



Università  
degli Studi  
di Ferrara



# INFORMATICA

Prof. Giorgio Poletti  
giorgio.poletti@unife.it

Laurea Triennale in Economia  
a.a. 2018 – 2019

# Sviluppo del corso

Modulo II e Modulo III



Google è un globale test di Rorschach.  
Vediamo in esso ciò che vogliamo vedere.

John Battelle *(giornalista)*

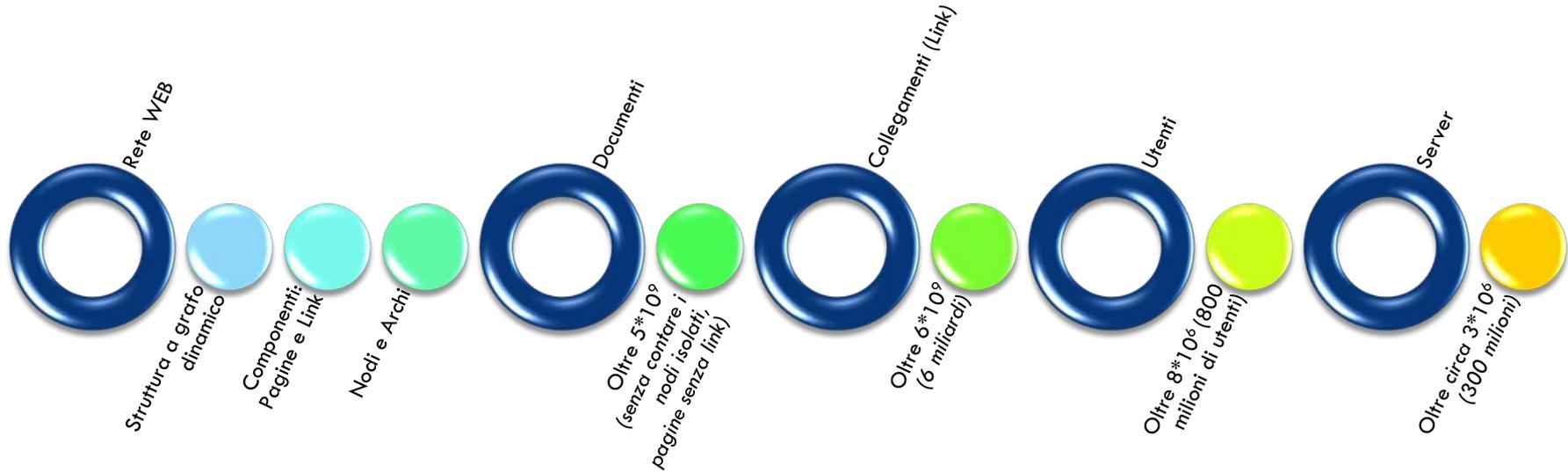
## ▣ Modulo II

- Dato e informazione: capire per comprendere – le relazioni
- IoT (Internet Of Things), BYOB e BYOT
- Data WareHouse e Big Data
- Tecnologia: cambiamenti organizzativi e sociali

## ▣ Modulo III

- Algoritmi di ricerca e posizionamento
- Distribuzione e ricerca di Informazioni
- Posizionarsi in Rete, Rete e i Social come luogo di Marketing
- Blockchain: significato, importanza e rischi

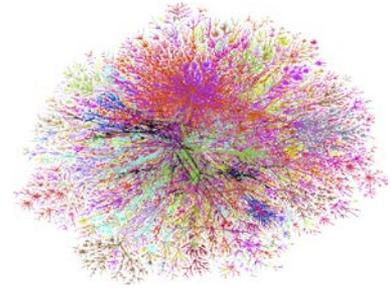
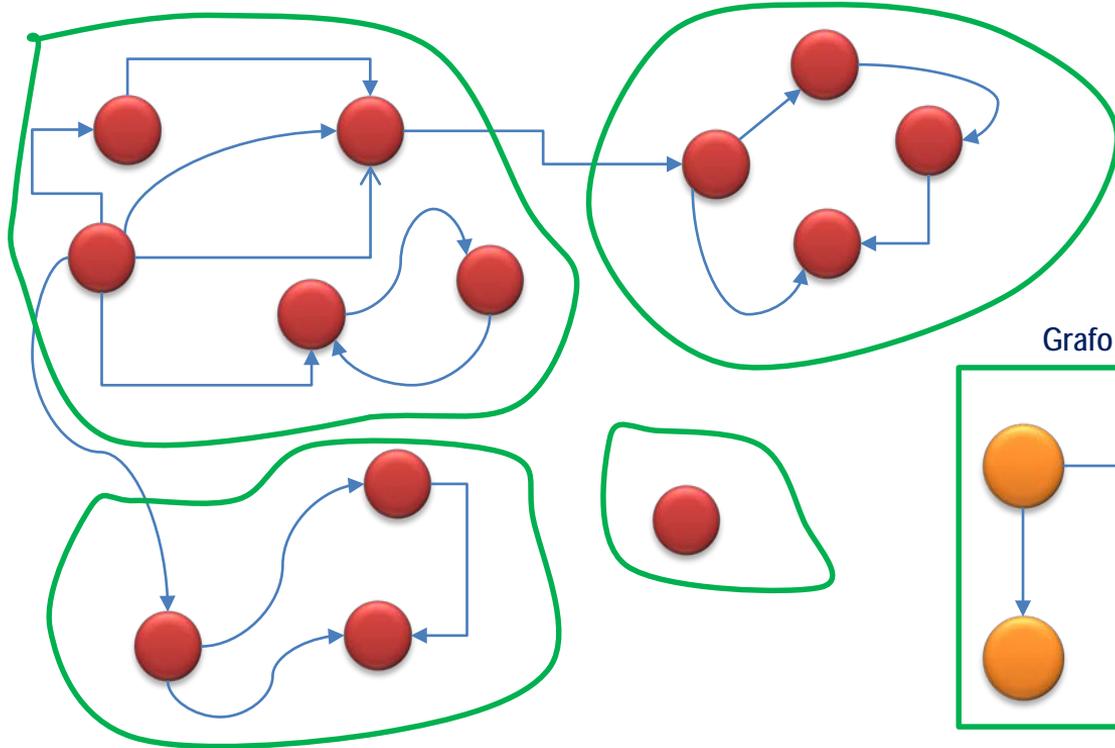
# Algoritmi di ricerca sul WEB



**DEFINIZIONE: GRAFO RIDOTTO  $G^*$** , i nodi sono le classi e due classi  $C_1$  e  $C_2$  sono connesse se esiste un nodo in  $C_1$  collegato a un nodo in  $C_2$ , esiste un arco da  $C_1$  a  $C_2$

# Algoritmi di Ricerca sul WEB

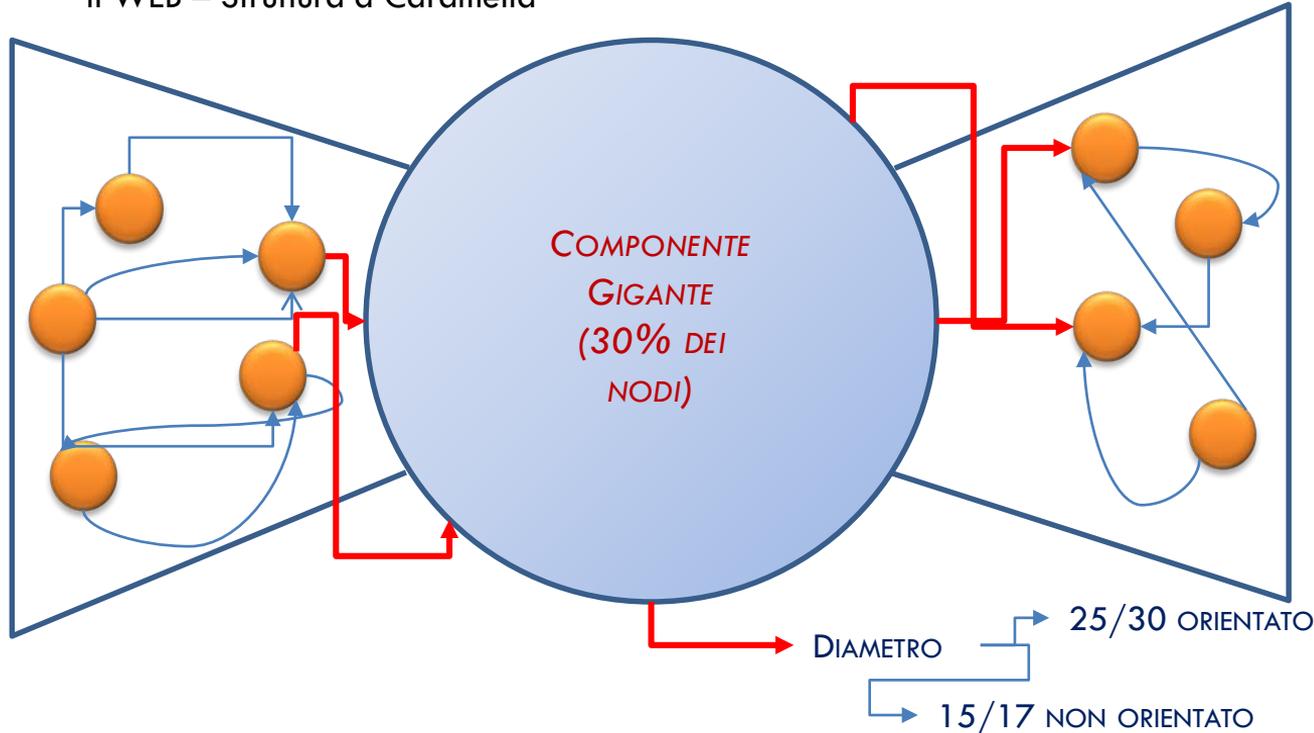
## Grafo Ridotto - Esempio



Grafo ridotto  $G^*$

# Algoritmi di Ricerca sul WEB

Il WEB – Struttura a Caramella

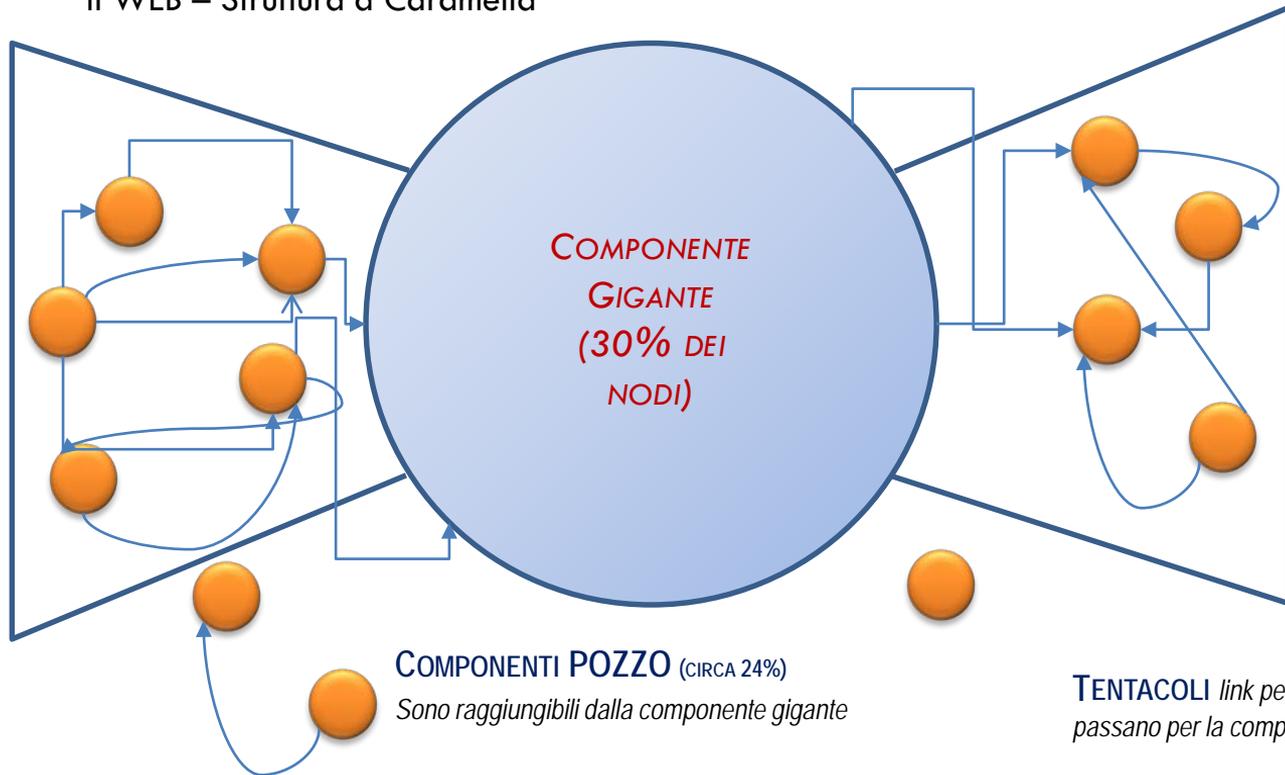


Teoria dei piccoli mondi  
(*SMALL WORLD THEORY*)

La teoria dei **piccoli mondi** sostiene che tutte le reti complesse presenti in natura sono tali che due qualunque nodi possono essere collegati da un percorso costituito da un numero relativamente piccolo di collegamenti. (6 gradi di separazione)

# Algoritmi di Ricerca sul WEB

Il WEB – Struttura a Caramella



**COMPONENTI SORGENTE**  
(CIRCA 24%)

*Puntano in maniera diretta o indiretta la componente gigante*

**PAGINE «REIETTE»**

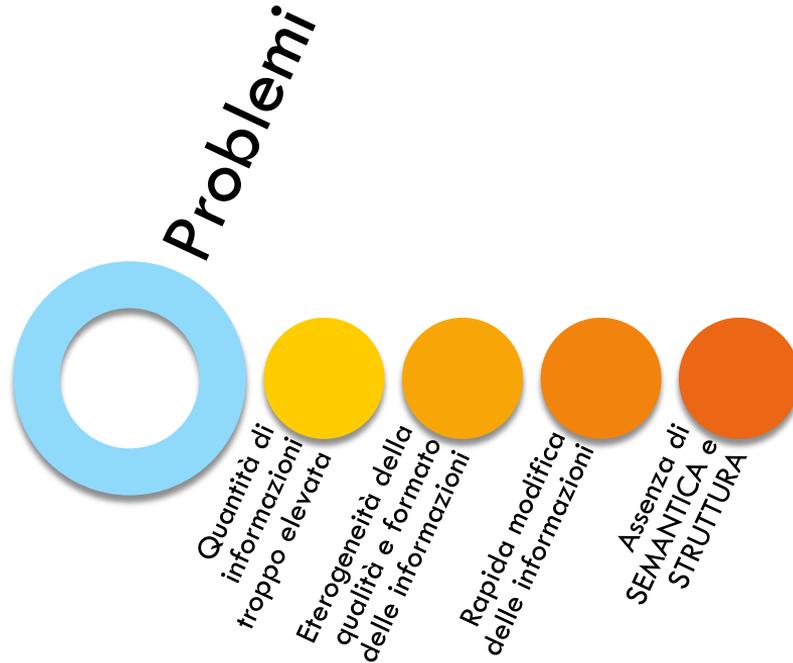
*Non sono raggiungibili dalla componente gigante*

**COMPONENTI POZZO** (CIRCA 24%)  
*Sono raggiungibili dalla componente gigante*

**TENTACOLI** *link per sorgenti e pozzi che non passano per la componente gigante*

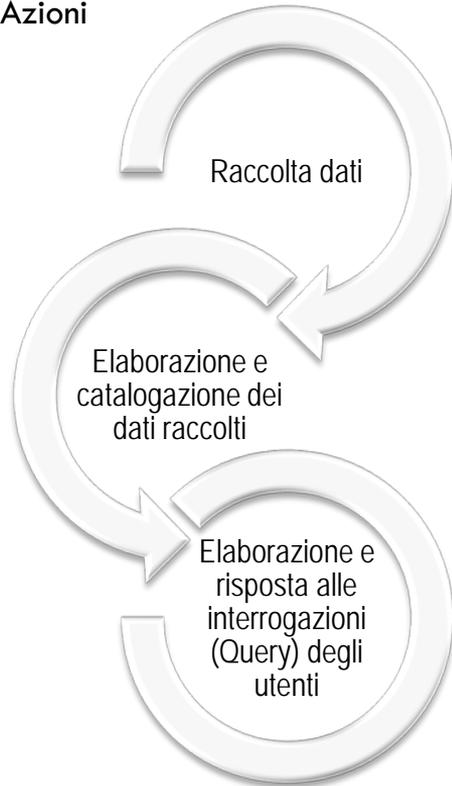
# Algoritmi di Ricerca sul WEB

Motori di Ricerca (Oltre 80% degli utenti li usa)



# MOTORI DI RICERCA

Azioni



```
<META NAME="title">  
<META NAME="description">  
<META NAME="keywords">
```

<http://www.altavista.com>  
<http://www.yahoo.com>  
<http://www.excite.com>  
<http://www.lycos.com>  
<http://www.hotbot.com>  
<http://www.virgilio.it>  
<http://www.arianna.it>

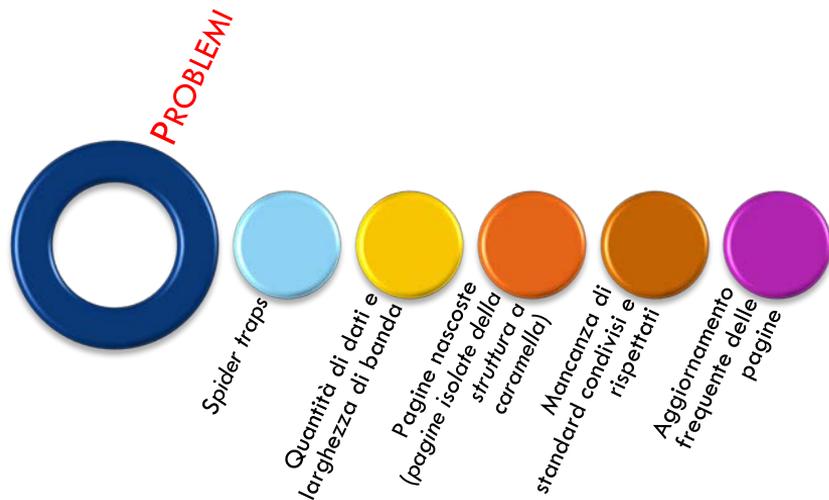
**SEO** (Search Engine Optimization /*optimizer*)

# MOTORI DI RICERCA

Raccolta dati

Raccolta del contenuto delle pagine Web (informazioni di tipo testuale ma anche immagini (google immagini e documentazione), robot.txt

Si usa uno spider o crawler o robot



## SPIDER

googlebot  
fast  
scooter  
mercator  
Ask Jeeves  
teoma\_agent  
ia\_archiver  
Yahoo! Slurp  
Romilda

## MOTORE DI RICERCA

Google  
Fast - Alltheweb  
Altavista  
Altavista  
Ask Jeeves  
Teoma  
Alexa - Internet Archive  
Yahoo  
Facebook

# MOTORI DI RICERCA

Elaborazione e catalogazione dei dati raccolti

Elaborazione e  
catalogazione dei  
dati raccolti

Rilevazione di  
presenza di  
SPAMMING

Rilevazione delle  
ridondanze  
(presenza di  
MIRRORING)

Reperimento e analisi  
delle informazioni  
per il calcolo del  
**RANKING**

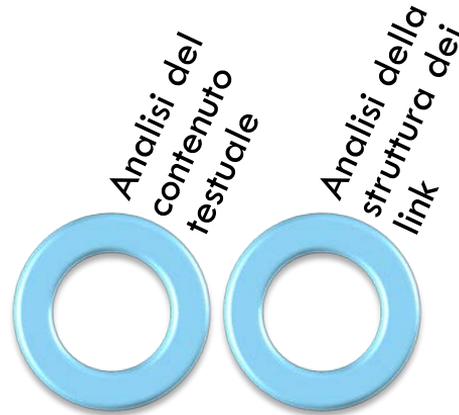


# ALGORITMO di Pagerank

**RANKING:** Dato un insieme di pagine **P** e una query **Q** il ranking è definito da una funzione:

$$R_q: P \rightarrow \mathbf{R} \text{ (insieme dei numeri reali)}$$

che associa ad ogni pagina un numero reale che indica la «**RILEVANZA**» di quella pagina nel contesto di quella query.

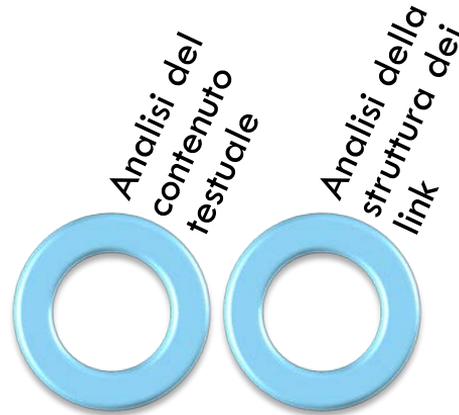


# ALGORITMO di Pagerank

**RANKING:** Dato un insieme di pagine **P** e una query **Q** il ranking è definito da una funzione:

$$R_q: P \rightarrow \mathbf{R} \text{ (insieme dei numeri reali)}$$

che associa ad ogni pagina un numero reale che indica la «**RILEVANZA**» di quella pagina nel contesto di quella query.



# ALGORITMO di Pagerank

Analisi dei link. L'importanza di una pagina è direttamente proporzionale al numero di pagine che la puntano

Se  $R_j$  è il rank (rango) di una pagina, la sua importanza, la pagina trasmette l'importanza alle pagine che punta, distribuisce il suo rango in maniera uniforme

$$R_i = \sum_{j \rightarrow i} \frac{R_j}{N_j}$$

$j \rightarrow i$  esiste un link dalla pagina  $j$  alla pagina  $i$   
 $N_j$  numero di link presenti nella pagina  $j$



Lawrence "Larry" Page & Sergey Brin (Fondano Google 1998)



# ALGORITMO di Pagerank



$$\sum_i R_i^{t+1} = \sum_i \left[ (1-\alpha) \sum_{j \rightarrow i} \frac{R_j^t}{N_j} + \alpha \frac{1}{N} \right] = (1-\alpha) \sum_i \sum_{j \rightarrow i} \frac{R_j^t}{N_j} + \alpha$$

Per ogni nodo  $j$  con almeno un arco uscente, il fattore  $R_j^t/N_j$  viene sommato per ciascun arco uscente (ce ne sono  $N_j$  in tutto):

$$\sum_i R_i^{t+1} = (1-\alpha) \sum_{j \text{ con un arco uscente}} R_j^t + \alpha$$

**La storia di Google per capire**

<https://moz.com/google-algorithm-change>



Di fronte alla giungla di documenti che si aggiungono minuto per minuto, la domanda cruciale è piuttosto la seguente: se lancio un'informazione in rete, qualcuno la noterà? Per esser letti bisogna essere visibili: una banale verità che vale tanto per gli scrittori quanto per gli scienziati.

*(Albert-László Barabási in «Link»)*

