The Structure of the Web

Redes de Informação

- Nós são pedaços de informação
- Arestas ligam pedaços relacionados entre si
- Exemplos
    - Web
    - Redes de citação
    - Referências em uma enciclopédia
    - Wireless communication

Antes da Web ...

Memex [As we may think by V. Bush, 1945]

- Informações em livros é altamente linear
- Nossa memória associativa representa uma rede semântica
- Memex é um sistema que imita a rede semântica com conhecimentos e arestas entre os conceitos

1990: Quando a Web nasceu

- Criado por Tim Barners-Lee em 1990 no Cern
  - Grande banco de dados com hipertexto (links)

1990: Primeira Página Web

World Wide Web

The WorldWideWeb (W3) is a wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an executive summary of the project, Mailing lists , Policy , November's W3 news , Frequently Asked Questions .

What's out there?
Pointers to the world's online information, subjects , W3 servers, etc.

Help
on the browser you are using

Software Products
A list of W3 project components and their current state. (e.g. Line Mode ,X11 Viola , NeXTStep , Servers , Tools , Mail robot , Library)

Technical
Details of protocols, formats, program internals etc

Bibliography
Paper documentation on W3 and references.

People
A list of some people involved in the project.

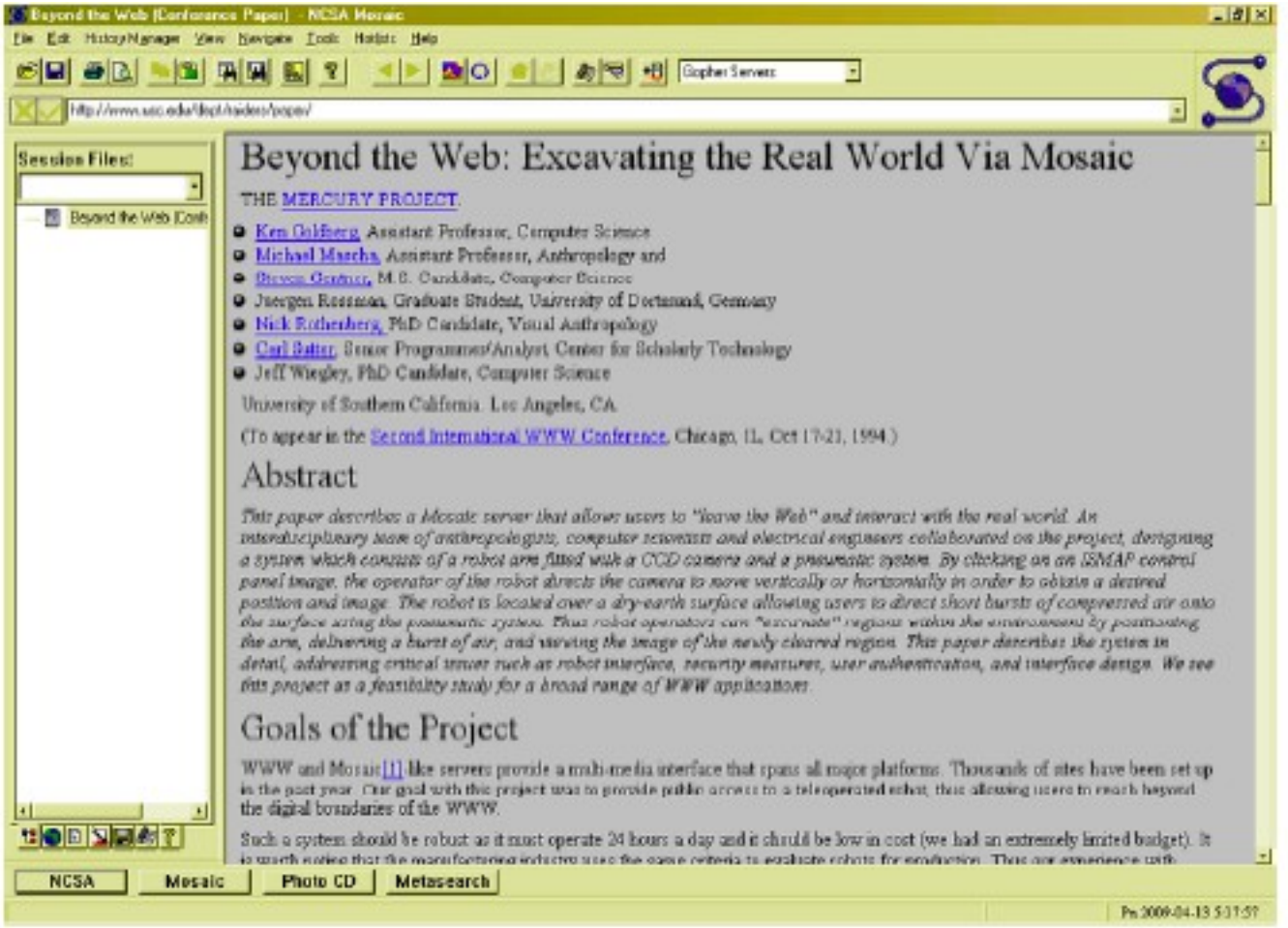
History
A summary of the history of the project.

How can I help ?
If you would like to support the web..

Getting code
Getting the code by anonymous FTP , etc.

http://www.w3.org/History/19921103-hypertext/hypertext/WWW/TheProject.html

1993: Primeiro Browser - Mosaic



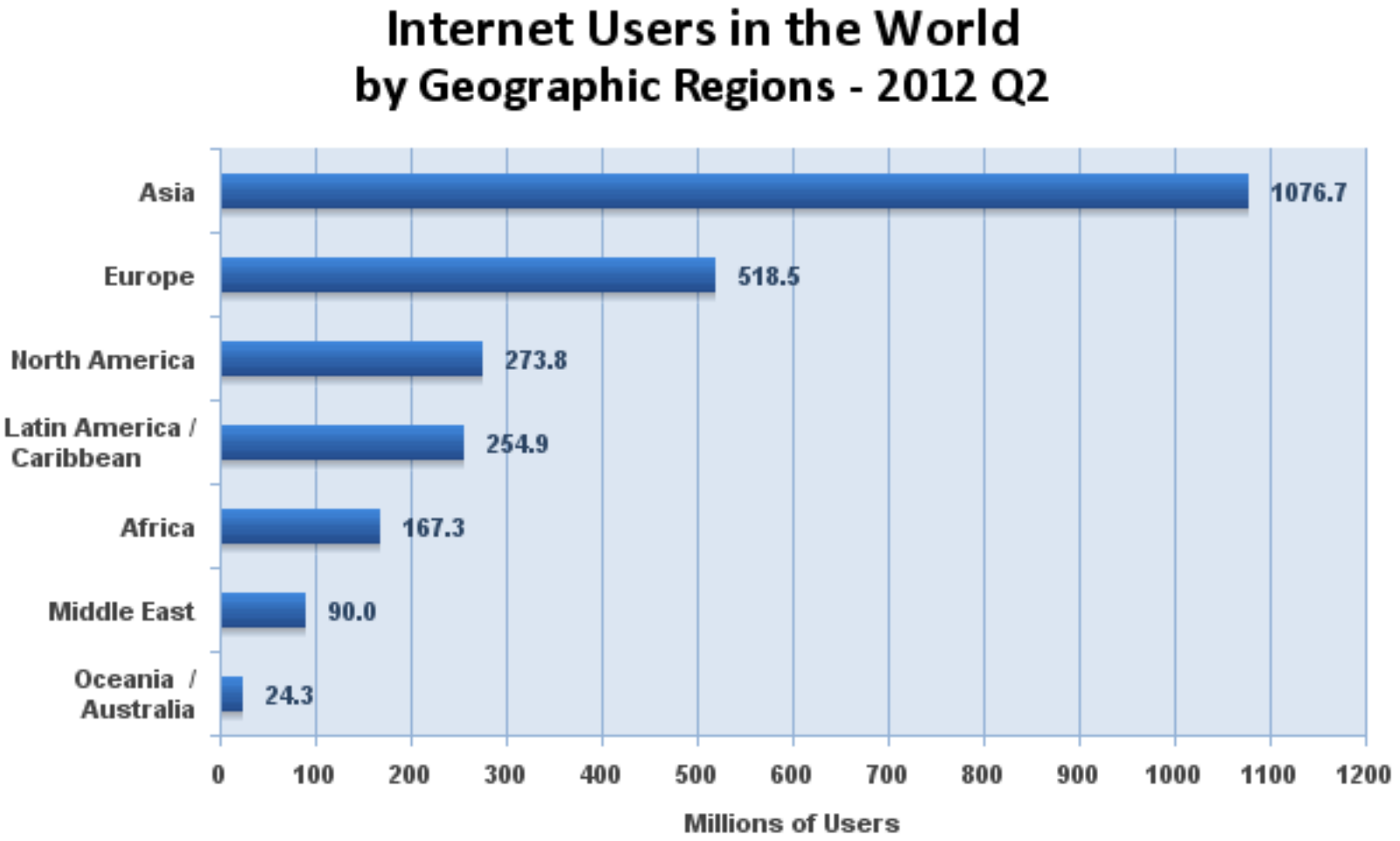
Breve História da Web

- 1989:
  - HTTP, HTML
- 1990, Christmas:
  - First web server at CERN
- 1992-95: Growth
  - 1993: first Unix graphical web browser – Mosaic
  - 1994: first WWW conference, W3C is formed
- 1996-98:
  - Commercialization of the WWW
- 1999-2001:
  - Dot-com boom
- 2002:
  - Web is ubiquitous
    - Web 2.0,
    - User generated content (blogs, rss)
    - Semantic Web

A Web

- É enorme!
  - 20 vezes Library of Congress
- É dinâmica!
  - 40% mudam em uma semana
  - 23% .com mudam diariamente
- É auto-organizável!
  - Sem padrões,heterogênea, sem processo de revisão
- É hyperlinked! (link analysis – máquinas de busca)

Web: Número de Usuários



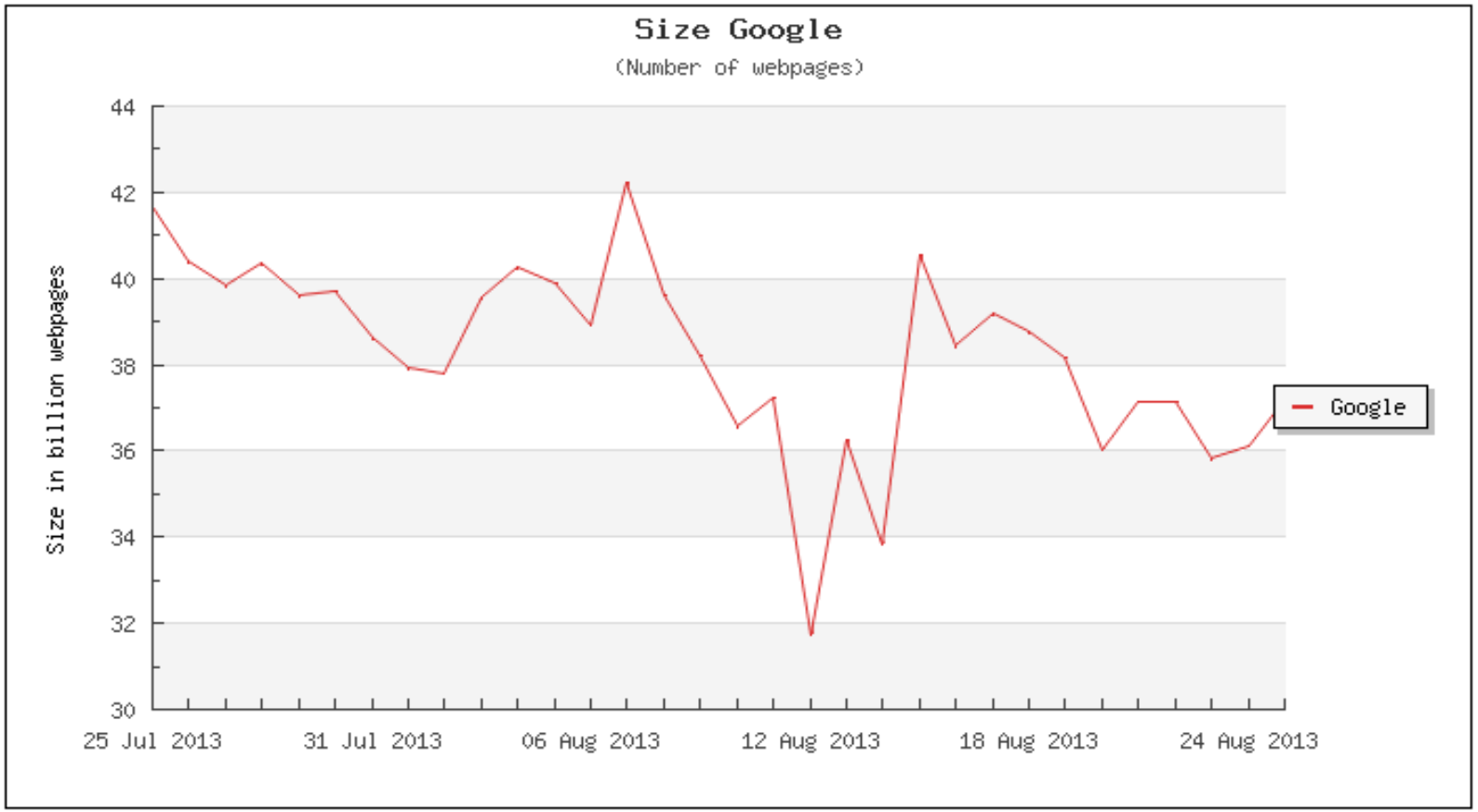
Source: Internet World Stats - www.internetworldstats.com/stats.htm
2,405,518,376 Internet users estimated for June 30, 2012

Web: Estimativa Número de Páginas

Google

The size of the World Wide Web:
Estimated size of Google's index

Last Month | Last Three Months | Last Year | Last Two Years

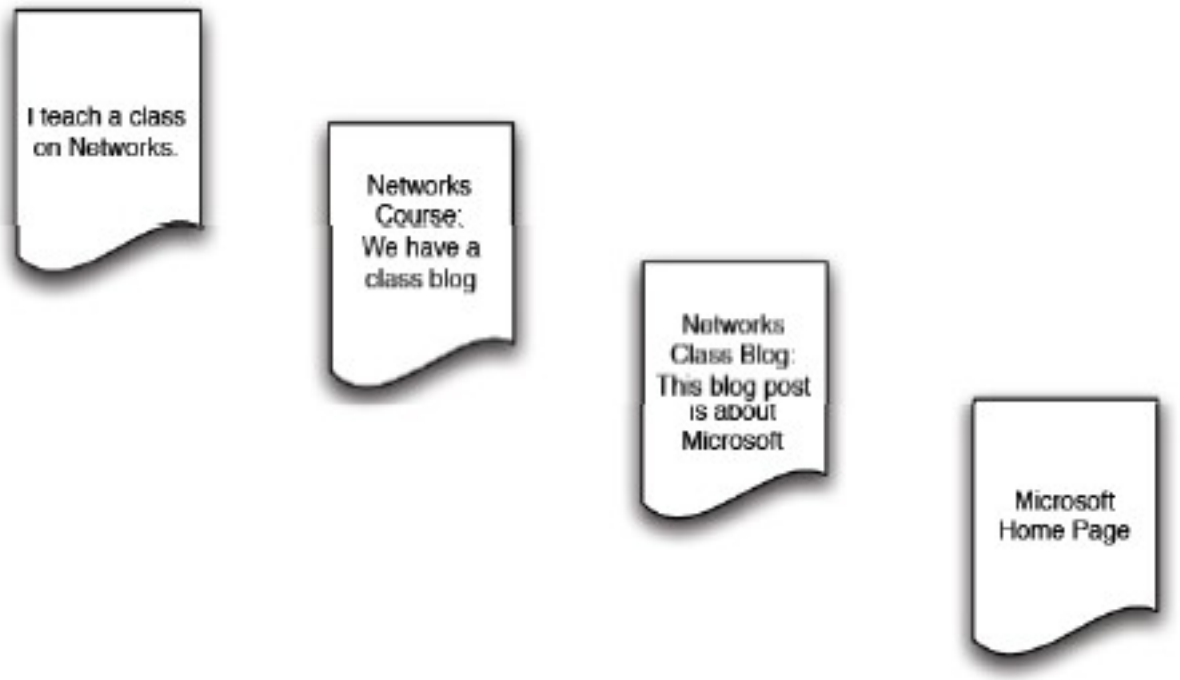


Web: Estimativa Número de Páginas

- Agosto de 2013
  - Google: 37 bilhões
  - Bong: 3 bilhões

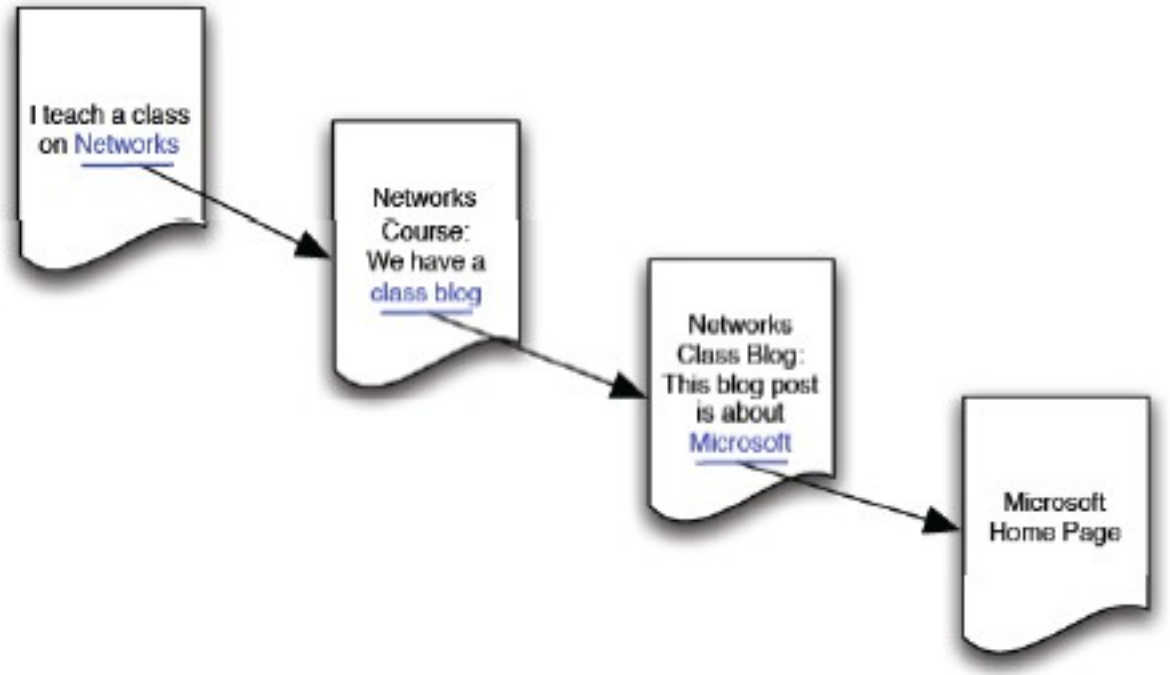
- Não é claro:
  - Como descobrir novas páginas
  - Documentos escondidos
  - Tipo de página:
    - Páginas geradas dinamicamente (Deep Web)
    - URLs (time stamped)

Web como um Grafo

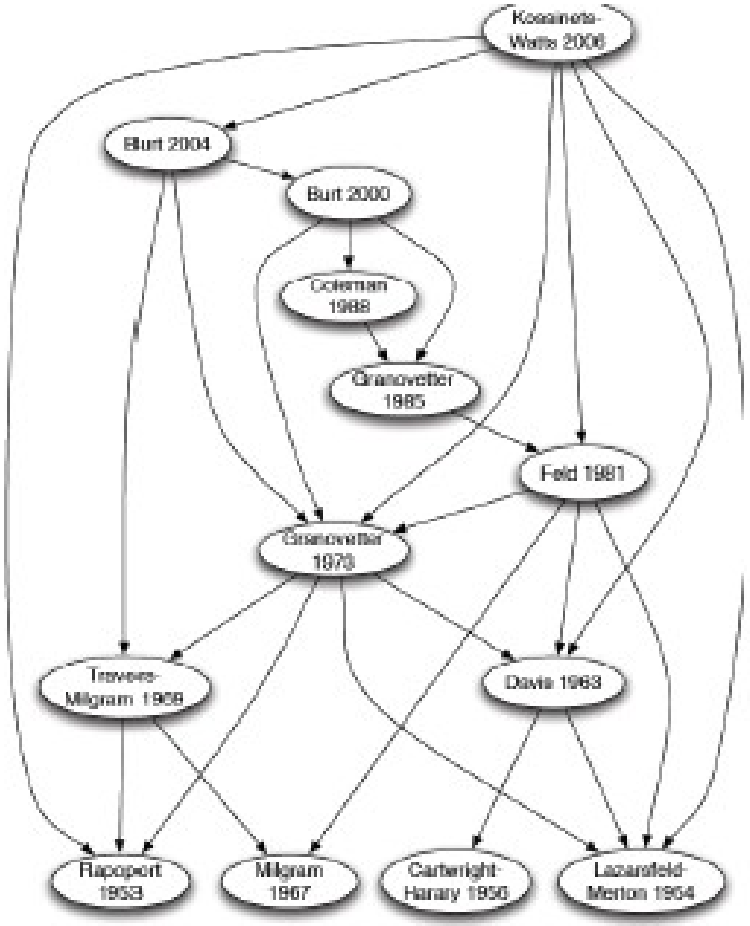


Web como um Grafo

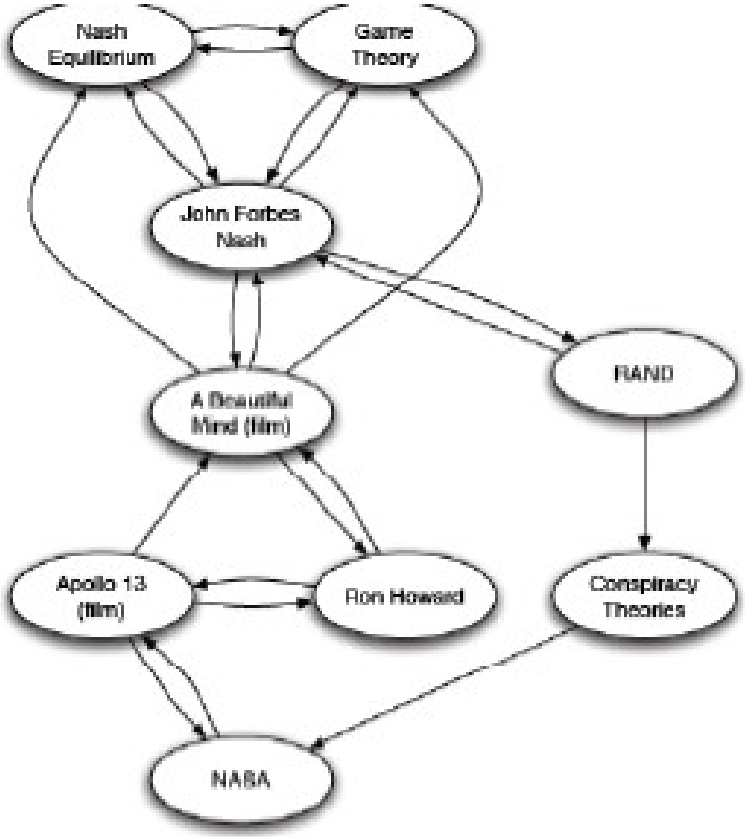
- No começo: navegação
- Atualmente: grande volume de transações



Outras Redes de Informação



Citations

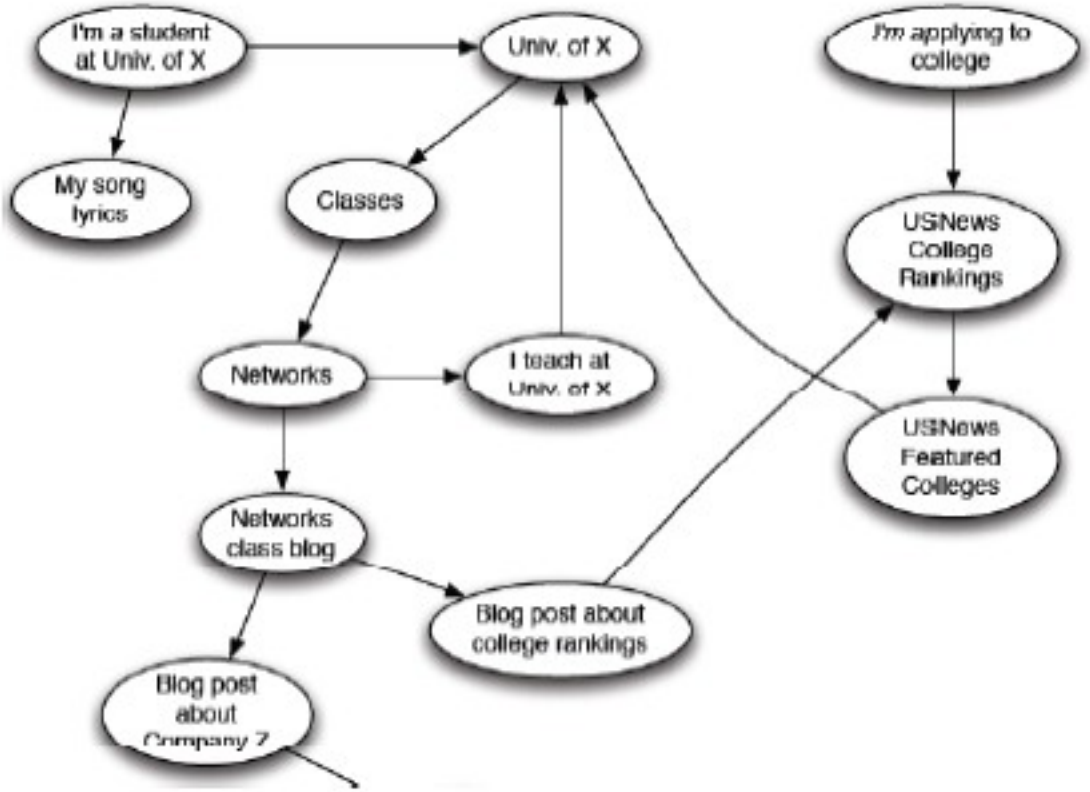


References in enciclopedia

Com o quê a Web se parece?

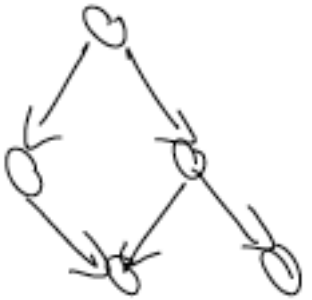
- Como a informação é organizada?
- Como a Web se interconecta?
- Como é o "mapa" da Web?

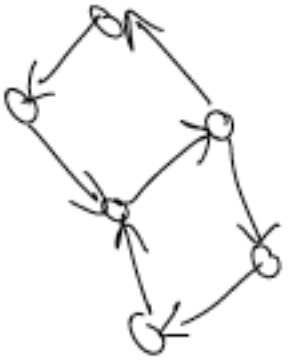
Web como um grafo direcionado



Grafos Direcionados

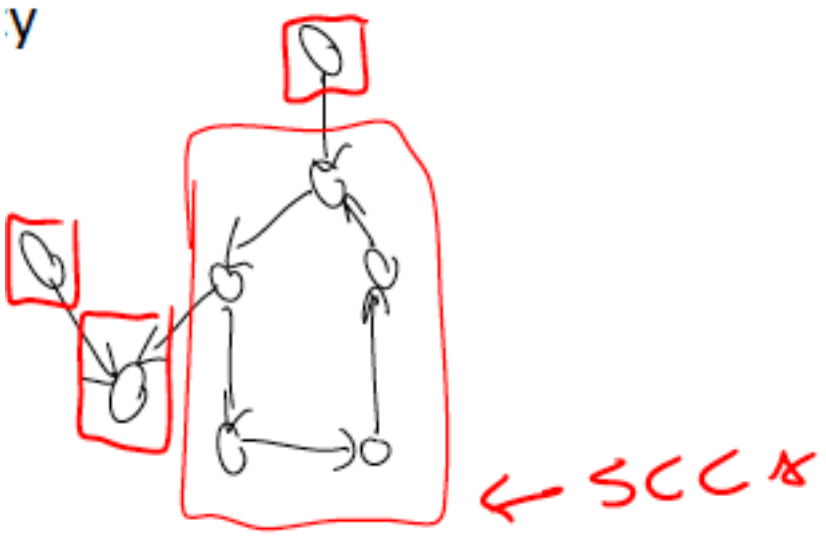
- Dois tipos:
  - Grafos diretos acíclicos
    - Se u pode alcançar v então v não alcança u
  - Fortemente conectados
    - Todo nó alcança outro nó via caminho direcionado





Componente Fortemente Conectada

- Conjunto de nós em um grafo tal que:
  - Todo par de nós em S pode alcançar uns aos outros
  - Não existe um grupo maior contendo S que possua esta propriedade



A estrutura em grafo da Web

- Existe uma componente gigante [Broder et al 2000] fortemente conectada (CGFC)

- Outros conjuntos de nós
  - IN: conjunto de nós que alcançam a CGFC e que não podem ser alcançados pela CGFC
  - OUT: conjunto de nós que são alcançados pela CGFC, mas não alcançam os nós

A estrutura em grafo da Web

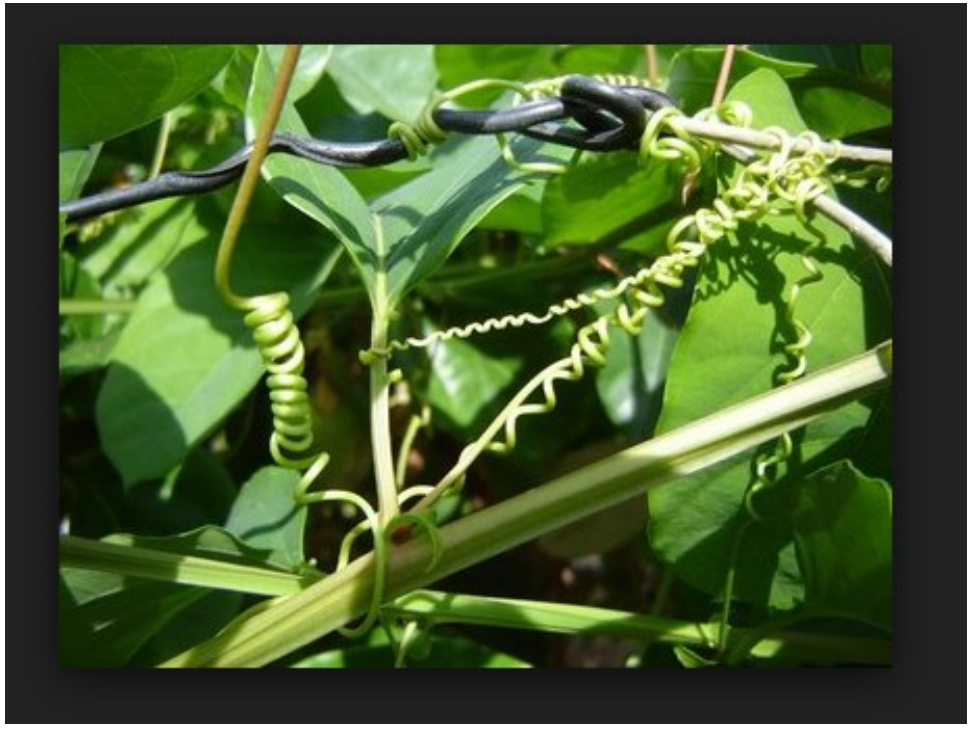
- Outros conjuntos de nós
  - Tendrils:

    (a) nós que são alcançados por IN, mas não alcançam CGFC;

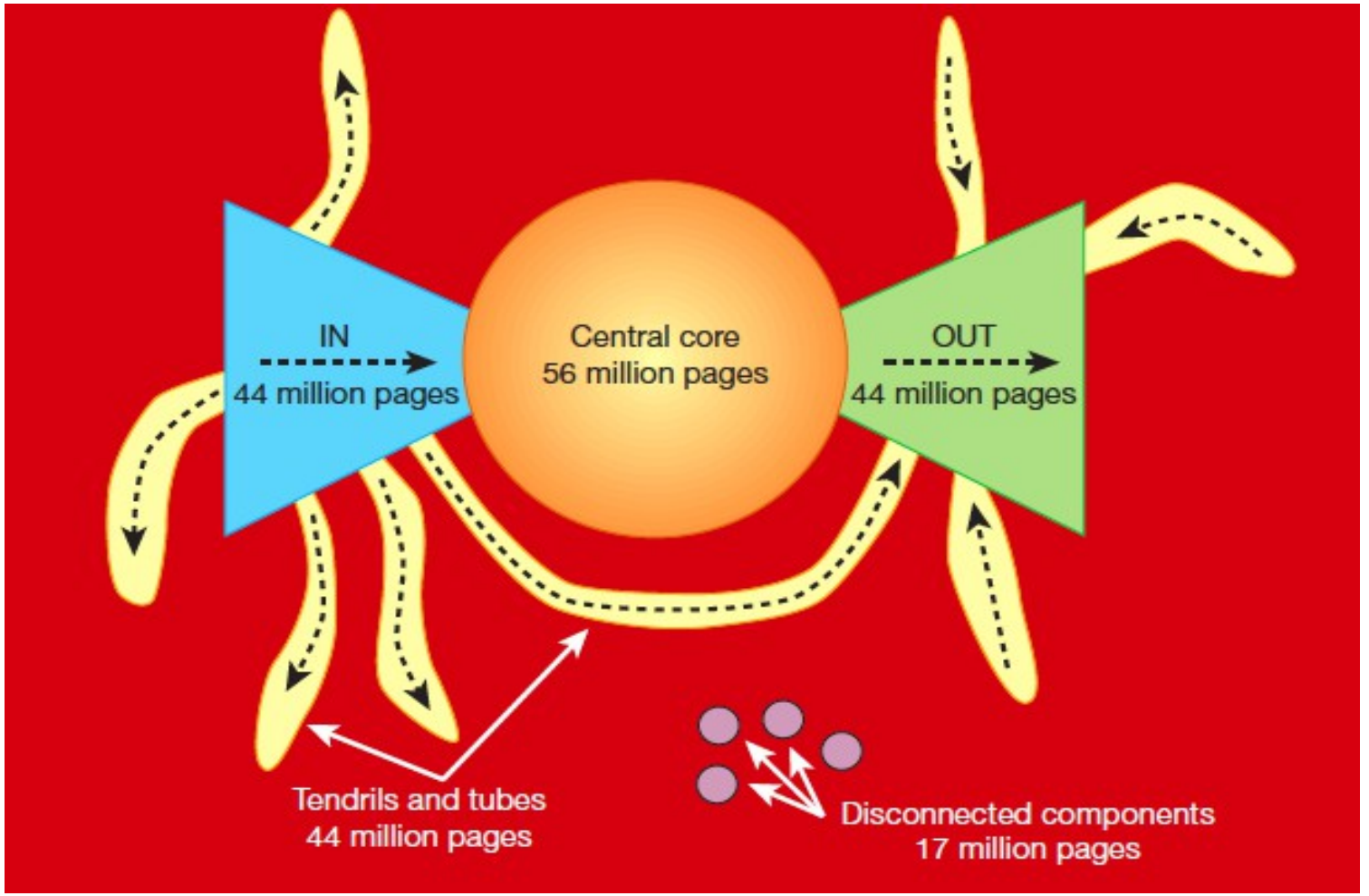
    (b) nós que podem alcançar OUT, mas não podem ser alcançados por CGFC

  - Desconectados: Nós sem caminhos para CGFC, mesmo ignorando direção

Tendrils = Gavinhas :)



A estrutura da Web (gravata borboleta)

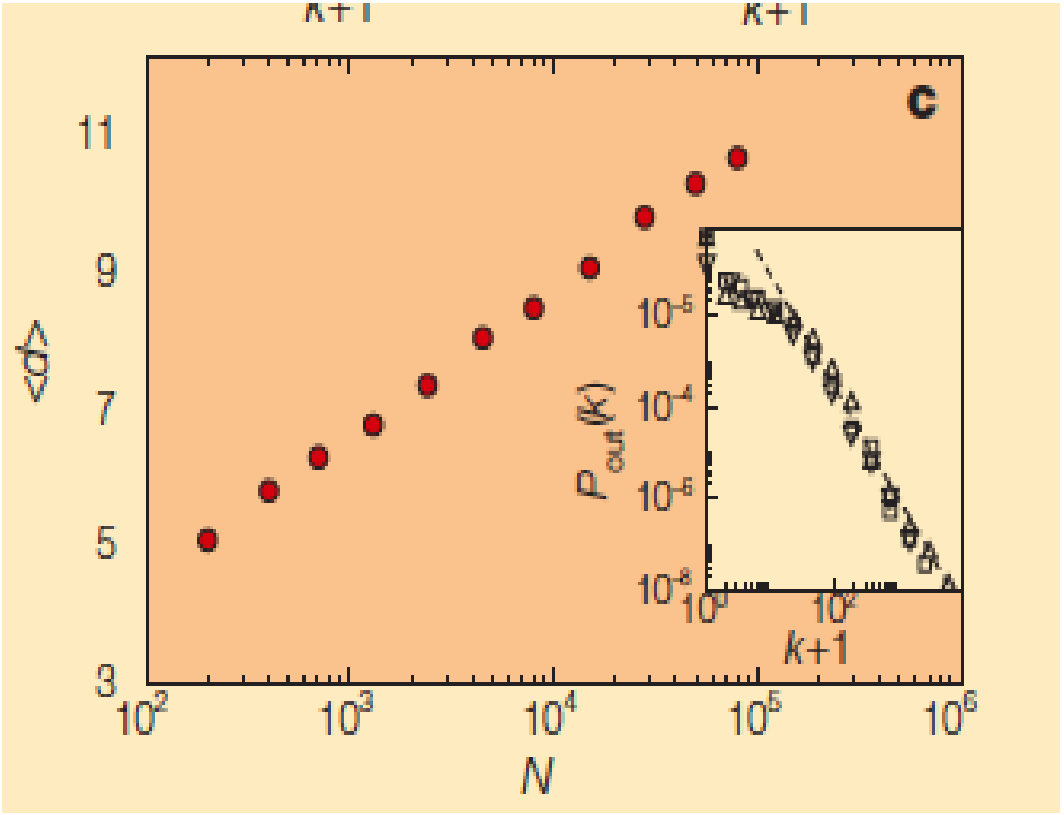


- 250 milhões de páginas, 1.5 bilhões de links [Altavista]

A estrutura da Web (gravata borboleta)

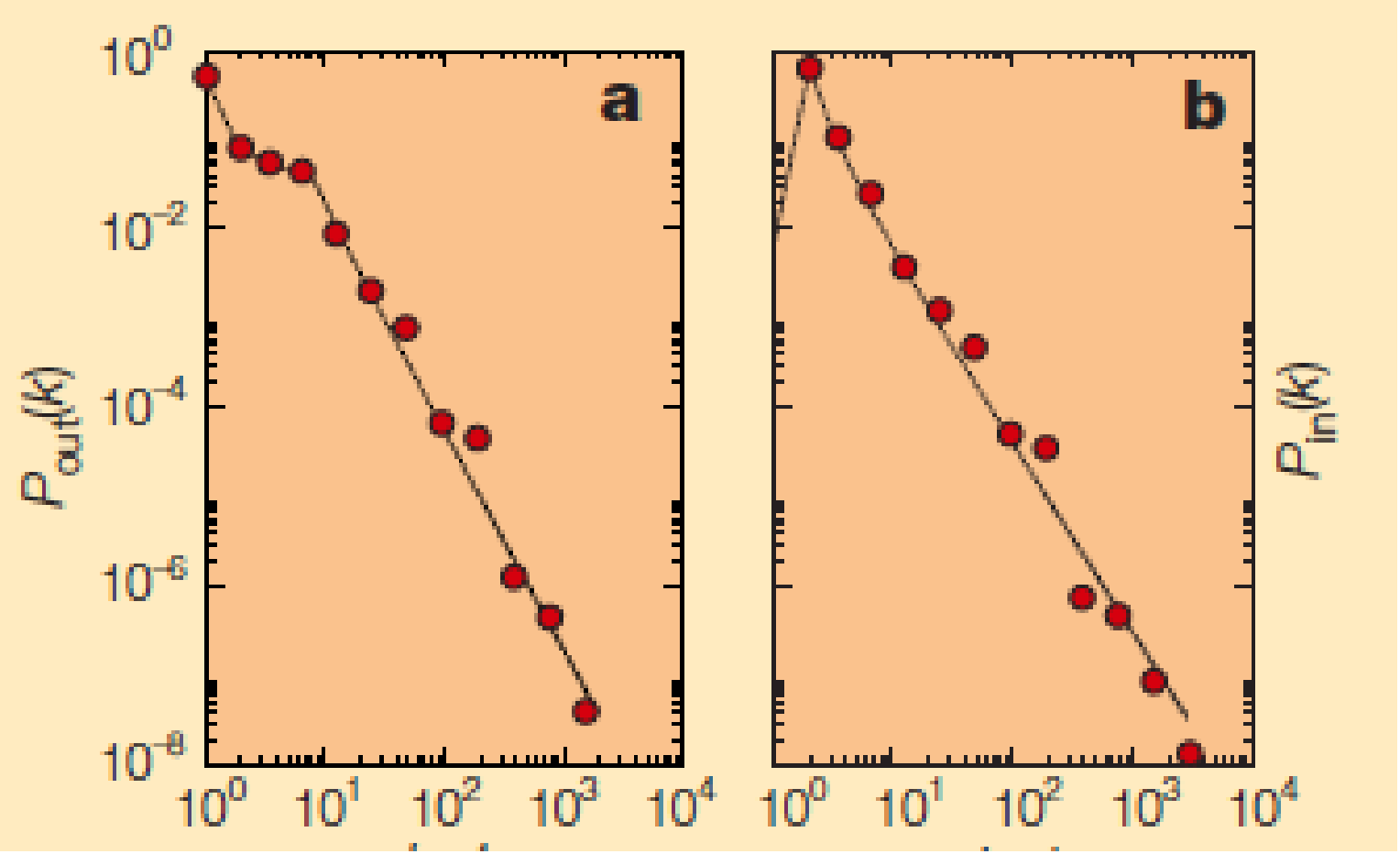
- Núcleo central, com as páginas mais importantes
- Páginas que chegam e saem, ou não interagem com este núcleo
- Dinâmica: Limites estão sempre mudando, com nós entrando, saindo da CGFC

Diâmetro [Albert et al - 1999]



- Diâmetro (média dos caminhos mínimos) é 19

Distribuição do Grau [Albert et al - 1999]



- Power Law

Decaimento Exponencial x Power-Law