

# Reti complesse

## Il grafo del Web

Matteo Dell'Amico  
dellamico@disi.unige.it



Applicazioni di rete 2  
A.A. 2006-07

# Outline

- 1 Storia
  - Memex
  - WWW
  
- 2 Caratteristiche
  - Web statico e dinamico
  - Il grafo del Web
  - Il Bow-Tie

# Memex

## Vannevar Bush, "As We May Think", 1945



- Previsioni sullo sviluppo della tecnologia dopo la fine della guerra.
- **MEMEX**: dispositivo foto-meccanico che memorizza documenti ed immagini in microfilm.
- **Tracce associative**: permettono di associare più pagine collegate di documenti diversi.
- Considerato il primo esempio di **ipertesto**.

# Nasce il World Wide Web

## Tim Berners-Lee (CERN)



- 1980 Scrive *enquire* (da *enquire upon within everything*), che permette di **collegare** documenti correlati.
- 1989 Produce un documento intitolato "Information Management: A Proposal".
- 1990 Prima comunicazione tra un server Web ed un browser Web.
- 1994 Fonda il **W3C** (*World Wide Web Consortium*).

## Effetto collaterale

<http://www.netvalley.com/intval2.html>

Ben,

It happened many times during history of science that the most impressive results of large scale scientific efforts appeared far away from the main directions of those efforts.

I hope you agree that **Web** was a **side effect** of the CERN's scientific agenda. [...]

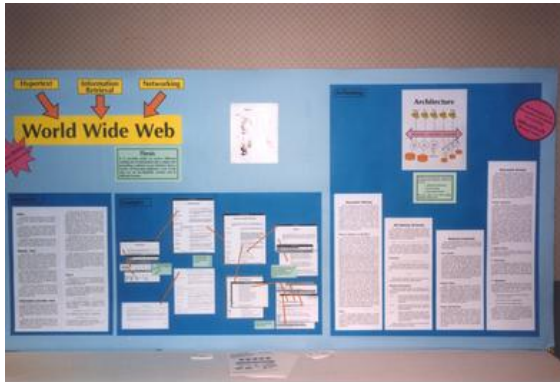
The Web, – **crucial point** of human's history, was born... Nothing could be compared to it. You wrote the *best* about it: “**synergy**, **serendipity** and **coincidence**...”

We can't imagine yet the real scale of the recent **shake**, because there has not been so fast growing *multi-dimension* social-economic processes in human history...

Gregory Gromov

P.S. It is quite remarkable to say that “Highlights of CERN History: 1949 – 1994” do **not** have a **word** about Web. So, it looks like a classic *side effect* that normally is not mentioned at the main text of *official* record...

## Basso interesse?



- 1991: il primo articolo sul WWW fu accettato solo come “poster”.
- Oggi, è considerato una delle **rivoluzioni** della storia recente.

# Web statico

- Vogliamo studiare le pagine che fanno **parte** del WWW.
- Alcune pagine non sono **persistenti**:
  - cataloghi
  - risultati di query
  - generate da script
- Ci limitiamo alle pagine **“statiche”**.

# Web statico e pubblico

## Statico

- Non è risultato di script lato server.
- Non ha “?” nell'URL.
- Non cambia molto spesso.

## Pubblico

- Non sono richieste password.
- Non c'è il file `robots.txt`.
- Non c'è il meta tag `noindex`.

## Non è così semplice. . .

- Pagine statiche possono venire costruite “al volo” (header, barre di navigazione, ecc. . . )
- Pagine dinamiche che appaiono statiche (cataloghi navigabili, siti di notizie. . . )
- Oggetti “a metà” (blog, wiki. . . )
- Server poco affidabili
- Spider trap o honeypot
- Pagine di “spam”

## ... cosa resta

### Web dinamico

- Detto anche “deep web” o “invisible web”.
- Formato da tutti i possibili risultati delle query.
- Contenuti rintracciabili solo tramite **query dirette**.
- Risultati **non persistenti** prodotti in **tempo reale**.

## Web dinamico e statico

- 2004: indicizzate più di **84 miliardi** di pagine dinamiche (**750 TB** di informazioni).
- La parte statica non comprende più di **10 miliardi** di pagine.
- Tasso di espansione **10 volte** superiore a quello della parte statica.
- La parte dinamica è in pratica **impossibile** da indicizzare ragionevolmente.

# Il grafo del Web

$$G = (V, E)$$

- $V$  è l'insieme delle pagine HTML **statiche**.
- $E$  è l'insieme degli hyperlink **statici**.
- $G$  è un grafo diretto detto “**grafo delle pagine**”.

## Il grafo del Web (2)

### Perché studiarlo?

- È il **più grande artefatto** mai costruito dall'uomo.
- Possiamo **sfruttarne la struttura**:
  - Strategie di crawling
  - Ricerche mirate
  - Riconoscere lo “spam”
  - Scoprire comunità

### Predire l'evoluzione

- Modelli matematici
- Studi sociologici

## Grafi correlati

### Siti

- $V$ : siti web
- $E$ : una pagina sul sito A ha un link verso una pagina sul sito B

### Co-citazioni

- $V$ : pagine web statiche
- $E$ :  $(x, y)$  numero di pagine che linkano sia  $x$  che  $y$

## Quanto è grande il Web?

- Grafo in **costante espansione**.
- Non sappiamo **che percentuale ne conosciamo**.

### Crawling

- Scrivere un **robot** per visitare più parti del web possibili.
- Immagine che può essere **distorta**:
  - Limiti di dimensione
  - Regole di crawling
  - Inconvenienti sulla rete (es: caduta di router)

# Crawling

## Idea naïve

- Si parte da una pagina qualsiasi
- Si seguono tutti i link incontrati finché non si incontrano più pagine

## Problemi

- Spam
- Mirror
- Non rispetto del protocollo (es: soft 404)

## Quanto è grande il web? (2)

### Secondo approccio

- 1 Prendiamo pagine **a caso**
- 2 Vediamo **quante** di esse sono indicizzate da un motore di ricerca (**copertura**)
- 3 Usiamo la copertura ed il **numero di pagine campionate** per stimare la dimensione del web

### Problemi

- Come campionare le pagine in maniera **uniforme**?
- Come **verificare** se una pagina è indicizzata da un motore di ricerca?

## Campionamento quasi-uniforme

- Generare indirizzi IP in modo casuale uniforme, crawl completo dei siti.
  - Problemi con **virtual hosting** e siti di dimensione **variabile**.
- 
- Effettuare un **cammino casuale** sul WWW, seguendo link a **caso**.
  - Tutte le pagine hanno la **stessa probabilità** di essere raggiunte? Torneremo su questo punto.

# Strong Query

- Abbiamo una pagina, vogliamo sapere se è indicizzata da un motore di ricerca.
- Prendiamo gli  $m$  termini meno frequenti, verifichiamo se il motore di ricerca risponde positivamente.
- Tre approcci:
  - 1 Accettare le **copie identiche**
  - 2 Accettare lo **stesso URL**
  - 3 Accettare le **risposte non nulle**

## Proprietà del grafo del Web

- Conosciamo **solo** il web scoperto dai crawler.
- Influenzate da:
  - politiche di crawling
  - limiti di dimensione
  - perturbazioni di rete

## Crawling su vasta scala

### Alexa

- 1997, 200 milioni di pagine.

### Altavista

- 1999, 500 milioni di pagine.

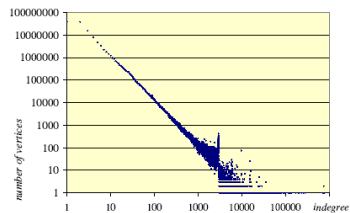
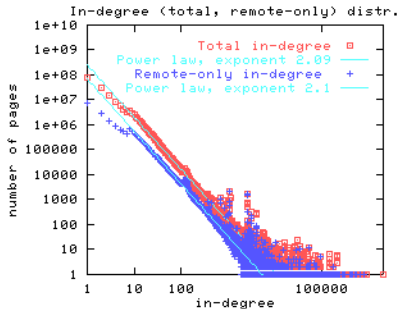
### Stanford WebBase

- 2001, 400 milioni di pagine.

## Proprietà interessanti

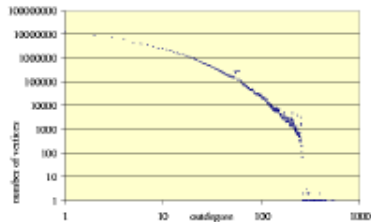
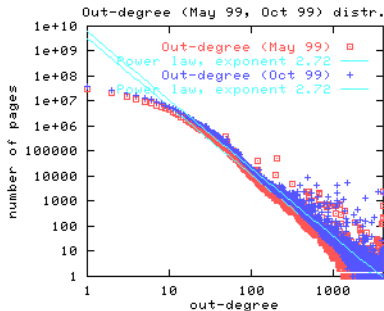
- Distribuzione dei degree
- Struttura globale
  - Che aspetto ha “visto da lontano”?
- Raggiungibilità
  - È possibile andare da qua a là? In quanti hop?
- Componenti connesse
  - Quali sono?
- Sottografi densi (cluster)
  - Ci sono pagine con link fitti tra di esse?

## In-degree



- Altavista (1999) e WebBase (2001)
- $P[\text{in-degree}(u) = x] \sim x^{-\gamma}, \gamma = 2.1$

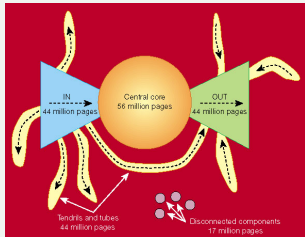
# Out-degree



- Secondo Altavista, è una power law.
- Secondo WebBase, le pagine con molti link uscenti sono meno frequenti.

## Il Bow-Tie

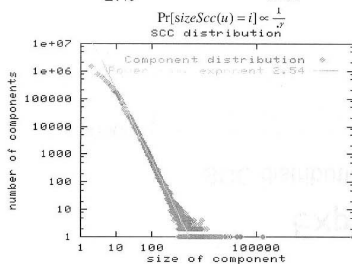
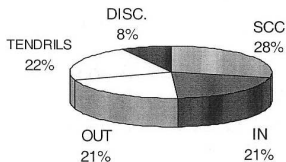
Broder et al., 2000



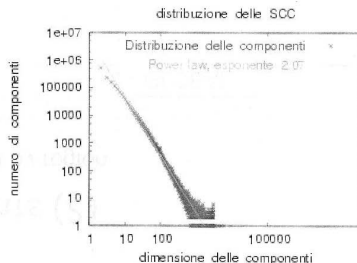
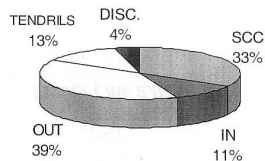
- **CORE:** la più grande componente fortemente connessa
- **IN:** nodi che possono raggiungere CORE
- **OUT:** nodi raggiungibili da CORE
- **TENDRILS:** nodi che non possono raggiungere ed essere raggiunti dal CORE
- **DISC:** nodi non connessi al bow-tie

# Esperimenti

## Altavista '99

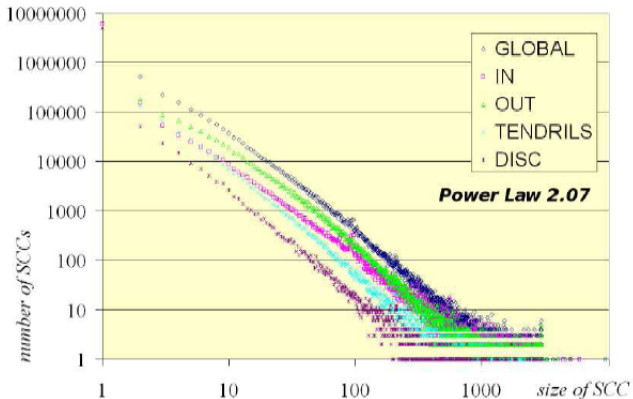


## WebBase '01



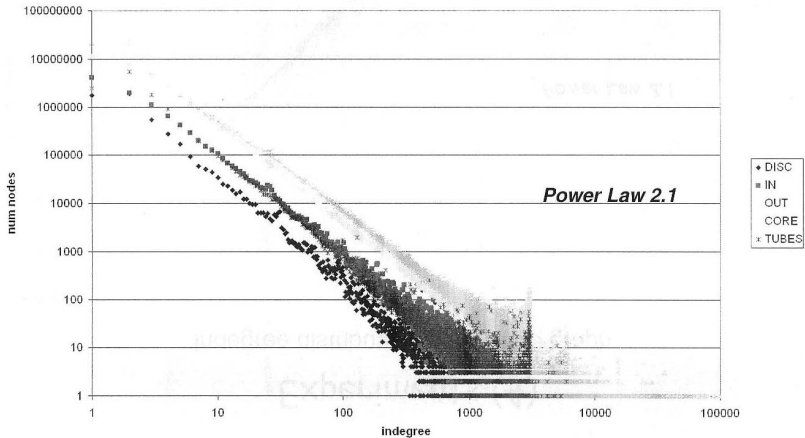
## Esperimenti (2)

SCC distribution region by region



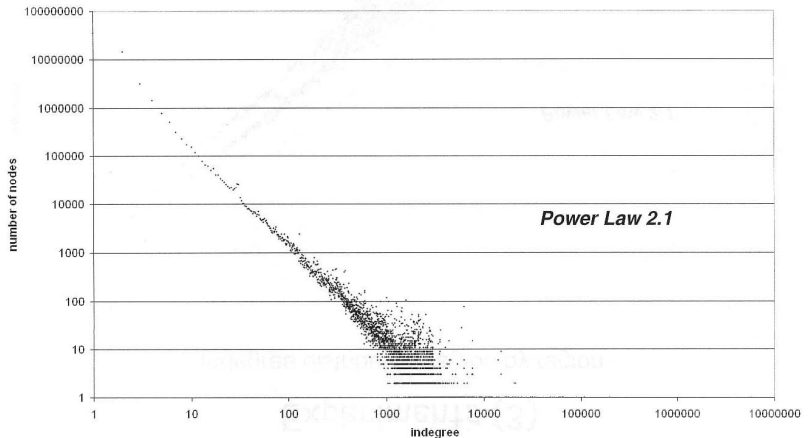
# Esperimenti (3)

## Indegree distribution region by region



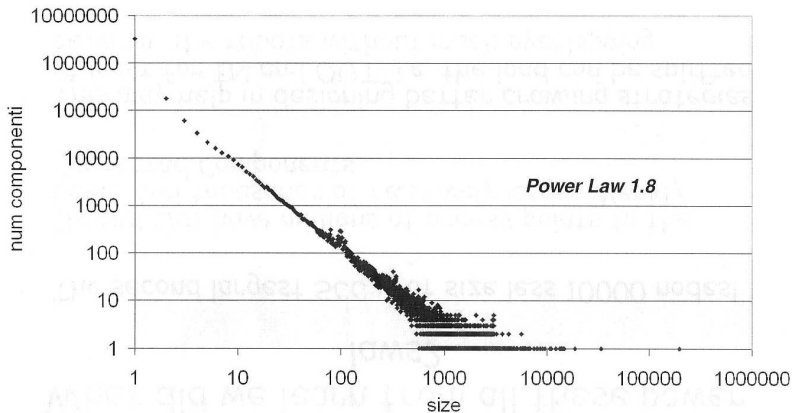
# Esperimenti (4)

## Indegree distribution in the SCC graph



# Esperimenti (5)

WCC distribution in IN



## In sintesi

- La seconda più grande componente del Web ha **meno di 10000 nodi**.
- IN ed OUT hanno milioni di punti di contatto con CORE e migliaia di componenti debolmente complesse **piuttosto grandi**.
- La power-law compare **ovunque**.

# Autosimilarità

- Chiamiamo **cluster tematicamente unificati (TUC)** insiemi di pagine raggruppate per
  - chiavi di ricerca
  - posizione geografica
  - nome dell'host
  - ecc. . .
- Tutti questi cluster presentano una **struttura di tipo bow-tie.**

## Autosimilarità (2)

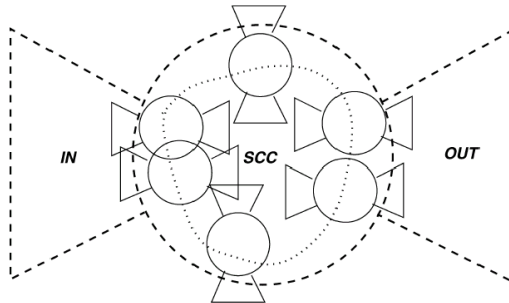


Fig. 4. TUCs connected by the navigational backbone inside the SCC of the Web graph.

- Collezione di strutture autosimili che formano lo “scheletro” del CORE.

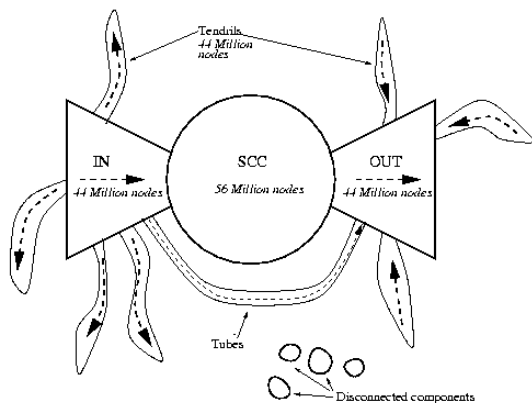
# Trovare CORE

- 1 Scegliere un vertice  $v$
- 2 Calcolare tutti i nodi raggiungibili da  $v$ :  $O(v)$
- 3 Calcolare tutti i nodi che raggiungono  $v$ :  $I(v)$
- 4 Calcolare  $SCC(v) = O(v) \cap I(v)$
- 5 Controllare se  $SCC$  è “sufficientemente grande”, altrimenti ripetere

## Esercizio

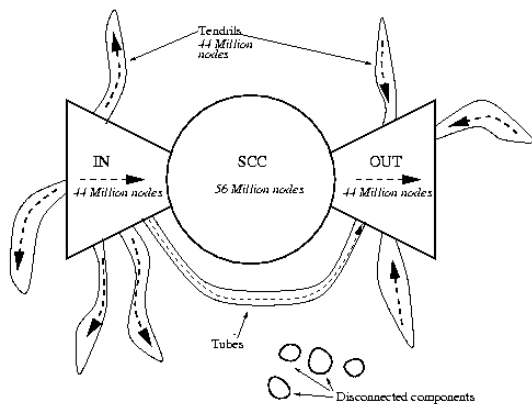
- Se  $SCC$  è il 25% del totale, qual è la probabilità di non trovarlo dopo 20 iterazioni?

## Trovare OUT



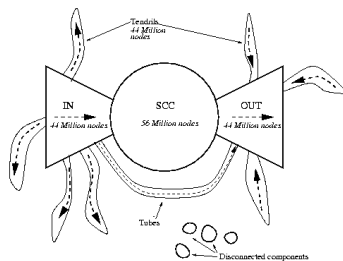
- SCC → OUT

# Trovare IN



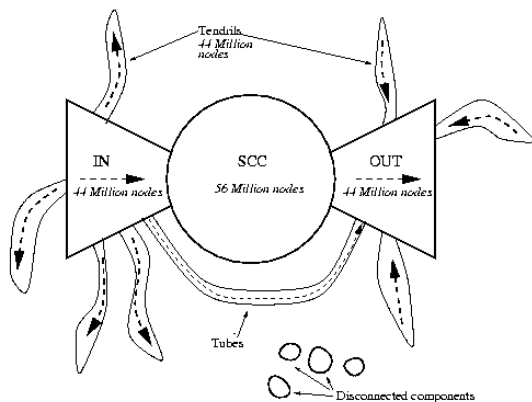
- IN → SCC

## Trovare TENDRILS e TUBES



- $IN \rightarrow TENDRILS\_IN$
- $TENDRILS\_OUT \rightarrow OUT$
- $TENDRILS\_IN \cup TENDRILS\_OUT \rightarrow TENDRILS$
- $TENDRILS\_IN \cap TENDRILS\_OUT \rightarrow TUBES$

## Trovare DISC



- Il resto...

## Il Web è small world?

- Parliamo di CORE, altrimenti non siamo sicuri ci siano i percorsi (circa **un quarto** dei nodi totali)

### Esperimenti sui crawling

- Distanza **massima**: maggiore di **28**.
- Distanza massima **diretta**: maggiore di **900**.
- Distanza **media**: circa **7**.
- Distanza media **diretta**: circa **16**.

# Riferimenti I



Vannevar Bush.

As We May Think.

*The Atlantic Monthly*, 1945.



K. Bharat e A. Broder.

A Technique for Measuring the Relative Size and Overlap of Public Web Search Engines.

*Proc. 7th International WWW Conference*, 1998.



M. Henzinger, A. Heydon, M. Mitzenmacher e M. Najork.

On Near-Uniform URL Sampling.

*Proc. 9th International WWW Conference*, 2000.



R. Albert, S. Jeong e A.-L. Barabási.

Diameter of the World Wide Web

*Nature*, 401, 130-131 (1999).

## Riferimenti II

-  A. Broder, R. Khumar, F. Maghoul, P. Raghavan, S. Rajahopalan, R. Stata, A. Tomkins e J. Wiener.  
Graph Structure in the Web.  
*Proc. 9th International WWW Conference, 2000.*
-  S. Dill, R. Kumar, K. McCurley, S. Rajagopalan, D. Sivakumar e A. Tomkins.  
Self-Similarity in the Web.  
*Proc. 27th International Conference on Very Large Data Bases, 2001.*