindi $S_i = S_{i-2} \cup \{i\}$. Questo, riflesso sulla *tabella delle soluzioni*, garantisce che se $DP[i] = DP[i-2] + D[i]$, i appartiene all'insieme S di indici;

NB. È possibile che risultino verificati entrambi i casi e, se succede, significa che esistono due soluzioni possibili.

Chiariti questi punti, possiamo ricostruire la soluzione fino a i in modo ricorsivo. Nello specifico, se $DP[i] = DP[i-1]$, si prende la soluzione valida per $i-1$, altrimenti si prende quella valida per $i-2$ e le si aggiunge i .

Frammento 88 - Implementazione della soluzione al secondo problema.

```
SET hateville(int[] D, int n)
    int[] DP = new int[0..n]                                % Tabella delle soluzioni
    DP[0] = 0
    DP[1] = D[1]
    for (i = 2 to n) do
        DP[i] = max(DP[i - 1], DP[i - 2] + D[i])
    return solution(DP, n)

% Funzione ausiliaria per la ricostruzione della soluzione
SET solution(int[] DP, int i)
    if (i == 0) then
        return {}
    if (i == 1) then
        return {1}
    if (DP[i] == DP[i - 1]) then
        return solution(DP, i - 1)
    else
        SET sol = solution(DP, i - 2)
        sol.insert(i)
        return sol
```

La *complessità temporale* di `solution` è $\Theta(n)$.

NB. È importante notare che non avremmo potuto ricostruire la soluzione senza l'intera *tabella delle soluzioni* e quindi non possiamo utilizzare la versione migliorata di `hateville`.

12.4 Problema dello zaino

Problema 7 - Problema dello zaino².

Dato un insieme di oggetti, ognuno caratterizzato da un peso e da un profitto, e uno “zaino” di capacità finita, individuare un sottoinsieme di oggetti tali per cui il loro peso totale sia al più uguale alla capacità dello zaino e il loro profitto totale sia massimale.

Questo problema ci chiede di realizzare un algoritmo che prenda in input un vettore w dei pesi, un vettore p dei profitti e la capacità C dello zaino e che restituisca un insieme $S \subseteq \{1, \dots, n\}$ tale che:

²Questo è uno dei problemi classici dell'informatica ed è noto con il nome di *Knapsack problem*

- il peso totale non superi la capacità, cioè $w(S) = \sum_{i \in S} w[i] \leq C$;
- il profitto totale sia massimale, cioè $\operatorname{argmax}_S p[S] = \sum_{i \in S} p[i]$;

Ad esempio, se $C = 12$, $w = [10, 4, 8]$ e $p = [20, 6, 12]$, l'algoritmo deve restituire l'insieme $S = \{1\}$.

Proviamo a dare un valore alla soluzione. Definiamo una funzione $DP[i][c]$ come il massimo profitto che può essere ottenuto dai primi $0 \leq i \leq n$ oggetti in uno zaino di capacità $c \leq C$. Il massimo profitto ottenibile nel problema originale è allora $DP[n][C]$.

Ora, consideriamo un generico oggetto i . Se non lo prendiamo, $DP[i][c] = DP[i-1][c]$ perché non cambiano né il profitto né la capacità. Se invece lo prendiamo, $DP[i][c] = DP[i-1][c - w[i]] + p[i]$ in quanto, il peso dell'oggetto i viene sottratto alla capacità rimanente e il profitto associato viene aggiunto a quello ottenuto con i precedenti $i-1$ oggetti.

La scelta migliore tra le due è quella che massimizza il profitto e la possiamo sintetizzare con la seguente:

$$DP[i][c] = \max \left(\overbrace{DP[i-1][c - w[i]] + p[i]}^{\text{Preso}}, \overbrace{DP[i-1][c]}^{\text{Non preso}} \right)$$

Per quanto riguarda i casi base, è facile dedurre che per $i = 0$ o $c = 0$, il profitto massimo possibile sia 0. C'è però un terzo caso: quando proviamo a prendere un oggetto, lo facciamo senza verificare che il peso non superi la capacità, quindi, potremmo ritrovarci ad avere un valore negativo per quest'ultima. Se questo succede, è ovvio che non avremmo potuto raccogliere quell'oggetto e, di conseguenza, decidiamo che in quel caso il profitto valga $-\infty$ in modo da assicurare che non venga selezionato dalla funzione max.

Detto questo, possiamo scrivere la formula nella sua forma completa:

$$DP[i][c] = \begin{cases} 0 & i = 0 \vee c = 0 \\ -\infty & c < 0 \\ \max(DP[i-1][c - w[i]] + p[i], DP[i-1][c]) & \text{altrimenti} \end{cases}$$

Equivalentemente, possiamo rimuovere il $-\infty$ e riscrivere la formula come segue:

$$DP[i][c] = \begin{cases} 0 & i = 0 \vee c = 0 \\ DP[i-1][c] & w[i] > c \\ \max(DP[i-1][c - w[i]] + p[i], DP[i-1][c]) & w[i] \leq c \end{cases}$$

12.4.1 Approccio basato su programmazione dinamica

Frammento 89 - Implementazione iterativa della soluzione.

```
int knapsack(int[] w, int[] p, int n, int C)
    int[][] DP = new int[0..n][0..C]                                % Tabella delle soluzioni
    for (i = 0 to n) do
        DP[i][0] = 0
    for (c = 0 to C) do
        DP[0][c] = 0
    for (i = 1 to n) do
        for (c = 1 to C) do
            if (w[i] ≤ c) then
                DP[i][c] = max(DP[i-1][c - w[i]] + p[i], DP[i-1][c])
```

```

else
    DP[i][c] = DP[i - 1][c]
return DP[n][C]

```

Esempio 29 - Esempio di esecuzione.

Dato uno zaino di capacità $C = 9$ e i seguenti vettori dei pesi e dei profitti:

$$w = [4, 2, 3, 4]$$

$$p = [10, 7, 8, 6]$$

al termine dell'esecuzione dell'algoritmo di cui sopra, la matrice DP corrisponde alla seguente tabella:

	c									
i	0	1	2	3	4	5	6	7	8	9
0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	10	10	10	10	10	10
2	0	0	7	7	10	10	17	17	17	17
3	0	0	7	8	10	15	17	18	18	25
4	0	0	7	8	10	15	17	18	18	25

Complessità Siccome l'algoritmo costruisce e popola una tabella di dimensione $n \cdot C$, la *complessità* è $T(n) = \Theta(nC)$. Ovviamente lo stesso vale anche per la *complessità spaziale*. Quindi, il costo dell'algoritmo dipende da due variabili, ma poiché C non rappresenta la dimensione dell'input e viene rappresentato mediante $k = \lceil \log C \rceil$ bit, la *complessità temporale* può essere riscritta come:

$$T(n) = \Theta(n2^k)$$

Ciò, rende questa implementazione dell'algoritmo *pseudo-polinomiale*³.

È possibile fare meglio di così?

Possiamo provare a realizzare una versione ricorsiva dell'algoritmo.

Frammento 90 - Implementazione ricorsiva della soluzione.

```

int knapsack(int[] w, int[] p, int n, C)
    return knapsackRec(w, p, n, C)

int knapsackRec(int[] w, int[] p, int n, c)
    if (c ≤ 0) then
        return -∞
    if (i == 0 or c == 0) then
        return 0
    int notTaken = knapsackRec(w, p, i - 1, c)
    int taken = knapsackRec(w, p, i - 1, c - w[i]) + p[i]
    return max(notTaken, taken)

```

³Parleremo meglio di cosa significhi più avanti nella trattazione

In questa implementazione abbiamo semplicemente tradotto in codice la prima forma della formula per DP e l'equazione di ricorrenza è:

$$T(n) = \begin{cases} 1 & n \leq 1 \\ 2T(n-1) + 1 & n > 1 \end{cases}$$

Per il *Teorema delle ricorrenze lineari di ordine costante* la forma chiusa associata è $T(n) = \Theta(2^n)$ e quindi la *complessità* è addirittura più alta di quella della soluzione iterativa.

Purtroppo infatti, secondo l'opinione della maggior parte degli informatici, non esistono soluzioni meno complesse di quelle che abbiamo esaminato. Ciò che possiamo comunque provare a fare, è realizzare delle implementazioni che, a parità di *complessità*, migliorino quelle già esistenti.

12.4.2 Approccio basato su memoization

Come già accennato, la *memoization* è una tecnica risolutiva che unisce l'utilizzo di una *tabella delle soluzioni* all'approccio top-down tipico del *divide-et-impera* classico.

In particolare, la *tabella delle soluzioni* viene inizializzata ad un valore arbitrario che usiamo per indicare che un certo sotto-problema non è ancora stato risolto. Quindi, quando è necessario risolvere un sotto-problema, se la tabella contiene già la soluzione, la si usa direttamente, altrimenti la si calcola. In questo modo, ogni sotto-problema viene risolto una sola volta.

Nel problema che stiamo affrontando, usiamo -1 come valore speciale e modifichiamo la soluzione ricorsiva in modo da riutilizzare le soluzioni.

Frammento 91 - Implementazione basata su memoization.

```
int knapsack(int[] w, int[] p, int n, int C)
    int[][] DP = new int[1...n][1...C]                % Tabella delle soluzioni
    for (i = 1 to n) do
        for (c = 1 to n) do
            DP[i][c] = -1
    return knapsackRec(w, p, n, C, DP)

int knapsackRec(int[] w, int[] p, int i, int c, int[][] DP)
    if (c < 0) then
        return -∞
    if (i == 0 or c == 0) then return 0
    if (DP[i][c] < 0) then
        int notTaken = knapsackRec(w, p, i - 1, c, DP)
        int taken = knapsackRec(w, p, i - 1, c - w[i], DP) + p[i]
        DP[i][c] = max(notTaken, taken)
    return DP[i][c]
```

Complessità La *complessità* rimane $T(n) = \Theta(nC)$ perché è il costo che paghiamo per inizializzare DP . Tuttavia, siamo comunque riusciti a ridurre le chiamate ricorsive che vengono generate. Addirittura, sostituendo la tabella con un *dizionario* implementato come *hash table* ci liberiamo anche della necessità fare l'inizializzazione e quindi il costo diventa $T(n) = \min(2^n, nC)$.

A conferma di quanto affermato circa la riduzione delle chiamate ricorsive, osserviamo un esempio di esecuzione sugli stessi dati dell'esempio precedente.

Esempio 30 - Esempio di esecuzione.

Dato uno zaino di capacità $C = 9$ e i seguenti vettori dei pesi e dei profitti:

$$w = [4, 2, 3, 4]$$

$$p = [10, 7, 8, 6]$$

al termine dell'esecuzione dell'algoritmo di cui sopra, la matrice DP corrisponde alla seguente tabella:

	c									
i	0	1	2	3	4	5	6	7	8	9
0										
1		-1	0	0	10	10	10	10	-1	10
2		-1	7	-1	-1	10	17	-1	-1	17
3		-1	-1	-1	-1	15	-1	-1	-1	25
4		-1	-1	-1	-1	-1	-1	-1	-1	25

Le chiamate ricorsive sono state soltanto 14, invece delle 36 che avevamo nella prima soluzione.

Sebbene il problema originale non lo richieda, può essere interessante provare a implementare l'algoritmo per la ricostruzione della soluzione a partire da DP .

Il principio è molto semplice. Considerando $DP[i][c]$, se $DP[i][c] \neq DP[i-1][c]$, l'elemento i è stato raccolto, quindi l'algoritmo continua ricorsivamente da $DP[i-1][c-w[i]]$. Se invece l'elemento i non è stato raccolto, l'algoritmo continua da $DP[i-1][c]$.

Frammento 92 - Implementazione algoritmo per la ricostruzione della soluzione.

```
SET solution(int[] w, int i, int c, int[][] DP)
    if (i == 0 or c == 0) then                                % Non si può raccogliere nulla
        return {}
    if (i == 1 and DP[i][c] > 0) then
        return {i}
    if (DP[i][c] == DP[i-1][c]) then                          % L'elemento i non è stato raccolto
        return solution(w, i - 1, c, DP)
    SET S = solution(w, i - 1, c - w[i], DP)
    S.insert(i)
    return S
```

NB. Il controllo $i == 1$ è necessario solo se per popolare DP è stato usato l'algoritmo realizzato mediante *memoization* perché, in quel caso, $DP[0][c]$ non esiste.

Per ricostruire l'intera soluzione è sufficiente invocare `solution(w, n, C, DP)` dove DP è stato precedentemente popolato.

12.5 Problema dello zaino senza limiti

Vediamo ora una variante al problema appena affrontato per studiarne possibili semplificazioni.

Problema 8 - Problema dello zaino senza limiti.

Dato un insieme di oggetti, ognuno caratterizzato da un peso e da un profitto, e uno “zaino” di capacità finita, individuare un sottoinsieme di oggetti tali per cui il loro peso totale sia al più uguale alla capacità dello zaino e il loro profitto totale sia massimale, ma senza porre limiti al numero di volte che un oggetto può essere selezionato.

La funzione che descrive il valore della soluzione può essere ottenuta modificando quella usata per il problema originale. In particolare, quando un oggetto viene selezionato, non è più necessario decrementare il valore dell'indice i , perché lo stesso potrebbe essere raccolto di nuovo. Applicando la modifica, otteniamo la seguente:

$$DP[i][c] = \begin{cases} 0 & i = 0 \vee c = 0 \\ -\infty & c < 0 \\ \max(DP[i][c - w[i]] + p[i], DP[i - 1][c]) & \text{altrimenti} \end{cases}$$

Possiamo semplificare di più?

La risposta è sì, infatti, se un oggetto può essere raccolto più volte, non è più necessario mantenere il valore i nella *tabella delle soluzioni*. Questo ci consente di ridefinire il valore della soluzione.

Nello specifico, dato uno zaino senza limiti di scelta, di capacità C e n oggetti caratterizzati da un peso w e un profitto p , definiamo $DP[c]$ il massimo profitto che può essere ottenuto da tali oggetti in uno zaino di capacità $c \leq C$.

Questa nuova descrizione ci porta a scrivere quanto segue:

$$DP[c] = \begin{cases} 0 & c = 0 \\ \max_{w[i] \leq c} \{DP[c - w[i]] + p[i]\} & c > 0 \end{cases}$$

Per la traduzione in codice utilizziamo la tecnica della *memoization*.

Frammento 93 - Implementazione della soluzione.

```
int knapsack(int[] w, int[] p, int n, int C)
    int[] DP = new int[0...C]                                % Tabella delle soluzioni
    for (c = 0 to C) do
        DP[c] = -1
    return knapsackRec(w, p, n, C, DP)

int knapsackRec(int[] w, int[] p, int n, int c, int[] DP)
    if (c == 0) then
        return 0
    if (DP[c] < 0) then                                       % Se è falso, la soluzione è già stata calcolata
        int maxSoFar = 0
        for (i = 1 to n) do
            if (w[i] ≤ c) then
                int value = knapsackRec(w, p, n, c - w[i], DP) + p[i]
                maxSoFar = max(maxSoFar, value)
        DP[c] = maxSoFar
    return DP[c]
```


Complessità Nel caso pessimo vengono riempite tutte le celle della tabella DP e, poiché per riempire una cella devono essere provati tutti gli n elementi, per ognuna si paga $\Theta(n)$. La *complessità temporale* totale è $T(n) = O(nC)$, quella *spaziale* è invece $S(n) = \Theta(C)$.

Rispetto al problema precedente abbiamo ridotto lo spazio utilizzato, ma la ricostruzione della soluzione a partire da DP è più difficile perché per capire da quale elemento deriva il profitto massimo dovremmo ispezionarli tutti. Per tanto, conviene memorizzare tale informazione nel momento stesso in cui costruiamo DP .

Frammento 94 - Implementazione con ricostruzione della soluzione.

```

LIST knapsack(int[] w, int[] p, int n, int C)
    int[] DP = new int[0...C]                                % Tabella delle soluzioni
    int[] pos = new int[0...C]    % Elemento raccolto per ogni valore di capacità
    for (c = 0 to C) do
        DP[c] = -1
        pos[c] = -1
    knapsackRec(w, p, n, C, DP, pos)
    return solution(w, C, pos)

int knapsackRec(int[] w, int[] p, int n, int c, int[] DP, int[] pos)
    if (c == 0) then
        return 0
    if (DP[c] < 0) then    % Se è falso, la soluzione è già stata calcolata
        DP[c] = 0
        for (i = 1 to n) do
            if (w[i] ≤ c) then
                int value = knapsackRec(w, p, n, c - w[i], DP, pos) + p[i]
                if (value > DP[c]) then
                    DP[c] = value
                    pos[c] = i
    return DP[c]

LIST solution(int[] w, int c, int[] pos)
    if (c == 0 or pos[c] < 0) then
        return List()
    LIST L = solution(w, c - w[pos[c]], pos)    % Lista degli indici selezionati
    L.insert(L.tail(), pos[c])
    return L

```

12.6 Ricerca della sottosequenza comune massimale

Definizione 112 - Sottosequenza.

Una sequenza P è una sottosequenza di T se P è ottenuto da T rimuovendo uno o più dei suoi elementi.

Alternativamente, possiamo definire P anche come l'insieme degli indici $\{1, \dots, n\}$ degli elementi di T che compaiono anche in P . I rimanenti elementi sono poi elencati nello stesso ordine, senza essere necessariamente contigui.

Ad'esempio, la stringa $P = \text{"AAATA"}$ è *sottosequenza* di $T = \text{"AAAAATATGA"}$ perché in T compaiono almeno tre **A**, seguite da almeno una **T**, seguita a sua volta da almeno un'altra **A**.

NB. La sequenza vuota $P = \emptyset$ è *sottosequenza* di qualunque altra stringa.

Definizione 113 - Sottosequenza comune.

Una sequenza X è *sottosequenza comune* (*Common Subsequence*) di due sequenze T ed U se è una sottosequenza di entrambe e scriviamo in simboli:

$$X \in \mathcal{CS}(T, U)$$

Definizione 114 - Sottosequenza comune massimale.

Una sequenza $X \in \mathcal{CS}(T, U)$ è una *sottosequenza comune massimale* (*Longest Common Subsequence*) di due sequenze T ed U se non esiste una sequenza comune $Y \in \mathcal{CS}(T, U)$ che sia più lunga di X ovvero tale per cui $|Y| > |X|$. Se è così, scriviamo in simboli:

$$X \in \mathcal{LCS}(T, U)$$

Date queste definizioni preliminari possiamo venire alla consegna del problema in questione:

Problema 9 - Ricerca della sottosequenza comune massimale.

Date due sequenze T ed U , trovare la più lunga sottosequenza comune di T ed U .

Ad esempio, se $T = \text{"AAAATTGA"}$ e $U = \text{"TAACGATA"}$, l'algoritmo deve restituire **AAATA**.

Ovviamente la prima soluzione che potrebbe venirci in mente sarebbe quella di ricercare tutte le *sottosequenze* e sceglierne una tra le più lunghe. Questo approccio ci porterebbe però ad una funzione di *complessità* $T(n) = \Theta(2^n(n + m))$ perché ricercare tutte le *sottosequenze* di una stringa T lunga n costa $\Theta(2^n)$ e verificare se una sequenza è *sottosequenza* di un'altra costa $O(m + n)$.

12.6.1 Approccio basato su programmazione dinamica

Per fare meglio di così possiamo provare a sfruttare la *programmazione dinamica* e il meccanismo della ricorsione. In particolare, definiamo una funzione che ci restituisca il *prefisso* di una sequenza:

Definizione 115 - Prefisso.

Data una sequenza T composta da n caratteri t_1, \dots, t_n , chiamiamo $T(i)$ la funzione che denota il prefisso di T dato dai primi i caratteri, ovvero tale che:

$$T(i) = t_1, \dots, t_i$$

NB. Ovviamente vale l'identità $T(0) = \emptyset$.

A questo punto, date le due sequenze T ed U lunghe rispettivamente n ed m caratteri, vogliamo scrivere la definizione di una formula ricorsiva $\mathcal{LCS}(T(i), U(j))$ che restituisca la *sottosequenza comune massimale* dei prefissi $T(i)$ e $U(j)$.

Il caso base è banale, perché se $i = 0$ o $j = 0$, l'unica *sottosequenza comune massimale* è la *sottosequenza* vuota \emptyset .

Per quanto riguarda i casi ricorsivi possiamo distinguerne due:

Caso 1: $t_i = u_j$ In questo caso, gli ultimi caratteri dei prefissi $T(i)$, $U(j)$ coincidono, questo significa che il carattere $t_i = u_j$ è comune ad entrambe le sequenze e quindi farà parte della *sottosequenza comune massimale*. La ricerca di tale *sottosequenza* può quindi continuare dai prefissi $T(i-1)$, $U(j-1)$.

Caso 2: $t_i \neq u_j$ In questo caso, gli ultimi caratteri non coincidono e quindi la soluzione è provare a cercare la *sottosequenza comune massimale* sia nella coppia di prefissi $T(i-1)$, $U(j)$ che nella coppia $T(i)$, $U(j-1)$.

Tutto ciò che abbiamo detto si traduce nella seguente formulazione:

$$\mathcal{LCS}(T(i), U(j)) = \begin{cases} \emptyset & i = 0 \vee j = 0 \\ \mathcal{LCS}(T(i-1), U(j-1)) \oplus t_i & i > 0 \wedge j > 0 \wedge t_i = u_j \\ \text{longest} \left(\begin{array}{l} \mathcal{LCS}(T(i-1), U(j)), \\ \mathcal{LCS}(T(i), U(j-1)) \end{array} \right) & i > 0 \wedge j > 0 \wedge t_i \neq u_j \end{cases}$$

NB. Il simbolo \oplus è l'operatore di concatenazione.

La dimostrazione della correttezza della formula si basa sul *Teorema della sottostruttura ottima*.

Definizione 116 - Sottostruttura ottima.

Date due sequenze $T = (t_1, \dots, t_n)$ e $U = (u_1, \dots, u_m)$, se $X \in \mathcal{LCS}(T(n), U(m))$ valgono i seguenti tre casi:

1. Se $t_n = u_m$, allora $x_k = t_n = u_m$ e $X(k-1) \in \mathcal{LCS}(T(n-1), U(m-1))$;
2. Se $t_n \neq u_m$ e $x_k \neq t_n$, allora $X \in \mathcal{LCS}(T(n-1), U(m))$;
3. Se $t_n \neq u_m$ e $x_k \neq u_m$, allora $X \in \mathcal{LCS}(T(n), U(m-1))$;

Dimostrazione. Procediamo dimostrando separatamente i tre punti.

Caso 1: $t_n = u_m$ Se per assurdo supponessimo che $x_k \neq t_n = u_m$, allora esisterebbe una *sottosequenza* $Y = X \oplus t_n$. Se così fosse però, varrebbero $Y \in \mathcal{CS}(T(n), U(m))$ e $|Y| > |X|$ generando un assurdo.

Venendo alla seconda parte, se supponessimo per assurdo che $X(k-1) \notin \mathcal{LCS}(T(n-1), U(m-1))$, allora esisterebbe un $Y \in \mathcal{LCS}(T(n-1), U(m-1))$ tale che $|Y| > |X(k-1)|$. Quindi, $Y \oplus t_n \in \mathcal{CS}(T(n), U(m))$ e $|Y| > |X(k-1) \oplus t_n| = |X|$ generando un assurdo.

Caso 2: $t_n \neq u_m$ e $x_k \neq t_n$ Se ipotizzassimo per assurdo che $X \notin \mathcal{LCS}(T(n-1), U(m))$, allora esisterebbe un $Y \in \mathcal{LCS}(T(n-1), U(m))$ tale che $|Y| > |X|$. Conseguentemente, sarebbe anche vero affermare che $Y \in \mathcal{LCS}(T(n), U(m))$ e quindi anche che $X \notin \mathcal{LCS}(T(n), U(m))$, generando un assurdo.

Caso 3: $t_n \neq u_m$ e $x_k \neq u_m$ Se ipotizzassimo per assurdo che $X \notin \mathcal{LCS}(T(n), U(m-1))$, allora esisterebbe un $Y \in \mathcal{LCS}(T(n), U(m-1))$ tale che $|Y| > |X|$. Conseguentemente, sarebbe anche vero affermare che $Y \in \mathcal{LCS}(T(n), U(m))$ e quindi anche che $X \notin \mathcal{LCS}(T(n), U(m))$, generando un assurdo. \square

Conclusa in questo modo la dimostrazione, possiamo passare alla scrittura della ricorrenza per il valore della soluzione. Trattandosi di sequenze di caratteri, il loro valore è dato dalla loro lunghezza. Vale quindi la seguente:

$$DP[i][j] = \begin{cases} 0 & i = 0 \vee j = 0 \\ DP[i-1][j-1] + 1 & i > j \wedge j > 0 \wedge t_i = u_j \\ \max(DP[i-1][j], DP[i][j-1]) & i > 0 \wedge j > 0 \wedge t_i \neq u_j \end{cases}$$

Chiaramente la soluzione al problema originale è $DP[n][m]$.

Frammento 95 - Implementazione della soluzione.

```
int lcs(ITEM[] T, ITEM[] U, int n, int m)
    int[][] DP = new int[0...n][0...m]                % Tabella delle soluzioni
    for (i = 0 to n) do
        DP[i][0] = 0
    for (j = 0 to m) do
        DP[0][j] = 0
    for (i = 1 to n) do
        for (j = 1 to m) do
            if (T[i] == U[j]) then
                DP[i][j] = DP[i - 1][j - 1] + 1
            else
                DP[i][j] = max(DP[i - 1][j], DP[i][j - 1])
    return DP[n][m]
```

Il problema non lo richiede, ma se volessimo ricostruire la soluzione, l'algoritmo sarebbe il seguente.

Frammento 96 - Ricostruzione della soluzione.

```
LIST solution(ITEM[] T, ITEM[] U, int i, int j, int[][] DP)
    if (i == 0 or j == 0) then
        return List()
    if (T[i] = U[i]) then
        LIST S = subsequence(T, U, i - 1, j - 1, DP)
        S.insert(S.tail(), T[i])
        return S
    if (DP[i - 1][j] > DP[i][j - 1]) then
        return subsequence(T, U, i - 1, j, DP)
    else
        return subsequence(T, U, i, j - 1, DP)
```

Complessità Per quanto riguarda la `subsequence`, siccome DP è una matrice $n \times m$ e per ricostruire la soluzione analizziamo la matrice partendo dalla posizione (n, m) e “salendo” verso $(0, 0)$, i possibili movimenti sono tre:

1. Da $DP[i][j]$ a $DP[i-1][j-1]$ se $T[i] = U[j]$;
2. Da $DP[i][j]$ a $DP[i-1][j]$ se $DP[i-1][j] > DP[i][j-1]$;

3. Da $DP[i][j]$ a $DP[i][j-1]$ se $DP[i-1][j] < DP[i][j-1]$;

Di conseguenza, nella situazione peggiore concludiamo l'esecuzione nella posizione $(0,0)$ arrivandoci non per la diagonale. Il costo in quella situazione è $O(n+m)$ perché vengono inseriti nella lista $n+m$ caratteri.

La *complessità* della lcs è $O(nm)$ perché dobbiamo riempire tutta la matrice DP che ha dimensione $n \times m$.

Se non siamo interessati a ricostruire la soluzione possiamo migliorare la *complessità spaziale* della lcs mantenendo in memoria soltanto due righe. Questo perché, quando calcoliamo il valore di $DP[i][j]$ utilizziamo $DP[i-1][j-1]$ oppure $DP[i-1][j]$ e $DP[i][j-1]$. Questi tre valori risiedono sempre su due sole righe, quindi le precedenti non sono necessarie.

Fatta questa riflessione, l'implementazione che ne risulta è la seguente:

Frammento 97 - Implementazione migliorata della soluzione.

```
int lcs(ITEM[] T, ITEM[] U, int n, int m)
    int[] DP = new int[0...m]                                % Riga corrente
    int[] DP' = new int[0...m]                               % Riga precedente
    for (j = 0 to m) do
        DP[j] = 0
    for (i = 1 to n) do
        DP  $\leftrightarrow$  DP'                                     % Swap delle righe
        DP[0] = 0
        for (j = 1 to m) do
            if (T[i] == U[j]) then
                DP[j] = DP'[j - 1] + 1                       % Pari a DP[i-1][j-1] + 1
            else
                DP[j] = max(DP'[j], DP[j - 1]) % Pari a max(DP[i-1][j], DP[i][j-1])
    return DP[m]
```

In questo modo la *complessità spaziale* si riduce a $O(\min(n, m))$ perché la dimensione di DP e DP' può essere scelta in base al minore tra n e m .

NB. Questo tipo di algoritmi sono molto usati nello studio del DNA e nella ricerca delle similitudini tra diverse sequenze di DNA.

12.7 Problema dello string matching approssimato

Definizione 117 - Occorrenza k -approssimata.

Siano:

- $P = p_1, \dots, p_m$ una stringa detta *pattern*;
- $T = t_1, \dots, t_n$ una stringa detta *testo*;

Un'occorrenza k -approssimata di P in T è una copia di P in T in cui sono ammessi k “errori” tra P e T del seguente tipo:

1. Sostituzione: I corrispondenti caratteri in P e T sono diversi;
2. Inserimento: un carattere di P non è incluso in T ;

3. Cancellazione: un carattere di T non è incluso in P ;

Ad esempio se $T = \text{"questo è un o s c e m p i o"}$ e $P = \text{"un e s e m p i o"}$, è possibile trovare un'occorrenza *approssimata* di P in T “correggendo due errori”, ovvero cancellando la c di $s c e m p i o$ e sostituendo la o di $u n o$ con una e .

Detto questo, possiamo introdurre il problema di questa sezione.

Problema 10 - String matching approssimato.

Date due stringhe P e T , trovare un'occorrenza k -approssimata di P in T tale che $0 \leq k \leq m$ sia minimo.

Questo problema ci chiede quindi di trovare il minimo valore di k per cui è possibile trovare un'occorrenza *approssimata* di P in T e, estendendolo, potremmo anche essere interessati a sapere dove si trovino gli errori e quali siano.

Come al solito, iniziamo definendo il valore della soluzione. Sia $DP[0 \dots m][0 \dots n]$ una tabella in cui $DP[i][j]$ contiene il valore minimo di k per cui esiste un'occorrenza k -approssimata di $P(i)$ in $T(j)$ che termina nella posizione j .

NB. La semantica delle funzioni $P(i)$ e $T(i)$ è la stessa vista nel problema della *sottosequenza comune massima*.

A questo punto, per ogni coppia di valori i e j diversi da zero, per $DP[i][j]$ valgono 4 possibili casistiche:

1. Se $P[i] = T[j]$ non c'è nessun errore e quindi $DP[i][j] = DP[i-1][j-1]$;
2. Se $P[i] \neq T[j]$ c'è un errore e le possibili correzioni sono tre:
 - (a) *Sostituzione*: $DP[i][j] = DP[i-1][j-1] + 1$;
 - (b) *Inserimento*: $DP[i][j] = DP[i-1][j] + 1$;
 - (c) *Cancellazione*: $DP[i][j] = DP[i][j-1] + 1$;

Quindi, per ogni coppia (i, j) , viene scelta la possibilità che restituisce il valore minore. Ciò è diretta conseguenza del fatto che anche se $P[i] = T[j]$ potrebbe essere più conveniente cercare il pattern nei caratteri precedenti del testo. La formulazione completa è dunque la seguente:

$$DP[i][j] = \begin{cases} 0 & i = 0 \\ i & j = 0 \\ \min \begin{pmatrix} DP[i-1][j-1] + \delta \\ DP[i-1][j] \\ DP[i][j-1] \end{pmatrix} & i > 0 \wedge j > 0 \wedge \delta = \begin{cases} 0 & P[i] = T[j] \\ 1 & P[i] \neq T[j] \end{cases} \end{cases}$$

Diversamente dagli esempi precedenti, la soluzione al problema non si trova necessariamente in $DP[m][n]$, in quanto, in quella posizione si trova il numero minimo di correzioni che devono essere apportate per ottenere un'occorrenza k -approssimata del pattern che termina nella posizione n del testo. Come abbiamo detto, il pattern potrebbe trovarsi più convenientemente cercandolo in altre porzioni del testo, quindi la soluzione è il minimo tra i valori $DP[m][j]$ con $0 \leq j \leq n$.

Frammento 98 - Implementazione della soluzione.

```
int stringMatching(ITEM[] P, ITEM[] T, int n, int m)
    int[][] DP = new int[0...m][0...n]                % Tabella delle soluzioni
    for (j = 0 to n) do
        DP[0][j] = 0                                    % Caso i = 0
    for (i = 0 to m) do
        DP[i][0] = i                                    % Caso j = 0
    for (i = 1 to m) do
        for (j = 1 to n) do
            DP[i][j] = min(
                DP[i - 1][j - 1] + iif(P[i] == T[j], 0, 1),
                DP[i - 1][j] + 1,
                DP[i][j - 1] + 1
            )
    int pos = 0                                          % Indice per cui DP[m][pos] è minimo
    for (j = 1 to n) do
        if (DP[m][j] < DP[m][pos]) then
            pos = j
    return DP[m][pos]
```

12.8 Problema del prodotto a catena di matrici

Problema 11 - Prodotto a catena di matrici.

Data una sequenza di matrici A_1, \dots, A_n a due a due compatibili per il prodotto matriciale, calcolarne il prodotto minimizzando il numero di prodotti scalari necessari.

NB. Il prodotto matriciale è associativo, ma non commutativo.

Ad esempio, date tre matrici A, B, C di dimensioni che sono rispettivamente 100×1 , 1×100 e 100×1 , il prodotto $A \cdot B \cdot C$ può essere calcolato come $(A \cdot B) \cdot C$ o $A \cdot (B \cdot C)$. Nel primo caso vengono calcolati $100 \cdot 1 \cdot 100 + 100 \cdot 100 \cdot 1 = 10000 + 10000 = 20000$ prodotto scalari. Nel secondo caso invece, i prodotti sono $1 \cdot 100 \cdot 1 + 100 \cdot 1 \cdot 1 = 100 + 100 = 200$.

Siamo riusciti a ridurre di un fattore 100 il numero di operazioni semplicemente cambiando la posizione delle parentesi, cioè modificando la *parentesizzazione*.

Definizione 118 - Parentesizzazione.

Una parentesizzazione $P_{i,j}$ del prodotto $A_i \cdot \dots \cdot A_j$ consiste nella matrice A_i se $i = j$ e nel prodotto di due parentesizzazioni $(P_{i,k} \cdot P_{k+1,j})$ altrimenti.

Definizione 119 - Parentesizzazione ottima.

La parentesizzazione che minimizza il numero di prodotti scalari è detta essere ottima.

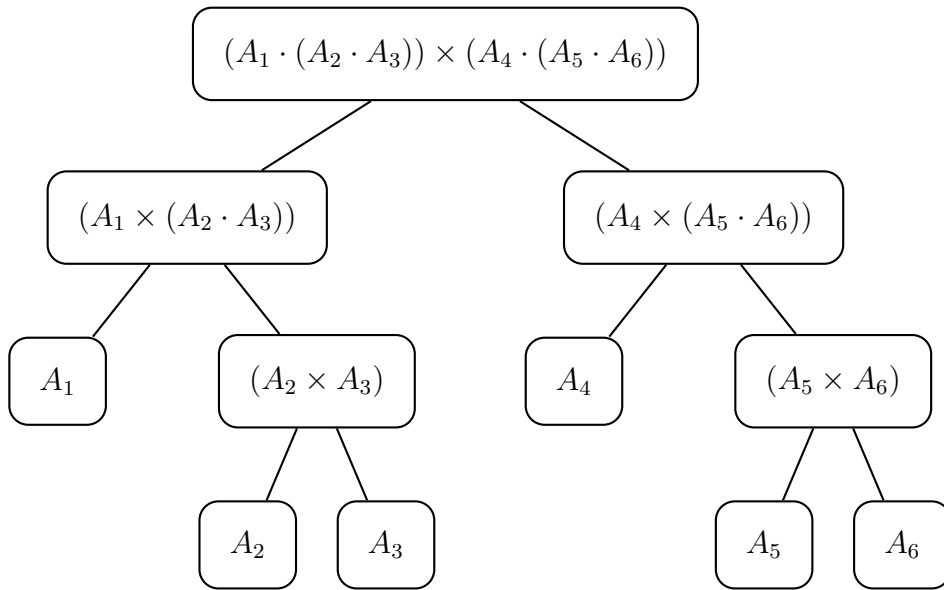
Esempio 31 - Possibile parentesizzazione.

Data la catena di prodotti:

$$A_1 \cdot A_2 \cdot A_3 \cdot A_4 \cdot A_5 \cdot A_6$$

una possibile parentesizzazione per $k = 3$ è la seguente:

$$(A_1 \cdot (A_2 \cdot A_3)) \times (A_4 \cdot (A_5 \cdot A_6))$$



NB. Il simbolo \times identifica il *prodotto finale* o *ultimo prodotto*.

Quante sono le *parentesizzazioni* possibili per una generica catena di n matrici?

Se $P(n)$ indica il numero di *parentesizzazioni* possibili per n matrici $A_1 \cdot \dots \cdot A_n$, l'*ultimo prodotto* può occorrere in $n - 1$ posizioni. Quindi, fissato l'indice k dell'*ultimo prodotto*, otteniamo $P(k)$ *parentesizzazioni* per $A_1 \cdot \dots \cdot A_k$ e altre $P(n - k)$ *parentesizzazioni* per $A_{k+1} \cdot \dots \cdot A_n$.

Ciò si traduce nella seguente formula:

$$P(n) = \begin{cases} 1 & n = 1 \\ \sum_{k=1}^{n-1} P(k) \cdot P(n - k) & n > 1 \end{cases}$$

Come esempio, per valori di n fino a 10, $P(n)$ vale:

n	1	2	3	4	5	6	7	8	9	10
$P(n)$	1	1	2	5	14	42	132	429	1430	4862

NB. I valori di $P(n)$ crescono seguendo la sequenza dei numeri di Catalan, che possono essere espressi anche con la seguente formulazione:

$$P(n) = C(n) = \frac{1}{n+1} \binom{2n}{n} = \frac{(2n)!}{(n+1)!n!} = \Theta\left(\frac{4^n}{n\sqrt{n}}\right)$$

È possibile dimostrare che tale sequenza cresce come $\Omega(2^n)$ e quindi ne deduciamo che algoritmi a forza bruta non possono essere utilizzati.

Prima di procedere oltre, fissiamo la notazione utilizzata.

Notazione	Significato
$A_1 \cdot \dots \cdot A_n$	Il prodotto di n matrici da ottimizzare
c_{i-1}	Il numero di righe della matrice A_i
c_i	Il numero di colonne della matrice A_i
$A[i \dots j]$	Il sottoprodotto di $A_i \cdot \dots \cdot A_j$
$P[i \dots j]$	Una <i>parentesizzazione</i> per $A[i \dots j]$

NB. Il numero di righe di una matrice coincide sempre con il numero di colonne della matrice che la precede nella catena e questo è dovuto alle regole di compatibilità del prodotto matriciale.

NB. $P[i \dots j]$ non indica necessariamente la *parentesizzazione* migliore.

Possiamo osservare che data una sottosequenza $A[i \dots j]$, se ne consideriamo una *parentesizzazione ottima*, esiste un *ultimo prodotto*, ovvero esiste un indice k tale che:

$$P[i \dots j] = P[i \dots k] \cdot P[k + 1 \dots j]$$

Quali sono le caratteristiche delle sue *sotto-parentesizzazioni*?

Definizione 120 - Teorema di sottostruttura ottima.

Se $P[i \dots j] = P[i \dots k] \cdot P[k + 1 \dots j]$ è una *parentesizzazione ottima* del prodotto $A[i \dots j]$, allora $P[i \dots k]$ e $P[k + 1 \dots j]$ sono rispettivamente le *parentesizzazioni ottime* dei prodotti $A[i \dots k]$ e $A[k + 1 \dots j]$.

Dimostrazione. Procediamo con una dimostrazione per assurdo considerando separatamente le due *sotto-parentesizzazioni*:

- Se supponessimo esistesse una *parentesizzazione ottima* $P'[i \dots k]$ di $A[i \dots k]$ con costo inferiore a $P[i \dots k]$, allora $P'[i \dots k] \cdot P[k + 1 \dots j]$ sarebbe una *parentesizzazione ottima* di $A[i \dots j]$ con costo inferiore a $P[i \dots j]$, generando un assurdo;
- Se supponessimo esistesse una *parentesizzazione ottima* $P'[k + 1 \dots j]$ di $A[k + 1 \dots j]$ con costo inferiore a $P[k + 1 \dots j]$, allora $P[i \dots k] \cdot P'[k + 1 \dots j]$ sarebbe una *parentesizzazione ottima* di $A[i \dots j]$ con costo inferiore a $P[i \dots j]$, generando un assurdo;

□

A questo punto, possiamo passare a ragionare sul valore della soluzione ottima. Definiamo una matrice DP in cui $DP[i][j]$ indica il numero minimo di prodotti scalari necessari per calcolare il prodotto $A[i \dots j]$.

Sicuramente, se $i = j$, non è necessario eseguire alcun prodotto e quindi $DP[i][j] = 0$. Altrimenti, esiste una *parentesizzazione ottima* $P[i \dots j] = P[i \dots k] \cdot P[k + 1 \dots j]$ e, sfruttando la ricorsione, otteniamo che $DP[i][j] = DP[i][k] + DP[k + 1][j] + c_{i-1} \cdot c_k \cdot c_j$. Il fattore $c_{i-1} \cdot c_k \cdot c_j$ è il costo per moltiplicare la matrice ottenuta dal prodotto $A_i \cdot \dots \cdot A_k$ che ha c_{i-1} righe e c_k colonne e la matrice ottenuta dal prodotto $A_{k+1} \cdot \dots \cdot A_j$ che ha c_k righe e c_j colonne.

Ma qual è il valore di k ?

Non lo sappiamo, ma poiché k può assumere valori compresi tra i e $j - 1$, possiamo provarli tutti e scegliere quello che comporta un numero minore di prodotti. La formula finale è quindi la seguente:

$$DP[i][j] = \begin{cases} 0 & i = j \\ \min_{i \leq k < j} \{DP[i][k] + DP[k + 1][j] + c_{i-1} \cdot c_k \cdot c_j\} & i < j \end{cases}$$

12.8.1 Approccio basato su divide-et-impera

Traducendo quella formula in codice otteniamo la seguente funzione:

Frammento 99 - Implementazione ricorsiva della soluzione.

```
int recPar(int[] c, int i, int j)
    if (i == j) then
        return 0
    int minSoFar = +∞
    for (k = i to j - 1) do
        int val = recPar(c, i, k) + recPar(c, k + 1, j)
        val = val + c[i - 1] · c[k] · c[j]
        if (val < minSoFar) then
            minSoFar = val
    return minSoFar
```

Il parametro c è un vettore con indici da 0 a n contenente le dimensioni di tutte le matrici. In particolare, $c[0]$ contiene il numero di righe della prima matrice, $c[i - 1]$ il numero di righe della matrice $A[i]$ e $c[i]$ il numero di colonne della stessa.

Complessità Ad ogni livello vengono eseguite due chiamate, quindi la *complessità* è $T(n) = \Theta(2^n)$ che non è migliore di un algoritmo a forza bruta. Il problema è che molti sotto-problemi vengono risolti più di una volta e il loro numero è $\frac{n(n+1)}{2}$.

12.8.2 Approccio basato su programmazione dinamica

Introduciamo due matrici $n \times n$: DP e $last$. $DP[i][j]$ contiene il numero minimo di prodotti scalari necessari per moltiplicare le matrici $A[i \dots j]$, mentre $last[i][j]$ contiene il valore k dell'ultimo prodotto che minimizza il costo del sotto-problema.

Frammento 100 - Implementazione iterativa della soluzione.

```
int computePar(int[] c, int n)
    int[][] DP = new int[1...n][1...n]
    int[][] last = new int[1...n][1...n]
    for (i = 1 to n) do
        DP[i][i] = 0                                     % Pone a 0 la diagonale principale
    for (h = 2 to n) do                                  % h: indice della diagonale
        for (i = 1 to n - h + 1) do                      % i: riga
            int j = i + h - 1                             % j: colonna
            DP[i][j] = +∞
            for (k = i to j - 1) do                       % k: indice ultimo prodotto
                int temp = DP[i][k] + DP[k + 1][j]
                temp = temp + c[i - 1] · c[k] · c[j]
                if (temp < DP[i][j]) then
                    DP[i][j] = temp
                    last[i][j] = k
    return DP[1][n]
```

Esempio 32 - Esempio di esecuzione.

Consideriamo il seguente vettore c :

$$c = [7, 8, 4, 2, 3, 5, 6]$$

Abbiamo quindi sei matrici di dimensioni:

$$A_1 = 7 \times 8, A_2 = 8 \times 4, A_3 = 4 \times 2, A_4 = 2 \times 3, A_5 = 3 \times 5, A_6 = 5 \times 6$$

Le tabelle DP e $last$ vengono popolate procedendo diagonalmente a partire dalla diagonale principale e muovendo verso destra. Al termine dell'esecuzione valgono quindi:

DP	1	2	3	4	5	6
1	0	224	176	218	276	350
2		0	64	112	174	250
3			0	24	70	138
4				0	30	90
5					0	90
6						0

$last$	1	2	3	4	5	6
1	0	1	1	3	3	3
2		0	2	3	3	3
3			0	3	3	3
4				0	4	5
5					0	5
6						0

Il valore restituito dall'algoritmo è 350: il numero minimo di prodotti scalari che devono essere eseguiti per calcolare $A[1 \dots 6]$.

Complessità Il costo computazionale di questa soluzione è $T(n) = \Theta(n^3)$ perché la matrice ha dimensione n^2 e ogni cella richiede $\Theta(n)$ per essere popolata.

12.9 Ricerca dell'insieme indipendente di peso massimo

Problema 12 - Ricerca dell'insieme indipendente di peso massimo.

Siano dati n intervalli distinti $[a_1, b_1], \dots, [a_n, b_n]$ della retta reale, aperti a destra e associati a un costo w_i per $1 \leq i \leq n$. Trovare un insieme indipendente di peso massimo, ovvero un sottoinsieme di intervalli, disgiunti tra loro, tale che la somma dei loro pesi sia massima.

Definizione 121 - Intervalli disgiunti.

Due intervalli i e j sono disgiunti se e solo se $b_j \leq a_i$ o $b_i \leq a_j$.

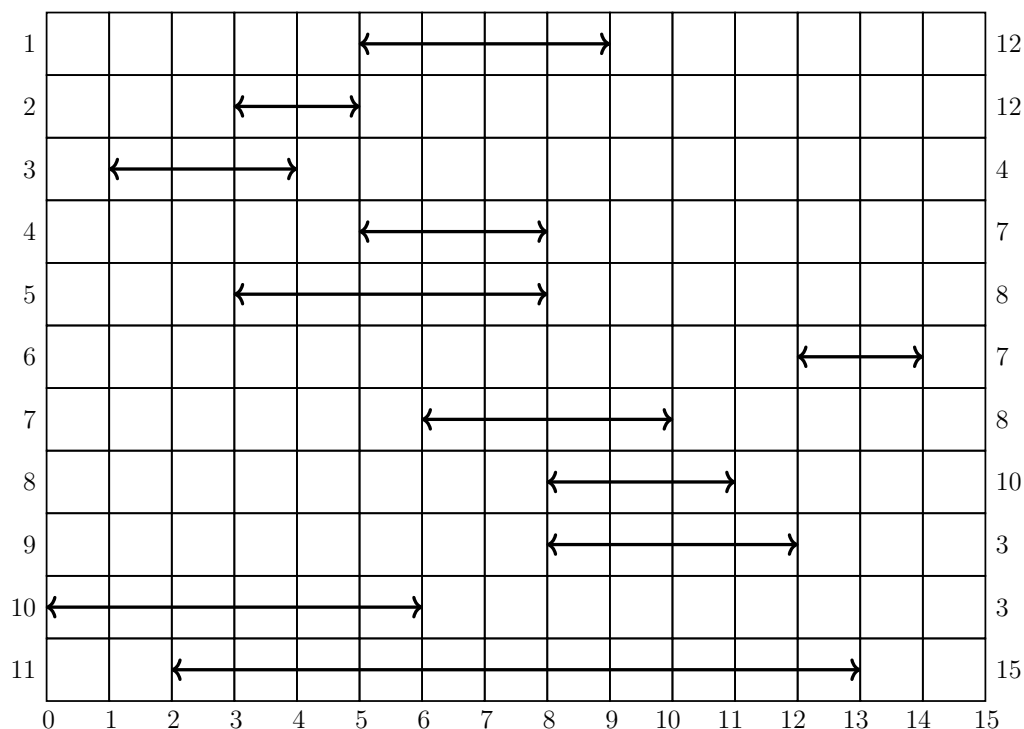
NB. Due intervalli sono *disgiunti* se non si intersecano.

Vediamo un esempio che possa aiutare a rendere più chiaro il problema.

Esempio 33 - Affitto di una sala conferenze.

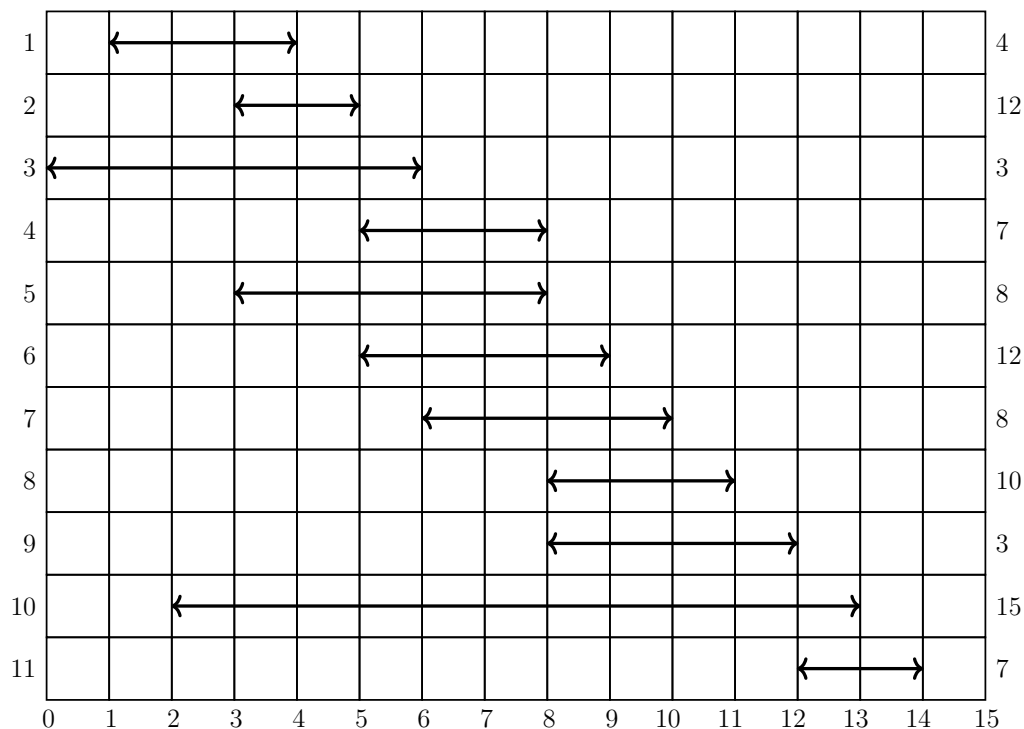
Consideriamo un hotel che deve decidere a quali clienti concedere in affitto la propria sala conferenze. Ogni cliente richiede la sala per un certo numero di ore ed è disposto a pagare una determinata cifra. L'hotel vuole scegliere l'insieme di clienti che gli permetta di ottenere il guadagno maggiore.

L'immagine di seguito schematizza l'insieme delle richieste ricevute dall'hotel e sul lato sinistro sono indicate le somme di denaro che i clienti hanno offerto.



Siccome non possono esserci sovrapposizioni, se l'hotel decidesse di concedere la sala al cliente 6, non potrebbe concederla al cliente 11 perché i due intervalli di tempo si intersecherebbero.

Il problema con questa disposizione è che preso un intervallo, per capire quali altri intervalli possono ancora essere considerati, siamo costretti ad analizzarli tutti. Tuttavia, in una situazione come questa può essere conveniente eseguire una pre-elaborazione dei dati, per esempio, potremmo ordinare gli intervalli per estremi di fine non decrescenti.



Ora, per decidere se tenere o meno l'ultimo intervallo, possiamo procedere come abbiamo fatto negli esempi precedenti, cioè considerando una qualche tabella DP associata all'intervallo di tempo che si conclude nell'istante in cui inizia l'intervallo considerato.

Per rispondere al quesito originale, l'hotel può massimizzare i profitti concedendo la sala ai clienti: 2, 4, 8 e 11.

Chiarita la necessità di effettuare una pre-elaborazione, dobbiamo decidere in che modo trattare i dati. L'idea già citata è quella di ordinare gli intervalli per estremi di fine non decrescenti, cioè ordinarli in modo che $b_1 \leq \dots \leq b_n$. A questo punto, definiamo una tabella DP tale che $DP[i]$ contenga il massimo profitto realizzabile con i primi i intervalli:

$$DP[i] = \begin{cases} 0 & i = 0 \\ \max(DP[i-1], \max\{DP[j] + w[i] : j < i \wedge b_j \leq a_i\}) & i > 0 \end{cases}$$

Con questa soluzione, cercare l'indice j costa $O(n)$ perché partendo dall'intervallo i è necessario scandire tutti gli altri. Poiché quest'operazione dev'essere ripetuto per ogni intervallo, il costo finale è $O(n^2)$.

È possibile trovare una pre-elaborazione migliore?

Potremmo provare a pre-calcolare il predecessore $pred_i = j$ di i tale che j sia il massimo valore minore di i per cui $b_j \leq a_i$. Se non esiste un tale valore, consideriamo $pred_i = 0$. Noto il predecessore, la definizione di DP diventa:

$$DP[i] = \begin{cases} 0 & i = 0 \\ \max(DP[i-1], DP[pred_i] + w[i]) & i > 0 \end{cases}$$

NB. È possibile escludere un intervallo i , se sceglierne un altro j con stesso tempo di fine, ma tempo di inizio precedente, ha un valore $DP[j] > DP[i]$.

Un modo per calcolare i predecessori è il seguente:

Frammento 101 - Funzione per il calcolo dei predecessori.

```
int[] computePredecessors(int[] a, int[] b, int n)
    int[] pred = new int[0...n]
    pred[0] = 0
    for (i = 1 to n) do
        int j = i - 1
        while (j > 0 and b[j] > a[i]) do
            j = j - 1
        pred[i] = j
    return pred
```

Il costo di questa funzione rimane $O(n^2)$, ma esistono delle implementazioni di *complessità* $O(n \log n)$.

Frammento 102 - Implementazione della soluzione.

```
SET maxInterval(int[] a, int[] b, int[] w, int n)
    % { Ordina gli intervalli per estremi di fine non decrescenti }
    int[] pred = computePredecessors(a, b, n)
    int[] DP = new int[0...n]
    DP[0] = 0
    for (i = 1 to n) do % Riempimento tabella delle soluzioni
        DP[i] = max(DP[i - 1], DP[pred[i]] + w[i])
    int i = n
    SET S = Set()
```

```

while (i > 0) do                                     % Ricostruzione della soluzione
    if (DP[i - 1] > DP[pred[i]] + w[i]) do
        i = i - 1
    else
        S.insert(i)
        i = pred[i]
return S

```

Complessità La *complessità* dell'intero algoritmo è $O(n \log n)$ in quanto paghiamo $O(n \log n)$ per ordinare gli intervalli e calcolare i predecessori, mentre per popolare la tabella e ricostruire la soluzione paghiamo $\Theta(n)$.

Capitolo Nr.13

Scelta della struttura dati

Abbiamo già discusso più volte su come la scelta di una *struttura dati* influisca sulla *complessità* degli algoritmi, tuttavia, ora riprendiamo l'argomento perché input diversi allo stesso algoritmo potrebbero far risultare più efficiente l'uso di una *struttura dati* piuttosto che un'altra.

Quanto detto emerge chiaramente nel caso degli algoritmi per la ricerca dei *cammini minimi* all'interno di *grafi*. È un problema che abbiamo già affrontato in precedenza, ma qui ne vediamo due versioni più complesse: la ricerca dei *cammini minimi* verso tutti i *nodi*, da sorgente singola e da sorgente multipla.

13.1 Ricerca dei cammini minimi da sorgente singola

Problema 13 - Ricerca dei cammini minimi da sorgente singola.

Dati un grafo orientato $G = (V, E)$, una funzione di peso $w = E \rightarrow \mathbb{R}$ e un nodo sorgente s , trovare un cammino da s a u , per ogni $u \in V$, il cui costo sia minimo, ovvero minore o uguale al costo di ogni altro cammino da s a u .

Definizione 122 - Costo di un cammino.

Dato un cammino $p = (v_1, \dots, v_k)$ con $k > 1$, il costo del cammino è dato dalla seguente sommatoria:

$$w(p) = \sum_{i=2}^k w(v_{i-1}, v_i)$$

Per quanto riguarda i pesi, possono essere sia numeri interi che reali, ma anche positivi e negativi e alcuni algoritmi potrebbero non funzionare per alcune tipologie di pesi.

Consideriamo i due seguenti *grafi*:

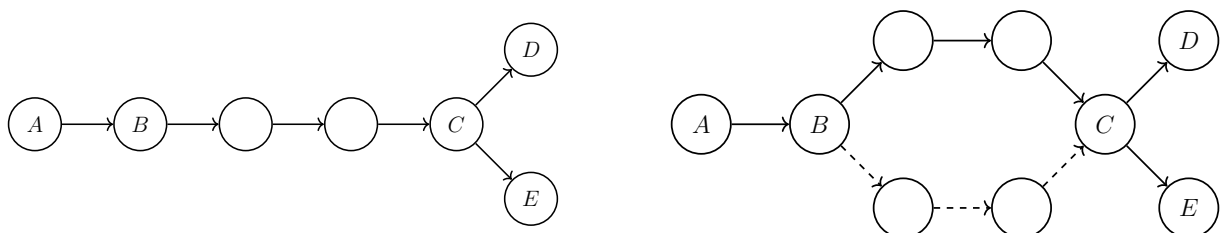


Fig. 13.1: Esempi di *cammini minimi*

Notiamo subito che *cammini minimi* verso *nod*i diversi potrebbero percorrere un tratto in comune, ma non potrebbero convergere su un *nodo* comune dopo aver percorso tratti diversi. Da ciò consegue che l'insieme dei *cammini minimi* da un *nodo* a tutti gli altri può essere rappresentato come un *albero di copertura*.

Definizione 123 - Albero di copertura.

Dato un grafo $G = (V, E)$ non orientato e connesso, un albero di copertura di G è un sottografo $T = (V, E_T)$ tale per cui T è un albero che contiene tutti i nodi di G ed $E_T \subseteq E$.

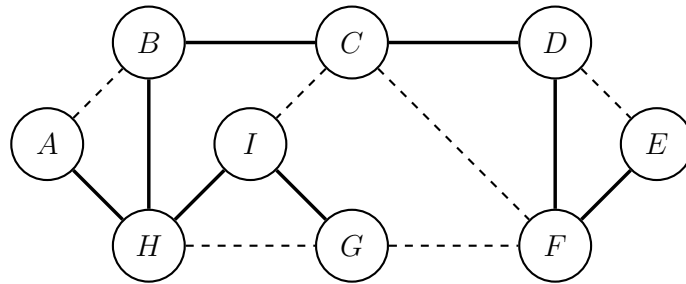


Fig. 13.2: Esempio di albero di copertura

Detto questo, possiamo dire che tutte le soluzioni che non generano un *albero di copertura* non sono ammissibili come soluzioni al problema.

Definizione 124 - Soluzione ammissibile.

Una soluzione ammissibile può essere descritta da un albero di copertura T radicato in s e da un vettore delle distanze d , i cui valori $d[u]$ rappresentano il costo del cammino da s a u in T .

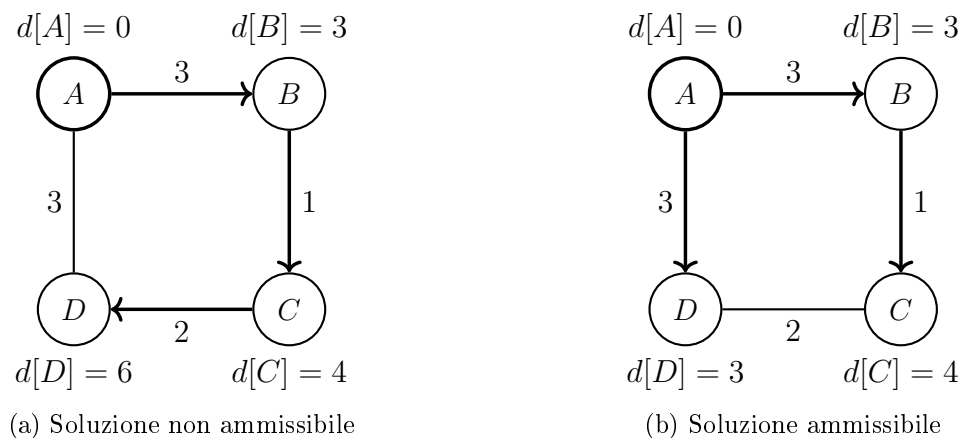


Fig. 13.3: Possibili *cammini* con sorgente nel *nodo* A

Nel primo caso la soluzione non è ammissibile perché la distanza del *nodo* D da A è 6, quando potrebbe essere 3.

Per rappresentare l'*albero di copertura* possiamo usare una versione modificata della `printPath` che avevamo usato per stampare il *cammino* tra due *nod*i.

Frammento 103 - printPath per la stampa dell'albero di copertura.

```
printPath(NODE s, NODE d, NODE[] T)
    if (s == d) then
        print s
    else if (T[d] == nil) then
        print "error"
    else
        printPath(s, T[d], T)
        print d
```

Siccome tra tutte le soluzioni possibili siamo interessati a quelle che includono solo *cammini minimi*, restringiamo il campo di ricerca alle *soluzioni ottime*, quelle cioè che rispettano il seguente teorema.

Definizione 125 - Teorema di Belman.

Dato un grafo $G = (V, E)$ e una soluzione $T = (V, E_T)$ in esso ammissibile, T è anche una soluzione ottima se e solo se:

$$\begin{aligned} d[v] &= d[u] + w(u, v) \quad \forall (u, v) \in E_T \\ d[v] &\leq d[u] + w(u, v) \quad \forall (u, v) \in E \end{aligned}$$

Nell'esempio di prima, la soluzione della prima figura non è *ottima* perché $d[D] > d[A] + w(A, D)$.

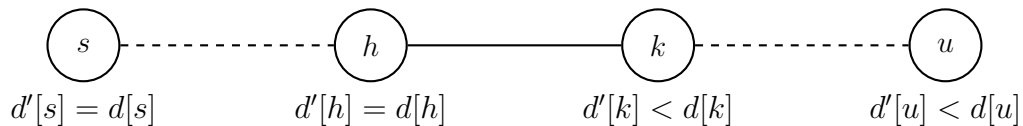
Dimostrazione. Dimostriamo separatamente le due parti del teorema.

Parte 1 Sia T una *soluzione ottima* e sia $(u, v) \in E$:

- Se $(u, v) \in E_T$, allora $d[v] = d[u] + w(u, v)$;
- Se $(u, v) \notin E_T$, allora $d[v] \leq d[u] + w(u, v)$, perché altrimenti esisterebbe nel *grafo* G un *cammino* da s a v più corto di quello in T , generando un assurdo.

Parte 2 Supponiamo per assurdo che T rispetti le condizioni del teorema, ma che non sia una *soluzione ottima*. Se così fosse, esisterebbe un *cammino non ottimo* C da s a un altro *nodo* u in T . Quindi, esisterebbe anche un *albero di copertura* T' , in cui il *cammino* C' da s a u abbia distanza $d'[u] < d[u]$, dove d' è il vettore delle distanze associato a T' .

Poiché $d'[s] = d[s] = 0$, ma $d'[u] < d[u]$, esiste un *arco* (h, k) in C' tale che $d'[h] = d[h]$ e $d'[k] < d[k]$. La situazione sarebbe dunque la seguente:



Per costruzione, $d'[h] = d[h]$ e $d'[k] = d'[h] + w(h, k)$. Per ipotesi invece, T rispetta le condizioni del teorema e quindi vale $d[k] \leq d[h] + w(h, k)$. Combinando queste due relazioni si ottiene:

$$d'[k] = d'[h] + w(h, k) = d[h] + w(h, k) \geq d[k]$$

Da ciò seguirebbe $d'[k] \geq d[k]$ che però contraddice la relazione $d'[k] < d[k]$ trovata in precedenza. \square

13.1.1 Prototipo di algoritmo

Vediamo quale potrebbe essere la struttura base di un algoritmo per la ricerca dei *cammini minimi*.

Frammento 104 - Prototipo di algoritmo.

```
<int[], int[]> minPathPrototype(GRAPH G, NODE s)
    % Inizializza T a una foresta di copertura composta da nodi isolati
    % Inizializza d con sovrastima della distanza, cioè  $d[s] = 0, d[x] = +\infty$ 
    while ( $\exists \langle u, v \rangle : d[u] + G.w(u, v) < d[v]$ ) do
         $d[v] = d[u] + G.w(u, v)$ 
        % Sostituisci il padre di v in T con u
    return <T, d>
```

Frammento 105 - Algoritmo generico.

```
<int[], int[]> shortestPath(GRAPH G, NODE s)
    int[] T = new int[1...G.size()]          %  $T[u]$  è il padre di u nell'albero T
    int[] d = new int[1...G.size()]          %  $d[u]$  è la distanza di u da s
    boolean[] b = new boolean[1...G.size()] %  $b[u]$  è true se  $u \in S$ 
    foreach ( $u \in G.V() - \{s\}$ ) do
        T[u] = nil
        d[u] =  $+\infty$ 
        b[u] = false
    T[s] = nil
    d[s] = 0
    b[s] = true
    DATASTRUCTURE S = DataStructure()
    S.add(s)
    while (not S.isEmpty()) do
        int u = S.extract()
        b[u] = false
        foreach ( $v \in G.adj(u)$ ) do
            if ( $d[u] + G.w(u, v) < d[v]$ ) then
                if (not b[v]) then
                    S.add(v)
                    b[v] = true
                else
                    % Azione da svolgere nel caso  $v \in S$ 
                    T[v] = u
                     $d[v] = d[u] + G.w(u, v)$ 
    return <T, d>
```

13.1.2 Algoritmo di Dijkstra

La prima implementazione vera e propria che vediamo è quella proposta da Dijkstra nel 1959. Si basa su *code a priorità* e funziona bene solo se i pesi sono positivi.

NB. Siccome gli *heap* furono introdotti solo nel 1964, la prima versione dell'algoritmo utilizzava *code a priorità* implementate usando un vettore.

Frammento 106 - Algoritmo di Dijkstra.

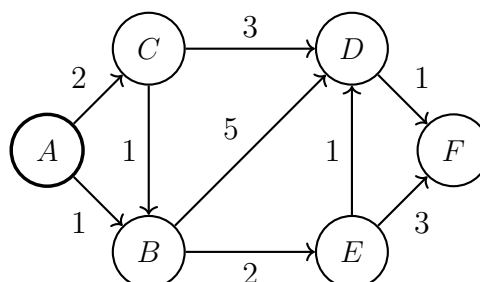
```
<int[], int[]> shortestPath(GRAPH G, NODE s)
  int[] T = new int[1...G.size()]           % T[u] è il padre di u nell'albero T
  int[] d = new int[1...G.size()]           % d[u] è la distanza di u da s
  boolean[] b = new boolean[1...G.size()]   % b[u] è true se u ∈ S
  foreach (u ∈ G.V() - {s}) do
    T[u] = nil
    d[u] = +∞
    b[u] = false
  T[s] = nil
  d[s] = 0
  b[s] = true
  PRIORITYQUEUE Q = MinPriorityQueue() % Quello che in precedenza era S
  Q.insert(s, 0)                        % La priorità di s è 0
  while (not Q.isEmpty()) do
    int u = Q.deleteMin()               % A ogni ciclo viene estratto il nodo più vicino
    b[u] = false
    foreach (v ∈ G.adj(u)) do
      if (d[u] + G.w(u, v) < d[v]) then
        if (not b[v]) then
          Q.insert(v, d[u] + G.w(u, v)) % Aggiungo v alla coda
          b[v] = true
        else
          Q.decrease(v, d[u] + G.w(u, v)) % Riduco la priorità di v da s
      T[v] = u
      d[v] = d[u] + G.w(u, v)
  return <T, d>
```

L'idea alla base di questa soluzione è quella di usare la distanza di un nodo da s come valore per la sua *priorità*, quindi ad ogni iterazione, estrarre il *nodo* con la *priorità* minore significa estrarre il *nodo* il cui cammino da s è di peso minore.

Quando viene estratto un *nodo* e l'esecuzione dell'algoritmo ricade nel ramo **else** del controllo sul valore $b[v]$, significa che dal *nodo* u è possibile raggiungere v con costo minore a quello che stiamo pagando attualmente. Di conseguenza, aggiorniamo la *priorità* di v indicando il costo del nuovo *cammino*.

Esempio 34 - Esempio d'esecuzione.

Consideriamo il seguente grafo:



Eseguendo l'algoritmo partendo da A otteniamo la seguente tabella:

		A	B	C	E	D	F
A	0	0	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset
B	$+\infty$	1	1	1	1	1	1
C	$+\infty$	2	2	2	2	2	2
D	$+\infty$	$+\infty$	6	5	4	4	4
E	$+\infty$	$+\infty$	3	3	3	3	3
F	$+\infty$	$+\infty$	$+\infty$	$+\infty$	6	5	5

Nella tabella, ogni colonna contiene lo stato del vettore d all'inizio di ogni iterazione del ciclo *while* (*not* $Q.isEmpty()$), mentre ogni riga traccia l'evoluzione dello stato del rispettivo nodo. I nodi barrati sono quelli che non sono presenti nella coda.

Prima dell'ingresso nel ciclo, la sorgente, ovvero il nodo A è a distanza 0 da sé stesso. Tutti gli altri sono a $+\infty$. Alla prima iterazione viene estratto A e vengono inseriti i nodi B e C ad esso adiacenti, indicando anche il peso del relativo arco.

Alla seconda iterazione viene estratto B perché il costo per raggiungerlo è minimo. Come prima, vengono inseriti nella coda i nodi D ed E che sono adiacenti a B . L'algoritmo continua fino all'estrazione di F .

Dimostrazione di correttezza per pesi positivi

Dimostrazione. La correttezza dell'algoritmo per pesi positivi si basa su due assunti:

1. Ogni *nodo* viene estratto una sola volta;
2. Al momento dell'estrazione il peso dal *cammino* dalla sorgente s è minimo;

Per la dimostrazione procediamo per induzione sul numero k di *nodi* estratti.

Caso base: $k = 0$ Il caso è verificato in quanto $d[s] = 0$ e non ci sono pesi negativi.

Passo induttivo: $k > 0$ Per ipotesi induttiva supponiamo che gli assunti siano corretti per i primi $k - 1$ elementi. Quando viene estratto il k -esimo elemento u , il peso $d[u]$ dipende esclusivamente dai *nodi* già estratti, quindi la sua distanza da s è minima. Siccome non ci sono pesi negativi e tutti gli altri *nodi* hanno una distanza da s almeno pari a quella di u , u non verrà mai più inserito nella *coda*. \square

Complessità Il costo per inizializzare i vettori T , d e b è $\Theta(n)$, la ricerca del minimo all'interno di una *coda a priorità* implementata come vettore costa $\Theta(n)$, mentre l'inserimento e la modifica della *priorità* di un elemento costano $\Theta(1)$. Complessivamente, dunque, l'algoritmo ha *complessità* $T(n) = \Theta(n^2)$ perché vengono inseriti e rimossi dalla *coda* tutti i *nodi*.

Esistono altre versioni dello stesso algoritmo che però fanno uso di implementazioni più efficienti delle *code a priorità*.

La versione di *Johnson* del 1977 si basa su *heap binari* e il costo delle operazioni sulla *coda* diventa $O(\log n)$, riducendo la *complessità* totale a $T(n) = O(m \log n)$. Una terza versione, quella proposta da *Fredman e Tarjan* nel 1987, usa gli *heap di Fibonacci*, per cui le operazioni di inserimento e rimozione costano $O(\log n)$, mentre la modifica di *priorità* ha un *costo ammortizzato* $O(1)$. La *complessità* complessiva è $O(m + n \log n)$.

13.1.3 Algoritmo di Bellman-Ford-Moore

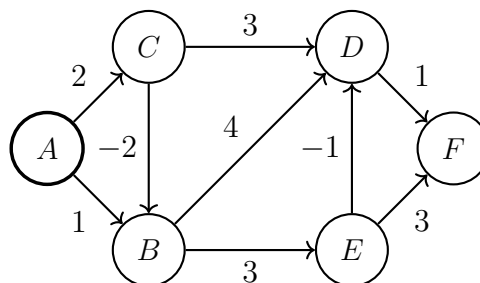
Questo algoritmo ha una struttura simile a quello appena visto, ma al posto di una *coda a priorità* viene usata una *coda* “classica”. Questa soluzione permette di trattare anche *grafi* con *archi* di peso negativo.

Frammento 107 - Algoritmo di Bellman-Ford-Moore.

```
<int[], int[]> shortestPath(GRAPH G, NODE s)
    int[] T = new int[1...G.size()]           % T[u] è il padre di u nell'albero T
    int[] d = new int[1...G.size()]           % d[u] è la distanza di u da s
    boolean[] b = new boolean[1...G.size()]   % b[u] è true se u ∈ S
    foreach (u ∈ G.V() - {s}) do
        T[u] = nil
        d[u] = +∞
        b[u] = false
    T[s] = nil
    d[s] = 0
    b[s] = true
    QUEUE Q = Queue()
    Q.enqueue(s)
    while (not Q.isEmpty()) do
        int u = Q.dequeue()
        b[u] = false
        foreach (v ∈ G.adj(u)) do
            if (d[u] + G.w(u, v) < d[v]) then
                if (not b[v]) then
                    Q.enqueue(v)
                    b[v] = true
                T[v] = u
                d[v] = d[u] + G.w(u, v)
    return <T, d>
```

Esempio 35 - Esempio d'esecuzione.

Consideriamo il seguente grafo:



Eseguendo l'algoritmo partendo da A otteniamo la seguente tabella:

		Passate											
Nodo	Inizio	0	1		2			3			4		5
		<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>B</i>	<i>F</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>D</i>	<i>F</i>
<i>A</i>	0	0	0	0	0	0	0	0	0	0	0	0	0
<i>B</i>	$+\infty$	1	1	0	0	0	0	0	0	0	0	0	0
<i>C</i>	$+\infty$	2	2	2	2	2	2	2	2	2	2	2	2
<i>D</i>	$+\infty$	$+\infty$	5	5	5	3	3	3	3	2	2	2	2
<i>E</i>	$+\infty$	$+\infty$	4	4	4	4	3	3	3	3	3	3	3
<i>F</i>	$+\infty$	$+\infty$	$+\infty$	$+\infty$	6	6	6	6	4	4	4	3	3
Coda	<i>A</i>	<i>BC</i>	<i>CDE</i>	<i>DEB</i>	<i>EBF</i>	<i>BFD</i>	<i>FDE</i>	<i>DE</i>	<i>EF</i>	<i>FD</i>	<i>D</i>	<i>F</i>	

Inizialmente, il peso dei cammini da A a tutti gli altri nodi è $+\infty$. Nella passata 0 vengono considerati tutti i nodi che possono essere raggiunti da A attraversando 0 archi. Ovviamente, viene considerato soltanto A che è anche l'unico nodo presente nella coda.

Quindi, A viene estratto dalla coda e ne vengono considerati i vicini: B e C. Siccome il peso dei cammini da A verso i suoi vicini è minore del valore attualmente noto (i.e. $+\infty$), B e C vengono aggiunti alla coda e i pesi dei cammini vengono aggiornati.

Nella passata 1 vengono considerati tutti i nodi che possono essere raggiunti da A attraversando un arco, e di nuovo ciò coincide con i nodi presenti nella coda. Da B è possibile raggiungere D ed E in modo conveniente, quindi questi vengono accodati e il costo per raggiungerli viene aggiornato. Da C invece, si possono raggiungere D e B. Per quanto riguarda D, il costo del cammino passante per C è $2 + 3 = 5$ che è uguale a quello del cammino passante per B, quindi D non viene aggiunto nuovamente alla coda. D'altra parte, il cammino che raggiunge B passando per C ha peso totale pari a 0, che è minore del peso del cammino diretto da A, quindi B viene riaggiunto in coda. L'algoritmo continua in questo modo fino a quando non viene svuotata la coda.

Possiamo definire in modo più formale ciò che abbiamo chiamato *passata*. Sfruttando la ricorrenza possiamo dire che per $k = 0$, la 0-esima passata consiste nell'estrazione dalla coda del solo nodo sorgente. Per $k > 0$ invece, la k -esima passata consiste nell'estrazione dalla coda di tutti i nodi presenti al termine della $k - 1$ -esima passata.

Questo ci permette anche di dire che al termine della k -esima passata, i vettori T e d descrivono i cammini minimi di lunghezza k . Conseguentemente, dopo la $n - 1$ -esima passata, dove n è il numero di nodi, sono noti tutti i cammini minimi del grafo. Ogni nodo può infatti essere inserito nella coda al massimo $n - 1$ volte.

Complessità L'inizializzazione dei vettori costa sempre $\Theta(n)$, mentre tutte le operazioni sulla coda hanno complessità $\Theta(1)$. Poiché ogni nodo può essere reinserito nella coda fino a $n - 1$ volte, la **dequeue** viene eseguita $O(n^2)$ volte. L'**enqueue** viene invece invocata tante volte quante sono i vicini di ogni nodo, ovvero $O(nm)$, dove m è il numero di archi del grafo. La complessità totale dell'algoritmo è dunque $O(nm)$.

13.1.4 Ricerca dei cammini minimi sui grafi DAG

I *grafi DAG* sono un caso particolare in quanto, non essendoci *cicli*, i cammini sono sempre definiti anche in presenza di pesi negativi. Di conseguenza, possiamo ricercare i cammini minimi semplicemente applicando l'algoritmo per l'*ordinamento topologico*.

Frammento 108 - Ricerca dei cammini minimi su grafi DAG.

```
<int[], int[]> shortestPath(GRAPH G, NODE s)  
  int[] T = new int[1...G.size()]  
  int[] d = new int[1...G.size()]  
  foreach (u ∈ G.V() - {s}) do  
    T[u] = nil  
    d[u] = +∞  
  T[s] = nil  
  d[s] = 0  
  STACK S = topsort(G)  
  while (not S.isEmpty()) do  
    NODE u = S.pop()  
    foreach (v ∈ G.adj(u)) do  
      if (d[u] + G.w(u, v) < d[v]) then  
        T[v] = u  
        d[v] = d[u] + G.w(u, v)  
  return <T, d>
```

13.1.5 Casi d'uso delle soluzioni

Algoritmo	Complessità	Caso d'uso
Dijkstra	$\Theta(n^2)$	Pesi positivi, grafi densi
Johnson	$O(m \log n)$	Pesi positivi, grafi sparsi
Fredman-Tarjan	$O(m + n \log n)$	Pesi positivi, grafi densi, dimensioni molto grandi
Bellman-Ford-Moore	$O(nm)$	Pesi negativi
topsort	$O(n + m)$	Grafi DAG
bfs	$O(n + m)$	Grafi non pesati

13.2 Ricerca dei cammini minimi da sorgente multipla

Un possibile approccio al problema è quello di invocare gli algoritmi appena visti una volta per ogni *nodo sorgente*. Il costo che si andrebbe a pagare sarebbe n volte il costo delle singole esecuzioni. È un'idea semplice, ma non necessariamente la migliore.

13.2.1 Algoritmo di Floyd-Warshall

L'intuizione alla base dell'algoritmo è quella dei *cammini k -vincolati*.

Definizione 126 - Cammini k -vincolati.

Sia $k \in \{0, \dots, n\}$. Diciamo che un cammino p_{xy}^k è un cammino minimo k -vincolato fra x e y se esso ha il costo minimo tra tutti i cammini minimi da x a y che non passano per nessun vertice in $\{v_{k+1}, \dots, v_n\}$.

NB. Stiamo assumendo, come in realtà abbiamo sempre fatto, che esista un ordinamento fra i nodi v_1, \dots, v_n del grafo.

NB. Seguendo la definizione data, p_{xy}^0 e p_{xy}^n corrispondono rispettivamente all'arco da x a y , se esiste, e al cammino minimo da x a y in tutto il grafo, cioè senza vincoli.

Se esistono i cammini k -vincolati, possiamo definire anche la distanza k -vincolata.

Definizione 127 - Distanza k -vincolata.

Denotiamo con $d^k[x][y]$ il costo del cammino k -vincolato da x a y :

$$d^k[x][y] = \begin{cases} w(p_{xy}^k) & \text{Se esiste} \\ +\infty & \text{altrimenti} \end{cases}$$

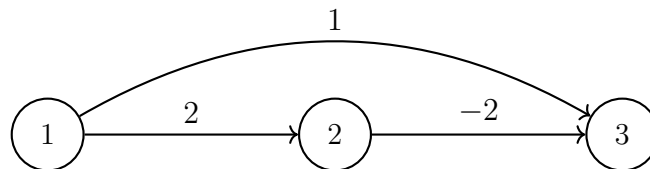
NB. Parallelamente a prima, $d^0[x][y]$ e $d^n[x][y]$ corrispondono rispettivamente al costo dell'arco tra x e y , se esiste, e al costo del cammino minimo da x a y in tutto il grafo.

I cammini k -vincolati ci permettono di ricercare i cammini minimi in modo ricorsivo partendo da $k = 0$ e incrementandolo. Inoltre, ci permettono anche di memorizzare di volta in volta il cammino di costo minore. Detto questo, possiamo esprimere la distanza k -vincolata con la seguente:

$$d^k[x][y] = \begin{cases} w(x, y) & k = 0 \\ \min(d^{k-1}[x][y], d^{k-1}[x][k] + d^{k-1}[k][y]) & k > 0 \end{cases}$$

Esempio 36 - Esempio del calcolo delle distanze k -vincolate.

Consideriamo il seguente grafo:



Le distanze k -vincolate tra 1 e 3 per valori crescenti di k sono:

$$\begin{aligned} d^0[1][3] &= 1 \\ d^1[1][3] &= 1 \\ d^2[1][3] &= \min(d^1[1][3], d^1[1][2] + d^1[2][3]) \\ &= \min(1, 0) = 0 \end{aligned}$$

Oltre alla matrice d dei costi, definiamo anche una matrice T dei padri, dove $T[x][y]$ rappresenta il predecessore di y nel cammino minimo tra x e y . Nell'esempio di prima avremmo avuto $T[1][2] = 1$, $T[2][3] = 2$ e $T[1][3] = 2$.

Frammento 109 - Algoritmo di Floyd-Warshall.

```
<int[][], int[][]> floydWarshall(GRAPH G)
    int[][] T = new int[1...G.size()][1...G.size()]
    int[][] d = new int[1...G.size()][1...G.size()]
```



```

foreach (u ∈ G.V()) do
  foreach (v ∈ G.V()) do
    d[u][v] = +∞           % Inizializza a +∞ i costi dei cammini
    T[u][v] = nil         % Inizializza a nil tutti i predecessori
  foreach (u ∈ G.V()) do
    foreach (v ∈ G.adj(u)) do % Inizializza i valori per nodi adiacenti
      d[u][v] = G.w(u, v)
      T[u][v] = u
  for (k = 1 to n) do
    foreach (u ∈ G.V()) do
      foreach (v ∈ G.V()) do
        if (d[u][k] + d[k][v] < d[u][v]) then
          d[u][v] = d[u][k] + d[k][v]
          T[u][v] = T[k][v]
  return ⟨T, d⟩

```

Complessità I tre cicli annidati fanno sì che la *complessità* dell'algoritmo risulti $\Theta(n^3)$.

Capitolo Nr.14

Programmazione greedy

Gli algoritmi realizzati per risolvere *problemi di ottimizzazione* eseguono una sequenza di decisioni che, nel caso della *programmazione dinamica*, sono prese in maniera bottom-up valutando tutte le possibilità, evitando però di ripetere decisioni già esaminate. Con la *programmazione greedy* invece, tra tutte le possibili decisioni ne viene scelta solo una, quella che sembra ottima, o meglio localmente ottima. Tuttavia, è necessario dimostrare che quella particolare scelta permette di ottenere un risultato ottimo anche a livello globale.

Proprio per questo motivo, la *programmazione greedy* è utilizzabile solo per quei problemi che presentano una *sottostruttura ottima* e per i quali è possibile dimostrare che esiste una scelta “ingorda”, cioè una scelta che porta ad una soluzione ottima.

14.1 Problema del resto

Problema 14 - Problema del resto.

Dati un insieme di “tagli” di monete, espressi come un vettore $t[1 \dots n]$ di interi positivi, e un valore R rappresentate un resto da dover restituire, trovare il più piccolo numero intero di pezzi necessari per dare un resto R di centesimi, utilizzando i tagli disponibili e assumendo di avere un numero illimitato di monete.

Formalmente, è necessario trovare un vettore x di interi non negativi tale per cui:

$$R = \sum_{i=1}^n x[i] \cdot t[i] \quad e \quad m = \sum_{i=1}^n x[i] \quad \text{ha valore minimo}$$

Questo problema possiede una *sottostruttura ottima*:

Definizione 128 - Sottostruttura ottima.

Siano $S(i)$ il problema di dare un resto pari ad i e $A(i)$ una soluzione ottima del problema $S(i)$ rappresentata da un multi-insieme. Sia poi $j \in A(i)$. Allora, $S(i - t[j])$ è un sotto-problema di $S(i)$, la cui soluzione è data da $A(i) - \{j\}$.

14.1.1 Approccio basato su programmazione dinamica

Prima di vedere come potrebbe essere realizzata una soluzione *greedy*, utilizziamo le proprietà della *sottostruttura ottima* per implementare una soluzione basata su *programmazione dinamica*.

Se $DP[i]$ rappresenta il numero minimo di monete necessario a risolvere il problema $S(i)$, vale la seguente funzione ricorsiva:

$$DP[i] = \begin{cases} 0 & i = 0 \\ \min_{1 \leq j \leq n} \{DP[i - t[j]] \mid t[j] \leq i\} + 1 & i > 0 \end{cases}$$

Frammento 110 - Soluzione basata su programmazione dinamica.

```
int[] moneyChange(int[] t, int n, int R)
    int[] DP = new int[0...R]                                % Tabella delle soluzioni
    int[] coins = new int[0...R]                             % Moneta da usare per uno specifico resto
    DP[0] = 0
    for (i = 1 to R) do
        DP[i] = +∞
        for (j = 1 to n) do
            if (i > t[j] and DP[i - t[j]] + 1 < DP[i]) then
                DP[i] = DP[i - t[j]] + 1
                coins[i] = j
    % Ricostruzione della soluzione
    int[] x = new int[1...n] = {0}                            % Contatore delle monete usate
    while (R > 0) do
        x[coins[R]] = x[coins[R]] + 1
        R = R - t[coins[R]]
    return x
```

Complessità La *complessità* di questa soluzione è $\Theta(nR)$ per la presenza dei due cicli annidati.

14.1.2 Approccio basato su programmazione greedy

Non è difficile immaginare che per minimizzare il numero di monete utilizzate sia sufficiente usare sempre il taglio più grande possibile. Ovvero, selezionare la moneta j più grande, tale per cui $t[j] \leq R$, e risolvere il sotto-problema $S(R - t[j])$.

Frammento 111 - Soluzione basata su programmazione greedy.

```
int[] moneyChange(int[] t, int n, int R)
    int[] x = new int[1...n]
    { Ordina le monete in modo decrescente }
    for (i = 1 to n) do
        x[i] = ⌊R / t[i]⌋
        R = R - x[i] · t[i]
    return x
```

Complessità Questa soluzione abbassa la *complessità* a $O(n \log n)$ se l'insieme di monete non è ordinato, $O(n)$ altrimenti.

Ciò che dobbiamo fare adesso è dimostrare che la scelta fatta sia corretta.

Dimostrazione. Supponiamo $t = [50, 10, 5, 1]$. Sia x una qualunque soluzione ottima, ovvero tale per cui:

$$R = \sum_{i=1}^n x[i] \cdot t[i] \quad \text{e} \quad m = \sum_{i=1}^n x[i] \quad \text{ha valore minimo}$$

Sappiamo che $t[k] \cdot x[k] < t[k-1]$ altrimenti basterebbe sostituire un certo numero di monete di taglio $t[k]$ con quelle del taglio $t[k-1]$. In questo caso, i limiti al numero di monete per ogni taglio sono:

$$\begin{aligned} t[2] \cdot x[2] &= 10 \cdot x[2] < t[1] = 50 \Rightarrow x[2] < 5 \\ t[3] \cdot x[3] &= 5 \cdot x[3] < t[2] = 10 \Rightarrow x[3] < 2 \\ t[4] \cdot x[4] &= 1 \cdot x[4] < t[3] = 5 \Rightarrow x[4] < 5 \end{aligned}$$

Sia ora m_k la somma delle monete di taglio inferiore a $t[k]$:

$$m_k = \sum_{i=k+1}^4 x[i] \cdot t[i]$$

Se riusciamo a dimostrare che $m_k < t[k] \forall k$, allora la soluzione proposta dall'algoritmo è proprio quella ottima. Per ogni valore di k otteniamo:

$$\begin{aligned} m_4 &= 0 < 1 = t[4] \\ m_3 &= x[4] \cdot 1 + m_4 \leq 4 \cdot 1 + m_4 < 4 + 1 = 5 = t[3] \\ m_2 &= x[3] \cdot 1 + m_3 \leq 1 \cdot 5 + m_3 < 5 + 5 = 10 = t[2] \\ m_1 &= x[2] \cdot 1 + m_2 \leq 4 \cdot 10 + m_2 < 40 + 10 = 50 = t[1] \end{aligned}$$

□

NB. Non tutti i tagli di monete permettono di utilizzare la *programmazione greedy*. Ad esempio, per $t[10, 8, 1]$ l'algoritmo *greedy* utilizzerebbe una moneta da 10 e 7 da 1, quando la scelta migliore sarebbe quella di prenderne 2 da 8 e una da 1.

NB. In ogni caso, l'insieme delle monete deve includere una moneta di taglio unitario.

14.2 Insieme indipendente massimale di intervalli

Vediamo un caso particolare del problema sulla **Ricerca dell'insieme indipendente di peso massimo** in cui il peso di tutti gli intervalli è 1. Cioè, vogliamo cercare l'insieme contenente il maggior numero di intervalli.

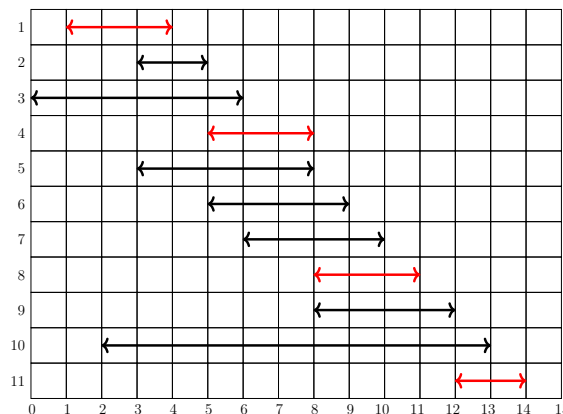


Fig. 14.1: Possibile soluzione

14.2.1 Approccio basato su programmazione dinamica

Iniziamo ricercando una *sottostruttura ottima*. Assumiamo che gli intervalli siano ordinati per tempo di fine:

$$b_1 \leq \dots \leq b_n$$

Definiamo il sotto-problema $S[i \dots j]$ come l'insieme di intervalli che iniziano dopo la fine di i e finiscono prima dell'inizio di j :

$$S[i \dots j] = \{k \mid b_i \leq a_k < b_k \leq a_j\}$$

Aggiungiamo due intervalli fittizi:

- *Intervallo 0*: $b_0 = -\infty$;
- *Intervallo $n+1$* : $a_{n+1} = +\infty$;

Il problema generale corrisponde quindi al problema $S[0 \dots n+1]$.

Definizione 129 - Sottostruttura ottima.

Supponiamo che $A[i \dots j]$ sia una soluzione ottimale di $S[i \dots j]$ e sia k un intervallo che appartiene ad $A[i \dots j]$. Allora:

- *Il problema $S[i \dots j]$ può essere suddiviso in due intervalli:*
 1. $S[i \dots k]$ contenente gli intervalli di $S[i \dots j]$ che finiscono prima di k ;
 2. $S[k \dots j]$ contenente gli intervalli di $S[i \dots j]$ che iniziano dopo k ;
- $A[i \dots j]$ contiene le soluzioni ottimali di $S[i \dots k]$ e $S[k \dots j]$:
 1. $A[i \dots j] \cap S[i \dots k]$ è la soluzione ottimale di $A[i \dots k]$;
 2. $A[i \dots j] \cap S[k \dots j]$ è la soluzione ottimale di $A[k \dots j]$;

Quanto detto ci permette di definire $A[i \dots j]$ come:

$$A[i \dots j] = A[i \dots k] \cup \{k\} \cup A[k \dots j]$$

Per determinare k proviamo tutte le possibilità. A questo punto, sia $DP[i][j]$ la dimensione del più grande sottoinsieme $A[i \dots j] \subseteq S[i \dots j]$ di intervalli indipendenti. Possiamo definire $DP[i][j]$ con la seguente funzione ricorsiva:

$$DP[i][j] = \begin{cases} 0 & S[i \dots j] = \emptyset \\ \max_{k \in S[i \dots j]} \{DP[i][k] + DP[k][j] + 1\} & \text{altrimenti} \end{cases}$$

Questa definizione ci permette di realizzare un algoritmo basato su *programmazione dinamica*, o su *memoization*, con *complessità* pari a $O(n^3)$. Questo perché bisogna risolvere tutti i problemi con $i < j$ e, nel caso peggiore, paghiamo $O(n)$ per ogni sotto-problema.

Possiamo fare meglio di così?

Potremmo semplicemente usare la soluzione al problema con pesi generici, che ha *complessità* $O(n \log n)$, ponendo a 1 tutti i pesi. Tuttavia, potremmo anche chiederci se sia davvero necessario analizzare tutti i possibili valori di k . Definiamo il seguente teorema:

Definizione 130 - Scelta ingorda.

Siano $S[i \dots j]$ un sotto-problema non vuoto e m l'intervallo di $S[i \dots j]$ con il minor tempo di fine. Allora, valgono i due seguenti punti:

1. Il sotto-problema $S[i \dots m]$ è vuoto;
2. m è compreso in una qualche soluzione ottima di $S[i \dots j]$, ovvero $m \in A[i \dots j]$;

Dimostrazione. Dimostriamo separatamente i due punti.

Punto 1 - $S[i \dots j] = \emptyset$ Per la definizione di intervallo sappiamo che $a_m < b_m$. Inoltre, poiché m è l'intervallo con il minor tempo di fine, $\forall k \in S[i \dots j] \ b_m \leq b_k$ e quindi, ne consegue che $\forall k \in S[i \dots j] \ a_m < b_k$. Se nessun intervallo in $S[i \dots j]$ termina prima di a_m , allora $S[i \dots m] = \emptyset$.

Punto 2 - $m \in A[i \dots j]$ Siano $A'[i \dots j]$ una soluzione ottima di $S[i \dots j]$ e m' l'intervallo con minor tempo di fine in $A'[i \dots j]$. Sia ora $A[i \dots j] = A'[i \dots j] - \{m'\} \cup \{m\}$ una nuova soluzione ottima ottenuta togliendo m' e aggiungendo m ad $A'[i \dots j]$. $A[i \dots j]$ è una soluzione ottima che contiene m in quanto ha la stessa cardinalità di $A'[i \dots j]$ e gli intervalli sono indipendenti. \square

Questa dimostrazione ci permette di selezionare sempre gli intervalli con tempo di fine minore e ignorare tutti quelli che si intersecano con esso, riducendo il problema alla ricerca dell'*insieme indipendente massimale* in $S[m \dots j]$.

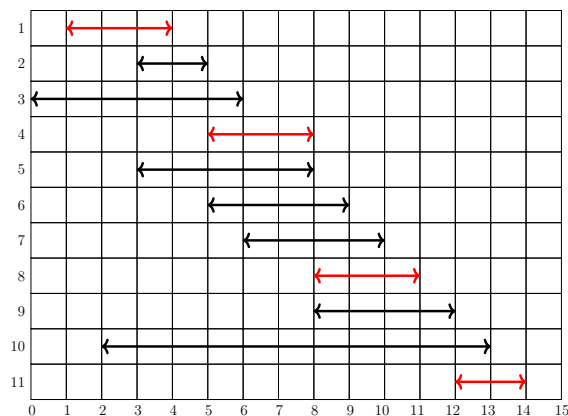


Fig. 14.2: Soluzione con *scelte ingorde*

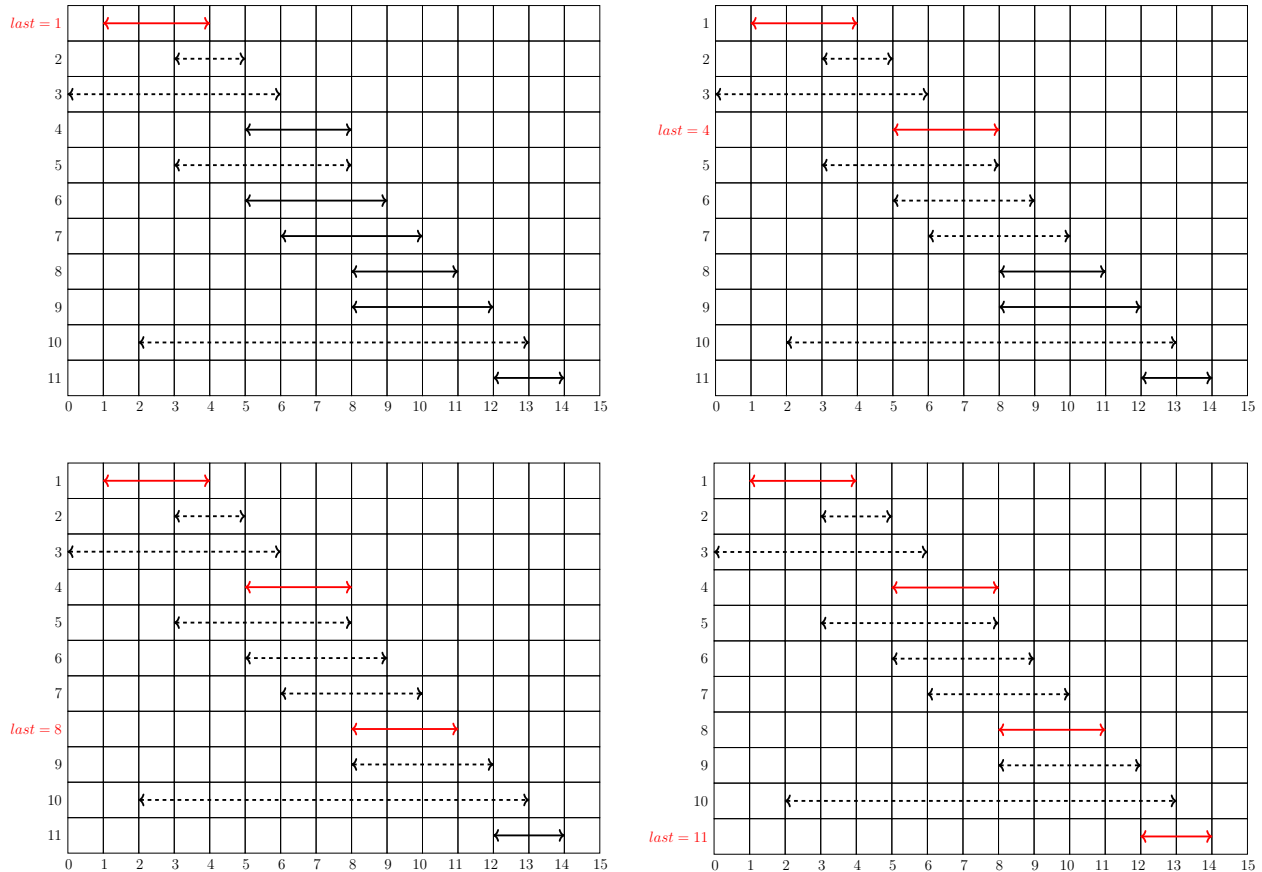
Frammento 112 - Soluzione basata su programmazione greedy.

```
SET indipendentSet(int[] a, int[] b)
{ Ordina i vettori a e b in modo che  $b_1 \leq \dots \leq b_n$  }
SET S = Set()
S.insert(1)
int last = 1
for (i = 2 to n) do
    if (a[i]  $\geq$  b[last]) then
        S.insert(i)
        last = i
return S
```

Complessità Questa soluzione abbassa la *complessità* a $O(n)$ se i vettori di input sono già ordinati, $O(n \log n)$ altrimenti.

Esempio 37 - Esempio d'esecuzione.

Il comportamento del nuovo algoritmo con gli stessi intervalli degli esempi precedenti è il seguente:



14.3 Problema dello scheduling

Problema 15 - Scheduling dei processi.

Dati un processore e n processi da eseguire su di esso. Se ogni processo i è caratterizzato da un tempo d'esecuzione $t[i]$ noto a priori, trovare una sequenza d'esecuzione che minimizzi il tempo di completamento medio.

Definizione 131 - Tempo di completamento.

Dato un vettore $A[1 \dots n]$ contenente una permutazione di $\{1 \dots n\}$, il tempo di completamento dell' h -esimo processo nella permutazione è calcolato come:

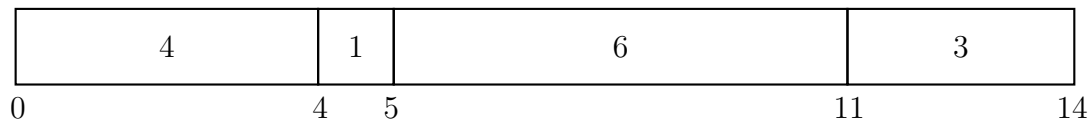
$$T_A(h) = \sum_{i=1}^h t[A[i]]$$

Esempio 38 - Possibili permutazioni.

Consideriamo i seguenti processi con relativo tempo d'esecuzione:

Processo	1	2	3	4
$t[i]$	4	1	6	3

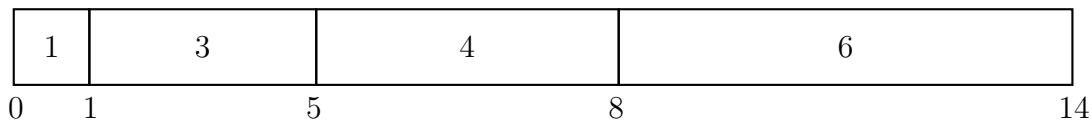
Se eseguiamo i processi nell'ordine $A[1, 2, 3, 4]$ la situazione sarebbe la seguente:



Il tempo di completamento medio sarebbe:

$$\frac{\sum_{i=1}^4 T_A(i)}{4} = \frac{4 + 5 + 11 + 14}{4} = \frac{34}{4} = 8.5$$

Se invece ordinassimo i processi per tempo d'esecuzione crescente, ponendo $A = [2, 4, 1, 3]$, la situazione diventerebbe:



Il tempo di completamento medio adesso è:

$$\frac{\sum_{i=1}^4 T_A(i)}{4} = \frac{1 + 5 + 8 + 14}{4} = \frac{27}{4} = 6.75$$

NB. Dall'esempio intuiamo che la scelta ottima potrebbe essere quella di prendere sempre il processo con minor tempo d'esecuzione.

Definizione 132 - Scelta ingorda.

Esiste una soluzione ottima A in cui il processo con minor tempo d'esecuzione si trova in prima posizione.

Definizione 133 - Sottostruttura ottima.

Sia A una soluzione ottima di un problema con n processi, in cui il processo con minor tempo d'esecuzione m si trova in prima posizione. La permutazione dei seguenti $n - 1$ processi è una soluzione al sotto-problema in cui m non viene considerato.

Dimostrazione. Consideriamo una permutazione ottima A :

$$A = [A[1], \dots, A[m], \dots, A[n]]$$

Siano m l'indice del processo con minor tempo d'esecuzione e A' una permutazione in cui i processi in posizione 1 e m vengono scambiati, ovvero:

$$A' = [A[m], \dots, A[1], \dots, A[n]]$$

Il *tempo di completamento medio* di A' è minore o uguale a quello di A . Questo è vero perché i processi nelle posizioni $1, \dots, m-1$ in A' hanno un *tempo di completamento* minore o uguale a quello dei processi in posizione $1, \dots, m-1$ in A . D'altra parte, i processi nelle posizioni m, \dots, n hanno lo stesso *tempo di completamento* sia in A che in A' .

Ora, poiché avevamo ipotizzato che A fosse una *permutazione ottima*, il suo *tempo di completamento medio* non può essere superiore a quello di A' , quindi anche A' è una *permutazione ottima*. \square

14.4 Problema dello zaino frazionario

Consideriamo una variante del *problema dello zaino* in cui è possibile prendere frazioni di oggetti.

Esempio 39 - Possibili ordinamenti degli oggetti.

Consideriamo uno zaino con capacità 70 e di avere i tre oggetti seguenti:

i	p_i	w_i
1	60	10
2	200	40
3	120	30

Per prendere gli oggetti possiamo provare a ordinarli in qualche modo così poter operare una scelta ponderata sul tipo di ordinamento applicato.

Ordinamento per profitto decrescente In questo modo andiamo a scegliere gli oggetti 2 e 3 ottenendo un guadagno totale di 320.

Ordinamento per peso crescente Questo approccio ci porta a scegliere la totalità degli oggetti 1 e 3, per un peso di 40. Per usare la rimanente capacità prendiamo i $\frac{3}{4}$ dell'oggetto 2. Il profitto che otteniamo sale quindi a 330.

Ordinamento per profitto specifico decrescente Definiamo il profitto specifico come il rapporto $\frac{p_i}{w_i}$. Per gli oggetti 1, 2, 3 il profitto specifico vale 6, 5 e 4 rispettivamente. Di conseguenza, prendiamo la totalità degli oggetti 1 e 2 e i $\frac{2}{3}$ dell'ultimo oggetto. Questa scelta ci assicura un guadagno di 340.

Frammento 113 - Implementazione della soluzione.

```
float[] zaino(float[] p, float[] w, float C, int n)
    float[] x = new float[1...n]           % Vettore delle quantità raccolte
    { Ordina p e w in modo che  $p[1]/w[1] \geq \dots \geq p[n]/w[n]$  }
    for (i = 1 to n) do
        x[i] = min(C / w[i], 1)
        C = C - x[i] * w[i]
    return x
```

Complessità La *complessità* è $O(n)$ se i vettori di input sono già ordinati, $O(n \log n)$ altrimenti.

Abbiamo definito l'algoritmo, ma in realtà non abbiamo ancora dimostrato che quel tipo di scelta sia corretta.

Dimostrazione. Assumiamo che gli oggetti siano ordinati per *profitto specifico decrescente*. Sia x una soluzione ottima e supponiamo che $x[1] < \min\left(\frac{C}{w[i]}, 1\right) \leq 1$. Allora, possiamo costruire una nuova soluzione in cui $x'[1] = \min\left(\frac{C}{w[i]}, 1\right)$ e la porzione di uno o più oggetti è ridotta di conseguenza. La soluzione x' è sicuramente di profitto uguale o maggiore rispetto ad x perché il *profitto specifico* del primo oggetto è massimo. \square

14.5 Problema della compressione

Quello della compressione dei dati è una problema classico dell'informatica. Una delle possibili tecniche per la compressione dei caratteri¹ è tramite una *funzione di codifica* del tipo $f : f(c) = x$ in cui c è un carattere preso da un alfabeto Σ e x è la sua rappresentazione binaria.

NB. Σ può essere definito come l'insieme dei caratteri usati all'interno di un file da comprimere.

Esempio 40 - Possibili compressioni.

Supponiamo di avere un file di n caratteri, che $\Sigma = \{a, b, c, d, e, f\}$ e di conoscere la frequenza di ogni carattere.

Caratteri	a	b	c	d	e	f	Dimensione
Frequenza	45%	13%	12%	16%	9%	5%	
ASCII	01100001	01100010	01100011	01100100	01100101	01100110	8n
Codifica 1	000	001	010	011	100	101	3n
Codifica 2	0	101	100	111	1101	1100	2.24n

Nella codifica 1 abbiamo codificato i caratteri usando il numero minimo di bit necessari per rappresentare 6 valori. Per quanto riguarda la codifica 2 invece, abbiamo calcolato la dimensione totale come:

$$(0.45 \cdot 1 + 0.13 \cdot 3 + 0.12 \cdot 3 + 0.16 \cdot 3 + 0.09 \cdot 4 + 0.05 \cdot 4) n = 2.24n$$

Definizione 134 - Codice a prefisso.

In un codice a prefisso, nessun codice è prefisso di un altro.

La caratteristica dei *codici a prefisso* è necessaria per consentire la decodifica.

Esempio 41 - Codici a prefisso e non.

Consideriamo un codice del tipo:

$$a \rightarrow 0, b \rightarrow 10, c \rightarrow 11$$

¹Per le immagini esistono tecniche migliori

La codifica per la stringa **babaca** è:

10 0 10 0 11 0

che può essere decodificata senza ambiguità percorrendola da sinistra, in quanto, appena viene riscontrata la corrispondenza con un carattere, si può sostituire quel carattere alla sua codifica sapendo che i bit successivi costituiranno la rappresentazione di un altro carattere.

Nel caso di un codice di questo tipo invece:

$a \rightarrow 0, b \rightarrow 1, c \rightarrow 11$

dovendo decodificare la sequenza di bit:

111111

si crea un'ambiguità, in quanto un singolo bit 1 rappresenta la *b*, mentre due bit rappresentano la *c*.

14.5.1 Algoritmo di Huffman

L'Algoritmo di Huffman utilizza rappresentazioni ad albero per definire il codice di codifica a partire dal testo e per ritornare al testo originale a partire dalla codifica. In particolare, viene usato un *albero binario* in cui al ramo sinistro di ciascun *nodo* è associato il bit 0, viceversa al ramo destro è sempre associato il bit 1. I caratteri si trovano nelle *foglie* e il *cammino* che si percorre per andare dalla *radice* a ciascuna *foglia* corrisponde alla codifica del carattere associato.

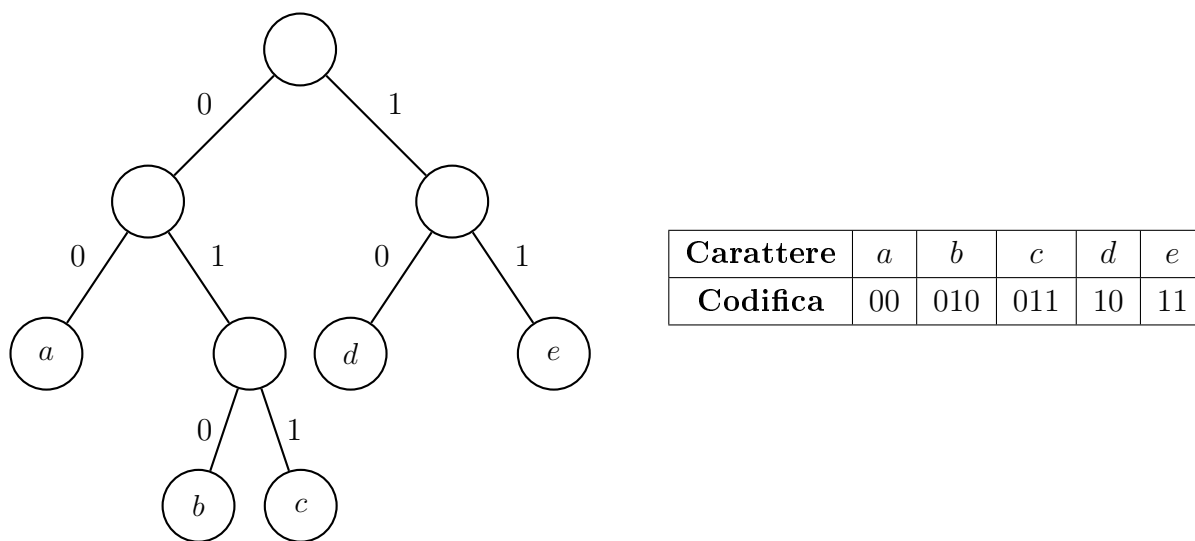


Fig. 14.3: Esempio di *albero binario di decodifica*

Questa rappresentazione ci permette di legare il numero di bit necessari per codificare un carattere alla *profondità* della *foglia* ad esso associata all'interno dell'*albero*.

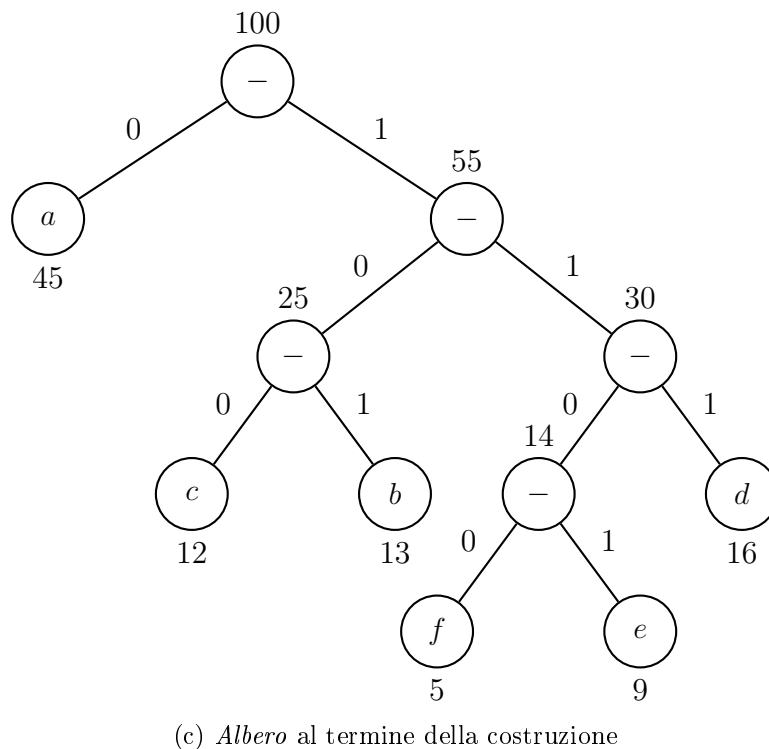
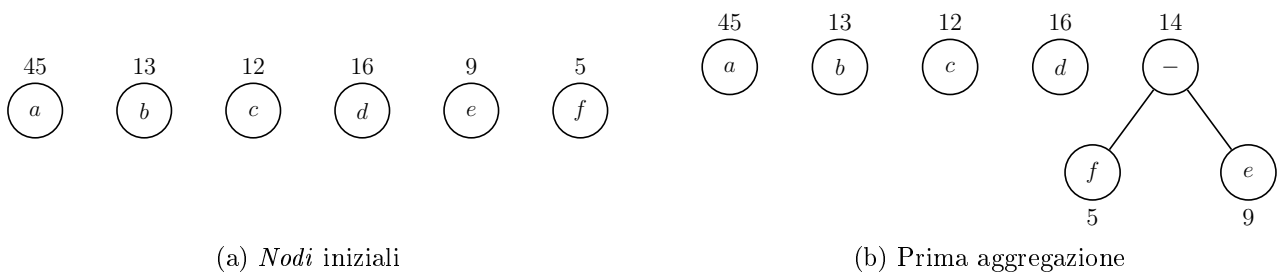
Numero di bit necessari a codificare un file In particolare, se Σ è l'alfabeto di un file F e T è un albero che rappresenta la codifica, per ogni carattere $c \in \Sigma$, $d_T(c)$ rappresenta la *profondità* della *foglia* associata a c in T . Dunque, la codifica di c richiede $d_T(c)$ bit e, se $f[c]$ è il numero di occorrenze di c in F , la dimensione della codifica per l'intero file è:

$$C(F, T) = \sum_{c \in \Sigma} f[c] \cdot d_T(c)$$

Principi dell'Algoritmo di Huffman L'idea alla base dell'*Algoritmo di Huffman* è quella di minimizzare la lunghezza delle codifiche per i caratteri che compaiono più frequentemente e, al contrario, assegnare ai caratteri con la frequenza minore i codici corrispondenti ai percorsi più lunghi all'interno dell'*albero*.

NB. Ogni codifica è associata a un solo file.

Ad ogni *nodo* dell'*albero* viene associato un carattere con la relativa frequenza. La costruzione dell'*albero* procede per aggregazione delle coppie di *nodi* con frequenza minima. Ad ogni iterazione, i due *nodi* x, y con frequenza f_x e f_y minori vengono resi figli di un nuovo *nodo* fittizio con frequenza $f_x + f_y$. Questo *nodo* viene quindi aggiunto all'insieme iniziale di caratteri. L'algoritmo continua in questo modo fino a quando non rimane un solo *nodo* e termina dopo aver etichettato gli *archi*.



La codifica ottenuta è quindi:

Carattere	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>
Codifica	0	101	100	111	1101	1100

Frammento 114 - Algoritmo di Huffman.

```

TREE huffman(int[] c, int[] f, int n)
    PRIORITYQUEUE Q = MinPriorityQueue()
    for (i = 1 to n) do
        % Inserisco nella coda il nodo c[i] con priorità f[i]
        Q.insert(f[i], Tree(f[i], c[i]))
    for (i = 1 to n - 1) do
        z1 = Q.deleteMin()
        z2 = Q.deleteMin()
        z = Tree(z1.f + z2.f, nil) % Crea il nodo fittizio
        z.insertLeft(z1)
        z.insertRight(z2)
        Q.insert(z.f, z) % Rimette il nodo fittizio nella coda
    return Q.deleteMin()

```

NB. In questo frammento ipotizziamo che ad ogni *nodo* sia possibile associare una frequenza oltre che un valore. La firma della funzione **Tree** è quindi **Tree(int frequency, ITEM value)**. Oltre a ciò, ipotizziamo che sia possibile accedere alla frequenza di un *nodo* mediante un attributo **f**.

Complessità La costruzione dell'*albero* attraverso l'uso di *code a priorità* costa $O(n \log n)$, quindi la *complessità* finale è $T(n) = O(n \log n)$.

Come al solito, dobbiamo dimostrare la correttezza dell'approccio scelto. In particolare, dobbiamo dimostrare che costruire un *albero* scegliendo sempre i due caratteri con frequenza minore conduce sempre ad una soluzione ottimale. Inoltre, dobbiamo dimostrare che questa soluzione gode di una *sottostruttura ottima*, per la quale, dato un problema sull'alfabeto Σ , è possibile costruire un sotto-problema con un alfabeto più piccolo aggregando i due caratteri con frequenza minore. Quanto detto è riassumibile nel seguente teorema:

Definizione 135 - Teorema per la correttezza dell'Algoritmo di Huffman.

Dato un alfabeto Σ e un vettore delle frequenze f , se x e y sono i due caratteri con frequenza minore, esiste un codice ottimo per Σ in cui x e y hanno la stessa profondità massima e i loro codici differiscono solo per l'ultimo bit.

Dimostrazione. Supponiamo che esista una codice ottimo T in cui i due caratteri a e b con *profondità massima* siano diversi da x e y . Possiamo assumere senza perdere di generalità che:

$$f[x] \leq f[y] \wedge f[a] \leq f[b]$$

Poiché le frequenze di x e y sono minime, vale:

$$f[x] \leq f[a] \wedge f[y] \leq f[b]$$

Se scambiamo x ed a otteniamo una soluzione T' e se poi scambiamo anche y e b otteniamo una terza soluzione T'' . A questo punto, dimostriamo che $C(f, T'') \leq C(f, T') \leq C(f, T)$:

$$\begin{aligned} C(f, T) - C(f, T') &= \sum_{c \in \Sigma} f[c] d_T(c) - \sum_{c \in \Sigma} f[c] d_{T'}(c) \\ &= (f[x] d_T(x) - f[a] d_{T'}(a)) - (f[x] d_{T'}(x) - f[a] d_{T'}(a)) \\ &= (f[x] d_T(x) - f[a] d_T(a)) - (f[x] d_T(a) - f[a] d_T(x)) \quad ^2 \\ &= (f[a] - f[x]) \cdot (d_T(a) - d_T(x)) \\ &\geq 0 \end{aligned}$$

Allo stesso modo possiamo dimostrare che $C(f, T') - C(f, T'') \geq 0$, ma poiché T è ottimo deve valere anche $C(f, T) \leq C(f, T'')$ e quindi anche T'' è una soluzione ottima. \square

14.6 Ricerca degli alberi di copertura di peso minimo

Problema 16 - Ricerca dell'albero di copertura di peso minimo.

Dati un grafo $G = (V, E)$ non orientato e una funzione di peso $w : V \times V \rightarrow \mathbb{R}$, determinare il modo di interconnettere tutti i suoi nodi che minimizzi il peso totale dei suoi archi.

NB. Questo tipo di alberi prende anche il nome di *minimum spanning tree* o *alberi di connessione di peso minimo*.

La funzione di peso w può essere meglio definita dicendo che se $(u, v) \in E$, allora $w(u, v)$ sarà pari al peso dell'arco da u a v . Altrimenti, assume valore $+\infty$. Inoltre, poiché il grafo è *non orientato*, $w(u, v) = w(v, u)$. Inoltre, siccome stiamo ricercando l'*albero di peso minimo*, il peso totale della soluzione T , definito come:

$$w(T) = \sum_{(u,v) \in E_T} w(u, v)$$

deve essere minimo rispetto al peso di tutti gli altri *alberi di copertura* definibili su quel *grafo*.

NB. L'*albero di copertura di peso minimo* non corrisponde necessariamente con l'*albero di copertura* associato ai *cammini minimi* da un *nodo* a tutti gli altri.

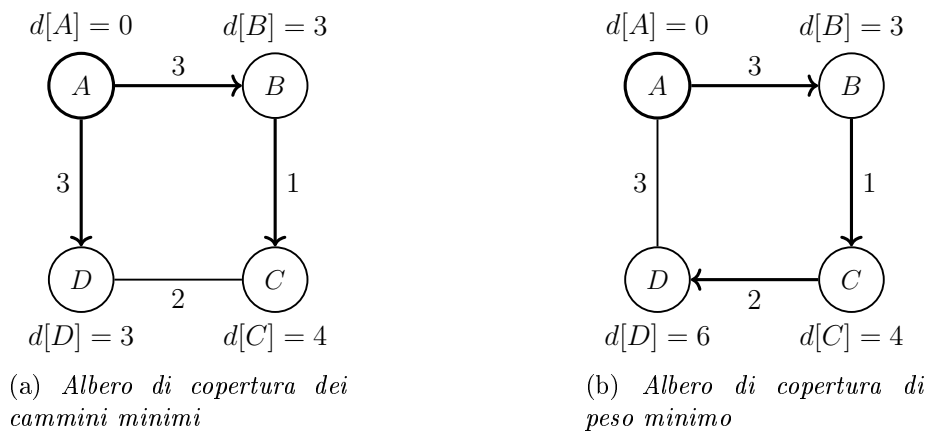


Fig. 14.4: Albero di copertura dei cammini minimi e di peso minimo

14.6.1 Approccio generico

L'approccio che intendiamo seguire è quello di partire da un sottoinsieme A di *archi* inizialmente vuoto, e di popolarlo in modo che ad ogni nuova aggiunta A corrisponda ad un sottoinsieme di un qualche *albero di copertura di peso minimo*.

Definizione 136 - Arco sicuro.

Dato un grafo $G = (V, E)$ e un insieme $A \subseteq E$, un arco $(u, v) \in E$ è detto essere sicuro per A se $A \cup \{(u, v)\}$ è ancora un sottoinsieme di un qualche albero di copertura di peso minimo.

²Siccome T' è ottenuto da T scambiando le codifiche di x e a , vale che $d_T(x) = d_{T'}(a)$ e $d_T(a) = d_{T'}(x)$

Frammento 115 - Algoritmo generico.

```
SET mst-generico(GRAPH G, int[] w)
  SET A =  $\emptyset$ 
  while ({ A non forma un albero di copertura }) do
    { Trova un arco sicuro  $(u,v)$  }
    A = A.insert(u, v)
  return A
```

Definizioni necessarie

Definizione 137 - Taglio.

Un taglio $(S, V - S)$ di un grafo non orientato $G = (V, E)$ è una partizione di V in due sottoinsiemi disgiunti.

Definizione 138 - Arco di attraversamento.

Un arco (u, v) attraversa il taglio $(S, V - S)$ se $u \in S$ e $v \in V - S$.

Definizione 139 - Taglio rispettoso.

Dati un grafo $G = (V, E)$ e un sottoinsieme $A \subseteq E$, un taglio rispetta A se nessun arco di A attraversa il taglio.

Definizione 140 - Arco leggero.

Un arco che attraversa un taglio è detto essere leggero, se il suo peso è minimo tra il peso di tutti gli archi che attraversano lo stesso taglio.

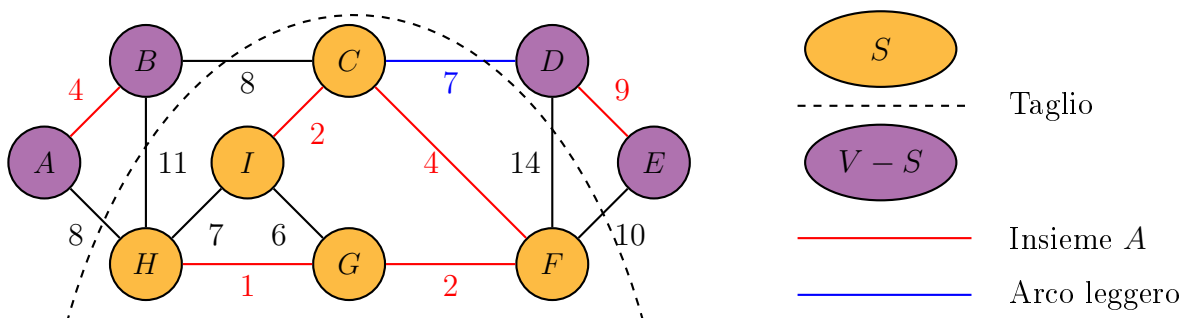


Fig. 14.5: Esempio di caratterizzazione

Definizione 141 - Teorema di caratterizzazione degli archi sicuri.

Dato un grafo $G = (V, E)$ non orientato e connesso sul quale è definita una funzione di costo $w : V \times V \rightarrow \mathbb{R}$, si considerino un insieme $A \subseteq E$ contenuto in qualche albero di copertura minimo di G e un taglio $(S, V - S)$ che rispetta A . Se (u, v) è un arco leggero che attraversa il taglio, allora è anche sicuro per A .

In parole povere, un *arco leggero* (u, v) è *sicuro* per A se $(u, v) \notin A$ e $w(u, v)$ è minimo tra i pesi di tutti gli altri *archi di attraversamento* che non appartengono ad A .

Dimostrazione. Se T è un *albero di copertura minimo* che contiene A , ci sono due casi:

1. $(u, v) \in T$: sicuramente (u, v) è sicuro per A ;
2. $(u, v) \notin T$: trasformiamo T in un *albero* T' contenente (u, v) e dimostriamo che T' è un *albero di copertura minimo*;

Procediamo quindi a dimostrare il secondo punto. Per la definizione di *albero*, u e v sono connessi da un *cammino* $C \subseteq T$ e poiché (u, v) è un *arco di attraversamento*, u e v stanno ai lati opposti del *taglio*.

Siccome T è un *albero di copertura* esiste un *arco* $(x, y) \in C$ che attraversa i due lati del *taglio*. Quindi, sia $T' = T - \{(x, y)\} \cup \{(u, v)\}$ un secondo *albero di copertura*. Poiché $w(u, v) \leq w(x, y)$, $w(T') \leq w(T)$, ma dato che T è un *albero di copertura minimo*, vale anche $w(T) \leq w(T')$. Ovvero, $w(T) = w(T')$ e $w(u, v) = w(x, y)$. \square

A questo punto, possiamo definire il seguente corollario:

Definizione 142 - Corollario.

Dato $G = (V, E)$ un grafo non orientato e connesso sul quale è definita una funzione di costo $w : V \times V \rightarrow \mathbb{R}$, si considerino un insieme $A \subseteq E$ contenuto in qualche albero di copertura minimo di G e una componente connessa della foresta $G_A = (V, A)$. Se (u, v) è un arco leggero che connette C a qualche altra componente connessa in G_A , l'arco (u, v) è sicuro per A .

14.6.2 Algoritmo di Kruskal

L'idea alla base dell'*Algoritmo di Kruskal* è di aumentare la dimensione di sottoinsiemi disgiunti di un *albero di copertura minimo* fino arrivare all'*albero* complessivo. In particolare, di volta in volta viene individuato un *arco sicuro* di peso minimo tra tutti gli *archi* che connettono due *componenti connesse* della foresta. Questo algoritmo è *greedy* perché ad ogni passo viene aggiunto all'*albero* l'*arco* di peso minimo.

Frammento 116 - Implementazione Algoritmo di Kruskal.

```
SET kruskal(EDGE[] A, int n, int m)
    SET T = Set()
    MFSET M = Mfset(n)
    { Ordina A[1,...,m] in modo che A[1].weight ≤ ... ≤ A[m].weight }
    int count = 0
    int i = 1
    % Termina quando l'albero ha n-1 archi o quando non ce ne sono più
    while (count < n - 1 and i ≤ m) do
        % Controlla se u e v sono in componenti connesse diverse
        if (M.find(A[i].u) ≠ M.find(A[i].v)) then
            M.merge(A[i].u, A[i].v)
            T.insert(A[i])
            count = count + 1
        i = i + 1
    return T
```


In pratica, l'algoritmo considera un *arco* del *grafo* alla volta, partendo da quello di peso minore e, se questo collega due *nodi* che sono in due *componenti connesse* separate, viene aggiunto agli *archi* dell'*albero di copertura minimo*. Altrimenti non viene più considerato.

Complessità La *complessità* dell'algoritmo dipende dall'implementazione usata per i *merge-find set*. Se ipotizziamo di usare la versione con *euristica sul rango* e *compressione dei cammini* le operazioni *find* e *merge* hanno *costo ammortizzato* costante.

Fase	Numero di ripetizioni	Costo
Inizializzazione	1	$O(n)$
Ordinamento	1	$O(m \log m)$
Operazioni <i>find</i> e <i>merge</i>	$O(m)$	$O(1)$

Il costo totale è:

$$T(n) = O(n + m \log m + m) = O(m \log m)$$

Poiché il numero di *archi* m è $O(n^2)$ vale:

$$O(m \log m) = O(m \log n^2) = O(m \log n)$$

14.6.3 Algoritmo di Prim

Questo algoritmo si basa sull'idea di mantenere in A un singolo *albero*. Quindi viene individuato casualmente un *nodo* r , che diventerà la *radice* dell'*albero*, a partire dal quale vengono raggiunti tutti gli altri *nodi*. Ad ogni passo viene aggiunto ad A un *arco leggero* che collega un *nodo* in V_A con un *nodo* in $V - V_A$, dove V_A è l'insieme di *nodi* raggiunti dagli *archi* in A .

L'algoritmo viene implementato facendo uso di una *coda a priorità* Q nella quale vengono mantenuti tutti i *nodi* non ancora raggiunti e sono ordinati in base a una *priorità*, definita come il peso minimo di un *arco* che collega un *nodo* dell'*albero* a un *nodo* in Q o $+\infty$ se tale *arco* non esiste. Inoltre, implementando l'*albero* usando il *vettore dei padri*, ogni *nodo* v mantiene un puntatore al proprio *padre* $p[v]$ e quindi l'insieme A è mantenuto implicitamente:

$$A = \{(v, p[v]) \mid v \in V - Q - \{r\}\}$$

Frammento 117 - Implementazione Algoritmo di Prim.

```

int[] prim(GRAPH G, NODE r)
    PRIORITYQUEUE Q = MinPriorityQueue()
    PRIORITYITEM[] pos = new PRIORITYITEM[1...G.size()]
    int[] p = new int[1...G.size()] % Vettore dei padri
    foreach (u ∈ G.V() - {r}) do
        pos[u] = Q.insert(u, +∞)
    pos[r] = Q.insert(r, 0)
    p[r] = 0
    while (not Q.isEmpty()) do
        NODE u = Q.deleteMin()
        pos[u] = nil
        foreach (v ∈ G.adj(u)) do
            if (pos[v] ≠ nil and G.w(u, v) < pos[v].priority) then
                Q.decrease(pos[v], G.w(u, v))
                p[v] = u
    return p

```

Complessità Implementando la *coda a priorità* utilizzando un *heap binario*, otteniamo la seguente tabella dei costi:

Fase	Numero di ripetizioni	Costo
Inizializzazione	1	$O(n \log n)$
deleteMin	$O(n)$	$O(\log n)$
decrease	$O(m)$	$O(\log n)$

La *complessità* totale è:

$$T(n) = O(n + n \log n + m \log n) = O(m \log n)$$

Gli algoritmi di *Prim* e *Kruskal* sono quindi asintoticamente uguali.

Ricerca locale

15.1 Introduzione

La *ricerca locale* è una tecnica risolutiva che, partendo da una soluzione ammissibile qualunque, ricerca una soluzione migliore nelle “vicinanze” di quella nota. Questo processo continua fino ad arrivare a una soluzione localmente ottima. La particolarità di questo approccio è che la soluzione trovata potrebbe non essere la migliore possibile.

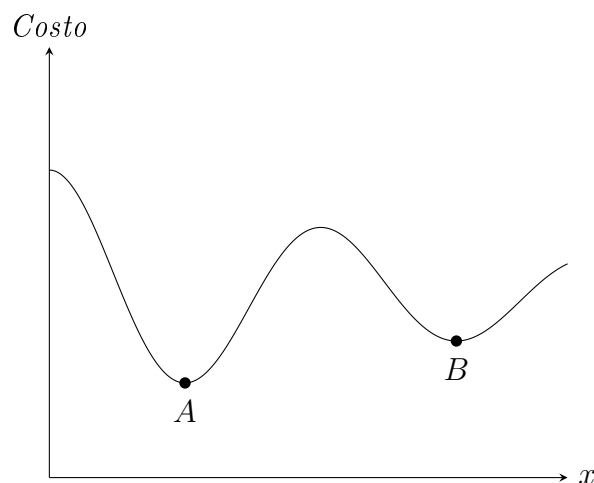


Fig. 15.1: Possibili soluzioni locali

Nel grafico di cui sopra, se la ricerca partisse da una soluzione “vicina” a B , un algoritmo basato su *ricerca locale* convergerebbe sul punto B , sebbene esso non rappresenti una soluzione globalmente ottima.

Frammento 118 - Implementazione generica di un algoritmo basato su ricerca locale.

```
SOL ricercaLocale()  
  SOL sol = { Una qualsiasi soluzione ammissibile }  
  while ( $\exists S \in I(\text{Sol})$  migliore di sol) do  
    sol = S  
  return sol
```

15.2 Problema della ricerca del flusso massimo

Definizione 143 - Rete di flusso.

Una rete di flusso $G = (V, E, s, t, c)$ è data da un grafo orientato $G = (V, E)$, da due nodi $s, t \in V$, detti rispettivamente sorgente e pozzo, e da una funzione di capacità $c : V \times V \rightarrow \mathbb{R}^+$ tale per cui $c(x, y) = 0 \ \forall (x, y) \notin E$.

NB. Assumiamo che per ogni nodo $x \in V$ esista un cammino da s in t passante per esso e ignoriamo qualsiasi nodo che non goda di questa proprietà.

Definizione 144 - Flusso.

Un flusso in G è una funzione $f : V \times V \rightarrow \mathbb{R}$ che soddisfa le seguenti proprietà:

1. Vincolo sulla capacità: $f(x, y) \leq c(x, y) \ \forall x, y \in V$;
2. Antisimmetria: $f(x, y) = -f(y, x) \ \forall x, y \in V$;
3. Conservazione del flusso: $\sum_{y \in V} f(x, y) = 0 \ \forall x \in V - \{s, t\}$;

La prima proprietà stabilisce che un flusso non può mai eccedere la capacità di un arco. In particolare, se la capacità di un arco tra due nodi A e B è 7, e $f(A, B) = 5$, nei grafi useremo la seguente notazione:

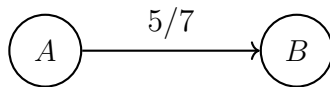


Fig. 15.2: Vincolo sulla capacità

La proprietà di *antisimmetria* ci dice invece che il flusso che attraversa un arco in direzione (x, y) è l'opposto del flusso che attraversa lo stesso arco in direzione (y, x) .

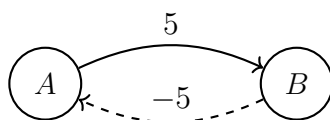


Fig. 15.3: Antisimmetria

Infine, la proprietà di *conservazione del flusso* ci dice che per ogni nodo diverso dalla sorgente e dal pozzo, la somma dei flussi entranti ed uscenti è sempre pari a 0.

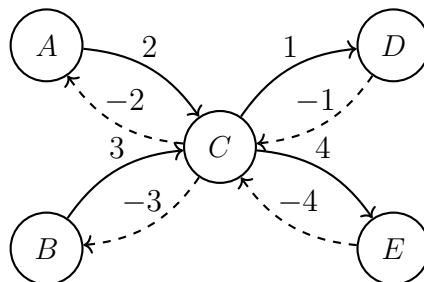


Fig. 15.4: Conservazione del flusso

Definizione 145 - Valore del flusso.

Il valore di un flusso f è definito come:

$$|f| = \sum_{(s,x) \in E} f(s,x)$$

ovvero come la quantità di flusso uscente da s .

Date tutte queste definizioni possiamo finalmente enunciare il problema in questione.

Problema 17 - Problema della ricerca del flusso massimo.

Data una rete $G = (V, E, s, t, c)$, trovare un flusso che abbia valore massimo fra tutti i flussi associabili alla rete. In particolare, ricercare il flusso f^* tale per cui $|f^*| = \max\{|f|\}$.

15.2.1 Metodo delle reti residue

L'algoritmo che intendiamo realizzare è detto *metodo delle reti residue* perché è basato su un approccio additivo. In particolare, partendo da un flusso f inizialmente nullo, si ripetono le seguenti operazioni:

1. "Sottrazione" del flusso f alla rete attuale e acquisizione di una *rete residua*;
2. Ricerca di un secondo flusso g all'interno della *rete residua*;
3. Somma di g ad f ;

fino a quando g non diventa un flusso nullo.

NB. Questo particolare algoritmo permette sempre di ottenere il *flusso massimo*, quindi non soffre del problema al quale avevamo accennato inizialmente.

Diamo ora delle altre definizioni:

Definizione 146 - Flusso nullo.

È definito *flusso nullo* la funzione $f_0 : V \times V \rightarrow \mathbb{R}^+$ tale che:

$$f_0(x,y) = 0 \quad \forall x,y \in V$$

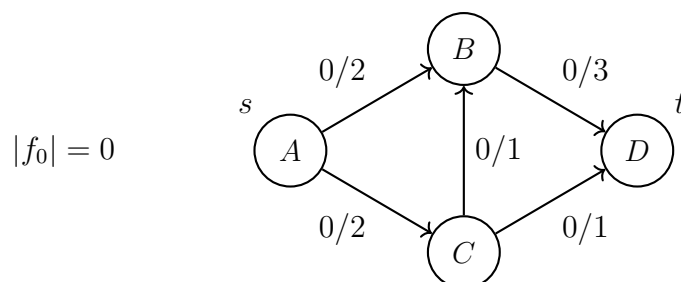


Fig. 15.5: *Flusso nullo*

Definizione 147 - Somma di flussi.

Per ogni coppia di flussi f_1 e f_2 in G , definiamo il flusso somma $g = f_1 + f_2$ come un flusso tale che:

$$g(x, y) = f_1(x, y) + f_2(x, y) \quad \forall x, y \in V$$

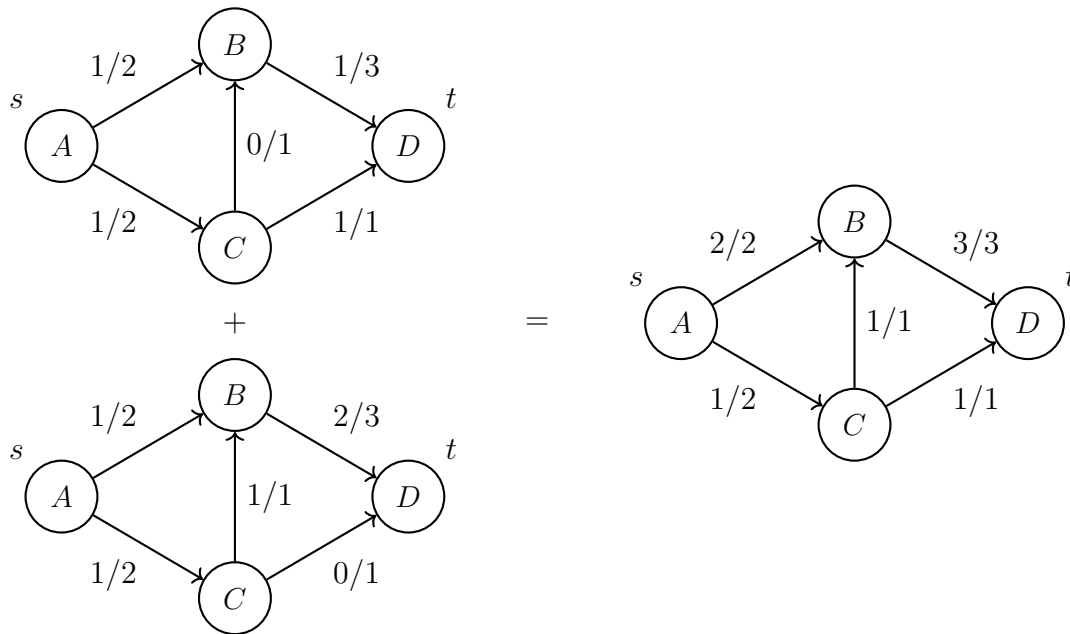


Fig. 15.6: Somma di flussi

Definizione 148 - Capacità residua.

È definita capacità residua di un flusso f in una rete $G = (V, E, s, t, c)$, la funzione $c_f : V \times V \rightarrow \mathbb{R}^+$ tale che:

$$c_f(x, y) = c(x, y) - f(x, y) \quad \forall x, y \in V$$

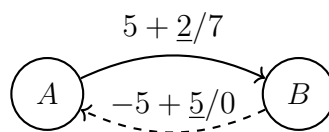


Fig. 15.7: Capacità residua

Possiamo notare che gli *archi all'indietro*, che avevamo visto enunciando la proprietà di *antisimmetria* dei flussi, si originano dalla definizione di *capacità residua*. Infatti, poiché l'arco (B, A) non esiste, $c(B, A) = 0$ e quindi, vale la seguente catena di uguaglianze:

$$\begin{aligned} c_f(B, A) &= c(B, A) - f(B, A) && \text{Dalla definizione di capacità residua} \\ &= c(B, A) - (-f(A, B)) && \text{Per la proprietà di } \textcolor{red}{\textit{antisimmetria}} \\ &= 0 - (-5) = 5 \end{aligned}$$

Definizione 149 - Rete residua.

Data una rete di flusso $G = (V, E, s, t, c)$ e un flusso f su G , una rete $G_f = (V, E_f, s, t, c_f)$ è una rete residua di G rispetto ad f , se è tale per cui:

$$c_f(x, y) > 0 \quad \forall x, y \in V : (x, y) \in E_f$$

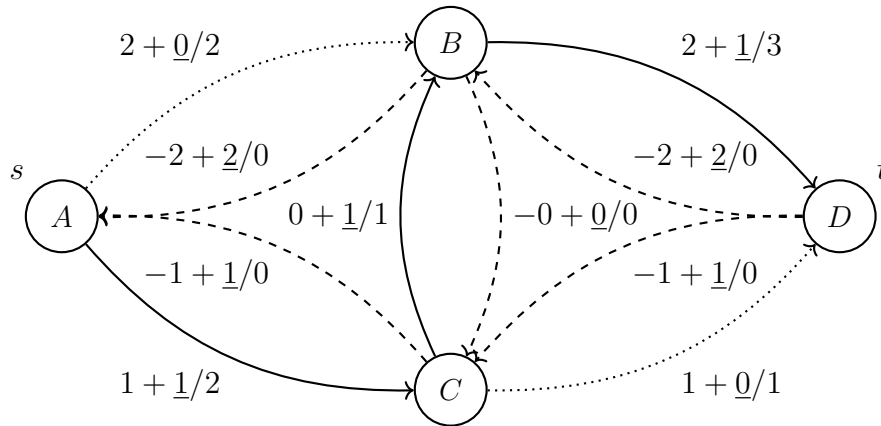


Fig. 15.8: Rete residua

Gli archi (A, B) e (C, D) sono stati saturati, ovvero $f(A, B) = c(A, B)$ e $f(C, D) = c(C, D)$. Di conseguenza, la loro capacità residua è 0 e quindi non fanno parte della rete residua.

Frammento 119 - Schema generale dell'algoritmo.

```
int[] [] maxFlow(GRAPH G, NODE s, NODE t, int[] [] c)
    FLOW f = f0                                     % Inizializza f a un flusso nullo
    int[] [] cf = c   % Inizialmente la capacità residua è pari a quella iniziale
    FLOW g
    do
        g = { trova un flusso in cf tale che |g| > 0, altrimenti f0 }
        f = f + g
        cf = { capacità della rete residua di G rispetto ad f }
    while (g ≠ f0)
    return f
```

Dimostrazione Per dimostrare la correttezza dell'algoritmo dobbiamo dimostrare il seguente lemma:

Definizione 150 - Lemma della somma di flussi.

Se f è un flusso in G e g è un flusso in G_f , allora anche $f + g$ è un flusso in G .

Dimostrazione. Dobbiamo dimostrare che $f + g$ rispetta tutte e tre le proprietà dei flussi.

Vincolo sulla capacità Per ogni coppia di nodi $x, y \in V$ vale la seguente catena di disuguaglianze:

$$\begin{aligned} g(x, y) &\leq c_f(x, y) && \text{È vero perché } g \text{ è un flusso in } G_f \\ f(x, y) + g(x, y) &\leq c_f(x, y) + f(x, y) \\ (f + g)(x, y) &\leq (c(x, y) - f(x, y)) + f(x, y) && \text{Definizione di capacità residua} \\ (f + g)(x, y) &\leq c(x, y) && \text{Semplificazione algebrica} \end{aligned}$$

Antisimmetria Per ogni coppia di *nodi* $x, y \in V$ vale la seguente catena di uguaglianze:

$$\begin{aligned} f(x, y) + g(x, y) &= -f(y, x) - g(y, x) \\ f(x, y) + g(x, y) &= -(f(y, x) + g(y, x)) \\ (f + g)(x, y) &= -(f + g)(y, x) \end{aligned}$$

Conservazione del flusso Per ogni coppia di *nodi* $x, y \in V$ vale la seguente catena di uguaglianze:

$$\begin{aligned} \sum_{y \in V} (f + g)(x, y) &= \sum_{y \in V} (f(x, y) + g(x, y)) \\ &= \sum_{y \in V} f(x, y) + \sum_{y \in V} g(x, y) \\ &= 0 + 0 \\ &= 0 \end{aligned}$$

□

15.2.2 Metodo dei cammini aumentanti

Quindi l'algoritmo che abbiamo descritto è corretto, tuttavia rimane un problema: come facciamo a trovare il *flusso* g ?

Una soluzione è quella proposta da Ford e Fulkerson nel 1956, e raffinata poi da Edmonds e Karp nel 1972, che fa uso dei cosiddetti *cammini aumentanti*. In particolare, viene identificato nella *rete residua* G_f un *cammino* $C = (v_0, \dots, v_n)$, tale per cui $v_0 = s$ e $v_n = t$. Quindi, si definisce la *capacità* del *cammino* come la minore tra le *capacità* degli *archi* attraversati, ovvero:

$$c_f(C) = \min_{i \in \{2, \dots, n\}} c_f(v_{i-1}, v_i)$$

NB. La *capacità* di un *cammino* è definibile anche come la *capacità* del suo “collo di bottiglia”.

Infine, il *flusso* *addizionale* g è definito in modo che $g(v_{i-1}, v_i) = c_f(C)$, $g(v_i, v_{i-1}) = -c_f(C)$ per la proprietà *antisimmetrica* e $g(x, y) = 0$ per tutte le coppie (x, y) in cui $x \notin C$ o $y \notin C$.

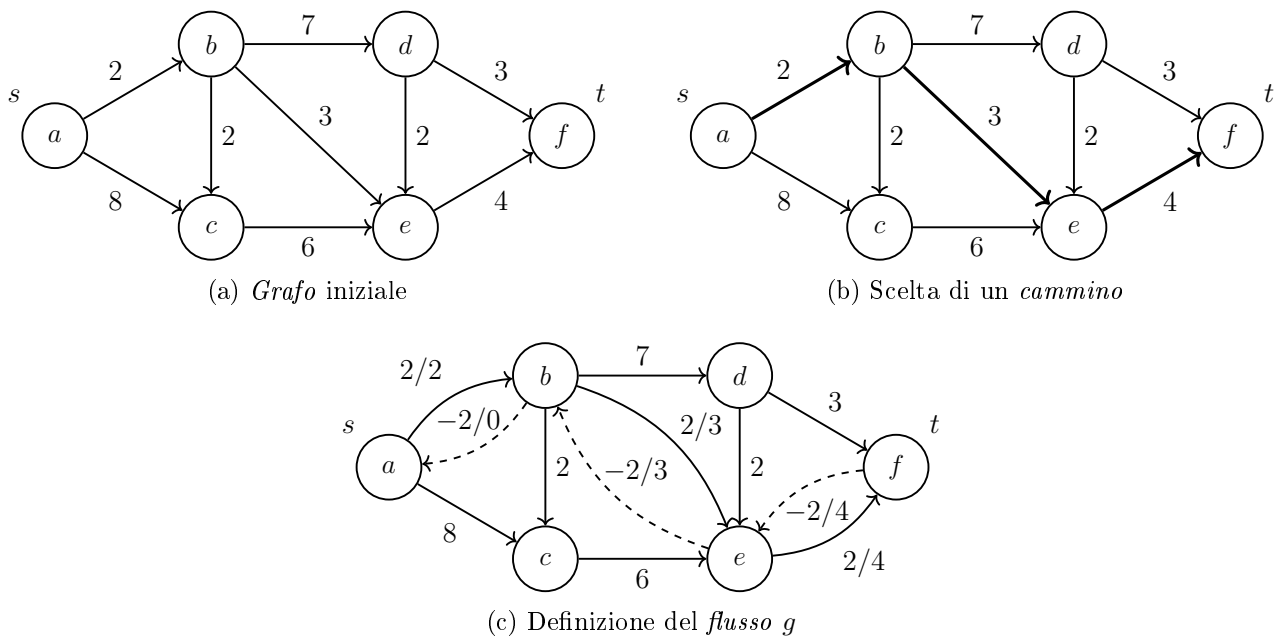


Fig. 15.9: *Cammini aumentanti* in generale

Frammento 120 - Schema generale dell'algoritmo.

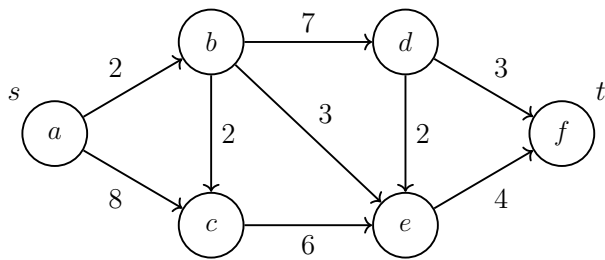
```

int[] [] maxFlow(GRAPH G, NODE s, NODE t, int[] [] c)
    int[] [] f = new int[1...G.size()][1...G.size()]           % Flusso parziale
    int[] [] g = new int[1...G.size()][1...G.size()]           % Flusso da cammino aumentante
    int[] [] cf = new int[1...G.size()][1...G.size()]           % Capacità residua
    foreach (x,y ∈ G.V()) do
        f[x][y] = 0                                           % Inizializza un flusso nullo
        cf[x][y] = c[x][y]
    do
        g = { flusso associato ad un cammino aumentante in cf, oppure f0 }
        foreach (x,y ∈ G.V()) do
            f[x][y] = f[x][y] + g[x][y]                       % Somma dei flussi: f = f + g
            cf[x][y] = c[x][y] - f[x][y]                     % Aggiorna la capacità residua
        while (g ≠ f0)
    return f

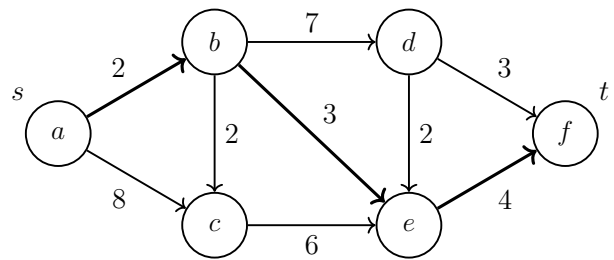
```

Esempio 42 - Esempio d'esecuzione.

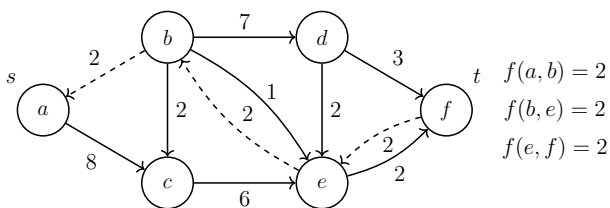
Supponiamo di eseguire l'algoritmo sullo stesso grafo di prima:



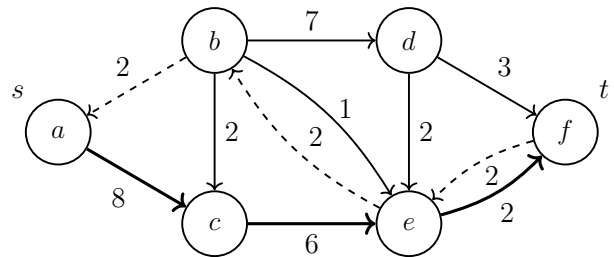
(a) Grafo iniziale e $f = f_0$



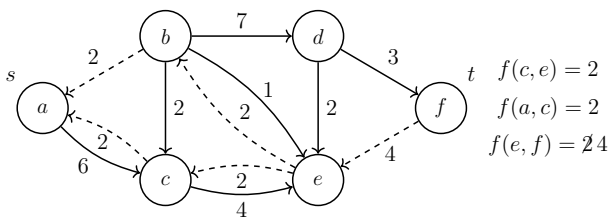
(b) Scelta di un cammino in G_f



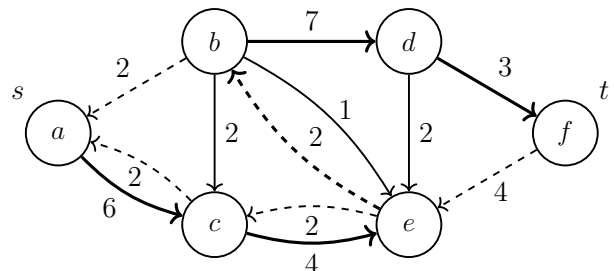
(c) Aggiornamento di f



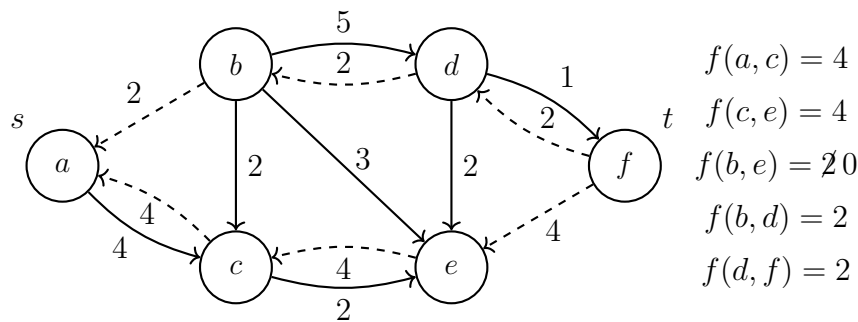
(d) Scelta di un cammino in G_f



(e) Aggiornamento di f



(f) Scelta di un cammino in G_f



(g) Aggiornamento di f

Nel passaggio (f), il cammino scelto passa per l'arco (e, b) che però è un arco all'indietro, quindi $f(e, b) = -2$. Il motivo per il quale possiamo usarlo è che $c_f(e, b) = c(e, b) - f(e, b) = 2$, tuttavia, usarlo significa “annullare” il precedente flusso attraverso (b, e) e infatti al passo (g) $f(b, e) = 0$. Questo “tornare indietro” è caratteristico degli algoritmi basati su ricerca locale.

Rimane comunque il problema di ricercare un *cammino*, ma per quello possiamo usare una *visita in ampiezza* o in *profondità*.

Frammento 121 - Implementazione maxFlow con metodo dei cammini aumentanti.

```
int[] [] maxFlow(int[] [] c, int n, NODE s, NODE t)
    int[] [] f = new int[1...n][1...n]                                % Matrice del flusso
    boolean[] visited = new int[1...n] = {false}
    % Rimane nel ciclo fino a quando esistono cammini da s a t
    while (dfs(c, f, n, s, t, visited, +∞) > 0) do
        for (i = 1 to n) do
            visited[i] = false    % Resetta il vettore per la visita successiva
    return f

% Versione modificata della dfs
int dfs(int[] [] c, int[] [] f, int n, int s, int t, boolean[] visited, int minCap)
    if (s == t) then
        return min
    visited[s] = true
    for (i = 1 to n) do
        if (c[s][i] > 0 and not visited[i]) then
            int val = dfs(c, f, n, i, t, visited, min(minCap, c[s][i]))
            if (val > 0) then
                c[s][i] = c[s][i] - val    % Aggiorna la capacità dell'arco (s,i)
                c[i][s] = c[i][s] + val    % Aggiorna la capacità dell'arco (i,s)
                f[s][i] = f[s][i] + val    % Aggiorna la capacità del flusso in (s,i)
                f[i][s] = f[i][s] - val    % Aggiorna la capacità del flusso in (i,s)
            return val
    return 0    % Se l'esecuzione arriva qui non è stato trovato un cammino
```

Note sull'implementazione Il grafo è definito esclusivamente mediante la *matrice delle capacità*. In particolare, esiste un arco (x, y) solo se $c[x][y] > 0$. Il parametro n rappresenta il numero di *nodi*. Nella *dfs*, il parametro *minCap* indica il più basso valore di *capacità* incontrato nella ricerca del *cammino*.

Dimostrazione La dimostrazione di correttezza per l'algoritmo ci richiede di dare delle altre definizioni.

Definizione 151 - Taglio.

Un taglio (S, T) della rete di flusso $G = (V, E, s, t, c)$ è una partizione di V in due sottoinsiemi disgiunti tali per cui $S = V - T$, $s \in S$ e $t \in T$.

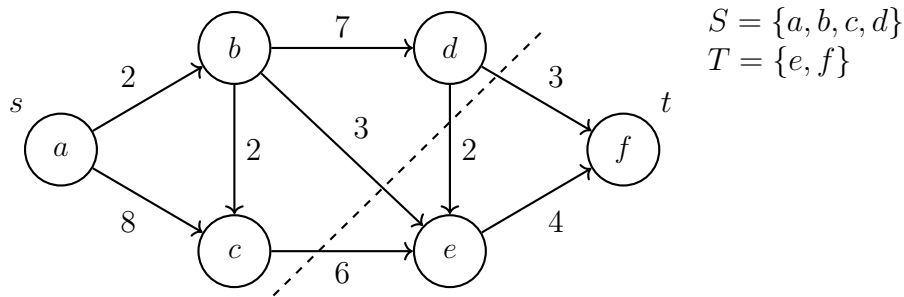


Fig. 15.9: Taglio di una rete di flusso

Definizione 152 - Capacità di un taglio.

La capacità $C(S, T)$ attraverso il taglio (S, T) è definita come:

$$C(S, T) = \sum_{x \in S, y \in T} c(x, y)$$

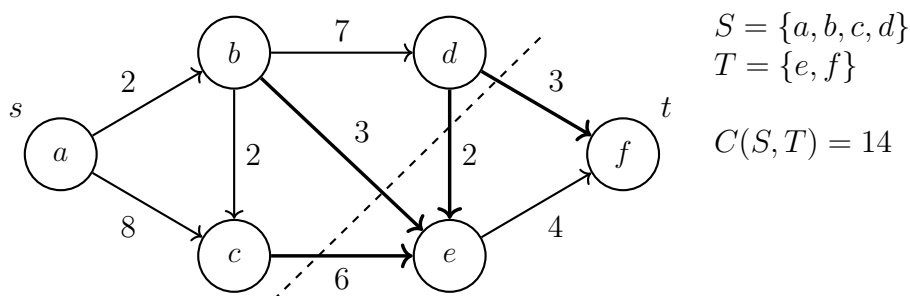


Fig. 15.10: Capacità di un taglio

Definizione 153 - Flusso netto di un taglio.

Se f è un flusso in G , il flusso netto $F_f(S, T)$ attraverso (S, T) è:

$$F_f(S, T) = \sum_{x \in S, y \in T} f(x, y)$$

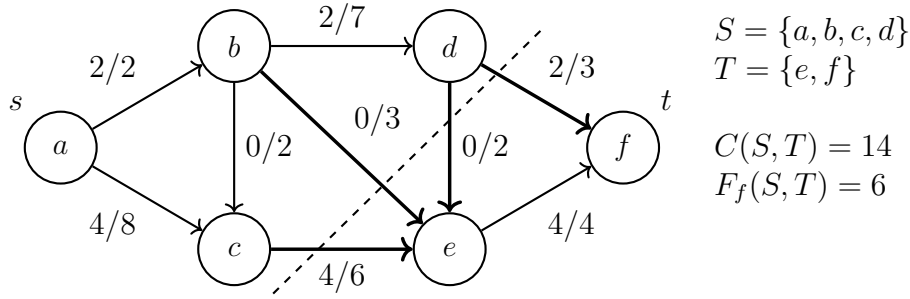


Fig. 15.11: *Flusso netto di un taglio*

Definizione 154 - Lemma del valore del flusso di un taglio.

Dati un flusso f e un taglio (S, T) , il flusso netto $F_f(S, T)$ che attraverso il taglio è uguale a $|f|$.

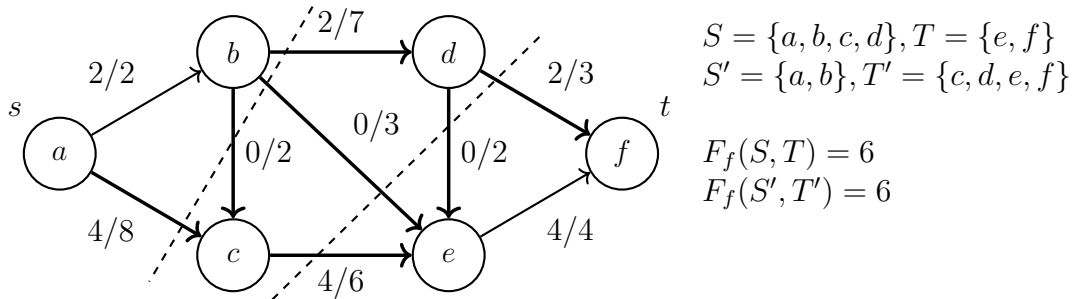


Fig. 15.12: *Valore del flusso di un taglio*

Dimostrazione. Vale la seguente catena di uguaglianze:

$$\begin{aligned}
 F_f(S, T) &= \sum_{x \in S, y \in T} f(x, y) \\
 &= \sum_{x \in S, y \in V} f(x, y) - \sum_{x \in S, y \in S} f(x, y) && T = V - S \\
 &= \sum_{x \in S, y \in V} f(x, y) - 0 && \text{Antisimmetria} \\
 &= \sum_{x \in S - \{s\}, y \in V} f(x, y) + \sum_{y \in V} f(s, y) && s \in S \\
 &= \sum_{x \in S - \{s\}} \left(\sum_{y \in V} f(x, y) \right) + \sum_{y \in V} f(s, y) \\
 &= \sum_{x \in S - \{s\}} 0 + \sum_{y \in V} f(s, y) && \text{Conservazione del flusso} \\
 &= \sum_{y \in V} f(s, y) = |f| && \text{Definizione di valore del flusso}
 \end{aligned}$$

□

Definizione 155 - Lemma del limite superiore al flusso massimo.

Il flusso massimo è limitato superiormente dalla capacità del taglio minimo, cioè del taglio con minor capacità.

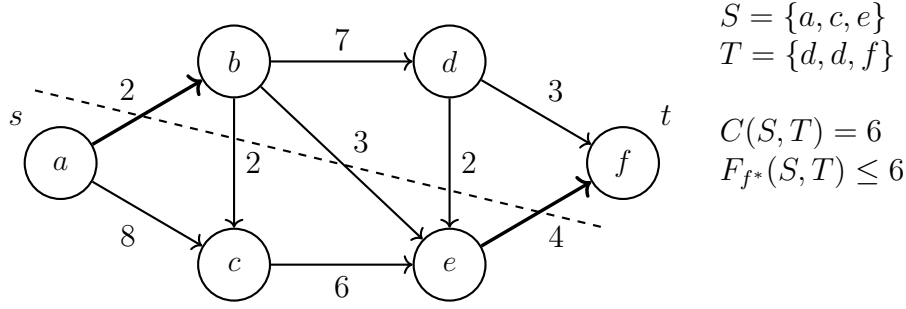


Fig. 15.13: Limite superiore al *flusso massimo*

Dimostrazione. Dobbiamo dimostrare che nessun *flusso netto* supera la *capacità del taglio* attraversato. Ovvero, dobbiamo verificare la seguente disuguaglianza:

$$F_f(S, T) \leq C(S, T) \quad \forall f \wedge \forall (S, T)$$

Per la proprietà del *vincolo sulla capacità* vale quanto segue:

$$F_f(S, T) = \sum_{x \in S, y \in T} f(x, y) \leq \sum_{x \in S, y \in T} c(x, y) = C(S, T)$$

Inoltre, il *flusso* che attraversa un *taglio* è uguale al *valore del flusso*:

$$|f| = F_f(S, T) \quad \forall f \wedge \forall (S, T)$$

quindi il *valore del flusso* è limitato superiormente dalla capacità di tutti i possibili *tagli*:

$$|f| \leq C(S, T) \quad \forall f \wedge \forall (S, T)$$

□

Definizione 156 - Teorema del taglio minimo-flusso massimo.

Le seguenti tre affermazioni sono equivalenti:

1. f è un *flusso massimo*;
2. Non esiste alcun *cammino aumentante* nella rete G_f ;
3. Esiste un *taglio minimo* (S, T) tale per cui $C(S, T) = |f|$;

Dimostrazione. Procediamo dimostrando circolarmente che ogni punto implica gli altri.

(1) \Rightarrow (2) Se esistesse un *cammino aumentante*, il *flusso* potrebbe essere aumentato e quindi non sarebbe massimo, generando un assurdo.

(2) \Rightarrow (3) Poiché non esiste alcun *cammino aumentante* nella *rete residua* G_f , non esiste un *cammino* da s a t . Se S è l'insieme dei *nodi* raggiungibili da s e $T = V - S$, (S, T) è un *taglio* perché $s \in S$ e $t \in T$, ma siccome t non è raggiungibile da s , tutti gli *archi* (x, y) con $x \in S$ e $y \in T$ sono saturi, ovvero $f(x, y) = c(x, y)$. Per il **Lemma del valore del flusso di un taglio** vale:

$$|f| = \sum_{x \in S, y \in T} f(x, y)$$

Ne segue che:

$$|f| = \sum_{x \in S, y \in T} f(x, y) = \sum_{x \in S, y \in T} c(S, T) = C(S, T)$$

Poiché $|f| = C(S, T)$ e grazie al *Lemma del limite superiore al flusso massimo* sappiamo che $|f| \leq C(S', T')$ per ogni *taglio* (S', T') , possiamo affermare che (S, T) è un *taglio* di *capacità* minima.

(3) \Rightarrow (1) Poiché per ogni *flusso* f e ogni *taglio* (S, T) vale la relazione $|f| \leq C(S, T)$, il *flusso* che soddisfa $|f| = C(S, T)$ deve essere massimo. \square

15.2.3 Complessità

Limite superiore nell'Algoritmo di Ford-Fulkerson Nell'*Algoritmo di Ford-Fulkerson* la ricerca del *flusso massimo* parte da un *flusso nullo* che viene di volta in volta aumentato fino ad arrivare al *flusso massimo*. Poiché ogni incremento può essere unitario, il numero di incrementi è $O(|f^*|)$. Ogni incremento richiede una *visita del grafo*, la *somma tra due flussi* e la determinazione della *rete residua*, e ognuna di queste operazioni costa $O(V + E)$, quindi la *complessità* finale del caso peggiore è $O((V + E)|f^*|)$.

NB. Dato che $E = \Omega(V) = O(V^2)$, possiamo esprimere la *complessità* come $O(E|f^*|)$ o $O(V^2|f^*|)$.

Limite superiore nell'Algoritmo di Edmonds-Karp Edmonds e Karp suggerirono di usare una *visita in ampiezza* invece di una *visita generica*. Gli *aumenti di flusso* vengono eseguiti $O(VE)$ volte e per ognuno di essi si paga il costo $O(V + E)$ di una *visita in ampiezza*. La *complessità* finale è quindi $O(VE(V + E))$, ma siccome $E = \Omega(V)$ perché ogni *nodo*, ad eccezione di s e t , ha almeno un *arco entrante* e un *arco uscente*, possiamo semplificare scrivendo che la *complessità* è $O(VE^2)$ o $O(V^5)$.

Dimostrazione

Definizione 157 - Monotonia.

Siano $\delta_f(s, x)$ la distanza minima da s a x in una rete residua G_f e $f' = f + g$ un *flusso* nella rete iniziale con g non nullo e derivante da un cammino aumentante. Allora, $\delta_{f'}(s, x) \geq \delta_f(s, x)$.

Dimostrazione. Questo comportamento è facile da dimostrare in quanto ad ogni aumento di *flusso* alcuni archi si “spengono” perché la loro *capacità residua* scende a 0. Poiché quegli archi erano utilizzati nei cammini minimi, visite successive non potranno individuare cammini più corti. \square

Dimostrata la *monotonia*, possiamo dimostrare formalmente che il numero di aumenti di *flusso* è limitato superiormente da $O(VE)$.

Dimostrazione. Siano G_f una rete residua e C un cammino aumentante di G_f . (x, y) è un arco critico, o un collo di bottiglia, in C se:

$$c_f(x, y) = \min_{(u, v) \in C} \{c_f(u, v)\}$$

Ogni cammino contiene almeno un arco critico e ogni volta che il *flusso* associato a C viene aggiunto, tutti gli archi critici scompaiono dalla rete residua.

Poiché i *cammini aumentanti* sono *minimi*, abbiamo che:

$$\delta_f(s, y) = \delta_f(s, x) + 1$$

Una volta che l'*arco* (x, y) sarà stato *saturato*, potrà essere riutilizzato soltanto se il *flusso* attraverso quell'*arco* diminuisce. Ovvero, solo se verrà utilizzato l'*arco* (y, x) . Se g è il *flusso* quando ciò accade, vale:

$$\delta_g(s, x) = \delta_g(s, y) + 1$$

Per la *monotonia* sappiamo anche che $\delta_g(s, y) \geq \delta_f(s, y)$, quindi vale quanto segue:

$$\begin{aligned} \delta_g(s, x) &= \delta_g(s, y) + 1 \\ &\geq \delta_f(s, y) + 1 \\ &= \delta_f(s, x) + 1 + 1 \end{aligned}$$

Questo significa che dal momento in cui un *arco critico* viene *saturato* a quando torna *critico*, il *cammino minimo* passante per esso si è allungato di almeno due passi. Di conseguenza, la lunghezza *massima* del *cammino* fino a x , considerando che poi si deve attraversare l'*arco* (x, y) , è al massimo $V - 2$. Ne consegue che un *arco* può diventare *critico* un massimo di $\frac{V-2}{2} = \frac{V}{2} - 1$ volte e, poiché ci sono $O(E)$ *archi* che possono diventare *critici* $O(V)$ volte, il numero massimo di *flussi aumentanti* è $O(VE)$. \square

Riassunto delle complessità

Nome	Complessità	Note
Ford-Fulkerson	$O(E f^*)$	Converge con valori razionali
Edmonds-Karp	$O(VE^2)$	Specializzazione basata su BFS
Dinic, blocking flow	$O(V^2E)$	In alcune reti particolari $O(\min(V^{2/3}, E^{1/2})E)$
MPN	$O(V^3)$	Solo su DAG
Dinic	$O(VE \log V)$	Usa Dynamic tree
Goldberg e Rao	$O(\min(V^{2/3}E^{1/2})E \log(\frac{V^2}{E} + 2) \log C)$	$C = \max_{(x,y) \in E} c(x, y)$
Orlin, King Rao, Tarjan	$O(VE)$	Pubblicato nel 2013

Capitolo Nr.16

Backtracking

16.1 Introduzione

La tecnica del *backtrack* è usata per risolvere problemi in cui è necessario esplorare l'intero spazio delle soluzioni o quando la soluzione va ricercata in un insieme molto ampio. Andando maggiormente nello specifico, il *backtracking* può essere utilizzato nelle seguenti categorie di problemi:

- *Enumerazioni*: problemi che richiedono di elencare tutte le soluzioni ammissibili;
- *Conteggio*: problemi che richiedono di contare tutte le soluzioni ammissibili;
- *Ricerca*: problemi che richiedono di trovare una soluzione ammissibile in uno spazio delle soluzioni molto grande;
- *Ottimizzazione*: problemi che richiedono di trovare una delle soluzioni ammissibili ottime rispetto ad un certo criterio di valutazione;

NB. Il *backtracking* è paragonabili alla tecnica del *brute force*.

Il problema che sorge quando si sceglie di analizzare l'intero spazio delle soluzioni è che questo potrebbe avere una dimensione *superpolinomiale* e richiedere un tempo inaccettabile per essere esaminato. Per questo motivo, è consigliato usare il *backtracking* solo quando è l'unica via possibile.

Ciò che ci interessa però, è che la tecnica del *backtrack* ci fornisce un approccio sistematico all'esplorazione di uno spazio di ricerca, utilizzando la ricorsione per memorizzare le scelte fatte e consentendoci di "ritornare sui nostri passi" qualora quelle scelte si rivelassero errate o non ottimali.

NB. Per alcuni problemi è possibile realizzare anche soluzioni iterative unendo approcci *greedy* alla possibilità di annullare le scelte fatte.

16.2 Problema dell'enumerazione dei sottoinsiemi

Problema 18 - Problema dell'enumerazione dei sottoinsiemi.

Dato un insieme $\{1, \dots, n\}$ elencarne tutti i sottoinsiemi.

Rappresentiamo una soluzione come un vettore di scelte $S[1 \dots n]$ nel quale il contenuto di $S[i]$ è preso tra i valori dell'insieme originale non ancora scelti. In particolare, l'indice i rappresenta l'indice della prossima decisione da prendere e la soluzione parziale $S[1 \dots i - 1]$ contiene tutte le decisioni prese finora.

Procedendo in modo ricorsivo possiamo dire che nel caso base $S[1 \dots i - 1]$ è una soluzione ammissibile e quindi può essere processata facendo terminare la ricorsione. Altrimenti, calcoliamo l'insieme C delle scelte possibili, quindi, per ogni $c \in C$ scriviamo c nella scelta $S[i]$ e continuiamo la ricerca per $i + 1$.

16.2.1 Approccio generale

Come precedentemente accennato, il *backtracking* ci fornisce un modo sistematico di analizzare uno spazio delle soluzioni. Questo si traduce nella possibilità di definire, per ogni classe di problema, uno schema generale per l'algoritmo risolutivo. Il problema dell'*enumerazione dei sottoinsiemi* appartiene alla classe dei *problemi di enumerazione*, e l'algoritmo risolutivo segue grosso modo il seguente schema:

Frammento 122 - Schema generale per algoritmi di enumerazione.

```
enumeration(<dati problema>, ITEM[] S, int i, <dati parziali>)
% Verifica se  $S[1 \dots i - 1]$  contiene una soluzione ammissibile
if (accept(<dati problema>, S, i, <dati parziali>))
    % "Processa" la soluzione (e.g. stampa, conta, ...)
    processSolution(<dati problema>, S, i, <dati parziali>)
else
    % Calcola l'insieme delle scelte in funzione di  $S[1 \dots i - 1]$ 
    SET C = choices(<dati problema>, S, i, <dati parziali>)
    % Itera sull'insieme delle scelte
    foreach ( $c \in C$ ) do
         $S[i] = c$ 
        % Chiamata ricorsiva
        enumeration(<dati problema>, S, i + 1, <dati parziali>)
```

Albero delle decisioni L'uso della ricorsione per esaminare lo spazio delle soluzioni, ci consente di rappresentare quest'ultimo come un *albero*, in cui la *radice* rappresenta la soluzione parziale vuota, ciascun *nodo* interno è associato a una soluzione parziale e, infine, le *foglie* sono le soluzioni ammissibili.

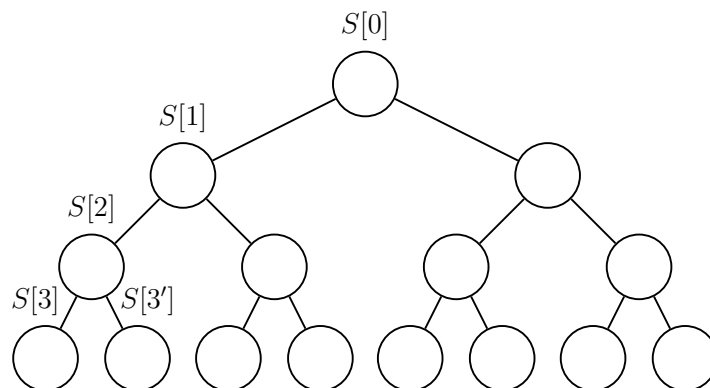


Fig. 16.1: Esempio di *albero delle decisioni*

Implementazione ricorsiva

Frammento 123 - Implementazione ricorsiva della soluzione.

```
subset(int n)
    int[] S = new int[1...n]                % Vettore delle scelte
    subsetRec(n, S, 1)

subsetRec(int n, int[] S, int i)
    % Una soluzione è ammissibile dopo n scelte
    if (i > n) then
        print S
    else
        SET C = {0, 1}                      % Ogni valore può essere preso o non preso
        foreach (c ∈ C) do
            S[i] = c
            subsetRec(n, S, i + 1)
```

Complessità Come richiesto dal problema, l'algoritmo esplora tutto lo spazio delle soluzioni e quindi la *complessità* è $T(n) = \Theta(n2^n)$.

Implementazione iterativa Per questo particolare problema, a parità di *complessità*, possiamo scrivere anche una versione iterativa della soluzione.

Frammento 124 - Implementazione iterativa della soluzione.

```
subset(int n)
    for (i = 0 to 2n - 1) do
        print "{"      % Usiamo le parentesi graffe per distinguere i sottoinsiemi
        for (j = 0 to n - 1) do
            if (i && 2j ≠ 0) do                % &&: operatore di bitwise and
                print j, " "
        println }
```

16.3 Problema dell'enumerazione delle permutazioni

Problema 19 - Problema dell'enumerazione delle permutazioni.

Dato un insieme A, stamparne tutte le permutazioni.

Questo è un altro *problema di enumerazione* quindi segue lo stesso schema appena visto.

Frammento 125 - Prima implementazione della soluzione.

```
permutations(SET A)
    int n = size(A)
    int[] S = new int[1...n]
    permRec(A, S, 1)
```

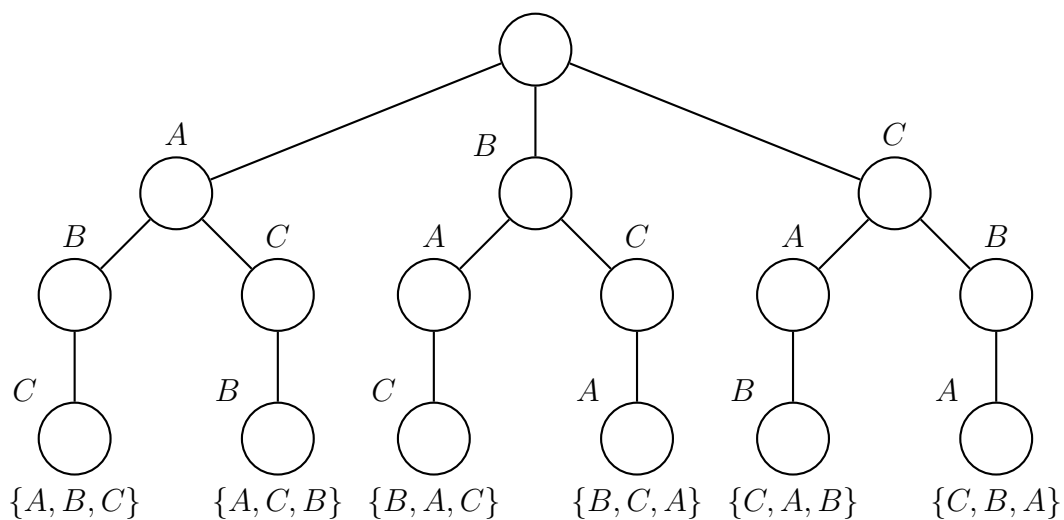
```

permRec(SET A, ITEM[] S, int i)
    if (A.isEmpty()) then                                     % Se A è vuoto, S è ammissibile
        print S
    else
        SET C = copy(A)                                       % Nel ciclo usiamo una copia di A
        foreach (c ∈ C) do
            S[i] = c
            A.remove(c)
            permRec(n, S, i + 1)
            A.insert(C)

```

Esempio 43 - Esempio d'esecuzione.

Se $A = \{1, 2, 3\}$, l'albero delle soluzioni associato è il seguente:



Complessità Il costo delle copie dell'insieme lungo i cammini radice-foglia è $\sum_{i=1}^n O(i) = O(n^2)$ e poiché il numero di foglie è $n!$, la complessità totale è $T(n) = O(n^2 n!)$.

Versione migliorata dell'algoritmo

Frammento 126 - Seconda implementazione della soluzione.

```

permutations(ITEM[] S, int n)
    permRec(S, n)

permRec(ITEM[] S, int i)
    if (i == 1) then                                         % C'è solo una permutazione
        println S
    else
        for (j = 1 to i) do
            swap(S, i, j)
            permRec(S, i - 1)
            swap(S, i, j)

```

In questa soluzione, ad ogni iterazione del ciclo **for** vengono scambiati di posizione gli elementi $S[i]$ e $S[j]$, quindi viene continuata ricorsivamente l'esecuzione sul sottoinsieme $S[1 \dots i - 1]$. Infine, vengono di nuovo scambiati gli elementi $S[i]$ e $S[j]$.

Complessità Per ogni permutazione vengono eseguite $2n$ swap e poiché le permutazioni sono $n!$, la *complessità* di questa versione è $T(n) = O(n!n)$.

16.4 Problema dell'enumerazione dei k -sottoinsiemi

Problema 20 - Problema dell'enumerazione dei k -sottoinsiemi.

Dato un insieme $\{1, \dots, n\}$, elencarne tutti i sottoinsiemi di dimensione k .

Frammento 127 - Prima implementazione della soluzione.

```
kSubset(int n, int k)
    int[] S = new int[1...n]
    kssRec(n, k, S, 1)

% Conta il numero di elementi del sottoinsieme
int count(int[] S, int n)
    int count = 0
    for (i = 1 to n) do
        count = count + S[i]
    return count

kssRec(int n, int k, int[] S, int i)
    int size = count(S, n)
    if (i > n and size == k) then                % La soluzione trovata è ammissibile
        processSolution(S, n)
    else if (i > n and size ≠ k) then % La soluzione trovata non è ammissibile
        return
    else
        foreach (c ∈ {0, 1}) do
            S[i] = c
            kssRec(n, k, S, i + 1)
```

La *complessità* è $O(2^n n)$ perché ogni elemento può essere incluso oppure no in un sottoinsieme e la `count` costa sempre $O(n)$. Ciò che è diverso dai problemi precedenti è che potremmo arrivare ad una soluzione non ammissibile e, di conseguenza, possiamo modificare lo **Schema generale per algoritmi di enumerazione** introducendo un controllo sull'inammissibilità di una soluzione.

Frammento 128 - Schema generale completo per algoritmi di enumerazione.

```
enumeration(<dati problema>, ITEM[] S, int i, <dati parziali>))
    % Verifica se  $S[1 \dots i - 1]$  contiene una soluzione ammissibile
    if (accept(<dati problema>, S, i, <dati parziali>))
        % "Processa" la soluzione (e.g. stampa, conta, ...)
```

```

    processSolution(<dati problema>, S, i, <dati parziali>)
% Verifica se  $S[1..i-1]$  contiene una soluzione non ammissibile
else if (reject(<dati problema>, S, i, <dati parziali>)) then
    return
else
    % Calcola l'insieme delle scelte in funzione di  $S[1..i-1]$ 
    SET C = choices(<dati problema>, S, i, <dati parziali>)
    % Itera sull'insieme delle scelte
    foreach ( $c \in C$ ) do
        S[i] = c
        % Chiamata ricorsiva
        enumeration(<dati problema>, S, i + 1, <dati parziali>)

```

Ritornando all'implementazione della soluzione, possiamo eliminare l'invocazione alla `count` e ridurre a $O(2^n)$ la *complessità*.

Frammento 129 - Seconda implementazione della soluzione.

```

kssRec(int n, int missing, int[] S, int i)
    if (i > n and missing == 0) then % Se missing = 0 la soluzione è ammissibile
        processSolution(S, n)
    else if (i > n or missing < 0) then
        return
    else
        foreach ( $c \in \{0, 1\}$ ) do
            S[i] = c
            kssRec(n, missing - c, S, i + 1)

```

NB. L'implementazione della funzione wrapper `kSubset` rimane invariata.

Il parametro `missing` tiene traccia del numero di elementi che devono ancora essere inseriti nel sottoinsieme per renderlo una soluzione.

Albero delle decisioni Eseguendo l'algoritmo con input $n = 3$ e $k = 1$, l'*albero delle decisioni* risultante è il seguente:

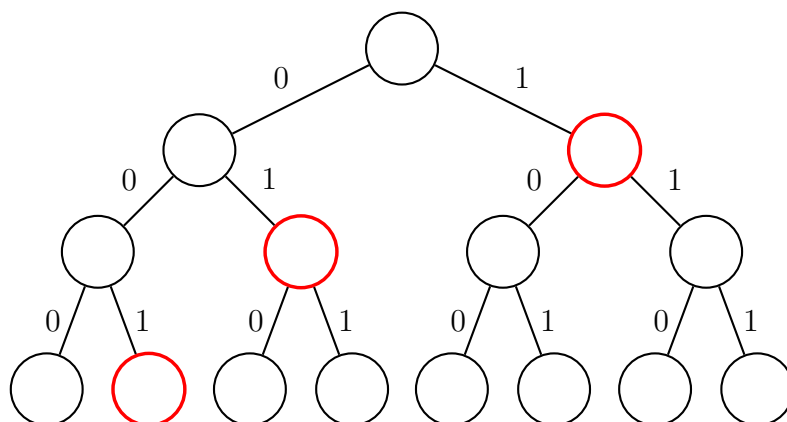


Fig. 16.2: *Albero delle decisioni*

I *nodi* rossi rappresentano il punto in cui il valore di `missing` arriva a 0 per la prima volta nel *cammino radice-foglia*. Continuare la ricerca dopo quei *nodi* non ha senso, perché ogni

nuovo elemento aggiunto al sottoinsieme porterebbe ad una soluzione inammissibile. Questo ci suggerisce che l'implementazione potrebbe essere migliorata ulteriormente “cancellando” quei *sottoalberi* dall'insieme delle soluzioni. Questa operazione prende il nome di *pruning*.

Frammento 130 - Terza implementazione della soluzione con pruning totale.

```
kssRec(int n, int missing, int[] S, int i)
    if (missing == 0) then                % Se missing = 0 la soluzione è ammissibile
        processSolution(S, n)
    else if (i > n or missing < 0 or n - (n - 1) < missing) then
        return
    else
        foreach (c ∈ {0, 1}) do
            S[i] = c
            kssRec(n, missing - c, S, i + 1)
```

Oltre a rimuovere i *sottoalberi* in eccedenza, possiamo togliere anche quelli che di sicuro non porteranno ad alcuna soluzione ammissibile. Per esempio, se $k = 2$, nell'*albero* precedente avremmo potuto eseguire il *pruning* del *sottoalbero* radicato nel primo *nodo* da sinistra del secondo *livello*. Questo perché, anche scegliendo di inserire nel sottoinsieme tutti i valori successivi, non si sarebbe potuto arrivare ad annullare il valore di *missing*.

NB. Quando valutiamo se eseguire o meno il *pruning* di un certo *sottoalbero*, prendiamo la decisione in base alla soluzione parziale associata al *nodo radice* del *sottoalbero*.

Frammento 131 - Implementazione ripulita della soluzione.

```
kssRec(int n, int missing, int[] S, int i)
    if (missing == 0) then                % Se missing = 0 la soluzione è ammissibile
        processSolution(S, n)
    else if (i ≤ n and 0 < missing ≤ n - (n - 1)) then
        foreach (c ∈ {0, 1}) do
            S[i] = c
            kssRec(n, missing - c, S, i + 1)
```

Complessità Sebbene l'introduzione del *pruning* non riduca la *complessità*, permette comunque di migliorare l'efficienza per valori di k vicini agli estremi 1, n . Per quei valori di k infatti, la porzione di *albero* che viene eliminata è sostanziale. Viceversa il guadagno è soltanto marginale per valori di k vicini a $\frac{n}{2}$.

16.5 Problema del subset sum

Problema 21 - Problema del subset sum.

Dati un vettore A contenente n interi positivi ed un intero positivo k , dire se esiste un sottoinsieme di indici $S \subseteq \{1 \dots n\}$ tale per cui $\sum_{i \in S} A[i] = k$.

Ad esempio, se $A = [1, 4, 3, 12, 7, 2, 21, 55]$ e $k = 23$, due possibili soluzioni sono $S_1 = \{2, 4, 5\}$ perché $4 + 12 + 7 = 23$, e $S_2 = \{6, 7\}$ perché $2 + 21 = 23$.

NB. Questo problema può essere risolto in tempo $O(kn)$ sfruttando la *programmazione dinamica*, tuttavia qui ne vediamo una soluzione tramite *backtracking* di tempo $O(2^n)$ in cui però interrompiamo l'esecuzione alla prima soluzione trovata.

16.5.1 Approccio generale

Lo schema generale è simile a quello usato per l'*enumerazione*, ma alla prima soluzione ammissibile trovata, l'algoritmo termina restituendo `true`.

Frammento 132 - Schema generale completo con interruzione.

```
boolean enumeration(<dati problema>, ITEM[] S, int i, <dati parziali>)
% Verifica se  $S[1...i-1]$  contiene una soluzione ammissibile
if (accept(<dati problema>, S, i, <dati parziali>))
    % "Processa" la soluzione (e.g. stampa, conta, ...)
    processSolution(<dati problema>, S, i, <dati parziali>)
    return true
% Verifica se  $S[1...i-1]$  contiene una soluzione non ammissibile
else if (reject(<dati problema>, S, i, <dati parziali>))
    return false
else
    % Calcola l'insieme delle scelte in funzione di  $S[1...i-1]$ 
    SET C = choices(<dati problema>, S, i, <dati parziali>)
    % Itera sull'insieme delle scelte
    foreach ( $c \in C$ ) do
        S[i] = c
        % Chiamata ricorsiva
        if (enumeration(<dati problema>, S, i + 1, <dati parziali>)) then
            return true
    return false
```

Implementazione

Frammento 133 - Implementazione della soluzione.

```
boolean subsetSum(int[] A, int n, int k)
    int[] S = new int[1...n]
    return ssRec(A, n, k, S, 1)

boolean ssRec(int[] A, int n, int missing, int[] S, int i)
    if (missing == 0) then
        return true
    else if (i > n or missing < 0) then
        return false % I valori sono terminati o la somma è eccessiva
    else
        foreach ( $c \in \{0, 1\}$ ) do
            S[i] = c
            if (ssRec(A, n, missing - A[i] · c, S, i + 1)) then
                return true
        return false
```

16.6 Problema dell'involuppo convesso

Definizione 158 - Poligono convesso.

Un poligono nel piano è detto essere convesso se ogni segmento di retta che congiunge due punti del poligono rimane sempre dentro il poligono stesso.

Problema 22 - Problema dell'involuppo convesso.

Dati n punti p_1, \dots, p_n nel piano, con $n \geq 3$, determinare l'involuppo convesso, ovvero il poligono convesso di superficie minima contenente tutti gli n punti.

Intuitivamente, poiché stiamo cercando il *poligono convesso* di superficie minima, i vertici di quel poligono dovranno coincidere con alcuni o tutti i punti del piano. In particolare, presi due punti, il segmento che li congiunge è un lato del poligono se tutti i rimanenti $n - 2$ punti stanno dalla “stessa parte”. Per verificare la condizione usiamo la seguente funzione:

Frammento 134 - Implementazione sameSide.

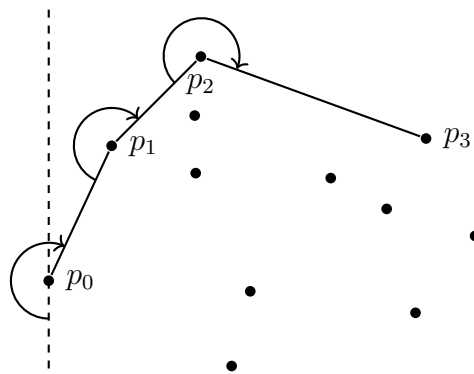
```
boolean sameSide(POINT p1, POINT p2, POINT p, POINT q)
    float dx = p2.x - p1.x
    float dy = p2.y - p1.y
    float dx1 = p.x - p1.x
    float dy1 = p.y - p1.y
    float dx2 = q.x - p2.x
    float dy2 = q.y - p2.y
    % Restituisce true se p e q stanno dalla stessa parte della retta tra p1 e p2
    return ((dx · dy1 - dy · dx1) · (dx · dy2 - dy · dx2) ≥ 0)
```

Questa soluzione è intuitiva e facile da implementare, ma la *complessità* è $O(n^3)$, perché le possibili coppie di punti sono $O(n^2)$ e per ognuna bisogna verificare che i rimanenti $n - 2$ punti stiano dalla “stessa parte”.

16.6.1 Algoritmo di Jarvis

Anche detto *Gift Packing*, l'*Algoritmo di Jarvis* parte dal punto più a sinistra dell'insieme, ovvero dal punto p_0 con ordinata minore. Quindi, viene calcolato l'angolo d'inclinazione delle rette passanti per p_0 e ogni altro punto rispetto alla verticale.

Il punto p_1 associato alla retta con angolo minore viene selezionato come vertice del poligono. Successivamente, si considera la retta r passante per p_{i-1} e p_{i-2} e si confrontano gli angoli che le rette passanti per p_{i-1} e tutti i punti rimanenti formano con r . Ad essere selezionato come vertice è il punto p_i associato all'angolo minore. Il tutto viene ripetuto fino a quando ad essere selezionato non è di nuovo p_0 .



Complessità La selezione del punto p_0 costa $O(n)$ e la ricerca dello spigolo successivo costa di nuovo $O(n)$ perché viene considerato ogni punto rimanente. Alla fine, se i vertici del poligono sono h , l'algoritmo ha una *complessità* totale di $O(nh)$.

16.6.2 Algoritmo di Graham

L'*Algoritmo di Graham* inserisce nell'*inviluppo convesso* il punto di ordinata minima e ordina i rimanenti in base all'angolo formato dalla retta passante per ciascun punto e il punto di ordinata minima e la retta orizzontale. Il primo punto dell'ordinamento viene aggiunto all'*inviluppo*.

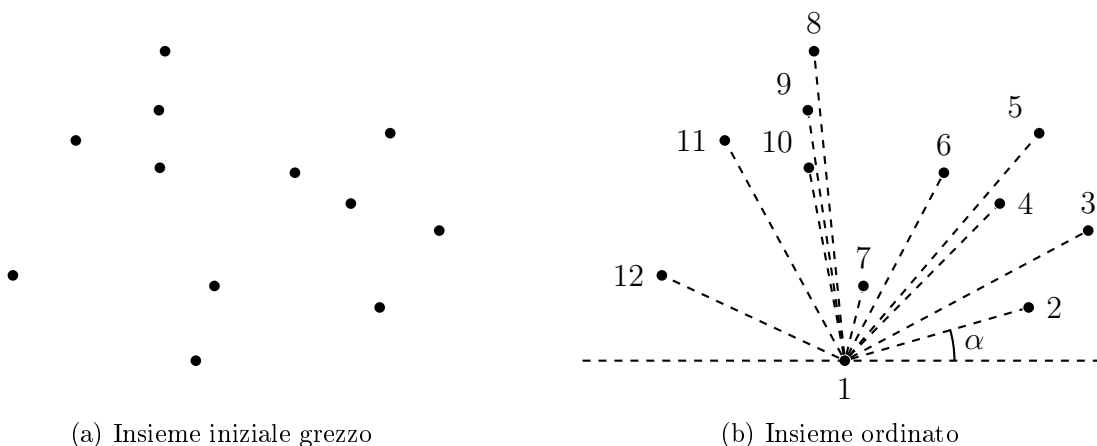
A questo punto, siano p_j e p_{j-1} l'ultimo e il penultimo punto dell'*inviluppo*. Per tutti i punti p_i , con $i \in [3, n]$, se p_1 e p_i non si trovano dalla stessa parte rispetto alla retta passante per p_j e p_{j-1} , il punto p_j viene rimosso dall'*inviluppo*. Se invece la scansione dei punti termina senza che p_j debba essere rimosso, p_i viene aggiunto all'*inviluppo*.

Frammento 135 - Implementazione Algoritmo di Graham.

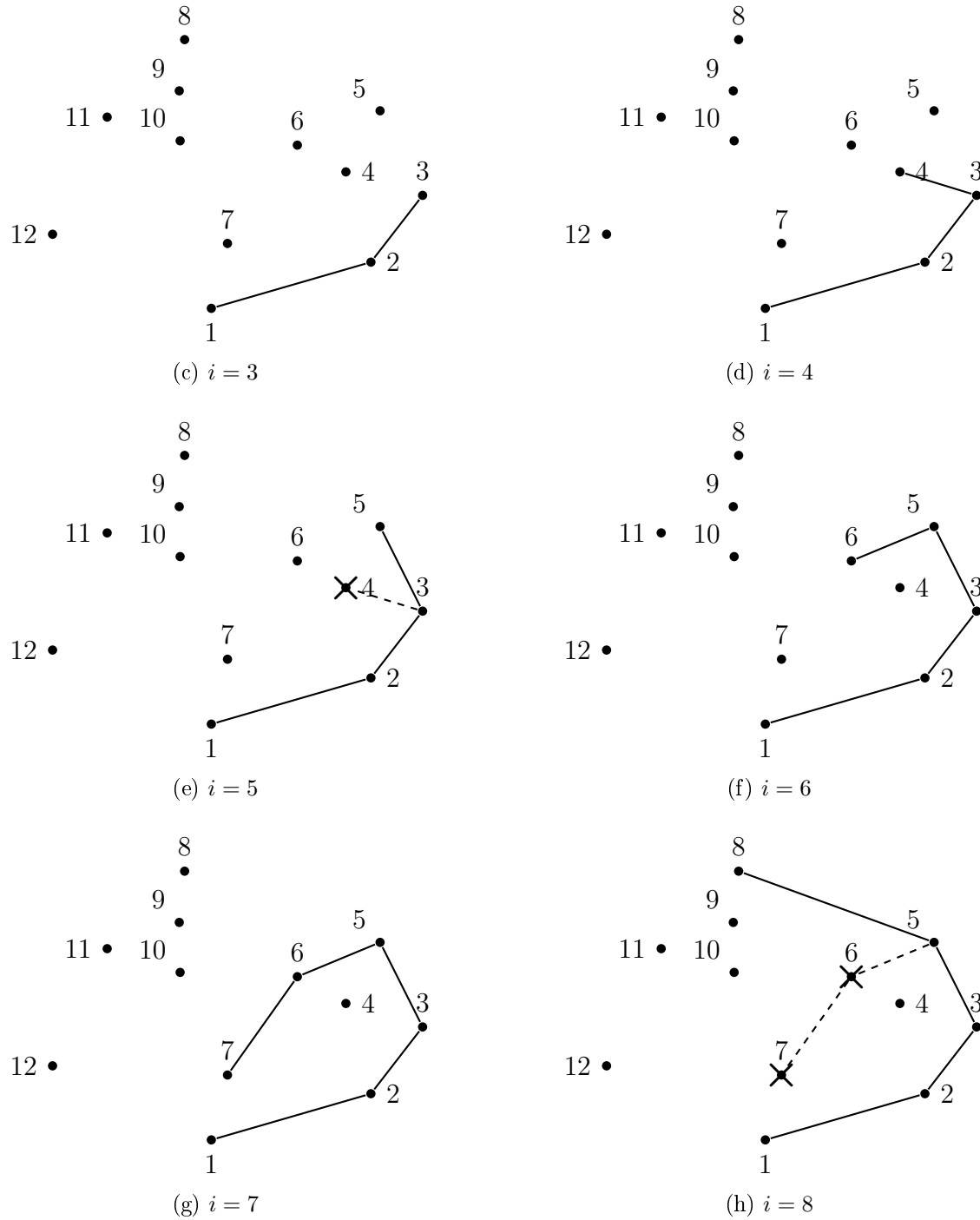
```
STACK graham(POINT[] p, int n)
    int min = 1
    for (i = 2 to n) do
        if (p[i].y < p[min].y) then
            min = i
    p[1] ↔ p[min]                % Sposta all'inizio il punto di ordinata minima
    { Ordina p[2,...,n] in base all'angolo formato con all'asse orizzontale }
    { Elimina i punti allineati mantenendo solo quello più distante da p1 }
    { Aggiorna n riducendolo del numero di punti eliminati }
    STACK S = Stack()
    S.push(p[1])
    S.push(p[2])
    for (i = 3 to n) do
        while (not sameSide(S.top(), S.top2(), p[1], p[i])) do
            S.pop()
        S.push(p[i])
    return S
```

Esempio 44 - Esempio d'esecuzione.

Consideriamo il seguente insieme di punti:



L'*Algoritmo di Graham* inserisce nell'*inviluppo convesso* i punti 1 e 2. Poi, a partire dal terzo, controlla che ogni punto stia dalla stessa parte del primo rispetto all'ultimo lato dell'*inviluppo*.



L'algoritmo continua in questo modo provando tutti i punti.

Complessità Il calcolo degli angoli tra le rette e la rimozione dei punti allineati costa $O(n)$ e l'ordinamento costa $O(n \log n)$. Una volta ordinati, i punti vengono aggiunti e rimossi dall'*inviluppo* al massimo una volta ciascuno, quindi la *complessità* finale è $T(n) = O(n \log n)$.

Capitolo Nr.17

Problemi intrattabili e teoria dell'NP-completezza

17.1 Introduzione

Finora, con l'unica eccezione della sezione sul *backtracking*, abbiamo trattato unicamente problemi con soluzioni in tempo *polinomiale*, ovvero problemi le cui soluzioni sono eseguibili in tempo $O(n^k)$ con $k \in \mathbb{R}^+$. Esistono anche problemi che richiedono un tempo *esponenziale* o che addirittura non sono risolvibili (e.g. halting problem).

Ciò che andremo ad introdurre in questo capitolo invece, sono una serie di problemi per i quali non è chiaro se esista o meno una soluzione *polinomiale*. Vedremo anche come tutti questi problemi siano in realtà legati tra loro, in modo che se esiste una soluzione *polinomiale* per uno di essi, allora ne esiste una anche per gli altri. Viceversa, se si riesce a dimostrare che uno di essi non è risolvibile in tempo *polinomiale*, allora non lo è nessuno.

Procediamo dando alcune definizioni.

Definizione 159 - Problema astratto.

Un problema astratto è una relazione binaria $R \subseteq I \times S$ tra un insieme I di istanze del problema e un insieme S di soluzioni.

NB. Ad esempio, nel problema della *ricerca del cammino minimo tra due nodi*, un'istanza del problema è la tupla (V, R, u, v) , mentre una soluzione è una sequenza di *nodi* (v_1, \dots, v_n) .

Come visto nei capitoli scorsi, esistono varie tipologie di problema. Le principali sono tre: *ottimizzazione*, *ricerca* e *decisione*.

Definizione 160 - Problema di ottimizzazione.

Data un'istanza, trovare la soluzione ottima secondo un insieme di criteri prestabiliti.

Definizione 161 - Problema di ricerca.

Data un'istanza, trovare una possibile soluzione tra quelle esistenti.

Definizione 162 - Problema di decisione.

Data un'istanza, verificare se soddisfa o meno una data proprietà.

NB. Nei problemi di decisione, R è una funzione del tipo $R : I \rightarrow \{0, 1\}$.

Informalmente, possiamo dimostrare che i *problemi di ottimizzazione* e di *decisione* sono equivalenti.

Dimostrazione. Se è possibile risolvere efficientemente un *problema di ottimizzazione*, allora è possibile usare la soluzione a quel problema per verificare efficientemente la proprietà interessata dal *problema di decisione* associato.

Da questa affermazione possiamo derivare la seguente: se non è possibile risolvere efficientemente un *problema di decisione*, allora non è nemmeno possibile risolvere efficientemente il *problema di ottimizzazione* associato. \square

NB. Ad esempio, nel problema della *ricerca del cammino tra due nodi*, se si conosce il *cammino minimo* tra essi, è possibile risolvere efficientemente un *problema di decisione* nel quale ci si chiede se esista un *cammino* di lunghezza non superiore a un qualche valore k .

Questa dimostrazione, sebbene informale, ci permette di proseguire la trattazione concentrandoci unicamente su *problemi decisionali*, che sono più facili sia da definire che da elaborare.

17.2 Riduzioni

Definizione 163 - Riduzione polinomiale.

Dati due problemi decisionali $R_1 \subseteq I_1 \times \{0, 1\}$ e $R_2 \subseteq I_2 \times \{0, 1\}$, diciamo che R_1 è riducibile polinomialmente a R_2 , e scriviamo in simboli $R_1 \leq_p R_2$, se esiste una funzione $f : I_1 \rightarrow I_2$ che sia calcolabile in tempo polinomiale e tale per cui, per ogni istanza x del problema R_1 e ogni soluzione $s \in \{0, 1\}$, sia vero che $(x, s) \in R_1 \Leftrightarrow (f(x), s) \in R_2$.

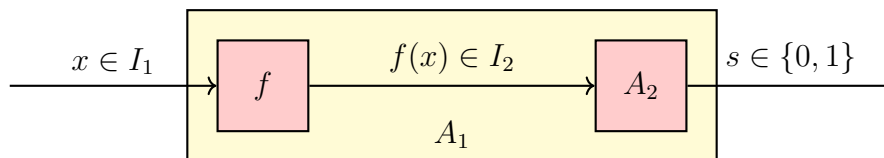


Fig. 17.1: Riduzione polinomiale

Proseguiamo la trattazione con una carrellata di problemi che, come vedremo, possono essere legati tra loro mediante *riduzioni polinomiali*.

17.2.1 Colorazione di grafi

Definizione 164 - Colorazione di grafi.

Dati un grafo non orientato $G = (V, E)$ e un insieme di colori C , una colorazione dei vertici è una funzione $f : V \rightarrow C$ che assegna ad ogni nodo uno dei colori in C in maniera tale che nessuna coppia di nodi adiacenti abbia lo stesso colore.

Problema 23 - Colorabilità di un grafo (GRAPH-COLORING).

Dato un grafo non orientato $G = (V, E)$ e un valore k , determinare se esiste una colorazione di G con k colori.

17.2.2 Sudoku

Problema 24 - Risolvibilità di un sudoku (SUDOKU).

Data una matrice $n^2 \times n^2$ con alcuni numeri già inseriti, determinare se esiste un modo per assegnare i numeri restanti in modo coerente con le regole dei Sudoku.

Il problema sulla *risolvibilità di un sudoku* può essere ridotto *polinomialmente* al problema sulla *colorabilità di un grafo*. In particolare, è possibile tradurre la matrice del sudoku in un *grafo* in cui ogni valore della matrice diventa un *nodo* del grafo e in cui tra due *nodi* esiste un *arco* soltanto se, all'interno della matrice, quei *nodi* sono sulla stessa riga, sulla stessa colonna o sulla stessa diagonale. Formalmente gli insiemi dei *nodi* e degli *archi* sono definiti come segue:

$$V = \{(x, y) : 1 \leq x \leq n^2, 1 \leq y \leq n^2\}$$

$$E = \left\{ [(x, y), (x', y')] : x = x' \vee y = y' \vee \left(\left\lceil \frac{x}{n} \right\rceil = \left\lceil \frac{x'}{n} \right\rceil \wedge \left\lceil \frac{y}{n} \right\rceil = \left\lceil \frac{y'}{n} \right\rceil \right) \right\}$$

L'insieme dei colori è $C = \{1, \dots, n\}$.

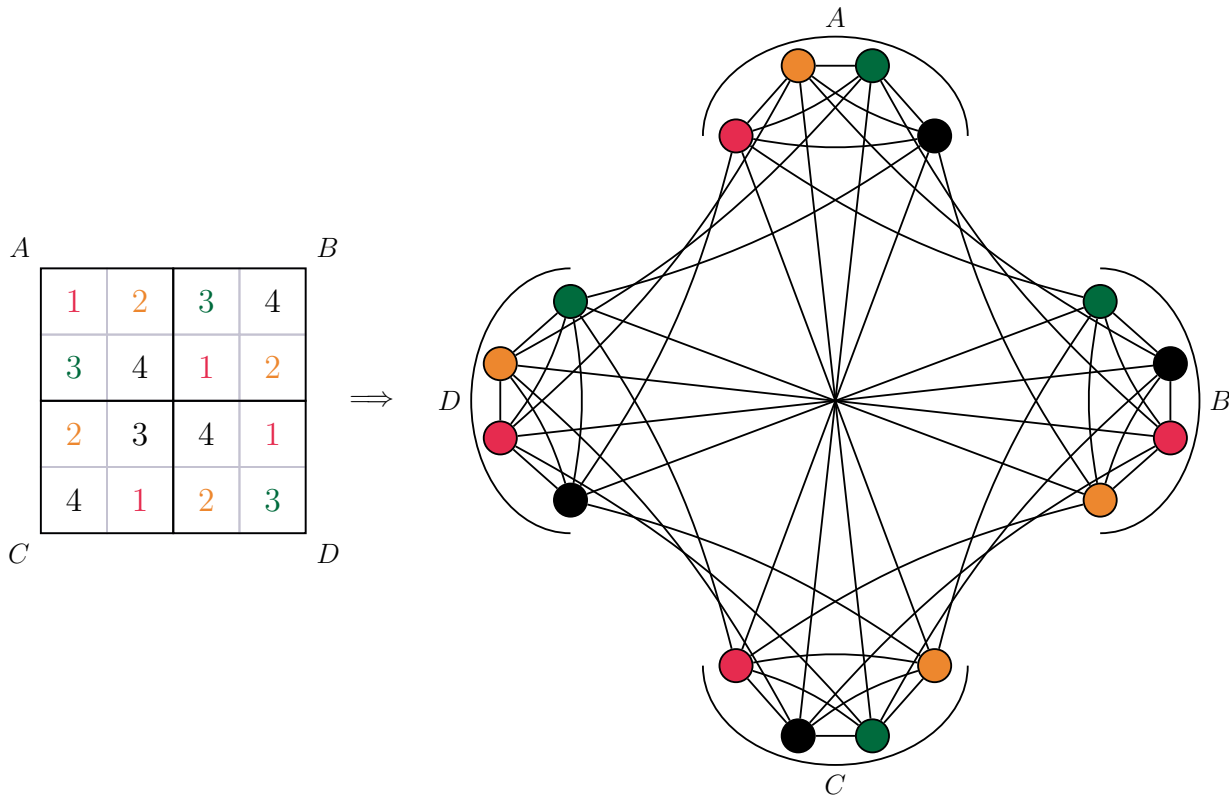


Fig. 17.2: Trasformazione *matrice-grafo*

Questo significa che $\text{SUDOKU} \leq_p \text{GRAPH-COLORING}$, quindi una soluzione al problema della colorazione può essere usata per risolvere il problema del sudoku.

17.2.3 Insieme indipendente

Definizione 165 - Insieme indipendente.

Dato un grafo non orientato $G = (V, E)$, un insieme $S \subseteq V$ è un insieme indipendente se e

solo se nessun arco in E unisce due nodi in S . Ovvero, S è un insieme indipendente se vale la seguente relazione:

$$x \notin S \vee y \notin S \quad \forall (x, y) \in E$$

Problema 25 - Esistenza di un insieme indipendente (INDIPENDENT-SET).

Dati un grafo non orientato $G = (V, E)$ e un valore k , dire se esiste un insieme indipendente di dimensione almeno k .

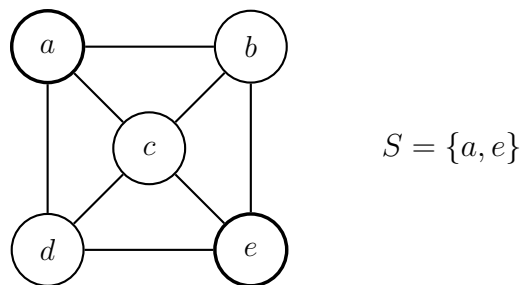


Fig. 17.3: Esempio di *insieme indipendente* di dimensione $k = 2$

NB. In questo esempio, avremmo anche potuto scegliere $S = \{b, d\}$.

17.2.4 Copertura di vertici

Definizione 166 - Copertura di vertici.

Dato un grafo non orientato $G = (V, E)$, un insieme $S \subseteq V$ è una *copertura di vertici* se e solo se ogni arco in E ha almeno un vertice in S . Ovvero, S è una *copertura di vertici* se vale la seguente relazione:

$$x \in S \vee y \in S \quad \forall (x, y) \in E$$

Problema 26 - Esistenza di una copertura di vertici (VERTEX-COVER).

Dati un grafo non orientato $G = (V, E)$ e un valore k , dire se esiste una copertura di vertici di dimensione al massimo k .

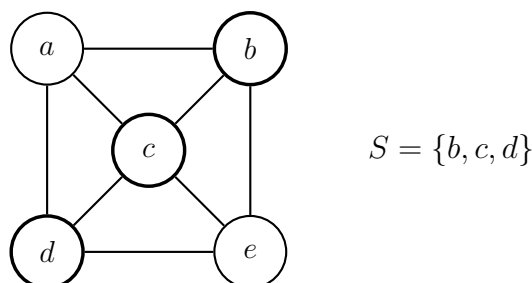


Fig. 17.4: Esempio di *copertura di vertici* di dimensione $k = 3$

NB. In questo esempio, avremmo anche potuto scegliere $S = \{a, c, e\}$.

Dagli ultimi due esempi fatti risulta particolarmente evidente il legame di dualità che sussiste tra un *insieme indipendente* e una *copertura di vertici* di uno stesso *grafo*. Infatti, se $S \subseteq V$ è un *insieme indipendente*, $V - S$ è una *copertura di vertici*, e viceversa.

La dimostrazione di quanto detto è banale:

Dimostrazione. Dimostriamo separatamente le due implicazioni:

S insieme indipendente $\Rightarrow V - S$ copertura di vertici Se S è un *insieme indipendente*, per definizione, non esistono *archi* $(x, y) \in E$ che abbiano entrambi i *vertici* in S . Di conseguenza, almeno uno dei due *vertici* deve appartenere a $V - S$.

$V - S$ copertura di vertici $\Rightarrow S$ insieme indipendente Se supponiamo per assurdo che S non sia un *insieme indipendente*, allora dovrebbe esistere un *arco* (x, y) tale per cui $x, y \in S$. Se così fosse però, né x , né y apparterrebbero a $V - S$ e quindi $V - S$ non potrebbe essere una *copertura di vertici*, generando un assurdo. \square

Con questa dimostrazione abbiamo dimostrato che i due problemi sono equivalenti, e quindi valgono le seguenti relazioni di *riducibilità polinomiale*:

$$\text{VERTEX-COVER} \leq_p \text{INDIPENDENTE-SET}$$

$$\text{INDIPENDENTE-SET} \leq_p \text{VERTEX-COVER}$$

17.2.5 Soddisfacibilità di formule booleane

Definizione 167 - Forma normale congiuntiva per formule booleane.

Dato un insieme V contenente n variabili booleane si definisce *letterale* una variabile v o il suo negato \bar{v} . Una *disgiunzione* di letterali, ovvero un insieme di letterali separati dall'operatore **or**, costituisce una *clausola*. Infine, una formula booleana si dice essere in *forma normale congiuntiva* se è costituita da una *congiunzione* di *clausole*, ovvero da un insieme di *clausole* separate dall'operatore **and**.

Problema 27 - Soddisfacibilità di formule booleane (SAT).

Data un'espressione in forma normale congiuntiva, stabilire se esiste un'assegnazione di valori di verità alle variabili che renda vera l'espressione.

Esempio 45 - Esempio.

Se $V = \{w, x, y, z\}$ e l'espressione da valutare è la seguente:

$$(x \vee \bar{y} \vee z) \wedge (\bar{x} \vee w) \wedge y$$

Alcune assegnazioni che rendono vera l'espressione sono:

- $w = x = y = z = \text{true}$;
- $w = x = y = \text{true}, z = \text{false}$;

Problema 28 - Soddisfacibilità di formule booleane con 3 letterali (3-SAT).

Data un'espressione in forma normale congiuntiva in cui ciascuna clausola ha esattamente 3 letterali, stabilire se esiste un'assegnazione di valori di verità alle variabili che renda vera l'espressione.

Ciò che vogliamo fare ora è dimostrare che $3\text{-SAT} \leq_p \text{INDIPENDENT-SET}$. Per farlo, *riduciamo polinomialmente* 3-SAT a INDIPENDENT-SET trasformando l'espressione booleana in un *grafo*, nel quale, per ogni *clausola* viene aggiunto un trio di *nodi* collegati tra loro da *archi*. Inoltre, viene aggiunto un *arco* anche tra ogni *letterale* e il suo negato.

Ad esempio, il *grafo* associato all'espressione $(x \vee \bar{y} \vee z) \wedge (\bar{x} \vee \bar{y} \vee \bar{z}) \wedge (\bar{x} \vee y \vee z)$ è il seguente:

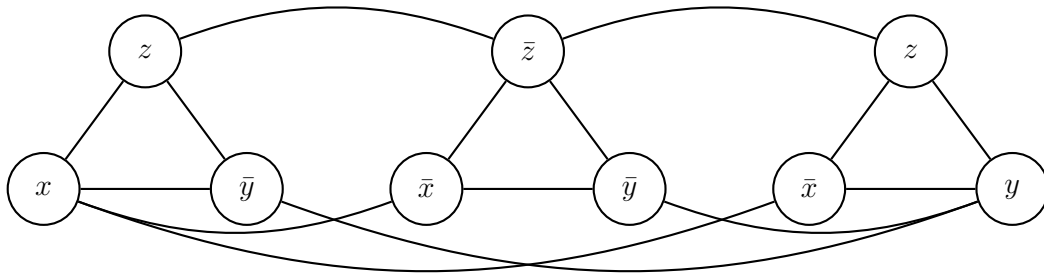


Fig. 17.5: Grafo associato a una formula 3-SAT

Costruito il *grafo*, la formula è soddisfacibile se e solo se è possibile trovare un *insieme indipendente* composto da tanti elementi quante sono le *clausole* dell'espressione booleana.

Nell'esempio che stiamo considerando, la soluzione è la seguente:

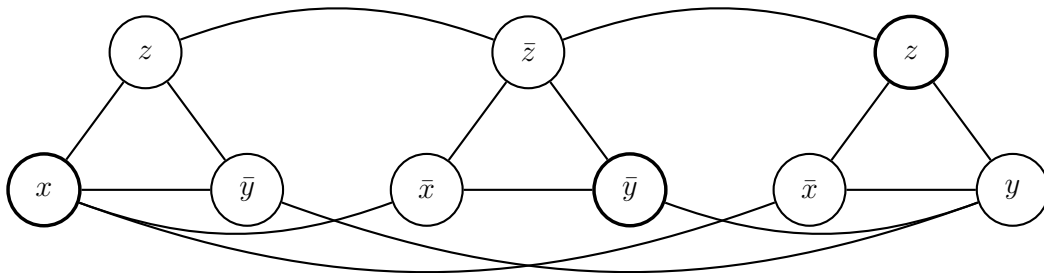


Fig. 17.6: Possibile soluzione

L'insieme che abbiamo individuato ci dice quali *letterali* devono essere posti a **true**.

Relazione tra SAT, 3-SAT e INDIPENDENT-SET Ovviamente $3\text{-SAT} \leq_p \text{SAT}$ perché il primo è un caso particolare del secondo. Tuttavia, è possibile dimostrare anche la relazione inversa, in quanto, ogni espressione SAT può essere trasformata in una 3-SAT applicando delle semplici trasformazioni:

- Se la *clausola* contiene più di 3 *letterali*, si introduce una nuova variabile e si spezza la precedente *clausola*:

$$(a \vee b \vee c \vee d) \equiv (a \vee b \vee z) \wedge (\bar{z} \vee c \vee d)$$

- Se la *clausola* è contiene meno di 3 *letterali*, si fa il “padding”, ovvero si ripete più volte uno stesso *letterale* all’interno della stessa *clausola*:

$$(a \vee b) \equiv (a \vee a \vee b)$$

Riassumendo i rapporti di *riducibilità polinomiale* riscontrati finora, otteniamo la seguente catena:

$$\text{SAT} \leq_p \text{3-SAT} \leq_p \text{INDIPENDENT-SET} \leq_p \text{VERTEX-COVER}$$

17.3 Classi di problemi

Definizione 168 - Algoritmo.

Dati un problema decisionale R e un algoritmo A (scritto in un modello di calcolo Turing-equivalente) che lavora in tempo $f_t(n)$ e spazio $f_s(n)$, diciamo che A risolve R se restituisce una soluzione s su un’istanza x se e solo se $(x, s) \in R$.

Definizione 169 - Classi di complessità.

Data una qualunque funzione $f(n)$, chiamiamo $\text{TIME}(f(n))$ e $\text{SPACE}(f(n))$ l’insieme dei problemi decisionali risolvibili da un algoritmo che lavora rispettivamente in tempo e spazio $O(f(n))$.

Definizione 170 - Classe \mathbb{P} .

La classe \mathbb{P} comprende tutti i problemi decisionali che sono risolvibili in tempo polinomiale rispetto alla dimensione n dell’istanza d’ingresso. Ovvero, l’insieme \mathbb{P} è definito come:

$$\mathbb{P} = \bigcup_{c=0}^{+\infty} \text{TIME}(n^c)$$

Definizione 171 - Classe PSPACE .

La classe PSPACE comprende tutti i problemi decisionali che sono risolvibili in spazio polinomiale rispetto alla dimensione n dell’istanza d’ingresso. Ovvero, l’insieme PSPACE è definito come:

$$\text{PSPACE} = \bigcup_{c=0}^{+\infty} \text{SPACE}(n^c)$$

NB. La classe \mathbb{P} è un sottoinsieme della classe PSPACE , ovvero $\mathbb{P} \subseteq \text{PSPACE}$.

Definizione 172 - Certificato.

Dati un problema decisionale R e un’istanza x di input tale per cui $(x, \mathbf{true}) \in R$, un certificato è un insieme di informazioni che permette provare che $(x, \mathbf{true}) \in R$.

Ad esempio, nel problema SAT, un *certificato* è un assegnamento di valori di verità alle variabili della formula, in GRAPH-COLORING è un’associazione *nodo-colore* $f : V \rightarrow C$ e in INDIPENDENT-SET è un sottoinsieme $S \subseteq V$ di k elementi.

NB. Tutti questi *certificati* hanno dimensione *polinomiale* rispetto alla dimensione dell’input.

È facile convincersi che la verifica di tutti i *certificati* citati sia possibile in tempo *polinomiale* e questa caratteristica è distintiva dei problemi appartenenti alla *classe* \mathbf{NP} .

Definizione 173 - Classe \mathbf{NP} .

La classe \mathbf{NP} comprende tutti i problemi decisionali che ammettono un certificato verificabile in tempo polinomiale.

NB. Non è detto che tutti i *certificati* siano verificabili in tempo *polinomiale*.

Definizione 174 - Trasitività dell'appartenenza alla classe \mathbb{P} .

Dati due problemi decisionali R_1 e R_2 tali per cui $R_1 \leq_p R_2$, se $R_2 \in \mathbb{P}$ allora anche $R_1 \in \mathbb{P}$.

Dimostrazione. Sia $T_f(n) = O(n^{k_f})$ il tempo necessario per trasformare un input di R_1 in un input di R_2 tramite una funzione f e sia $T_2(n) = O(n^{k_2})$ il tempo necessario per risolvere R_2 .

Poiché la funzione f può prendere un input di dimensione n e trasformarlo in un input di dimensione $O(n^{k_f})$ per R_2 , il tempo necessario per risolvere R_1 attraverso R_2 è $T_1(n) = O(n^{k_f k_2})$ che è un tempo *polinomiale*. \square

Definizione 175 - Problema \mathbf{NP} -arduo.

Un problema decisionale R si dice essere \mathbf{NP} -arduo se ogni problema $Q \in \mathbf{NP}$ è riducibile polinomialmente ad R , ovvero se $Q \leq_p R \quad \forall Q \in \mathbf{NP}$.

Definizione 176 - Problema \mathbf{NP} -completo.

Un problema decisionale R si dice essere \mathbf{NP} -completo se appartiene a \mathbf{NP} ed è \mathbf{NP} -arduo.

Se ogni problema decisionale \mathbf{NP} -completo appartenesse alla classe \mathbb{P} , allora risulterebbe $\mathbb{P} = \mathbf{NP}$. Tuttavia, ad oggi non siamo ancora riusciti a dimostrare se quell'uguaglianza sia vera o falsa.

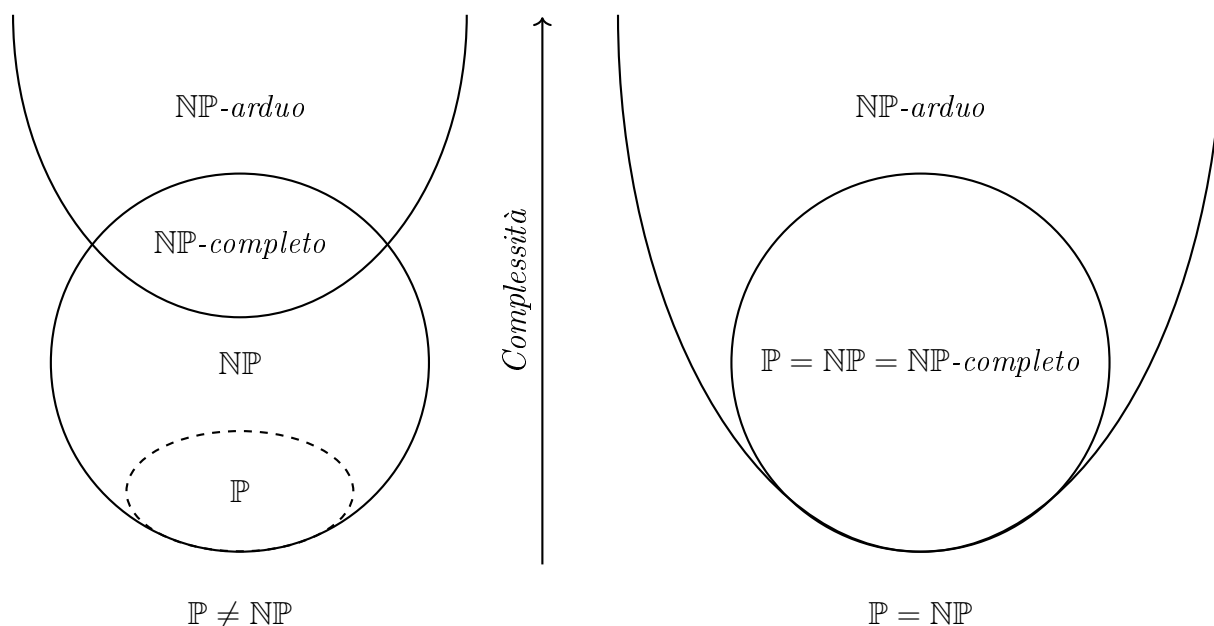


Fig. 17.7: \mathbb{P} VS \mathbf{NP}

Ora però ci troviamo di fronte a un problema: come facciamo a dimostrare che un problema è NP-completo se non conosciamo tutti i problemi in NP ?

La dimostrazione che $Q \leq_p R \quad \forall Q \in \text{NP}$ sembra apparentemente impossibile in quanto non conosciamo l'intero insieme NP , tuttavia, con una dimostrazione molto complessa, si è riuscito a dimostrare il *teorema di Cook-Levin* secondo il quale SAT è un problema NP-completo . Di conseguenza, partendo dalle *riduzioni* viste in precedenza e utilizzando il *teorema di Cook-Levin* otteniamo:

$$\text{SAT} \leq_p 3\text{-SAT} \leq_p \text{INDIPENDENT-SET} \leq_p \text{VERTEX-COVER} \leq_p \text{SAT}$$

Da qui, deriviamo che anche 3-SAT, INDIPENDENT-SET e VERTEX-COVER sono problemi NP-completi .

17.4 Alcuni problemi NP-completi

Problema 29 - Cricca (CLIQUE).

Dati un grafo non orientato ed un valore $k \in \mathbb{N}$, dire se esiste un sottoinsieme di almeno k nodi mutualmente adiacenti.

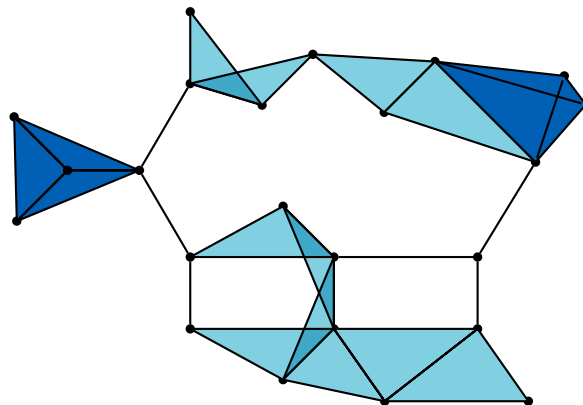


Fig. 17.8: Esempio di soluzione per CLIQUE

Problema 30 - Programmazione lineare 0/1.

Data una matrice A di elementi interi e di dimensione $m \times n$, ed un vettore b di m interi, dire se esiste un vettore x di valori 0/1 tale per cui $Ax \leq b$.

Esempio 46 - Esempio di soluzione.

Per A e b definiti come segue:

$$\begin{array}{rcl} \overbrace{x_1 + x_2 + x_3 + x_4}^A & \leq & \overbrace{2}^b \\ x_1 - x_2 - x_3 + x_4 & \leq & 0 \\ x_1 & + & x_3 + x_4 \leq 1 \end{array}$$

il sistema è verificato per $x_1 = x_2 = 1$ e $x_3 = x_4 = 0$.

Problema 31 - Commesso viaggiatore (TSP).

Date n città, le distanze tra esse ed un valore $k \in \mathbb{N}$, dire se è possibile definire un percorso che parta da una città, visiti tutte le altre esattamente una volta e ritorni alla città di partenza, che abbia una lunghezza totale non superiore a k .

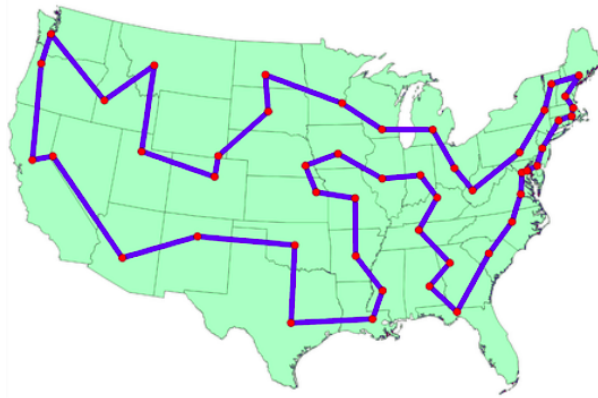


Fig. 17.9: Esempio di soluzione per TSP

Problema 32 - Copertura esatta di insiemi (EXACT-COVER).

Dati un insieme X e una collezione $\mathcal{Y} = \{Y_1, \dots, Y_n\}$ di sottoinsiemi di X , dire se esiste una sottocollezione $\mathcal{Z} \subseteq \mathcal{Y}$ che partizioni X .

Esempio 47 - Esempio di soluzione.

Siano X e \mathcal{Y} definiti come segue:

$$X = \{1, 2, 3, 4, 5, 6, 7\} \quad \mathcal{Y} = \{A, B, C, D, E, F\}$$

e siano gli elementi di \mathcal{Y} i seguenti sottoinsiemi:

$$A = \{1, 4, 7\} \quad B = \{1, 4\} \quad C = \{4, 5, 7\} \quad D = \{3, 5, 6\} \quad E = \{2, 3, 6, 7\} \quad F = \{2, 7\}$$

Una partizione di X è data da $\mathcal{Z} = \{B, D, F\}$. Un insieme \mathcal{Z} è una partizione di X in quanto $B \cap D \cap F = \emptyset$ e $B \cup D \cup F = X$.

Problema 33 - Partizione (PARTITION).

Dato un insieme A contenente n interi positivi, dire se esiste un sottoinsieme $S \subseteq \{1, \dots, n\}$ tale per cui $\sum_{i \in S} A[i] = \sum_{i \notin S} A[i]$.

Problema 34 - Somma di sottoinsieme (SUBSET-SUM).

Dati un vettore A contenente n interi positivi ed un intero positivo k , dire se esiste un sottoinsieme $S \subseteq \{1, \dots, n\}$ tale per cui $\sum_{i \in S} A[i] = k$.

Problema 35 - Zaino (KNAPSACK).

Dati un intero positivo C e un insieme di n oggetti tali per cui ciascun oggetto i è caratterizzato da un peso $w[i] \in \mathbb{Z}$ e da un profitto $p[i] \in \mathbb{Z}$, dire se esiste un sottoinsieme $S \subseteq \{1, \dots, n\}$ tale che il peso totale $w(S) = \sum_{i \in S} w[i] \leq C$ e il profitto totale $p(S) = \sum_{i \in S} p[i] \geq k$ con $k \in \mathbb{Z}$.

Problema 36 - Circuito hamiltoniano (HAMILTONIAN-CIRCUIT).

Dato un grafo non orientato G , dire se esiste un cammino che attraversi ogni nodo una sola volta.

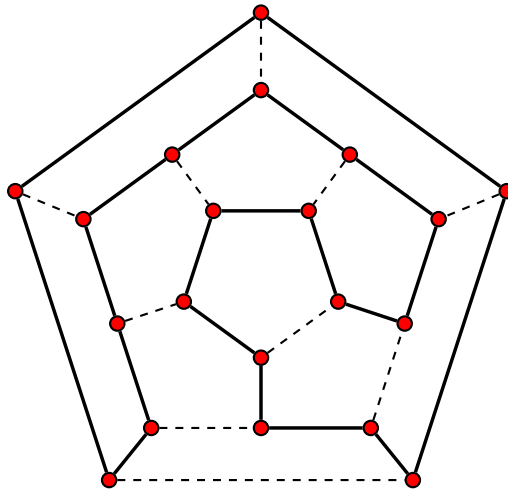


Fig. 17.10: Esempio di *circuito hamiltoniano*

Capitolo Nr.18

Soluzioni per problemi intrattabili

A questo punto è chiaro che per alcuni problemi non è sempre possibile trovare la soluzione ottima in tempo *polinomiale*. Per arrivare comunque ad una soluzione siamo dunque costretti a rinunciare a qualcosa, e questo significa perdere una o più delle seguenti caratteristiche:

- *Generalità*: potremmo accettare di definire un algoritmo che sia efficiente solo per alcuni casi dell'input;
- *Ottimalità*: potremmo accettare di ottenere soluzioni “vicine” a quella ottima;
- *Formalità*: potremmo accettare di definire un algoritmo che sperimentalmente dimostri fornire risultati buoni;
- *Efficienza*: potremmo accettare di perdere in efficienza, ma ottenere la soluzione globalmente ottima;

Come vedremo in questo capitolo, ciascuno di questi punti è associato ad una particolare categoria di algoritmi.

18.1 Algoritmi pseudo-polinomiali

Partiamo con un problema ormai noto: **SUBSET-SUM**. Abbiamo già risolto questo problema utilizzando il *backtracking*, ora però proviamo risolverlo di nuovo con una soluzione basata su *programmazione dinamica* per poi confrontare le *complessità* delle due soluzioni ottenute.

18.1.1 Problema del subset sum

Soluzione con programmazione dinamica Come al solito, definiamo una *tabella delle soluzioni* $DP[0 \dots n][0 \dots k]$ tale per cui $DP[i][r]$ è **true** se esiste un sottoinsieme dei primi i valori di A la cui somma sia esattamente r , **false** altrimenti:

$$DP[i][r] = \begin{cases} \text{true} & r = 0 \\ \text{false} & r > 0 \wedge i = 0 \\ DP[i-1][r] & r > 0 \wedge i > 0 \wedge A[i] > r \\ DP[i-1][r] \vee DP[i-1][r - A[i]] & r > 0 \wedge i > 0 \wedge A[i] \leq r \end{cases}$$

Le prime due clausole di questa definizione fanno sì che i valori della prima colonna della tabella siano tutti **true** e quelli della prima riga, ad eccezione del primo, siano tutti **false**. È giusto che sia così perché se $r = 0$ è sufficiente considerare l'insieme vuoto, mentre se $i = 0$ e $r > 0$ non è possibile definire alcun sottoinsieme diverso da quello vuoto.

Frammento 136 - Soluzione basata su programmazione dinamica.

```
boolean subsetSum(int[] A, int n, int k)
    boolean[][] DP = new boolean[0...n][0...k] = {false}
    for (i = 0 to n) do                                     % Prima colonna
        DP[i][0] = true                                    % Obiettivo raggiunto
    for (r = 1 to k) do                                     % Prima riga
        DP[0][r] = false                                   % Valori terminati
    for (i = 1 to n) do
        for (r = 1 to A[i] - 1) do                         %  $r < A[i]$ 
            DP[i][r] = DP[i - 1][r]
        for (r = A[i] to k) do                             %  $r \geq A[i]$ 
            DP[i][r] = DP[i - 1][r] or DP[i - 1][r - A[i]]
    return DP[n][k]
```

La *complessità* di questa soluzione è $\Theta(nk)$ perché DP ha quella dimensione.

Soluzione con backtracking

Frammento 137 - Soluzione basata su backtracking.

```
boolean ssRec(int[] A, int i, int r)
    if (r == 0) then                                       % Obiettivo raggiunto
        return true
    else if (i == 0) then                                  % Valori terminati
        return false
    else if (A[i] > r) then
        return ssRec(A, i - 1, r)
    else
        return ssRec(A, i - 1, r) or ssRec(A, i - 1, r - A[i])
```

Poiché nel caso peggiore vengono eseguite 2 chiamate ricorsive per ogni livello, la *complessità* è $O(2^n)$.

Soluzione con memoization Possiamo scrivere una soluzione basata su *memoization* in cui, per non dover inizializzare l'intera tabella, memorizziamo le soluzioni in un *dizionario*.

Frammento 138 - Soluzione basata su memoization.

```
boolean ssRec(int[] A, int i, int r, DICTIONARY DP)
    if (r == 0) then                                       % Obiettivo raggiunto
        return true
    else if (i == 0) then                                  % Valori terminati
        return false
    else
        boolean res = DP.lookup(<i, r>)
        if (res == nil) then                               % La soluzione non è ancora stata calcolata
            res = ssRec(A, i - 1, r, DP)                  % Valore non preso
            if (A[i] ≤ r) then
                res = res or ssRec(A, i - 1, r - A[i], DP) % Valore preso
            DP.insert(<i, r>, res)
    return res
```

In questo caso l'algoritmo è limitato superiormente sia da $O(nk)$ che da $O(2^n)$. Il primo limite dipende dal fatto che, nel caso peggiore, il *dizionario* viene popolato con tutti gli nk valori come avviene nella soluzione con *programmazione dinamica* classica.

Se invece il vettore A fosse popolato interamente di 1, ad ogni livello della ricorsione si realizzerebbero due chiamate, portando quindi la complessità a $O(2^n)$. Di conseguenza, la *complessità* della soluzione con *memoization* è $O(\min(nk, 2^n))$.

La *complessità* $O(nk)$ è *polinomiale*?

Precedentemente abbiamo già risposto a questa domanda, dicendo che no, non è una *complessità polinomiale*, bensì *pseudo-polinomiale*. Il motivo è che k è parte dell'input e non una sua dimensione. In particolare, k viene rappresentato da $t = \lceil \log k \rceil$ cifre binarie, quindi $O(nk)$ può anche essere scritto come $O(n2^t)$ che è una *complessità esponenziale*.

18.1.2 Problemi fortemente e debolmente NP-completi

Per continuare la discussione sul tema della *pseudo-polinomialità* introduciamo i *problemi fortemente e debolmente NP-completi*.

Definizione 177 - Dimensioni del problema.

Dati un problema decisionale R e un'istanza I , chiamiamo d la lunghezza della stringa binaria che codifica I e definiamo il valore $\#$ come il più grande numero che appare in I .

Ad esempio, per i problemi SUBSET-SUM, **CLIQUE** e **TSP**, i valori d ed $\#$ sono i seguenti:

Problema	I	$\#$	d
SUBSET-SUM	$\{n, k, A\}$	$\max\{n, k, \max(A)\}$	$O(n \log \#)$
CLIQUE	$\{n, m, k, G\}$	$\max\{n, m, k\}$	$O(n + m + \log \#)$
TSP	$\{n, k, d\}$	$\max\{n, k, \max(d)\}$	$O(n^2 \log \#)$

L'idea dietro il valore $\#$ è che se tutti i valori sono codificati nello stesso modo (e.g. tutti i valori di A sono interi a 32 bit), il valore maggiore definisce la dimensione minima della stringa di codifica per ciascun valore di quell'insieme. Conseguentemente, poiché d descrive la dimensione della stringa di codifica dell'intero input, il suo valore dipende dalla numerosità dell'input e dalla dimensione di ciascuna sua componente.

Da qui, nel problema SUBSET-SUM, d è definito come $O(n \log \#)$ perché l'input è composto da $n + 2$ valori: n , k e gli n valori del vettore A . Ciascuno di quegli elementi poi, richiede $\#$ bit per essere rappresentato, e nel *criterio di costo uniforme* tutto ciò si traduce in una *complessità* di $O(n \log \#)$.

Definizione 178 - Problema fortemenete NP-completo.

Sia R_{pol} il problema R ristretto ai dati di input per i quali $\#$ è limitato superiormente da $T_p(d)$, con T_p funzione polinomiale di d . R è fortemente NP-completo se e solo se R_{pol} è NP-completo.

Definizione 179 - Problema debolmente NP-completo.

Se un problema NP-completo non è fortemente NP-completo, allora è debolmente NP-completo.

Dimostrazione che SUBSET-SUM è debolmente NP-completo

Dimostrazione. Supponiamo che $A[i] \leq k \quad \forall i \in \{1, \dots, n\}$, perché in ogni caso valori più grandi di k non potrebbero essere parte della soluzione. Se $k = O(n^c)$, allora il valore $\#$ è definito come $\# = \max\{n, k, a_1, \dots, a_n\} = O(n^c)$. Ora, la soluzione basata su *programmazione dinamica* ha complessità $O(nk) = O(n^{c+1})$ che è *polinomiale*, e quindi appartenente a \mathbb{P} , quindi SUBSET-SUM non è *fortemente* NP-completo. \square

Definizione 180 - Complessità pseudo-polinomiale.

Un algoritmo ha complessità pseudo-polinomiale se risolve un certo problema R , per qualsiasi input I , in tempo $T_p(\#, d)$, con T_p funzione con tempo polinomiale non costante in $\#$.

Definizione 181 - Legame tra NP-completezza e pseudo-polinomialità.

Nessun problema fortemente NP-completo è risolvibile da un algoritmo pseudo-polinomiale, a meno che non valga $\mathbb{P} = \text{NP}$.

NB. Gli algoritmi per SUBSET-SUM e **KNAPSACK** sono *pseudo-polinomiali*.

Dimostrazione che CLIQUE è fortemente NP-completo

Dimostrazione. Possiamo supporre $k \leq n$, perché altrimenti la risposta sarebbe sicuramente **false**. Se è così, il valore $\#$ è definito come $\# = \max\{n, m, k\} = \max\{n, m\}$. Conseguentemente, d vale $O(n + m + \log \#) = O(n + m)$. Ora, siccome $\#$ definito in questo modo è già limitato superiormente da $O(n + m)$, il problema ristretto è uguale alla versione completa di CLIQUE che è NP-completo, quindi CLIQUE è un problema *fortemente* NP-completo. \square

L'idea alla base di questo tipo di dimostrazioni è quella di considerare un sottoinsieme degli input, calcolare $\#$ e d a partire dall'input ridotto e, se la versione ridotta del problema ha *complessità polinomiale*, allora il problema originale è *debolmente* NP-completo e ammette una soluzione di costo *pseudo-polinomiale*.

NB. Ovviamente, qualsiasi problema che è riducibile ad un problema *debolmente* NP-completo è esso stesso *debolmente* NP-completo. Ad esempio, **PARTITION** può essere ridotto a SUBSET-SUM scegliendo come k la meta della somma di tutti i valori presenti, e siccome SUBSET-SUM è *debolmente* NP-completo, lo è anche PARTITION.

18.2 Algoritmi di approssimazione

Diversamente da quanto fatto finora, per la trattazione degli *algoritmi di approssimazione* consideriamo *problemi di ottimizzazione* invece di *problemi decisionali*.

Definizione 182 - Algoritmo di approssimazione.

Se è possibile dimostrare che esiste un limite superiore o inferiore al rapporto tra la soluzione trovata e la soluzione ottima, allora l'algoritmo usato per ricavare la soluzione non ottima è detto essere un algoritmo di approssimazione.

Definizione 183 - Algoritmo di $\alpha(n)$ -approssimazione.

Dato un problema di ottimizzazione sul quale è definita una funzione di costo non negativa c ,

un algoritmo si dice essere di $\alpha(n)$ -approssimazione se fornisce una soluzione ammissibile x , il cui costo $c(x)$ non si discosti dal costo $c(x^*)$ della soluzione ottima x^* per più di un fattore $\alpha(n)$, per qualunque input di dimensione n .

I problemi di ottimizzazione si dividono in due sotto-categorie: problemi di massimizzazione e problemi di minimizzazione. A seconda della sotto-categoria di appartenenza del problema, un algoritmo di approssimazione è tale se soddisfa una delle seguenti relazioni:

$$\begin{array}{llll} \alpha(n)c(x^*) \leq c(x) \leq c(x^*) & \alpha(n) < 1 & \text{Problema di massimizzazione} \\ c(x^*) \leq c(x) \leq \alpha(n)c(x^*) & \alpha(n) > 1 & \text{Problema di minimizzazione} \end{array}$$

NB. Il fattore $\alpha(n)$ può dipendere da n o essere una costante, in ogni caso, dimostrare che il valore scelto sia rispettato dall'algoritmo è ciò che rende quell'algoritmo un *algoritmo di approssimazione*.

18.2.1 Problema del bin packing approssimato

Problema 37 - Problema del bin packing.

Dato un vettore A contenente n interi positivi, rappresentanti ciascuno il volume di un oggetto, e un intero positivo k , rappresentante la capacità di una scatola e tale per cui $A[i] \leq k \quad \forall i \in \{1, \dots, n\}$, trovare una partizione dell'insieme di indici $\{1, \dots, n\}$ che minimizzi il numero di sottoinsiemi disgiunti tali che $\sum_{i \in S} A[i] \leq k$ per ogni insieme S della partizione.

NB. In pratica, si vuole capire come distribuire n oggetti in delle scatole in modo che il numero di scatole totali sia minimo.

Questo problema può essere approcciato in una quantità di modi, i primi che potrebbero venire in mente sono il *best fit* e il *first fit*. Il primo cerca di piazzare ogni oggetto nella scatola in cui la capacità residua dopo l'inserimento è minima. Il secondo invece, prende un oggetto alla volta e lo inserisce nella prima scatola con capacità sufficiente per contenerlo.

Consideriamo l'algoritmo *first fit* e proviamo a vedere se è un *algoritmo di approssimazione*.

Dimostrazione. Sia N il numero di scatole usate dall'algoritmo *first fit*. Il numero minimo di scatole utilizzabili N^* è limitato inferiormente da:

$$N^* \geq \frac{\sum_{i=1}^n A[i]}{k}$$

Poiché non possono esserci due scatole riempite per meno della metà¹, N è limitato superiormente da:

$$N \leq \frac{\sum_{i=1}^n A[i]}{k/2}$$

Se è così, vale quanto segue:

$$N \leq \frac{\sum_{i=1}^n A[i]}{k/2} = 2 \frac{\sum_{i=1}^n A[i]}{k} \leq 2N^*$$

Da qui risulta che il fattore $\alpha(n)$ vale 2. □

NB. È anche possibile dimostrare limiti più stretti, ma quello che ci importa è aver dimostrato con successo che *first fit* è un *algoritmo di approssimazione* per il problema del *bin packing*.

¹Non possono esserci due scatole riempite per meno della metà perché il contenuto di una di quelle due sarebbe stato messo nell'altra

18.2.2 Problema del commesso viaggiatore modificato

Problema 38 - Commesso viaggiatore con disuguaglianze triangolari (Δ -TSP).

Date n città e una matrice d delle distanze tra esse tale per cui:

$$d[i][j] \leq d[i][k] + d[k][j] \quad \forall i, j, k : 1 \leq i, j, k \leq n$$

trovare un percorso che parta da una città, visiti tutte le altre esattamente una volta e ritorni alla città di partenza, in modo che la distanza complessiva percorsa sia minima.

NB. Le distanze tra le città devono rispettare la cosiddetta “disuguaglianza triangolare”.

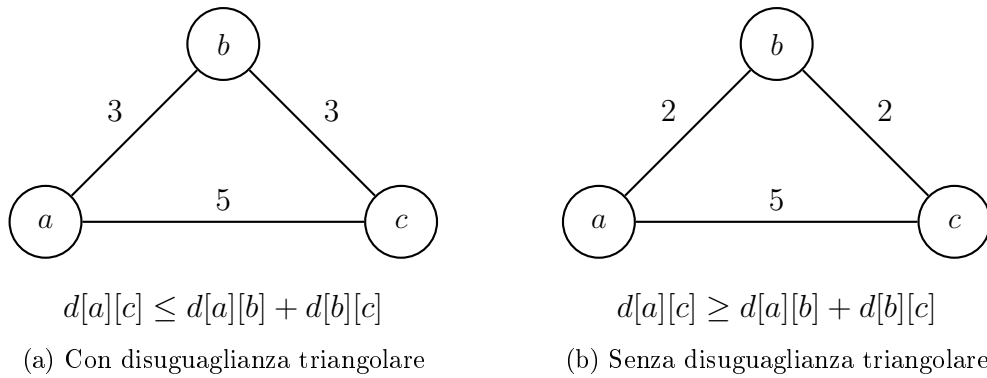


Fig. 18.1: Grafo con e senza disuguaglianza triangolare

NP-completezza di Δ -TSP Possiamo dimostrare che Δ -TSP è *riducibile polinomialmente* a **HAMILTONIAN-CIRCUIT** e poiché questo è NP-completo, lo è anche Δ -TSP.

Dimostrazione. Sia $G = (V, E)$ un grafo non orientato avente un nodo per ognuna delle n città, e definiamo a partire da esso una matrice delle distanze tale per cui:

$$d[i][j] = \begin{cases} 1 & (i, j) \in E \\ 2 & (i, j) \notin E \end{cases}$$

Poiché valgono le disuguaglianze triangolari, per ogni coppia di nodi i, j vale $d[i][j] \leq 2 \leq d[i][k] + d[k][j]$. A questo punto, G ha un circuito hamiltoniano se e solo se esiste un cammino in d di lunghezza n e passante per ogni nodo. \square

NB. La matrice d definisce il costo dell'arco tra ogni coppia di nodi, quindi ci permette di trattare G come se fosse un grafo completo.

In generale possiamo sempre interpretare TSP o Δ -TSP come il problema di ricercare il circuito hamiltoniano di peso minimo su un grafo pesato completo.

Algoritmo in generale Per definire un algoritmo di approssimazione per Δ -TSP partiamo proprio da questa interpretazione. In particolare, consideriamo un circuito hamiltoniano e ne cancelliamo un arco in modo da ottenere un albero di copertura.

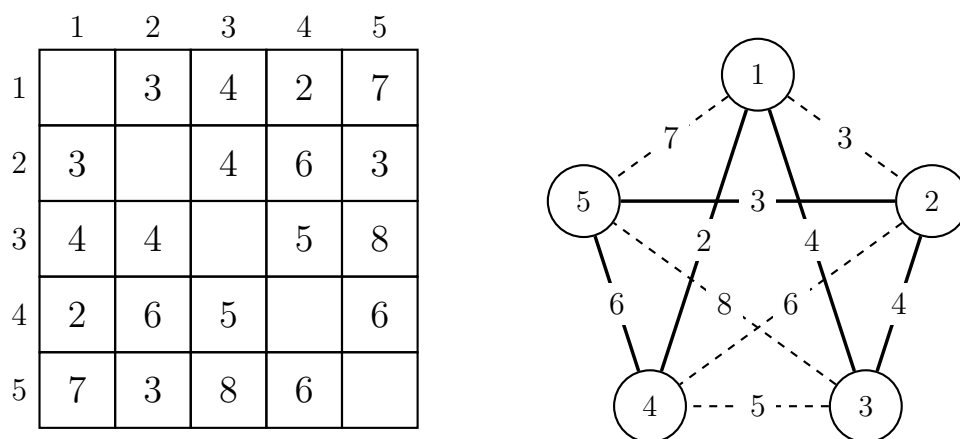


Fig. 18.2: Interpretazione di Δ -TSP come circuito hamiltoniano pesato

Definizione 184 - Lemma sul costo dell'albero di copertura di peso minimo.

Qualunque circuito hamiltoniano π ha un costo $c(\pi)$ che è superiore al costo mst di un albero di copertura di peso minimo dello stesso grafo. Ovvero, per ogni circuito hamiltoniano π vale $mst < c(\pi)$.

Dimostrazione. Supponiamo per assurdo che esista un circuito hamiltoniano π di costo $c(\pi) \leq mst$. Se eliminiamo un *arco* di π otteniamo un altro *albero di copertura* mst' di costo $mst' < c(\pi) \leq mst$. Questo genera una contraddizione in quanto avevo supposto che mst fosse il costo dell'*albero di copertura di peso minimo*. \square

Quindi, se individuiamo un *albero di copertura* di peso mst e ne percorriamo ogni *arco* due volte, prima in un senso e poi nell'altro, otteniamo un circuito di costo $2 \cdot mst$ nel quale ogni *nodo* viene visitato una o due volte. Ovviamente questo non è un circuito hamiltoniano, ma se saltiamo i *nodi* già visitati e passiamo direttamente a quelli non visitati, poiché valgono le disuguaglianze triangolare, il costo del circuito π così ottenuto è inferiore o uguale a $2 \cdot mst$.

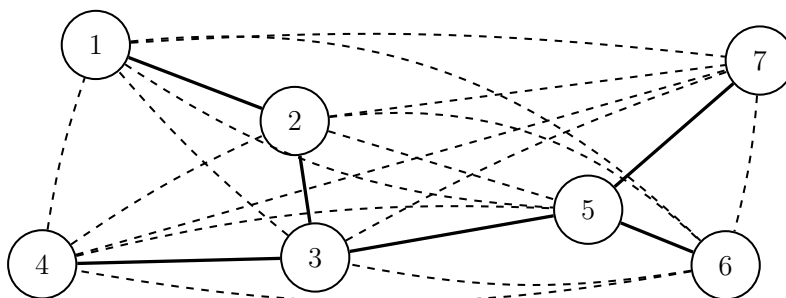
Di conseguenza, vale:

$$c(\pi) \leq 2 \cdot mst \leq 2 \cdot c(\pi^*)$$

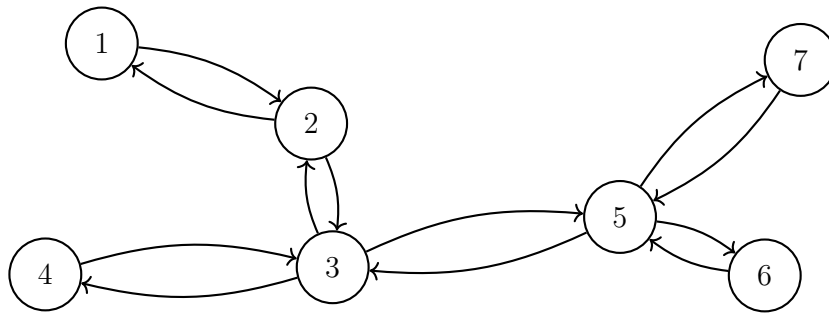
dove $c(\pi^*)$ è il costo del circuito hamiltoniano di peso minimo e 2 corrisponde anche al valore del fattore $\alpha(n)$.

Esempio 48 - Esempio d'esecuzione.

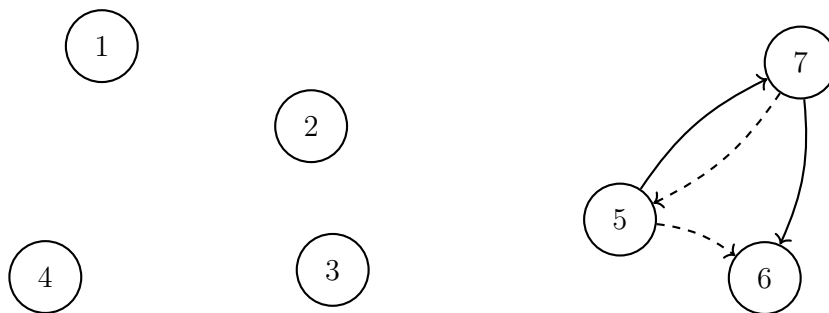
Consideriamo il seguente grafo completo sul quale abbiamo individuato un albero di copertura di peso minimo:



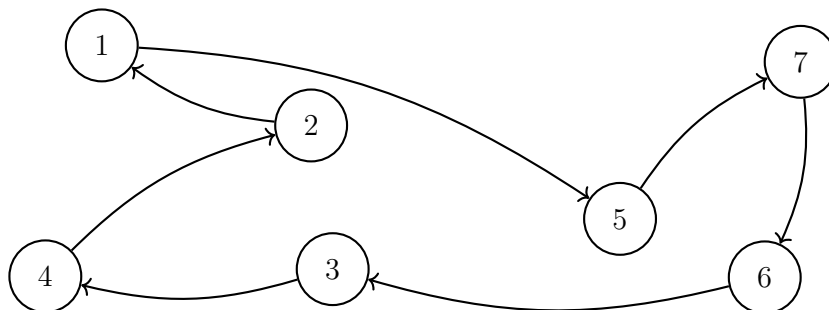
Se adesso percorressimo ogni arco dell'albero di copertura individuato in entrambi i sensi, otterremmo il seguente circuito:



A questo punto, sfruttando le disuguaglianze triangolari, otteniamo che per esempio, per passare dal nodo 7 al nodo 6 conviene usare l'arco (7,6) invece che tornare indietro sul nodo 5 già visitato:



Applicando la stessa nozione a tutti gli altri nodi, otteniamo finalmente un circuito hamiltoniano:



Complessità L'algoritmo proposto sfrutta l'Algoritmo di Kruskal per la ricerca dell'albero di copertura di peso minimo e una visita in profondità. L'Algoritmo di Kruskal costa $O(n^2 \log n)$ e la visita $O(n)$, quindi la complessità totale è $T(n) = O(n^2 \log n)$.

Definizione 185 - Teorema di non approssimabilità di TSP.

Non esiste alcun algoritmo di $\alpha(n)$ -approssimazione per TSP tale che $c(x') \leq s \cdot c(x^*)$, con $s \in \mathbb{N}$, a meno che non valga $\mathbb{P} = \mathbb{NP}$.

18.3 Algoritmi euristici

Come abbiamo già visto in precedenza, gli algoritmi euristici permettono di ottenere delle soluzioni ammissibili per problemi complessi, che però possono non essere né soluzioni ottime

né soluzioni approssimate. Spesso le soluzioni euristiche vengono definite a partire da intuizioni per *algoritmi greedy* oppure sfruttano di tecniche tipiche della *ricerca locale*.

Continuiamo la trattazione continuando a parlare di TSP e vediamo due soluzioni euristiche di tipo *greedy* e una di *ricerca locale*.

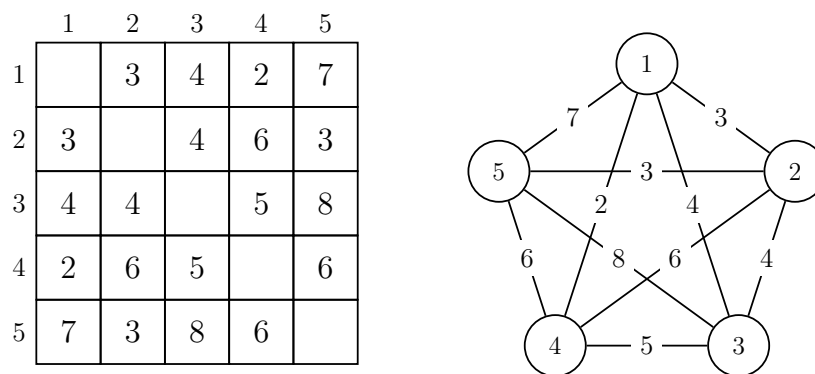
18.3.1 Prima soluzione greedy - Arco di costo minimo

Un'idea potrebbe essere quella di ordinare gli *archi* per pesi non decrescenti e di aggiungere un *arco* alla volta alla soluzione, prendendo di volta in volta quello di peso minore. Ovviamente la frase precedente non è formalmente corretta, in quanto non possiamo prendere un *arco* se uno dei suoi estremi è già raggiunto da due *archi* della soluzione. Inoltre, dobbiamo anche evitare che si formino circuiti, e per questo sfruttiamo i *merge-find set*. Infine, se n è il numero di *nod*i, l' n -esimo *arco* deve essere per forza l'*arco* che collega gli ultimi due *nod*i estremi della catena.

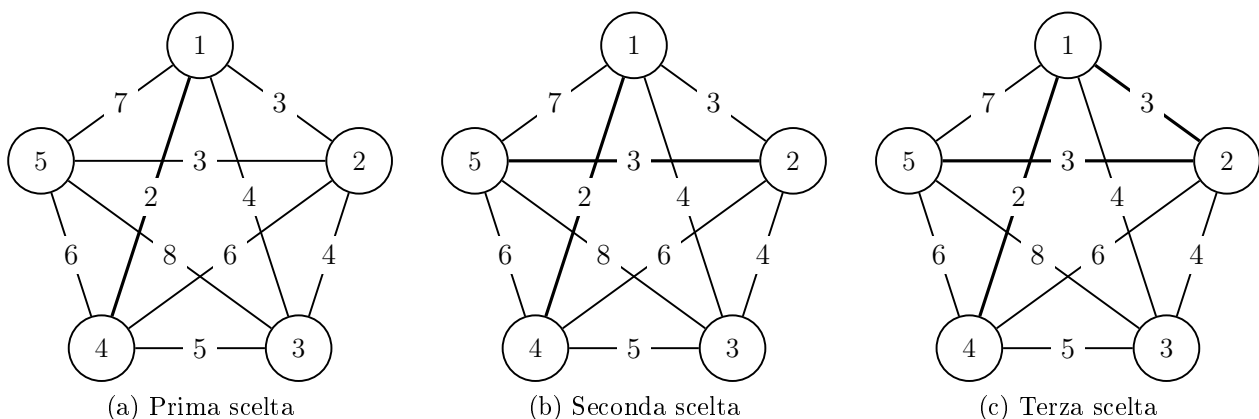
NB. Un circuito hamiltoniano privo di un *arco* è semplicemente una “catena” di *nod*i.

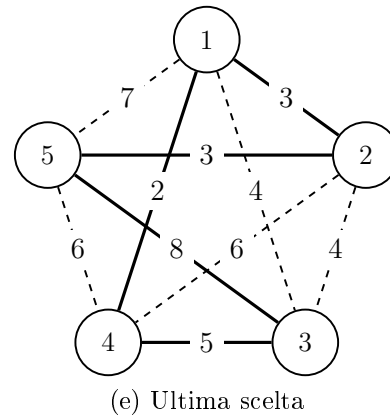
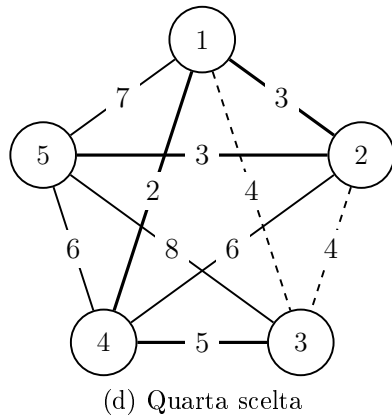
Esempio 49 - Esempio d'esecuzione.

Consideriamo lo stesso grafo completo dell'esempio precedente:



Passo dopo passo, l'algoritmo opera le seguenti scelte:





Dopo la terza scelta vengono estratti i due archi di peso 4, ma poiché i nodi 1 e 2 sono già collegati a due archi nell'insieme della soluzione, i due archi da 4 non possono essere selezionati. Alla fine invece, l'ultimo arco scelto è proprio quello più costoso per cui tutti gli i precedenti archi sono stati estratti e ignorati.

Il costo del circuito individuato è 21 e noi già sappiamo che quello di costo minimo ha un costo di 19, quindi, come ci aspettavamo, la soluzione trovata non è ottimale.

NB. Sarebbe possibile interrompere l'estrazione degli archi dopo l'($n - 1$)-esimo, ma questo tipo di accorgimenti non cambiano le dinamiche generali dell'algoritmo, e soprattutto non ne modificano la complessità.

Frammento 139 - Implementazione prima soluzione greedy.

```

SET greedyTsp(GRAPH G)
    SET result = Set()                                     % Insieme della soluzione
    MFSET M = Mfset(G.size())
    int[] edges = new int[1..G.size()] = {0}              % Archi collegati a ogni nodo
    int[] A = { Ordina gli archi per peso non decrescente }
    foreach (<u, v> ∈ A) do
        if (edges[u] < 2 and edges[v] < 2 and M.find(u) ≠ M.find(v)) then
            result.insert(<u, v>)
            edges[u] = edges[u] + 1
            edges[v] = edges[v] + 1
            M.merge(u, v)
    % Cerca le due estremità della catena per determinare l'arco mancante
    int u = 1
    while (edges[u] ≠ 1) do
        u = u + 1
    int v = u + 1
    while (edges[v] ≠ 1) do
        v = v + 1
    result.insert(<u, v>)
    return result

```

Complessità L'ordinamento degli archi costa $O(m \log m)$, ma poiché stiamo lavorare con grafi completi $m = O(n^2)$, quindi la complessità diventa $O(n^2 \log n)$.

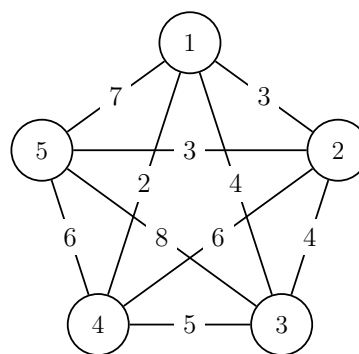
18.3.2 Seconda soluzione greedy - Vicino più vicino

Un'altra possibilità è quella di partire da un *nodo* e di muoversi di volta in volta verso il *nodo* *adiacente* più vicino, cioè quello collegato con l'*arco* di peso minore. Una visitati tutti i *nodi*, si seleziona l'*arco* di peso minimo verso il *nodo* di partenza.

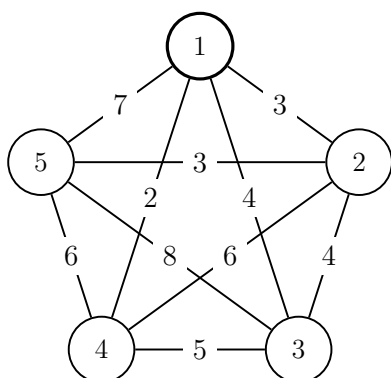
Esempio 50 - Esempio d'esecuzione.

Continuiamo a considerare sempre lo stesso grafo completo:

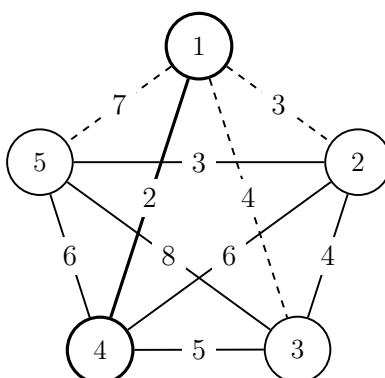
	1	2	3	4	5
1		3	4	2	7
2	3		4	6	3
3	4	4		5	8
4	2	6	5		6
5	7	3	8	6	



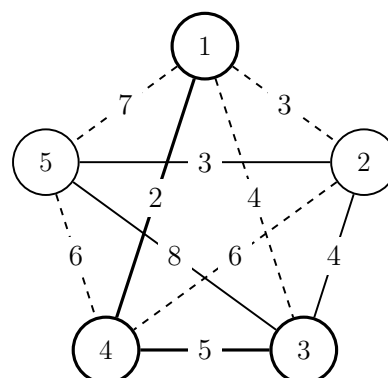
Passo dopo passo, l'algoritmo opera le seguenti scelte:



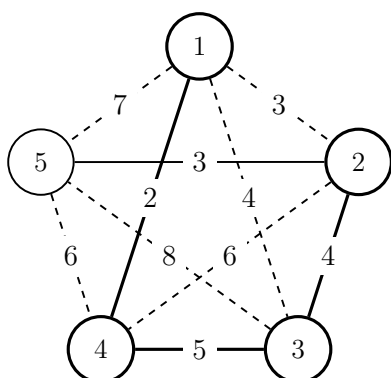
(a) Prima scelta



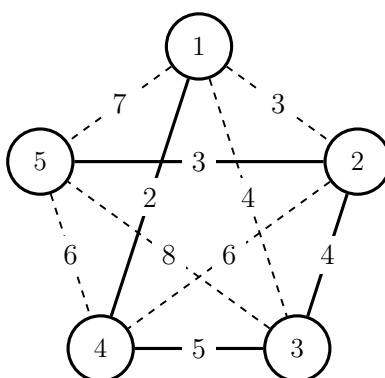
(b) Seconda scelta



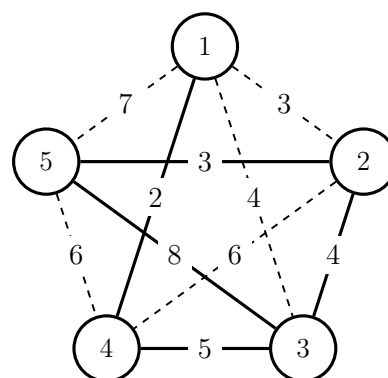
(c) Terza scelta



(d) Quarta scelta



(e) Quinta scelta



(f) Ultima scelta

Il circuito ottenuto costa 21 come il precedente.

Complessità Per quanto riguarda la *complessità*, questa soluzione si comporta meglio della precedente, in quanto scegliere l'*arco* uscente di peso minimo per $n-1$ *nodi* di un *grafo completo* costa $O(n^2)$.

18.3.3 Soluzione con ricerca locale

L'idea è quella che se π è un qualche circuito hamiltoniano del *grafo*, possiamo modificare π e ottenere un circuito di costo minore. In particolare, possiamo scegliere due *archi* non consecutivi in π e sostituirli con altri due *archi* non già inclusi in π . Quindi, il nuovo circuito π' ottenuto è un intorno di π nel quale abbiamo cambiato due *archi*.

Il processo di sostituzione termina quando non è più possibile ridurre il costo del circuito, ovvero quando si arriva ad un minimo locale, che come abbiamo già visto, potrebbe non coincidere con il minimo globale.

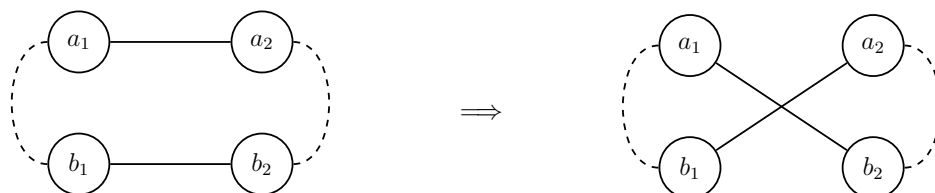
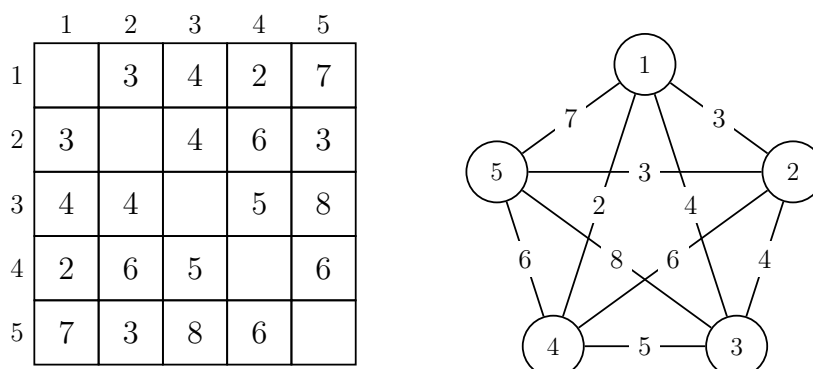


Fig. 18.2: Esempio di sostituzione

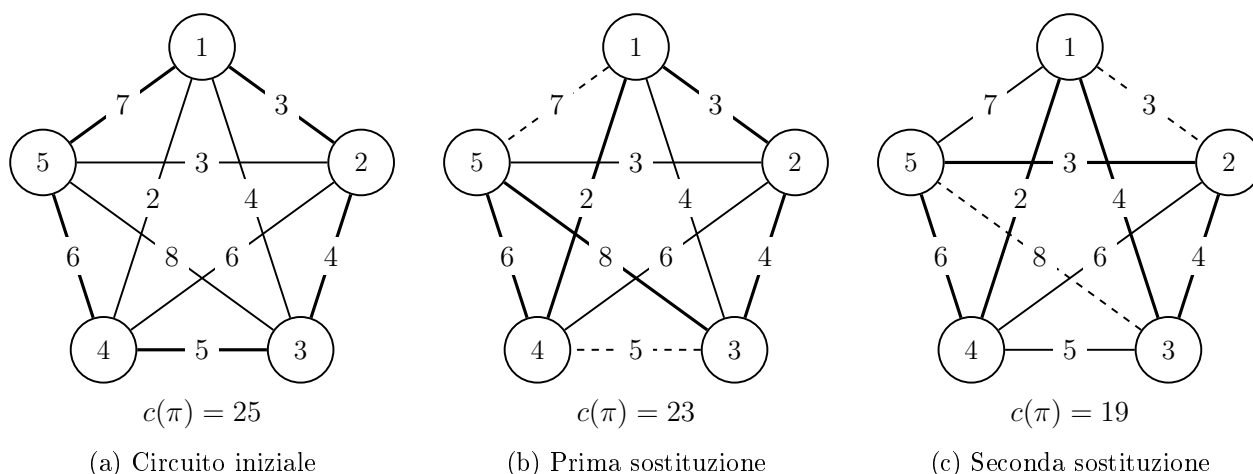
Scelti i due *archi* (a_1, a_2) e (b_1, b_2) da sostituire, l'unico modo per richiudere il circuito senza riutilizzare quegli stessi *archi*, è quello di sostituirli con gli *archi* (a_1, b_1) e (a_2, b_2) .

Esempio 51 - Esempio d'esecuzione.

Riprendiamo il solito *grafo completo*:



Partendo da un circuito hamiltoniano qualsiasi, l'algoritmo si comporta come segue:



Dopo due sostituzioni siamo arrivati a un minimo locale, che in questo caso è anche globale visto che il costo del circuito finale è minimo.

Complessità La *complessità* di questo algoritmo dipende sia dal costo che paghiamo per definire il circuito di partenza π che dal costo che paghiamo per ricercare una soluzione migliore nell'intorno di π . Nel primo caso possiamo usare uno degli *algoritmi greedy* appena descritti, mentre per esaminare l'intorno paghiamo $O(n^2)$ perché il numero di coppia di *archi* di un circuito è $n(n-1)/2$ e di questi n sono consecutivi. Di conseguenza, $|I_2(\pi)| = n(n-1)/2 - n = O(n^2)$.

18.4 Algoritmi branch-&-bound

Gli *algoritmi branch-&-bound* sono molto simili agli *algoritmi backtrack*, ma introducono delle accortezze che permettono di ottenere “potare” una parte più consistente dell'*albero delle soluzioni*.

NB. Per semplificare la trattazione, ipotizziamo che ogni sequenza di scelte abbia un costo non negativo e che quindi non sia possibile ridurre il costo di una soluzione parziale aggiungendo un'altra scelta alla sequenza. Inoltre, ci concentriamo solo su *problemi di minimizzazione*.

Lo strumento che questa tipologia di algoritmi usano per decidere che rami potare è costituito dalle funzioni di *upper bound* e *lower bound*. Nei *problemi di minimizzazione*, l'*upper bound* è definito come il costo della soluzione ammissibile nota migliore. Il *lower bound* è invece calcolato mediante una funzione *lb* con la seguente firma:

$$lb(\langle \text{dati problema} \rangle, S, i, \langle \text{dati parziali} \rangle)$$

Tale funzione deve essere definita in modo che il proprio valore dipenda dalla sequenza $S[1 \dots i]$ di scelte fatte e che il costo tutte le scelte ammissibili generabili da quella sequenza sia maggiore di quello calcolato.

Se il *lower bound* calcolato per una soluzione parziale è maggiore o uguale all'*upper bound*, l'*albero* radicato in quella soluzione può essere potato in quanto il suo costo sarà maggiore di quello della soluzione ottimale.

NB. È importante esplicitare che l'utilizzo di tecniche di potatura di questo tipo non abbassa la *complessità* della soluzione, ma ne riduce comunque il tempo d'esecuzione. Ovviamente, maggiore è l'accuratezza di *lb*, maggiore è la porzione di *albero* ad essere potato.

18.4.1 Schema generale

Frammento 140 - Schema generale per algoritmi branch-&-bound.

```
branch&bound(⟨dati problema⟩, ITEM[] S, int i, ⟨dati parziali⟩)
  SET C = choices(⟨dati problema⟩, S, i, ⟨dati parziali⟩)
  foreach (c ∈ C) do
    S[i] = c
    int lb = lb(⟨dati problema⟩, S, i, ⟨dati parziali⟩)
    if (lb < minCost) then                                     % Se false, pota l'albero
      if (i < n) then
        branch&bound(⟨dati problema⟩, S, i + 1, ⟨dati parziali⟩)
      else                                                     % Valuta il costo della soluzione ammissibile trovata
        if (cost(S, i) < minCost) then
          minSol = S                                           % Variabile globale
          minCost = cost(s, i)                                 % Variabile globale
```

18.4.2 Problema del commesso viaggiatore

Se n sono le città da visitare e $d[h][k]$ è la distanza tra h e k espressa come valore naturale, al passo i -esimo sono state fatte le scelte $S[1 \dots i]$, ovvero sono state visitate i città prese dall'insieme $\{1, \dots, n\}$.

Ciò che dobbiamo fare adesso è trovare un'implementazione di lb che ci permetta di potare il più possibile.

Lower bound come costo effettivo della soluzione parziale

Definizione 186 - Costo effettivo della soluzione parziale.

Definiamo $cost[i]$ come la somma del peso degli archi percorsi nei primi i passi, ovvero:

$$cost[i] = \sum_{h=2}^i d[S[h-1]][S[h]]$$

La cosa più facile è definire lb come:

$$lb(d, S, i) = cost[i]$$

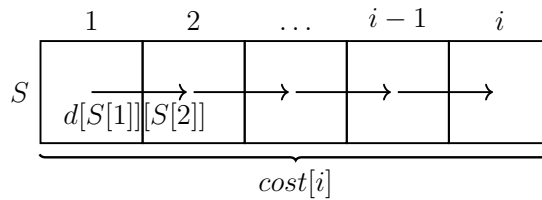


Fig. 18.3: Costo effettivo della soluzione parziale

Lower bound come costo per uscire dall'ultimo nodo

Definizione 187 - Costo per uscire dall'ultimo nodo.

Definiamo $out[S[i]]$ come il minimo fra i pesi degli archi che escono da $S[i]$ e vanno in uno dei nodi non ancora scelti, ovvero:

$$out[S[i]] = \min_{k \notin S[1 \dots i]} d[S[i]][k]$$

Noto il valore $out[S[i]]$, lb può essere definito come:

$$lb(d, S, i) = cost[i] + out[S[i]]$$

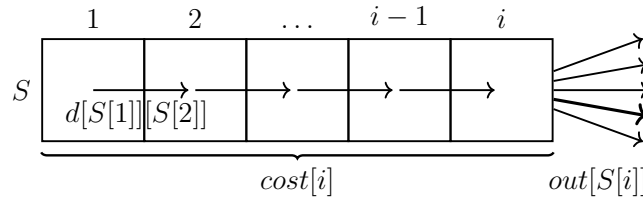


Fig. 18.4: Costo per uscire da un *nodo*

Lower bound come costo per uscire da tutti i nodi rimasti

Definizione 188 - Costo per uscire da tutti i nodi rimasti.

Per ogni nodo h non ancora scelto, sommiamo il minimo tra i pesi degli archi che escono da esso e vanno in $S[1]$ o in un nodo non ancora scelto, ovvero:

$$out[h] = \min_{k \notin S[2 \dots i]} d[h][k] \quad \forall k \notin S[1 \dots i]$$

Noto il valore $out[S[i]]$, lb può essere definito come:

$$lb(d, S, i) = cost[i] + out[S[i]] + \sum_{h \notin S[1 \dots i]} out[h]$$

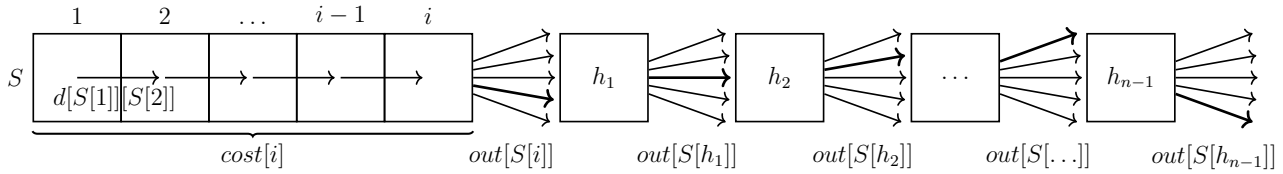


Fig. 18.5: Costo per uscire da tutti i *nodi* rimasti

Lower bound come costo per attraversare i nodi non ancora scelti

Definizione 189 - Costo per attraversare i nodi non ancora scelti.

Per ogni non h non ancora scelto, calcoliamo il costo di attraversamento $transfer[h]$ come la somma tra i pesi minimi di due archi (p, h) e (h, q) che entrano ed escono da h andando in $S[1]$ un in nodo non ancora scelto. La definizione di $transfer[h]$ è quindi la seguente:

$$transfer[h] = \min_{p \notin S[1 \dots i-1], q \notin S[2 \dots i] : p \neq q \neq h} \{d[p][h] + d[h][q]\}$$

Definizione 190 - Costo per tornare al primo nodo.

Definiamo il valore $last$ come il minimo tra i pesi degli archi che vanno da uno dei nodi non ancora scelti in $S[1]$, ovvero:

$$last = \min_{h \notin S[1 \dots n]} d[h][S[1]]$$

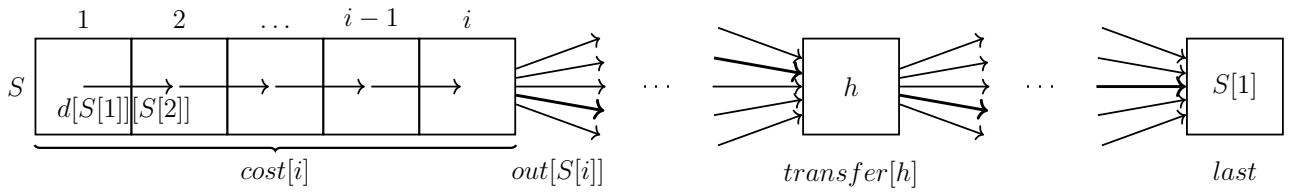


Fig. 18.6: Costo per attraversare tutti i *nod*i rimasti

A questo punto, definiamo *lb* come:

$$lb(d, S, i) = cost[i] + \left\lceil \frac{out[S[i]] + \sum_{h \notin S} transfer[h] + last}{2} \right\rceil$$

Stiamo dividendo per 2 perché il peso di ogni *arco* viene contato due volte: in ingresso e in uscita.

Frammento 141 - Implementazione della soluzione.

```

bbTsp(ITEM[] S, int cost, SET R, int n, int i)
    SET choices = copy(R)
    foreach (c ∈ choices) do
        S[i] = c
        R.remove(c)
        if (i < n) then
            { Calcola out, last e transfer[h] per ogni h ∈ R }
            int lb = cost[i] + ⌈(out + ∑h∉Stransfer[h] + last) / 2⌉
            if (lb < minCost) then
                bbTsp(S, cost + d[S[i - 1]][S[i]], R, n, i + 1)
        else
            cost = cost + d[S[i]][S[1]]
            if (cost < minCost) then
                minSol = S
                minCost = cost
    R.insert(c)

```

Invece di inizializzare la variabile globale `minCost` a $+\infty$, possiamo partire da una permutazione qualsiasi dei *nod*i. Inoltre, per evitare di generare più volte lo stesso circuito, possiamo decidere di iniziare sempre dallo stesso *nodo*.

Complessità Come accennato all'inizio della sezione, l'utilizzo delle funzioni *upper bound* e *lower bound* non migliorano la *complessità*, che rimane quella di una soluzione con *backtracking*.

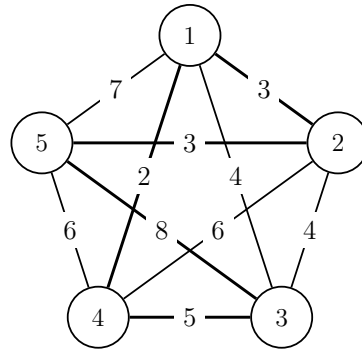
Esempio 52 - Esempio d'esecuzione.

Consideriamo lo stesso grafo completo degli esempi precedenti e scegliamo un circuito iniziale che parte da 1, per esempio:

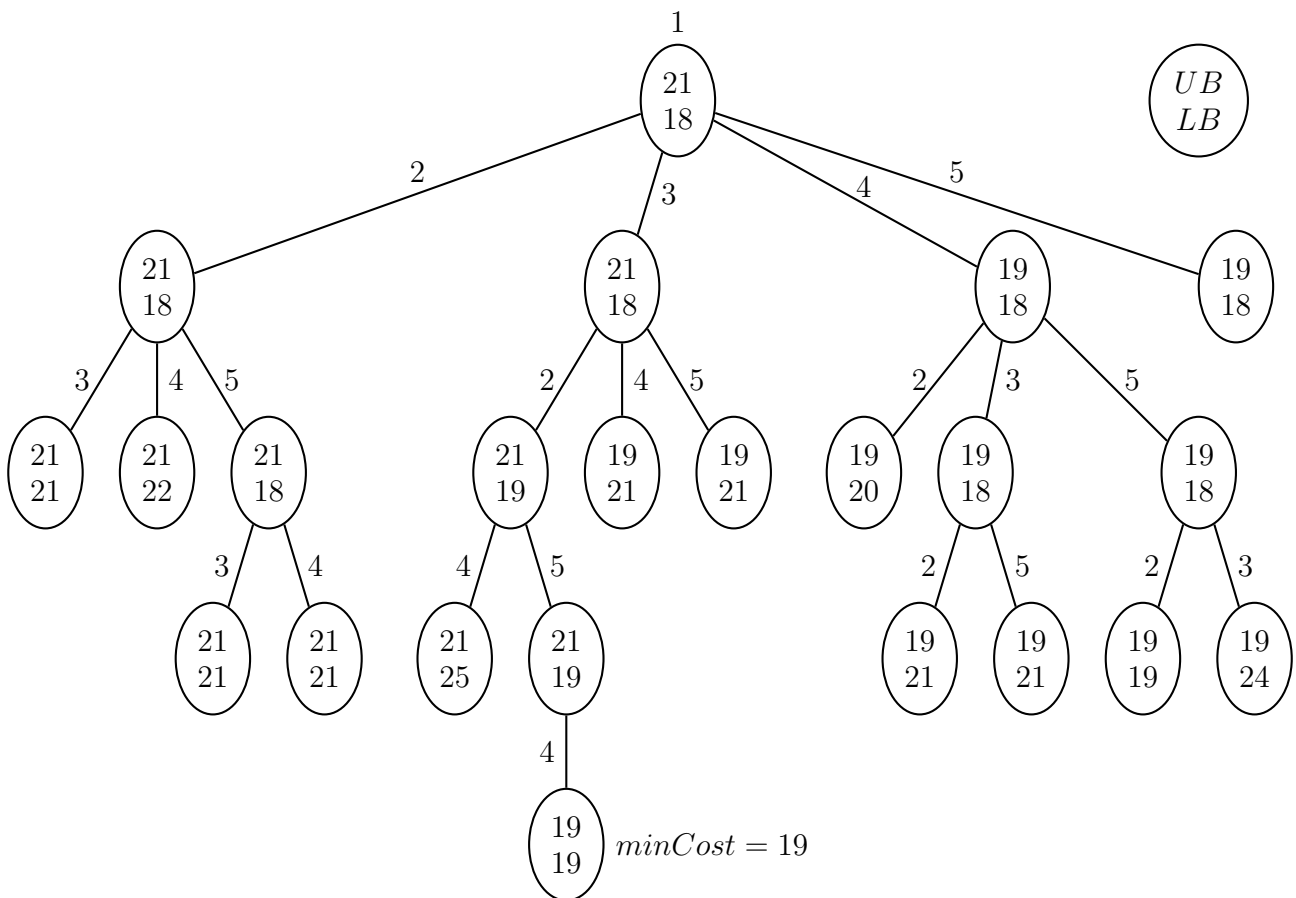
$$S = \{1, 2, 5, 3, 4\}$$

Il costo di questo percorso è 21, quindi $minCost = 21$ costituisce il nostro primo *upper bound*.

	1	2	3	4	5
1		3	4	2	7
2	3		4	6	3
3	4	4		5	8
4	2	6	5		6
5	7	3	8	6	



Proviamo a vedere quali rami dell'albero delle soluzioni vengono potati con questi parametri iniziali:



In questo caso siamo riusciti a potare 42 nodi su un totale di 65.

Algoritmi di ordinamento

Nel corso della trattazione abbiamo già parlato più volte di algoritmi di ordinamento, tuttavia, poiché il problema di ordinare in modo efficiente un insieme di valori è molto comune, è doveroso affrontare dedicare all'argomento una propria sezione.

19.1 Algoritmi basati su confronti

Tutti gli algoritmi di ordinamento che abbiamo visto finora si basano su confronti (i.e. $<$, $=$, $>$) tra coppie di valori e gli algoritmi migliori avevano una *complessità* pari a $O(n \log n)$.

Possiamo fare meglio di così?

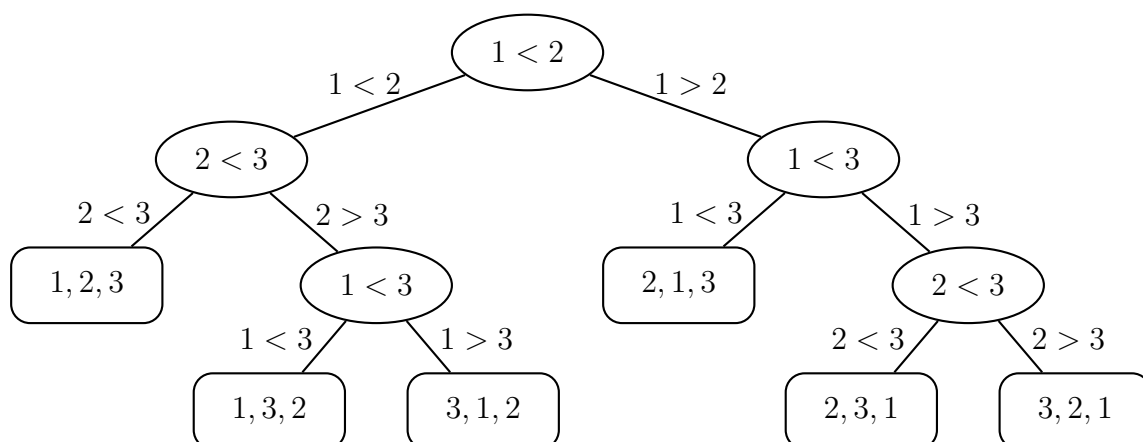
Sta volta la risposta è no, perché il limite inferiore alla *complessità* di qualsiasi algoritmo basato su confronti è $\Omega(n \log n)$. Questo limite deriva dal seguente teorema:

Definizione 191 - Limite inferiore alla complessità per soluzioni basate su confronti.

Dato un qualsiasi algoritmo di ordinamento basato su confronti, nel caso peggiore, ordinare n elementi costa $\Omega(n \log n)$.

NB. Per la dimostrazione del teorema assumiamo che tutti i valori dell'insieme siano distinti e che i confronti eseguiti dal generico algoritmo considerato siano rappresentabili tramite un *albero binario*.

Dimostrazione. Se supponiamo di dover ordinare 3 valori, l'*albero binario* dei confronti è il seguente:



In un *albero* di questo tipo, ciascun *cammino radice-foglia* rappresenta una sequenza di confronti eseguiti dall'algoritmo e l'*altezza* dell'*albero* corrisponde al numero di confronti eseguiti nel caso pessimo.

Se consideriamo tutti gli *alberi* ottenibili da algoritmi di ordinamento basati su confronti, notiamo che se n è il numero di elementi da ordinare, valgono le due seguenti proprietà:

1. L'*albero* possiede almeno $n!$ *foglie*;
2. Se ogni *nodo* dell'*albero* ha esattamente due *figli* e k è il numero di *foglie*, l'*altezza* dell'*albero* è almeno $\log k$, ovvero $\Omega(\log k)$;

Poiché $n!$ può essere approssimato a n^n e $\log n^n = n \log n$, nel caso pessimo, il numero di confronti eseguiti è $\Omega(n \log n)$. \square

19.2 Algoritmi non basati su confronti

19.2.1 Spaghetti sort

L'idea è molto semplice: ogni valore da ordinare viene associato ad uno spaghetti di lunghezza pari al valore stesso. Tutti gli spaghetti vengono poi appoggiati verticalmente su una superficie orizzontale e, uno per volta, vengono rimossi dal tavolo in ordine decrescente di altezza. Per ordinare i valori in senso crescente è sufficiente che l'altezza di ogni spaghetti rimosso venga scritta nell'ultima cella libera del vettore.

Se n sono i valori da ordinare, la *complessità* è $\Theta(n)$.

NB. Come risulterà ovvio, lo *spaghetti sort* è pensato unicamente come descrizione astratta per algoritmi di ordinamento in tempo lineare, e non può quindi essere utilizzato in pratica.

19.2.2 Counting sort

Senza perdere generalità, possiamo assumere che i valori da ordinare siano inclusi in un intervallo $[1, \dots, k]$. Il *counting sort* costruisce un vettore $B[1 \dots k]$ e lo usa per contare il numero di volte che un valore compreso in $[1, \dots, k]$ compare nel vettore originale. A quel punto, per ogni indice $i \in [1, \dots, k]$, i viene inserito nel vettore originale tante volte quanto vale $B[i]$.

Frammento 142 - Implementazione counting sort.

```
countingSort(int[] A, int n, int k)
    int[] B = new int[1...k]
    for (i = 1 to k) do
        B[i] = 0
    for (j = 1 to n) do
        B[A[j]] = B[A[j]] + 1
    int j = 1
    for (i = 1 to k) do
        while (B[i] > 0) do
            A[j] = i
            j = j + 1
            B[i] = B[i] - 1
```

% A viene popolato a partire dalla prima cella

Complessità La *complessità* è $O(n + k)$ che è una *complessità pseudo-polinomiale* perché k non descrive la dimensione dell'input. In ogni caso, se $k = O(n)$, il *counting sort* opera in tempo lineare. La *complessità spaziale* è $O(k)$ perché viene allocato un nuovo vettore.

19.2.3 Pigeonhole sort

Chiaramente *counting sort* può ordinare soltanto vettori di interi, quindi, per esempio, non lo potremmo usare per ordinare record associati ad una chiave. Il *pigeonhole sort* è una variante del *counting sort* in cui invece di contare il numero di occorrenze di ciascuna chiave, sfruttiamo delle *liste concatenate*.

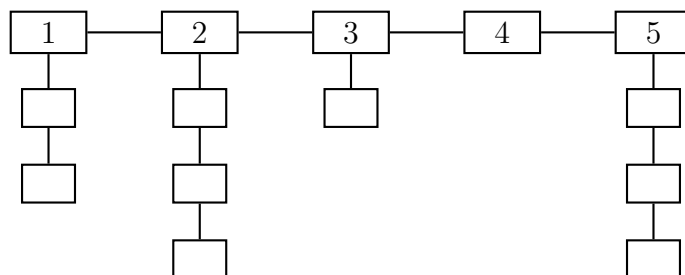


Fig. 19.1: *Pigeonhole sort*

Complessità Essendo una variante del *counting sort*, la *complessità temporale* rimane $O(n + k)$, mentre la *complessità spaziale* sale a $O(n + k)$.

19.2.4 Bucket sort

Il *bucket sort* è un'altra variante del *counting sort* in cui l'input viene normalizzato, ovvero viene manipolato in modo che ciascun valore venga associato univocamente ad un numero nell'intervallo $[0, 1)$. Quest'intervallo viene poi diviso in n sotto-intervalli di dimensione $1/n$, detti *bucket*. I valori da ordinare vengono distribuiti nei *bucket* allo stesso modo del *pigeonhole sort*, ma se i valori sono uniformemente distribuiti, per effetto della normalizzazione, ciascun *bucket* conterrà un solo valore. Proprio perché sono pochi, i valori all'interno di un *bucket* possono essere ordinati tra loro usando un algoritmo come l'*insertion sort*.

NB. L'operazione di normalizzazione garantisce sempre un certo livello di uniformità nella distribuzione, ma non è detto che sia perfetta.

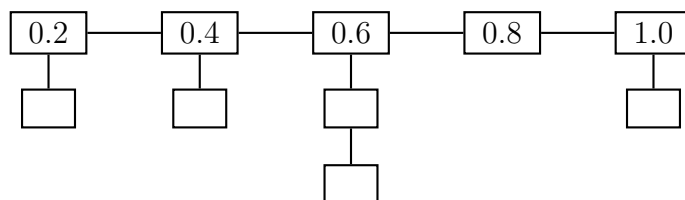


Fig. 19.2: *Bucket sort*

Complessità Il fatto che ciascun *bucket* contenga circa un solo valore rende la *complessità* media $O(n)$, anche se il costo del caso peggiore rimane $O(n + k)$. A livello di spazio, la *complessità* è $O(n)$.

19.3 Proprietà degli algoritmi di ordinamento

Definizione 192 - Stabilità.

Un algoritmo di ordinamento è detto essere stabile se preserva l'ordine iniziale tra due elementi con la stessa chiave.

Le definizioni appena enunciate affermano che nel caso in cui, ad esempio, si stia ordinando per anno di corso una lista di studenti già ordinata alfabeticamente, un metodo *stabile* produce una lista in cui gli alunni dello stesso anno sono ancora in ordine alfabetico, mentre un ordinamento *non stabile* probabilmente produrrà una lista senza più alcuna traccia del precedente ordinamento.

19.4 Confronto tra algoritmi di ordinamento

Algoritmo	Complessità	Caratteristiche
Insertion sort	$\Omega(n), O(n^2)$	Stabile, sul posto, iterativo e ottimo per sequenze quasi ordinate
Merge sort	$\Theta(n \log n)$	Stabile, ricorsivo, richiede $O(n)$ di spazio aggiuntivo, buona performance in cache e buona parallelizzazione
Heap sort	$\Theta(n \log n)$	Non stabile, sul posto, iterativo, cattiva performance in cache e cattiva parallelizzazione
Quick sort	$O(n \log n), O(n^2)$	Non stabile, ricorsivo, buona performance in cache, buona parallelizzazione e buoni fattori moltiplicativi
Counting sort	$\Theta(n + k)$	Richiede $O(k)$ di memoria aggiuntiva, iterativo e molto veloce quando $k = O(n)$
Pidgeonhole sort	$\Theta(n + k)$	Richiede $O(k)$ di memoria aggiuntiva, iterativo e molto veloce quando $k = O(n)$
Bucket sort	$O(n), O(n + k)$	Stabile e richiede $O(n)$ di spazio aggiuntivo
Shell sort	$O(n\sqrt{n})$	Stabile, adatto per piccoli valori e per sequenze quasi ordinate

Capitolo Nr.20

Algoritmi probabilistici

L'ultimo argomento che ci rimane da affrontare riguarda i cosiddetti *algoritmi probabilistici*, ovvero quel genere di algoritmi che applicano il calcolo delle probabilità ai dati di output. Possiamo distinguere due categorie: gli algoritmi la cui correttezza è probabilistica e gli algoritmi che sono corretti, ma hanno un tempo di funzionamento probabilistico. I primi vengono chiamati *Algoritmi Montecarlo*, i secondi *Algoritmi Las Vegas*.

20.1 Algoritmi Montecarlo

20.1.1 Test di primalità

Problema 39 - Test di primalità.

Dato un valore intero n , stabilire se è primo oppure no.

L'algoritmo naïf per questo problema prova tutti i numeri in $[2, \lfloor \sqrt{n} \rfloor]$ e se nessuno di essi è un divisore di n , allora n è certamente primo.

Frammento 143 - Algoritmo naïf.

```
boolean isPrime(int n)
  for (i = 2 to  $\lfloor \sqrt{n} \rfloor$ ) do
    if (n / i ==  $\lfloor n / i \rfloor$ ) then
      return false
  return true
```

Questa prima soluzione è certamente corretta, ma terribilmente inefficiente.

Soluzione basata sul Piccolo Teorema di Fermat

Definizione 193 - Piccolo Teorema di Fermat.

Se $n \in \mathbb{N}$ è primo, allora vale:

$$b^{n-1} \bmod n \equiv 1 \quad \forall b \in \mathbb{N} \text{ t.c. } 2 \leq b < n$$

Frammento 144 - Prima implementazione della soluzione.

```
boolean isPrime(int n)
    int b = random(2, n - 1)           % Restituisce un numero in [2, n - 1]
    if (bn-1 mod n ≠ 1) then
        return false
    return true
```

Il *Piccolo Teorema di Fermat* è tale per cui se esiste un valore di b per il quale la condizione non è verificata, n certamente non è primo. Tuttavia, esistono numeri composti per i quali esiste un $b \in [2, n - 1]$ tale per cui $b^{n-1} \bmod n \equiv 1$. Questi numeri si chiamano pseudo-primi e la loro esistenza fa sì che la soluzione che abbiamo realizzato possa restituire dei falsi positivi.

Per mitigare il rischio di falsi positivi è possibile verificare più volte la condizione del *Piccolo Teorema di Fermat* per diversi valori di b .

Frammento 145 - Seconda implementazione della soluzione.

```
boolean isPrime(int n)
    for (i = 1 to k) do
        int b = random(2, n - 1)       % Restituisce un numero in [2, n - 1]
        if (bn-1 mod n ≠ 1) then
            return false
    return true
```

Questa versione è migliore della precedente, ma per via dei *numeri di Carmichael*, numeri composti che soddisfano la condizione del *Piccolo Teorema di Fermat* per ogni b , la probabilità di ottenere falsi positivi rimane alta.

Algoritmo di Miller-Rabin Questo algoritmo si basa sul seguente teorema:

Definizione 194 - Teorema di Miller-Rabin.

Se $n \in \mathbb{N}$ è primo, allora per ogni intero b con $2 \leq b < n$ valgono entrambe le seguenti condizioni:

1. $\text{mcd}(n, b) = 1$
2. $b^m \bmod n \equiv 1 \vee \exists i \in \mathbb{N} : 0 \leq i < v \text{ per cui } b^{m2^i} \bmod n \equiv n - 1$

per m dispari e $n - 1 = m2^v$.

NB. La scrittura $n - 1 = m2^v$ ci dice che la rappresentazione di $n - 1$ in binario è pari a quella di m seguita da v zeri.

Quello che l'*Algoritmo di Miller-Rabin* fa è scegliere casualmente un valore di b e provare a vedere se almeno una delle proprietà risulta falsa. Se è così, il numero n sicuramente non è primo. Il problema di questa strategia è che per alcuni numeri composti esistono valori di b che verificano le due proprietà. I valori di b per i quali n risulta composto sono chiamati *testimoni* e Rabin ha dimostrato che per ogni numero composto n , ci sono almeno $\frac{3}{4}(n - 1)$ *testimoni* nell'intervallo $[2, n - 1]$. Di conseguenza, il *test di primalità* ha una probabilità inferiore a $\frac{1}{4}$ di restituire un falso positivo. Come per l'altro algoritmo, possiamo ridurre la probabilità di errore ripetendo più volte il test. Il risultato è un algoritmo come il seguente:

Frammento 146 - Implementazione della soluzione.

```
boolean isPrime(int n)
  for (i = 1 to k) do
    int b = random(2, n - 1)
    if (isComposite(n, b)) then           % Testa le condizioni di compostezza
      return false
  return true
```

Complessità L'Algoritmo di Miller-Rabin ha una probabilità di errore di $(\frac{1}{4})^k$ e una complessità pari a $O(k \log^2 n \log \log n \log \log \log n)$.

20.1.2 Identificazione di espressioni polinomiali nulle

Problema 40 - Identificazione di espressioni polinomiali nulle.

Data un'espressione algebrica polinomiale $p(x_1, \dots, x_n)$ in n variabili, stabilire se p è identicamente nulla oppure no.

Questo problema può essere risolto semplificando il polinomio, ma soluzioni di questo tipo sono molto complesse. Un'alternativa probabilistica è quella di generare una n -upla di valori v_1, \dots, v_n . Quindi, si risolve l'equazione $x = p(v_1, \dots, v_n)$ e, se $x \neq 0$, il polinomio certamente non è identicamente nullo. Altrimenti, potrebbe esserlo.

È dimostrato che se ogni v_i è un valore intero compreso tra 1 e $2d$, dove d è il grado del polinomio, la probabilità di falsi positivi non supera $\frac{1}{2}$. Quindi ripetendo l'algoritmo k volte, la probabilità di errore si riduce a $(\frac{1}{2})^k$.

20.2 Algoritmi Las Vegas

20.2.1 Problema della selezione

Problema 41 - Problema della selezione.

Dato un vettore A contenente n valori e un valore $1 \leq k \leq n$, trovare l'elemento che occuperebbe la posizione k se il vettore fosse ordinato.

NB. Il problema del calcolo della mediana è un sottoproblema della *selezione* con $k = \lfloor \frac{n}{2} \rfloor$.

Una soluzione banale è quella di ordinare il vettore pagando $O(n \log n)$ e cercare il valore in posizione k . Possiamo fare meglio di così?

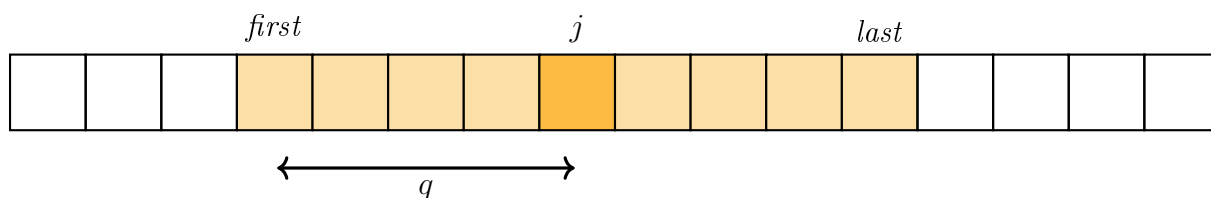
Approccio basato su Heap

Frammento 147 - Implementazione soluzione basata su Heap.

```
ITEM heapSelect(ITEM[] A, int n, int k)
  buildHeap(A)
  for (i = 1 to k - 1) do
    deleteMin(A, n)
  return deleteMin(A, n)
```

Complessità Le funzioni `heapBuild` e `deleteMin` costano rispettivamente $O(n)$ e $O(\log n)$ e, poiché la `deleteMin` viene invocata k volte, il costo totale è $O(n + k \log n)$. Se $k = O\left(\frac{n}{\log n}\right)$, la *complessità* si riduce a $O(n)$, ma se $k = \frac{n}{2}$, si attesta a $O(n + n \log n) = O(n \log n)$. Da questo deduciamo che la soluzione proposta è conveniente solo se i valori di k sono molto piccoli. Possono andare bene anche valori molto vicini a n , ma l'algoritmo va modificato in modo che utilizzi un *albero max-heap* e rimuova ogni volta il valore massimo.

Approccio basato su Quicksort Questo algoritmo si comporta come il *Quicksort* perché sceglie un valore detto *pivot* e sposta alla sua sinistra o alla sua destra tutti gli elementi del sottovettore con valore inferiore o superiore. Fatto questo, esprime k rispetto all'indice di inizio del sottovettore che sta considerando e, se il valore di k coincide con quello appena calcolato, allora il valore cercato si trova nella posizione del *pivot*. Se k è più piccolo, la ricerca continua in maniera ricorsiva nel sottovettore sinistro rispetto al *pivot*, oppure continua in quello destro se k è maggiore. In quest'ultimo, caso il valore di k viene ridotto in modo da mantenerlo coerente con i limiti del sottovettore.



Frammento 148 - Implementazione della soluzione basata su Quicksort.

```
ITEM selection(ITEM[] A, int first, int last, int k)
    if (first == last) then
        return A[first]
    int j = pivot(A, first, last)
    int q = j - first + 1 % Indice del valore ricercato rispetto al sottovettore
    if (k == q) then
        return A[j]
    else if (k < q) then
        return selection(A, first, j - 1, k)
    else
        return selection(A, j + 1, last, k - q)
```

Complessità La similitudine con il *Quicksort* si ripresenta anche nella *complessità* perché, se il vettore è ordinato e di volta in volta viene scelto come *pivot* il primo elemento del sottovettore, l'*equazione di ricorrenza* assume la seguente forma:

$$T(n) = \begin{cases} 1 & n \leq 1 \\ T(n-1) + n & n > 1 \end{cases}$$

Per il *Teorema delle ricorrenze lineari di ordine costante*, la forma chiusa è $T(n) = O(n^2)$.

Allo stesso modo, se ad ogni scelta del *pivot* il sottovettore viene diviso in due metà, rientriamo nel caso ottimo e l'*equazione di ricorrenza* vale:

$$T(n) = \begin{cases} 1 & n \leq 1 \\ T(n/2) + n & n > 1 \end{cases}$$

Per il *Teorema delle ricorrenze lineari con partizione bilanciata - Rid*, la forma chiusa è $T(n) = O(n)$.

Come si comporta l'algoritmo nel caso medio?

Per rispondere a questa domanda proviamo ad assumere che la funzione `pivot` restituisca con la stessa probabilità una qualsiasi posizione del vettore A . In questo modo, la *funzione di ricorrenza* diventa n più la media fra il costo di tutte le possibili scelte del *pivot*, ovvero:

$$\begin{aligned}
T(n) &= n + \frac{1}{n} \sum_{q=1}^n T(\max\{q-1, n-q\}) && \text{Media su } n \text{ casi} \\
&\leq n + \frac{1}{n} \sum_{q=\lfloor n/2 \rfloor}^{n-1} 2T(q) && n > 1 \\
&\leq n + \frac{1}{n} \sum_{q=\lfloor n/2 \rfloor}^{n-1} 2cq && \text{Sostituzione} \\
&\leq n + \frac{2c}{n} \sum_{q=\lfloor n/2 \rfloor}^{n-1} q && \text{Raccoglimento } 2c \\
&= n + \frac{2c}{n} \left(\sum_{q=1}^{n-1} q - \sum_{q=1}^{\lfloor n/2 \rfloor - 1} q \right) && \text{Sottrazione prima parte} \\
&= n + \frac{2c}{n} \left(\frac{n(n-1)}{2} - \frac{\lfloor n/2 \rfloor (\lfloor n/2 \rfloor - 1)}{2} \right) \\
&\leq n + \frac{2c}{n} \left(\frac{n(n-1)}{2} - \frac{(n/2-1)(n/2-2)}{2} \right) && \text{Rimozione limite inferiore} \\
&= n + \frac{2c}{n} \left(n^2 - n - \left(\frac{1}{4}n^2 - \frac{3}{2}n + 2 \right) \right) \\
&= n + \frac{c}{n} \left(n^2 - n - \left(\frac{1}{4}n^2 - \frac{3}{2}n + 2 \right) \right) && \text{Semplificazione del 2} \\
&= n + \frac{c}{n} \left(\frac{3}{4}n^2 + \frac{1}{2}n - 2 \right) \\
&\leq n + \frac{c}{n} \left(\frac{3}{4}n^2 + \frac{1}{2}n \right) && \text{Rimozione del } -2 \\
&= n + \frac{3}{4}cn + \frac{1}{2}c \\
&\leq cn
\end{aligned}$$

L'ultima disequazione è vera per $c > 4$ e $n \geq \frac{2c}{c-4}$, quindi la *complessità* del caso medio è $T(n) = O(n)$. Questa soluzione è probabilistica perché il tempo d'esecuzione dipende dalla selezione dei *pivot*.

NB. Per forzare la funzione `pivot` a scegliere equiprobabilisticamente tutti i valori compresi in $[1, n]$, possiamo forzare la cosa scambiando il valore in una posizione casuale con il primo: `A[random(first, last)] ↔ A[first]`¹.

Approccio deterministico Se supponessimo di avere a disposizione una funzione in grado di restituire in tempo $O(n)$ un valore che dista al più $\frac{3}{10}n$ dal mediano di un vettore, potremmo usare quella funzione per risolvere il *problema della selezione*.

L'idea sarebbe poi quella di dividere gli n valori in gruppi di 5. Quindi, per ogni gruppo S_i , con $i \in [1, \lceil n/5 \rceil]$, usare la suddetta funzione per ricavare il mediano M_i . A quel punto, usando la stessa funzione si può ricavare il mediano m delle mediane $[M_1, \dots, M_{\lceil n/5 \rceil}]$, e usare quell' m come *pivot* per poi richiamare ricorsivamente l'algoritmo sul sottovettore opportuno, seguendo lo stesso approccio della soluzione basata su *Quicksort*. Una volta raggiunto un sottovettore sufficientemente piccolo si potrebbe usare un algoritmo di ordinamento per determinare il valore ricercato.

1	4	7	10	13	2	5	8	11	14	3	6	9	12	15
---	---	----------	----	----	---	---	----------	----	----	---	---	----------	----	----

Fig. 20.1: La mediana delle mediane è 8 che è anche la mediana dell'intero vettore

¹L'implementazione di `pivot` sceglie sempre l'elemento in prima posizione

Frammento 149 - Implementazione soluzione deterministica.

```
ITEM select(ITEM[] A, int first, int last, int k)
    % Se la dimensione è inferiore a 10, ordina il vettore e restituisce
    % il  $k$ -esimo elemento di  $A[first...last]$ 
    if (last - first + 1 ≤ 10) then
        insertionSort(A, first, last)          % Ordina unicamente il sottovettore
        return A[first + k - 1]
    % Divide  $A$  in  $\lceil n/5 \rceil$  sottovettori di dimensione 5 e ne calcola la mediana
    ITEM[] M = new ITEM[1... $\lceil n / 5 \rceil$ ]
    for (i = 1 to  $\lceil n / 5 \rceil$ ) do
        M[i] = median5(A, first + (i - 1) · 5, last)
    % Individua la mediana delle mediane e la usa come pivot
    ITEM m = select(M, 1,  $\lceil n / 5 \rceil$ ,  $\lceil \lceil n / 5 \rceil / 2 \rceil$ )
    int j = pivot(A, first, last, m)           %  $m$  è il pivot e  $j$  ne è l'indice
    int q = j - first + 1 % Indice del valore ricercato rispetto al sottovettore
    if (k == q) then
        return m
    else if (k < q) then
        return select(A, first, j - 1, k)
    else
        return select(A, j + 1, last, k - q)
```

Complessità Il calcolo dei mediani M_i richiede al più $6\lceil n/5 \rceil$ confronti, la prima invocazione di `select` opera su $\lceil n/5 \rceil$ elementi, mentre la seconda su un massimo di $\frac{7}{10}n$ o, più esattamente, su $n - 3 \cdot \lceil \lceil n/5 \rceil / 2 \rceil$ elementi. Quindi, la `select` esegue $O(n)$ confronti nel caso pessimo. La *funzione di ricorrenza* è la seguente:

$$T(n) = T(n/5) + T(7n/10) + \frac{11}{5}n = O(n)$$

20.3 Bloom filter

Concludiamo il capitolo esaminando i *bloom filters*, una *struttura dati probabilistica* ottenuta unendo i *bit set* alle *tabelle hash*. Il risultato è quindi una *struttura dinamica* che utilizza pochi bit (e.g. 10 bit) per ogni oggetto, non memorizza i dati e quindi non consente cancellazioni e ha, come preannunciato, una risposta probabilistica nelle letture.

Specifica

Frammento 150 - Bloom filter.

```
% Inserisce l'elemento  $k$  nella struttura
insert(ITEM k)

% Controlla se  $k$  è presente o meno nella struttura
% Se restituisce false,  $k$  è certamente assente, altrimenti
% potrebbe essere presente oppure no
boolean container(ITEM k)
```


La probabilità di ottenere falsi positivi diminuisce all'aumentare dei bit usati per ogni oggetto. In ogni implementazione è quindi importante raggiungere un trade-off tra occupazione di memoria e probabilità di errore. In particolare, se ϵ è la probabilità di errore che si vuole garantire, il numero di bit che devono essere usati per ogni oggetto è $1.44 \log(\frac{1}{\epsilon})$.

Implementazione L'idea è che inizialmente la struttura sia costituita da un vettore di bit inizializzati 0. Quando un valore deve essere inserito, vengono calcolate le posizioni dei bit che rappresentano l'oggetto utilizzando H funzioni hash. Ciascuna funzione restituisce una posizione e il bit corrispondente viene impostato a 1. Similmente, per la verifica di esistenza, si controllano i bit nelle posizioni in cui dovrebbe trovarsi l'elemento e, se almeno uno è 0, l'elemento è certamente assente. Caso contrario, potrebbe essere presente o potrebbe essere un falso positivo. Quest'ultima possibilità è dovuto al fatto che elementi diversi potrebbero condividere parte dei bit.

Frammento 151 - Bloom filter.

```

insert(ITEM k)
  for (i = 1 to H) do
    A[hi(k)] = true
boolean contains(ITEM k)
  for (i = 1 to H) do
    if (A[hi(k)] == false) then
      return false
  return true

```

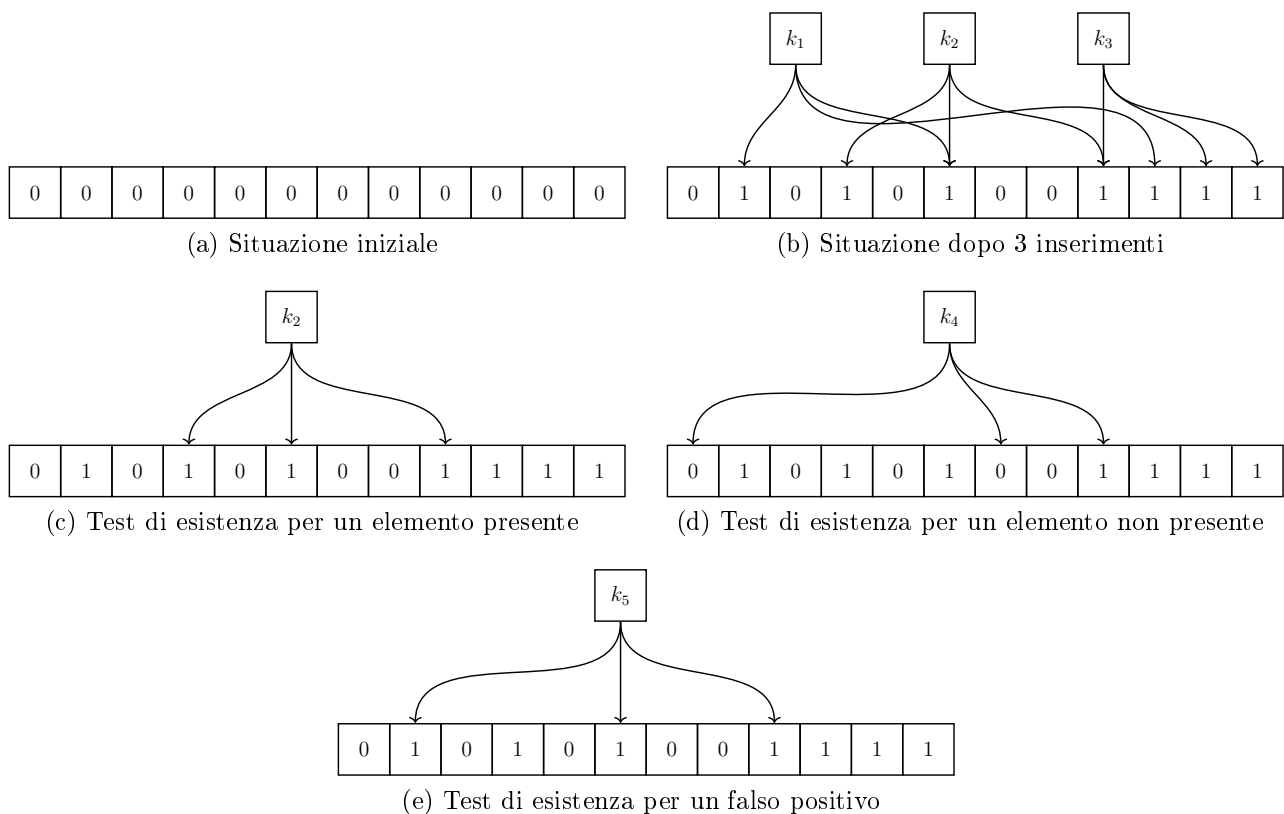


Fig. 20.2: Esempio di evoluzione di un *bloom filter*

Alcune formule

Definizione 195 - Probabilità di errore.

In un bloom filter costruito su m bit totali e H funzioni hash e che contiene n elementi, la probabilità ϵ di ottenere un falso positivo è:

$$\epsilon = \left(1 - e^{-\frac{Hn}{m}}\right)^H$$

Definizione 196 - Numero ottimale di funzioni hash.

In un bloom filter costruito su m bit totali e contenente n elementi, il numero ottimale di funzioni hash è:

$$H = \frac{m}{n} \ln 2$$

Definizione 197 - Numero di bit totali richiesti.

In un bloom filter contenente n elementi e caratterizzato da una probabilità d'errore ϵ , il numero di bit totali m richiesti è:

$$m = -\frac{n \ln \epsilon}{(\ln 2)^2}$$

Casi d'uso Questo tipo di *struttura dati* è molto utilizzato dai browser per memorizzare gli URL pericolosi. In particolare, ad ogni richiesta il browser verifica l'esistenza dell'URL cercato nella struttura e, se è presente, probabilmente si tratta di una destinazione pericolosa e quindi è necessario agire di conseguenza. Se invece non è presente, la destinazione è sicura.

A.1 Proprietà dei logaritmi

- | | |
|---|--|
| 1. $\log_a a = 1$ | <i>Proprietà fondamentale</i> |
| 2. $\log_a 1 = 0$ | <i>Proprietà fondamentale</i> |
| 3. $\log_a(b \cdot c) = \log_a(b) + \log_a(c)$ | <i>Teorema del prodotto</i> |
| 4. $\log_a\left(\frac{b}{c}\right) = \log_a(b) - \log_a(c)$ | <i>Teorema del rapporto</i> |
| 5. $\log_a(b^c) = c \cdot \log_a(b)$ | <i>Teorema della potenza</i> |
| 6. $\log_{a^n}(b^m) = \frac{m \cdot \log_a(b)}{n}$ | <i>Potenza alla base e all'argomento</i> |
| 7. $\log_{\frac{1}{a}}(b) = -\log_a(b)$ | <i>Base frazionaria</i> |
| 8. $\log_a\left(\frac{1}{b}\right) = -\log_b(a)$ | <i>Argomento frazionario</i> |
| 9. $\log_{\frac{1}{a}}\left(\frac{1}{b}\right) = \log_a(b)$ | <i>Base e argomento frazionario</i> |
| 10. $\log_a(b) = \frac{1}{\log_b(a)}$ | <i>Commutazione base e argomento</i> |
| 11. $\log_a(b) = \frac{\log_c(b)}{\log_c(a)}$ | <i>Cambio di base</i> |
| 12. $a^{\log_b(c)} = c^{\log_b(a)}$ | <i>Scambio base-argomento</i> |

A.2 Serie matematiche convergenti

- | | |
|--|--|
| 1. $\sum_{k=0}^{+\infty} k = \frac{k(k+1)}{2}$ | <i>Formula di Gauss</i> |
| 2. $\sum_{k=0}^n q^k = \frac{1-q^{n+1}}{1-q} = \frac{q^{n+1}-1}{q-1} \quad \forall q : q \geq 1$ | <i>Serie geometrica finita</i> |
| 3. $\sum_{k=0}^n q^k = \frac{1}{1-q} \quad \forall q : q < 1$ | <i>Serie geometrica finita</i> |
| 4. $\sum_{k=0}^{+\infty} q^k = \frac{1}{1-q} \quad \forall q < 1$ | <i>Serie geometrica infinita decrescente</i> |
| 5. $\sum_{k=0}^{+\infty} kq^k = \frac{q}{(1-q)^2} \quad \forall q < 1$ | <i>Serie geometrica infinita decrescente</i> |
| 6. $\sum_{k=1}^{+\infty} \frac{1}{k(k+1)} = \sum_{k=1}^{+\infty} \left(\frac{1}{k} - \frac{1}{k+1}\right) = 1$ | <i>Serie di Mengoli</i> |