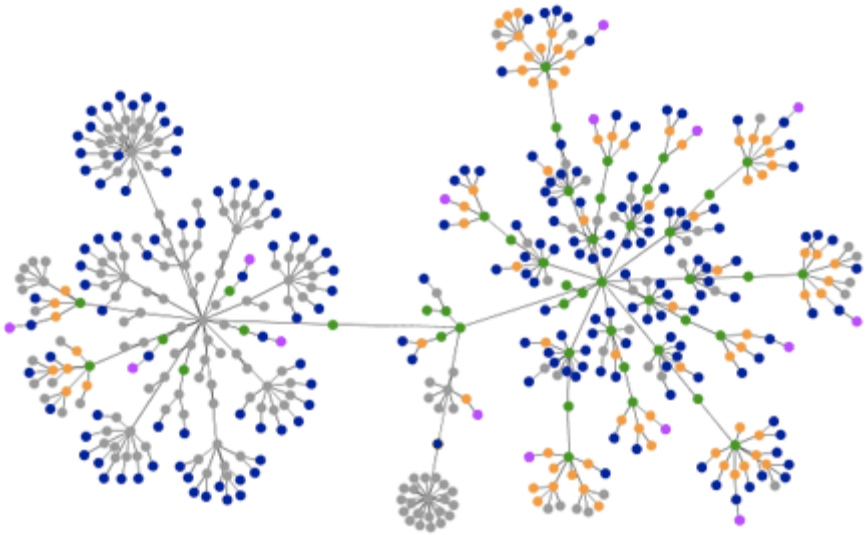
Laboratorio de Redes,
Recuperación de Información
y Estudios de la Web

Recuperación de Información en la Web y Motores de Búsqueda

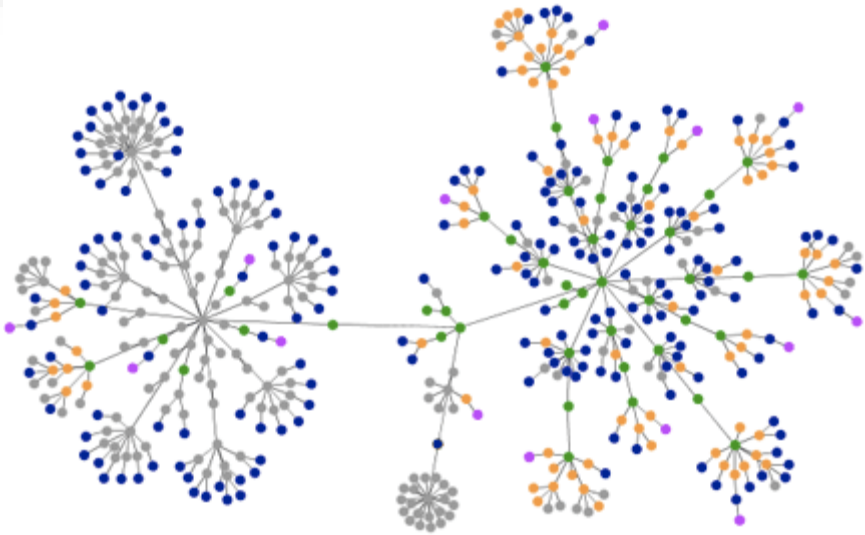
Gabriel H. Tolosa
tolosoft@unlu.edu.ar

- 2015 -



Estructura y Características de la Web

WWW



- Algunas preguntas:
  - ¿Qué es?
  - ¿Cuál es su estructura?
  - ¿Cuál es su tamaño?
  - ¿Cuántos sitios tiene?
  - ¿Y cuántas páginas?
  - ¿Cómo “cambia” una página web?

Qué es? (a los efectos de RI)

- Una “forma” de compartir información
  - Servidores independientes
  - Cada uno con recursos “propios”
  - Identificados por una URL
- Interface → Navegador
- Publicación abierta
- Multimedia

Hoy es una plataforma!!!

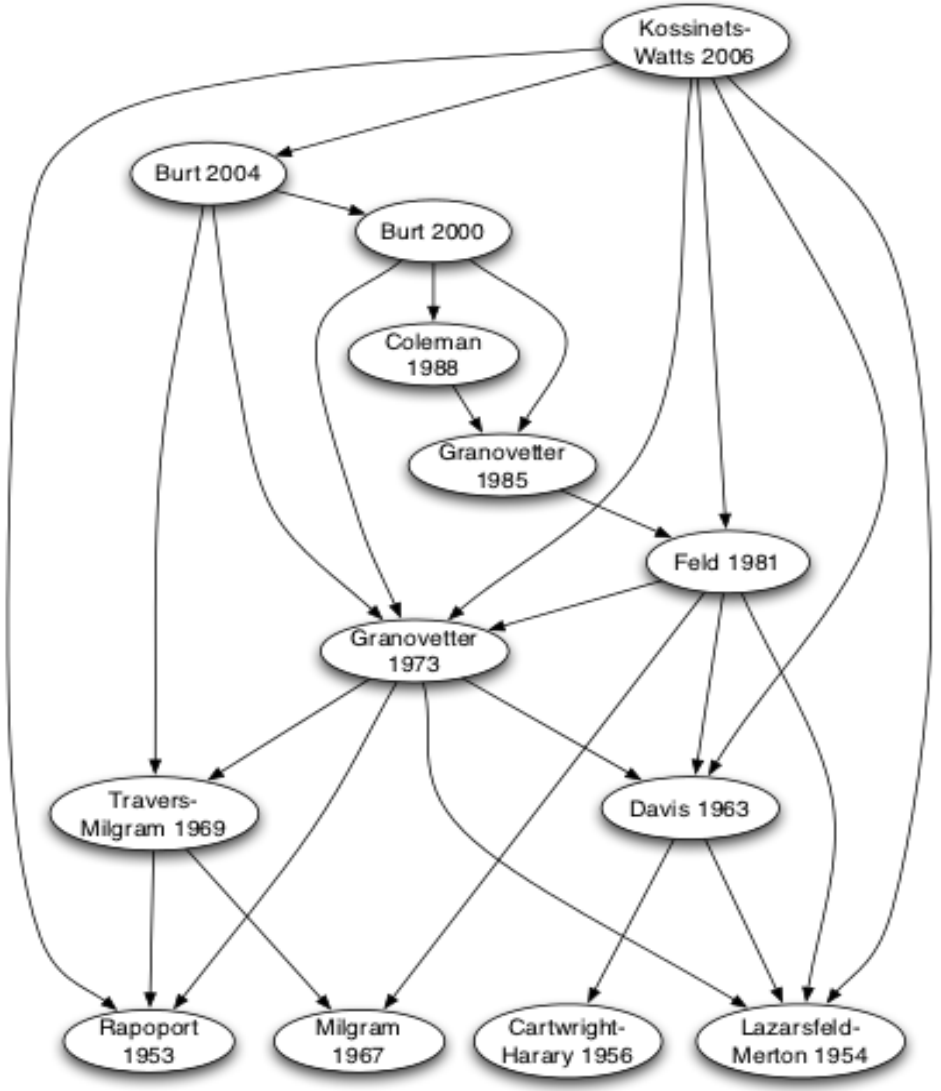
Qué es? (a los efectos de RI)

- Repositorio distribuido
  - Grafo dirigido masivo
- Complejo
- HTTP y HTML (básicamente)
- Hipertextual
- Hyperlinks
  - Estructura no-lineal
  - Relaciones lógicas
  - No “tan” obvia

Hyperlinks (no web)

- Citation networks
- Co-authorship
- Cross-references (enciclopedias)
- Cine (oob)

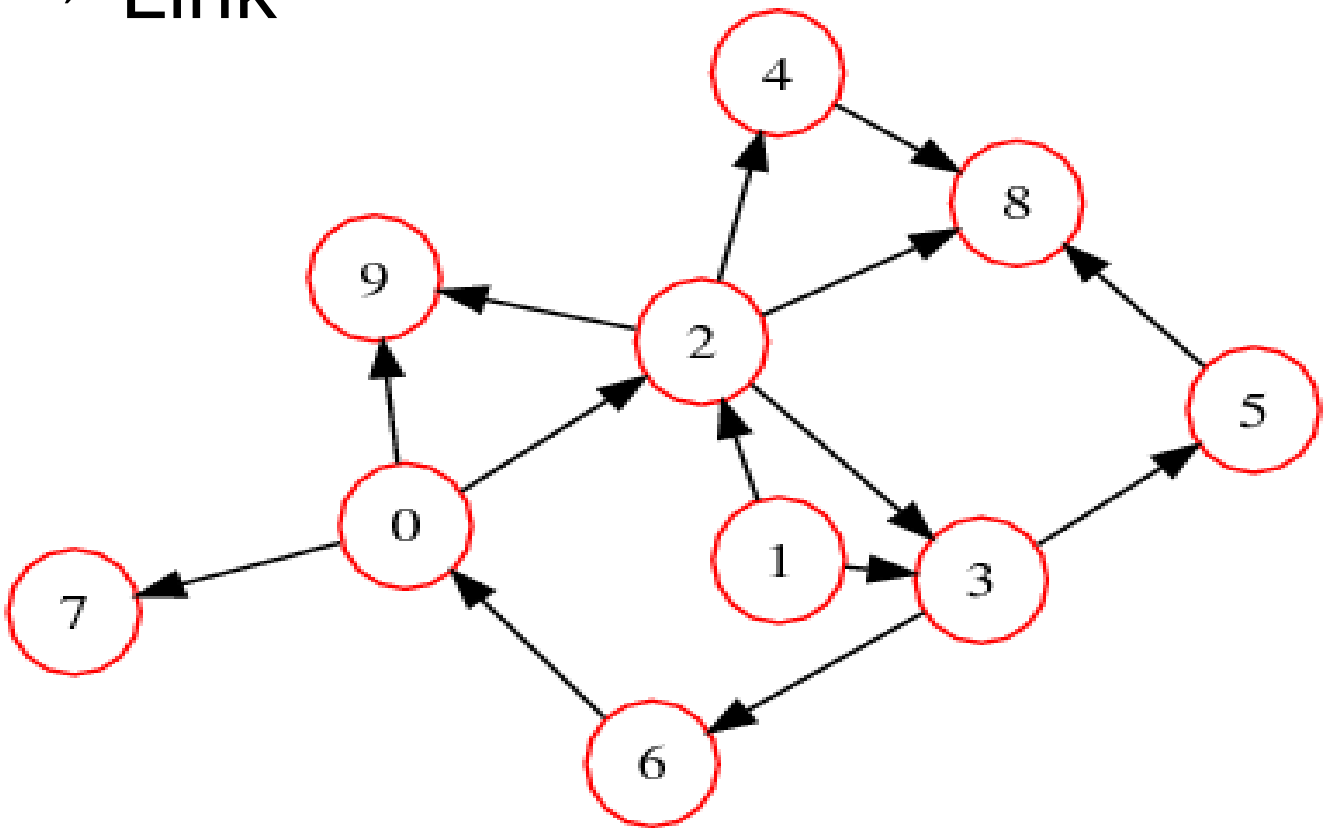
...



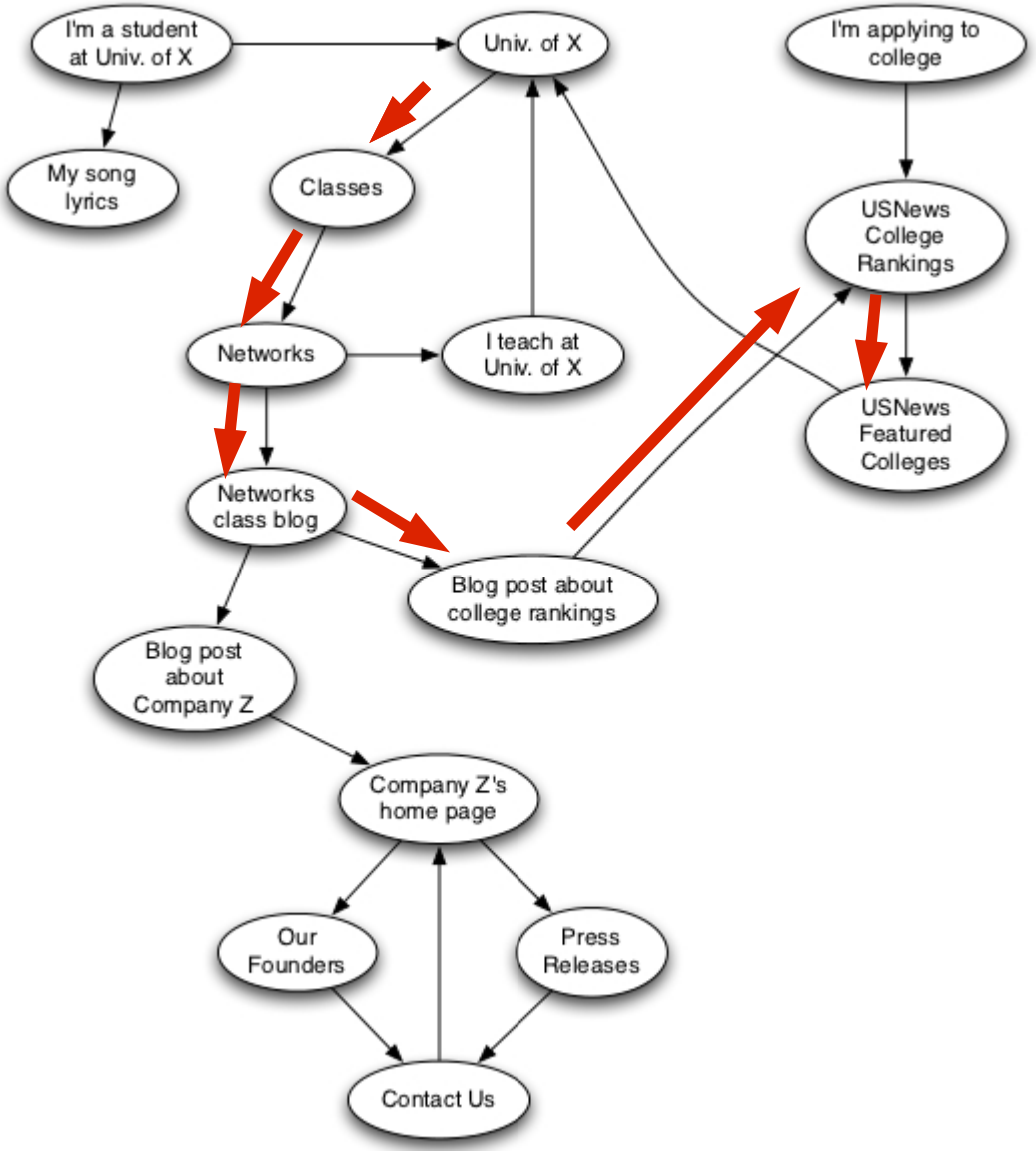
Networks, Crowds, and Markets: Reasoning About a Highly Connected World. D. Easley y J.Kleinberg http://www.cs.cornell.edu/home/kleinber/networks-book/

Estructura de grafo

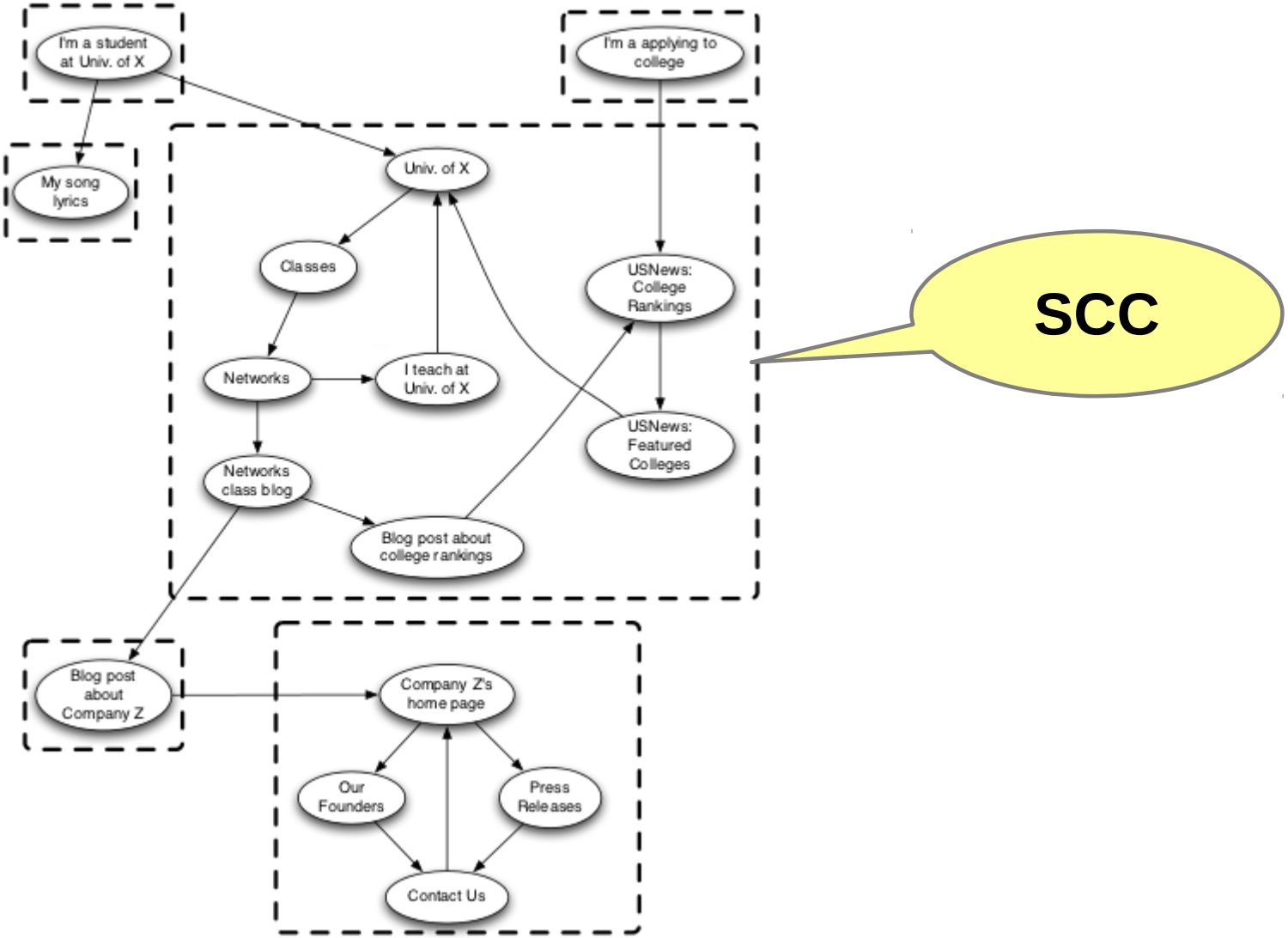
- Nodo → Página web
- Arco → Link



Estructura de grafo

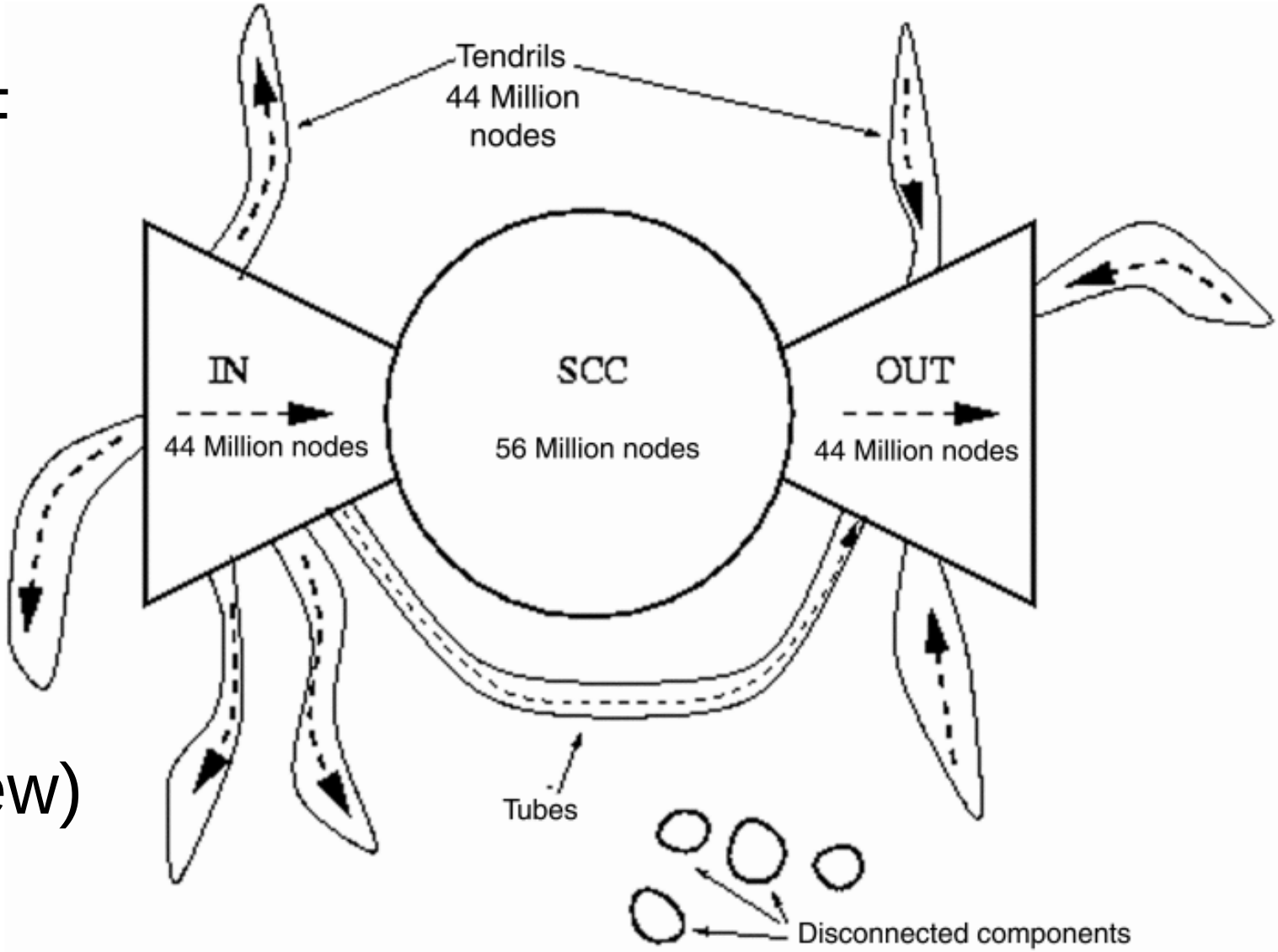


Estructura de grafo



Estructura de grafo

- Crawl BSF
- **203 M**
  de URLs
- **1,466 M**
  de links
- **Bow-Tie**
  (macro-view)

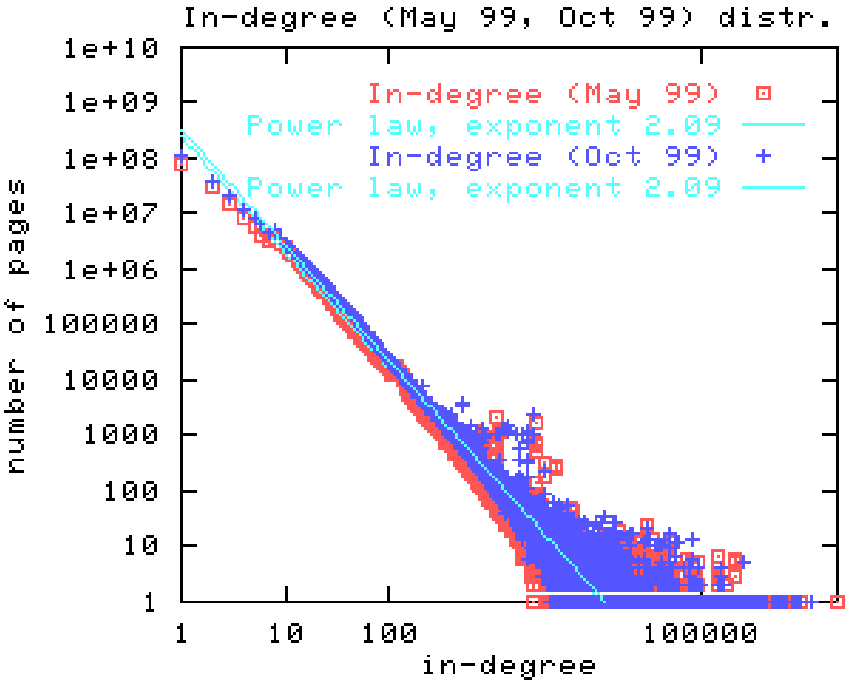


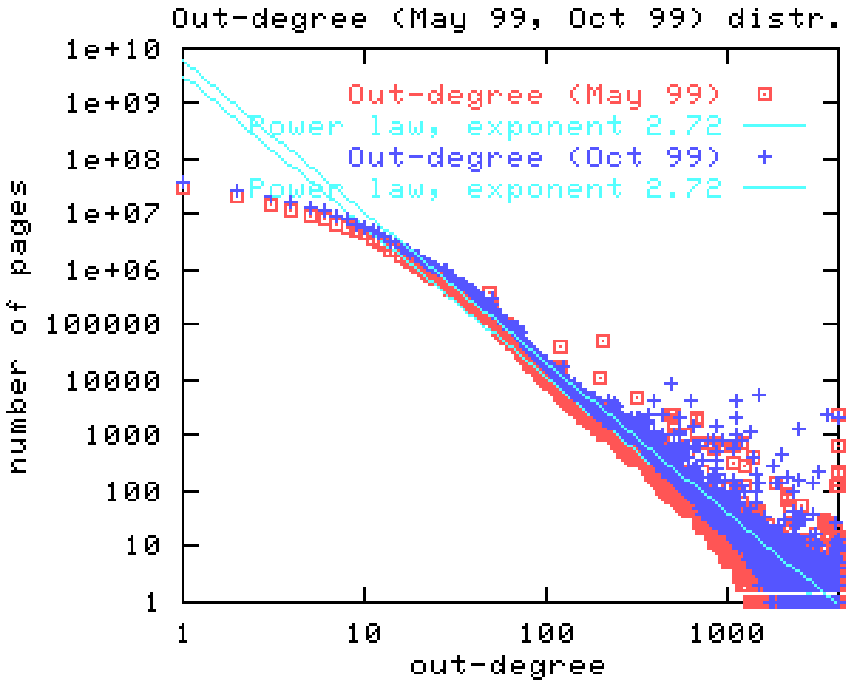
Estructura de grafo

- "*Graph Structure on the Web*" [Broder et al., 1999]
- Grado entrante/saliente → Distribuciones: Power-Law

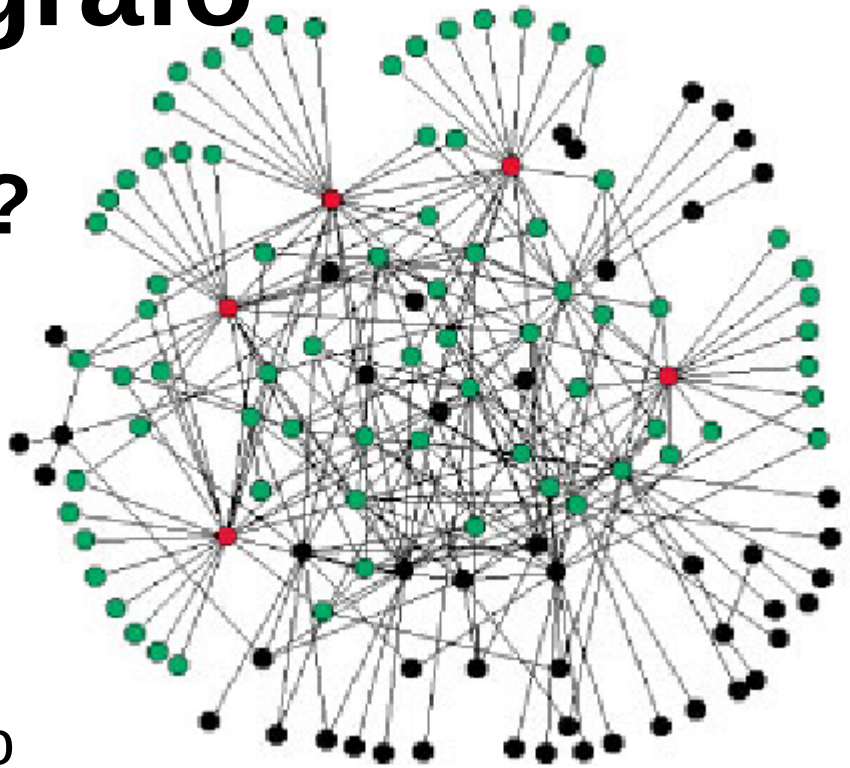
indegree : $\frac{1}{n^{2.1}}$

outdegree : $\frac{1}{n^{2.72}}$





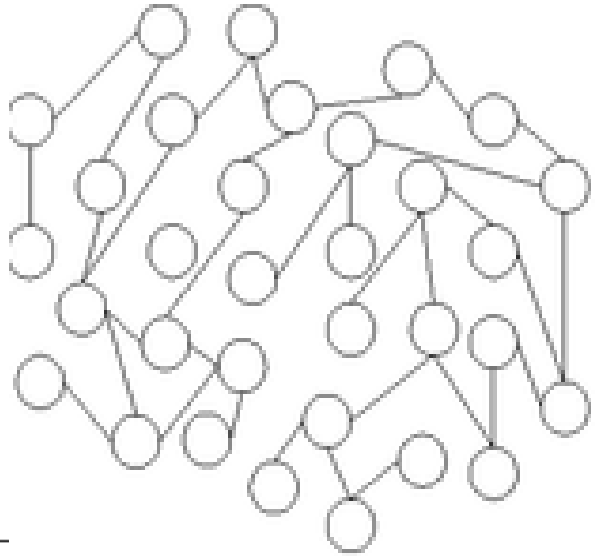
Estructura de grafo

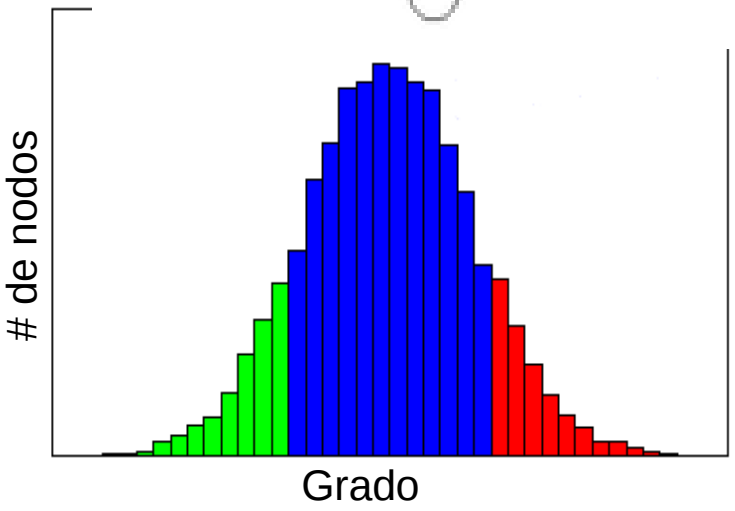


- **Por qué una “Power-Law”?**
- Efecto: **Richer-Get-Richer**
  - Un nuevo nodo se une a la red
  - Establece links con L de los existentes
    - El nodo ***X*** se conecta a un nodo ***Y*** con probabilidad proporcional al grado de ***Y***.
  - Entonces, los nodos con más enlaces tienden a “atraer” nuevas conexiones
- El efecto resultante: Red libre de escala (Scale-Free)

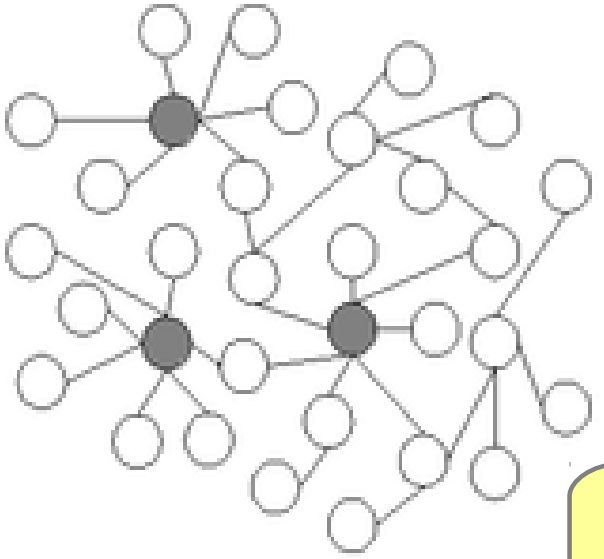
Estructura de grafo

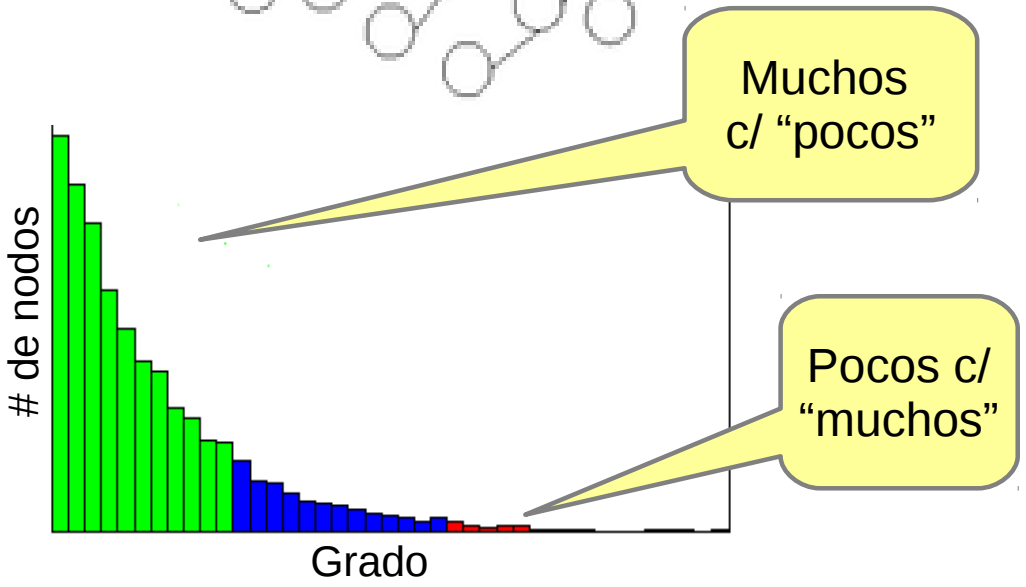
Random



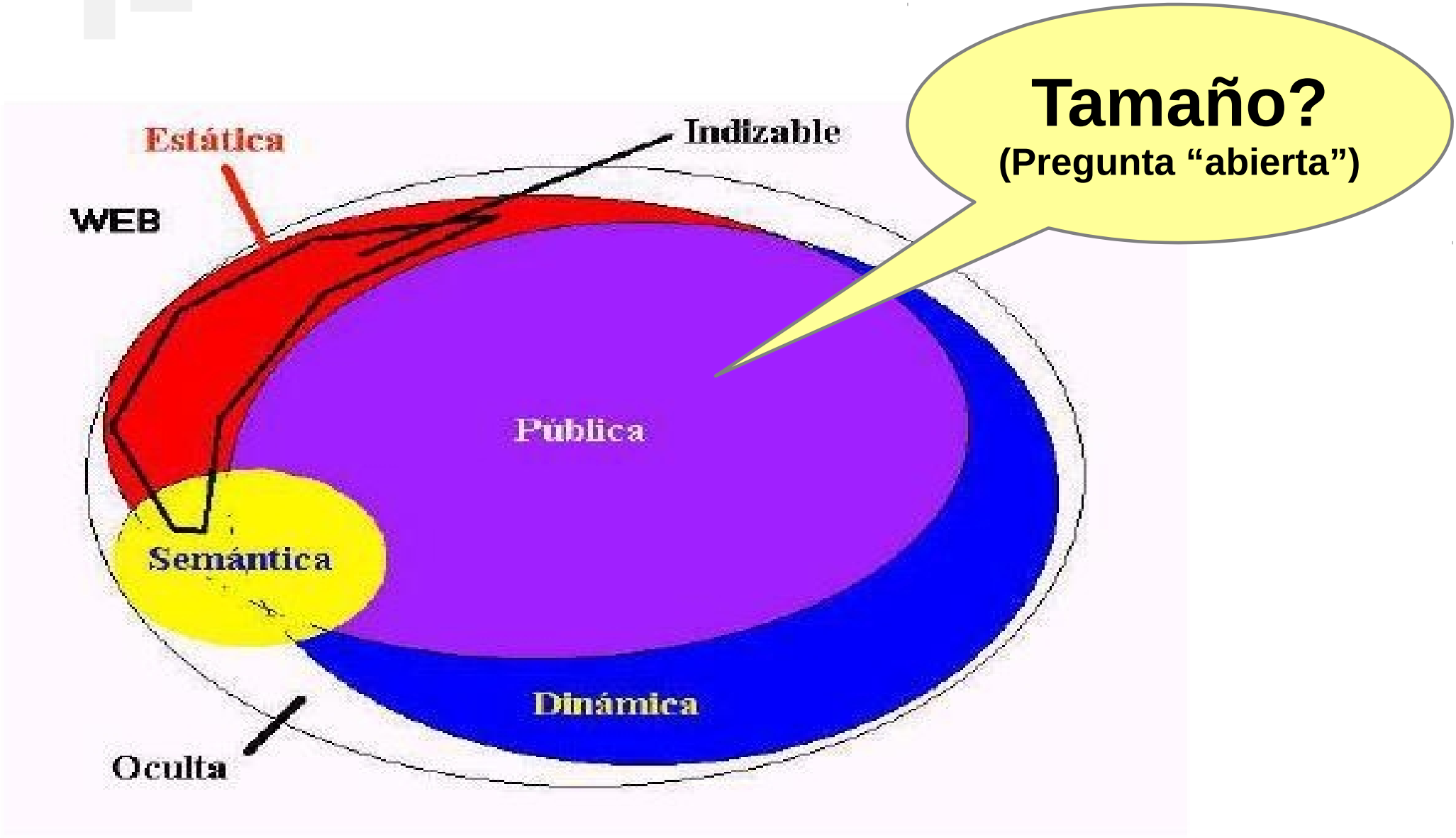


Scale-free

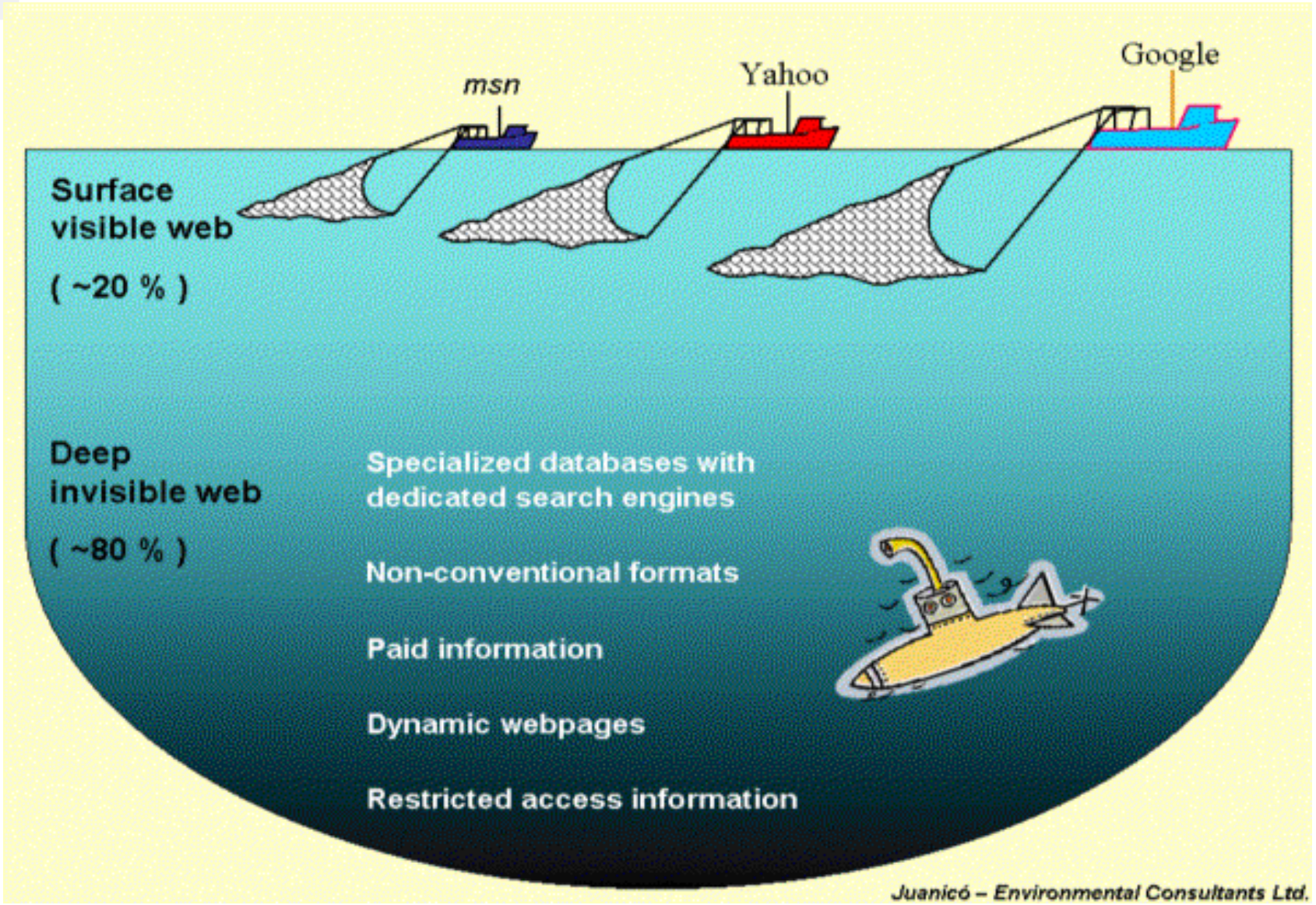




Otra vista [Baeza-Yates, 2003]



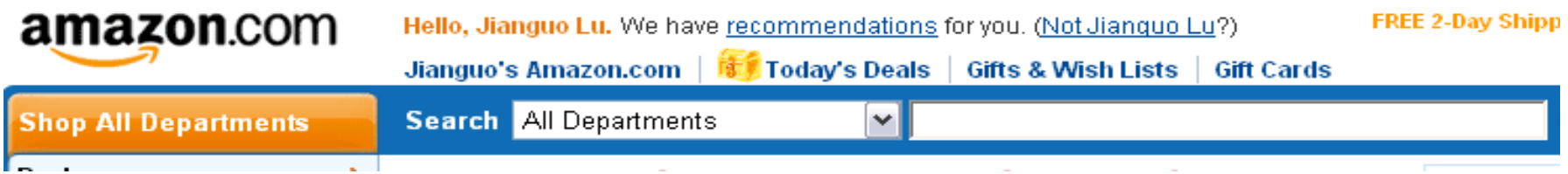
Web “profunda”



Web “profunda”

- No todo está en “superficie”, por qué?
  - Páginas “on the fly”
  - Datos históricos
  - Contenido con “derechos”
  - Contenido protegido por passwords
- Google “trata” de recorrer la web profunda

Madhavan, Jayant; David Ko, Łucja Kot, Vignesh Ganapathy, Alex Rasmussen, Alon Halevy. **Google's Deep-Web Crawl**. VLDB, 2008.



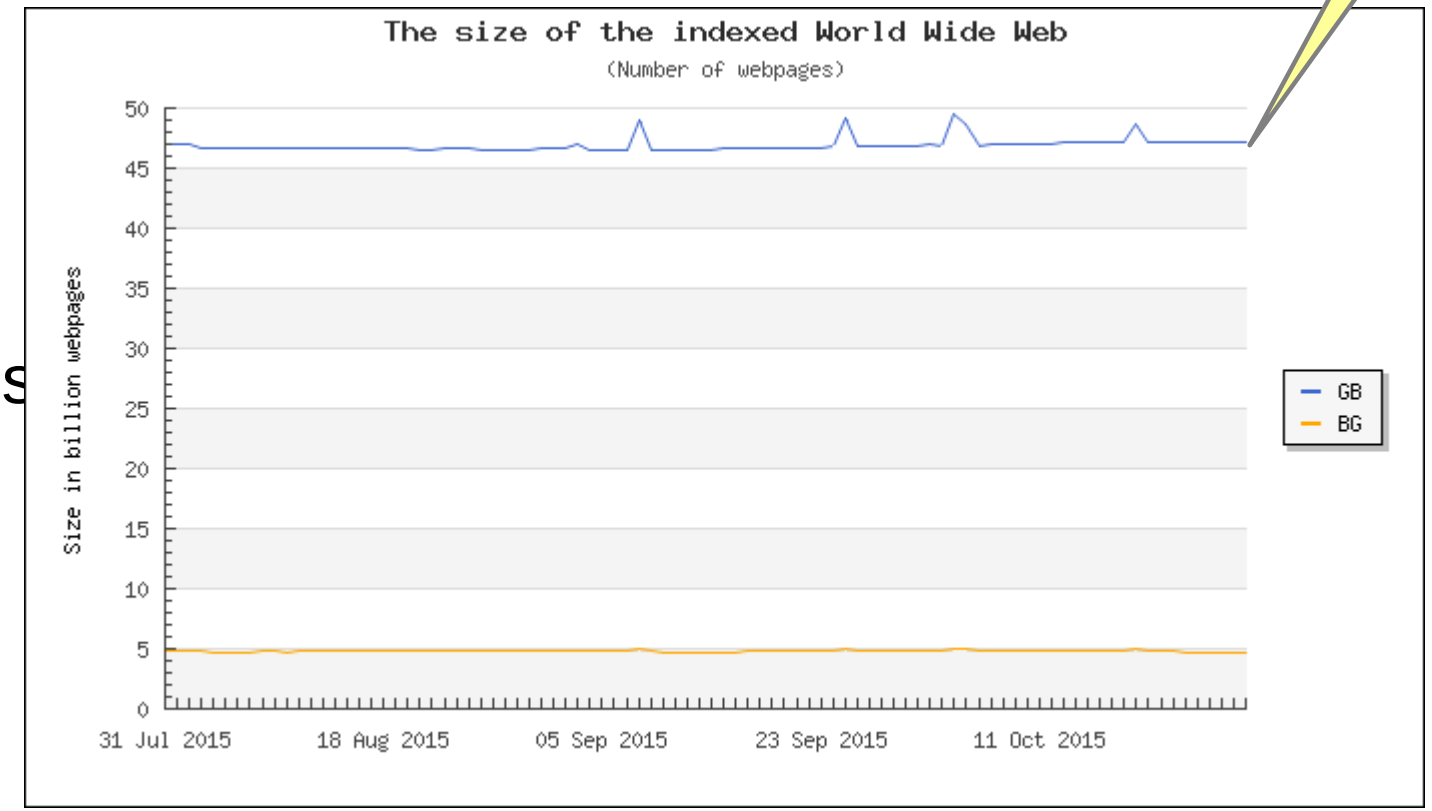
Tamaño

- Dificultades para definir “qué” medir
  - Nodos “temporales”: Su notebook con un web server personal, es parte de la web?
  - La porción dinámica es potencialmente infinita
    - Información del tiempo (climático)
    - Consultas a una base de datos
    - Blogs
  - Web “profunda”
    - Todos los artículos de un periódico
  - Duplicados (mirroring)
    - Se estiman en un 30% (antes del cross-posting)

Tamaño

En 2003, 24 mil millones

En 2005, se estimó en 11.5 miles de millones de páginas
[Gulli, et al., 2005]



¿Hoy?

http://www.worldwidewebsize.com/

[Gulli, et al., 2005] Gulli, A. & Signorini, A. **The Indexable Web is More tha 11.5 billion pages**. WWW, 2005.

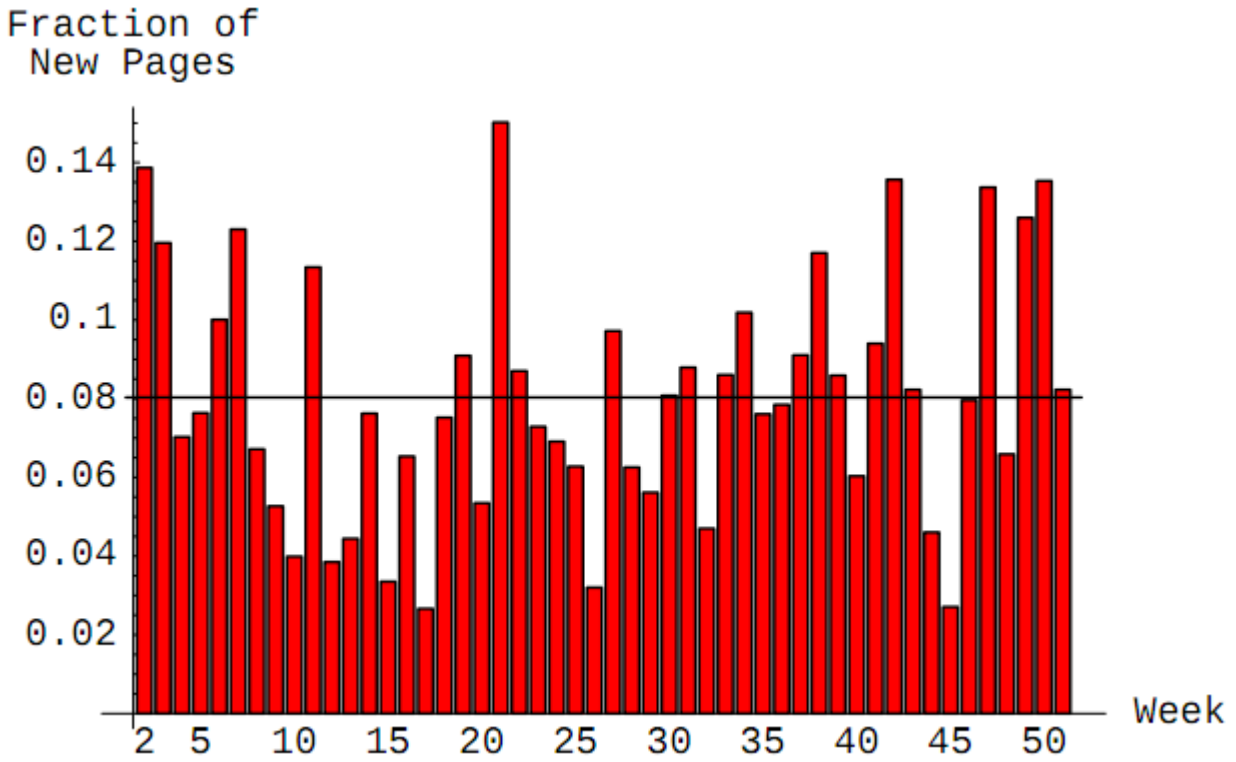
Dinámica [Ntoulas et al.]

- La web está constantemente evolucionando:
  - Las páginas aparecen, desaparecen, cambian. ¿Cómo?
- Experimento:
  - Crawling de 154 sitios durante 1 año (tomados del directorio de Google)
  - 4.4 millones/semana (65 GB/semana)

[Ntoulas et al.] What's New on the Web: The Evolution of the Web from a Search Engine Perspective. WWW04.

Dinámica [Ntoulas et al.]

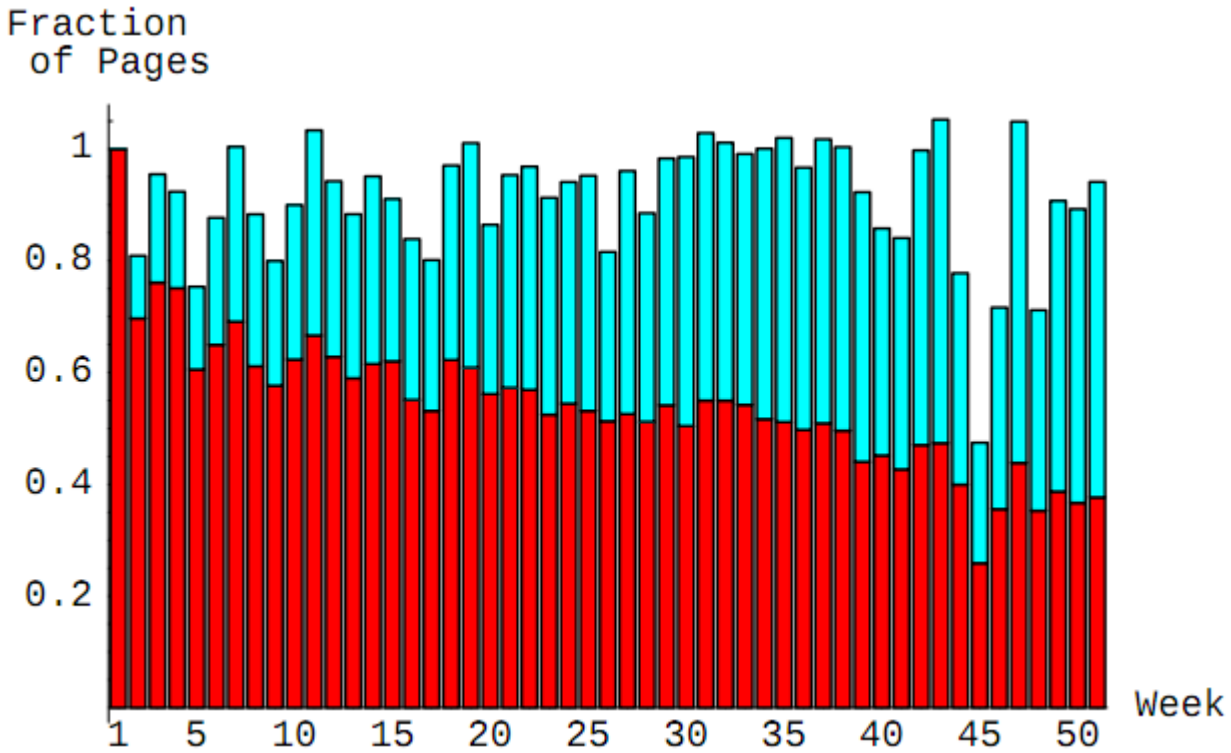
- Promedio de “nacimientos” semanales ~ 8%



[Ntoulas et al.] What's New on the Web: The Evolution of the Web from a Search Engine Perspective. WWW04.

Dinámica [Ntoulas et al.]

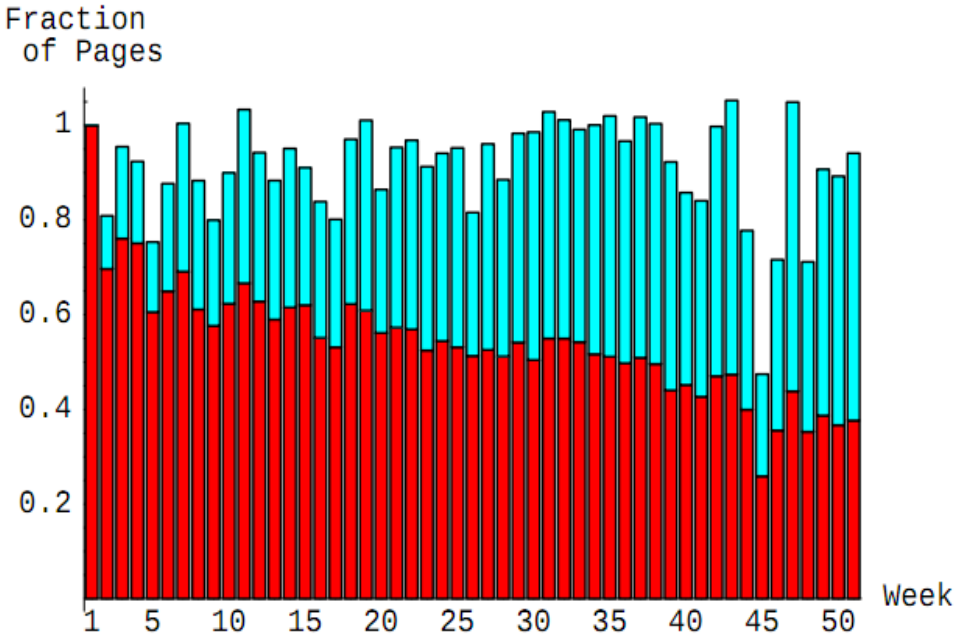
- Fracción de páginas desde el primer crawl que permanecen después de *n* semanas (rojo) y nuevas (celeste).



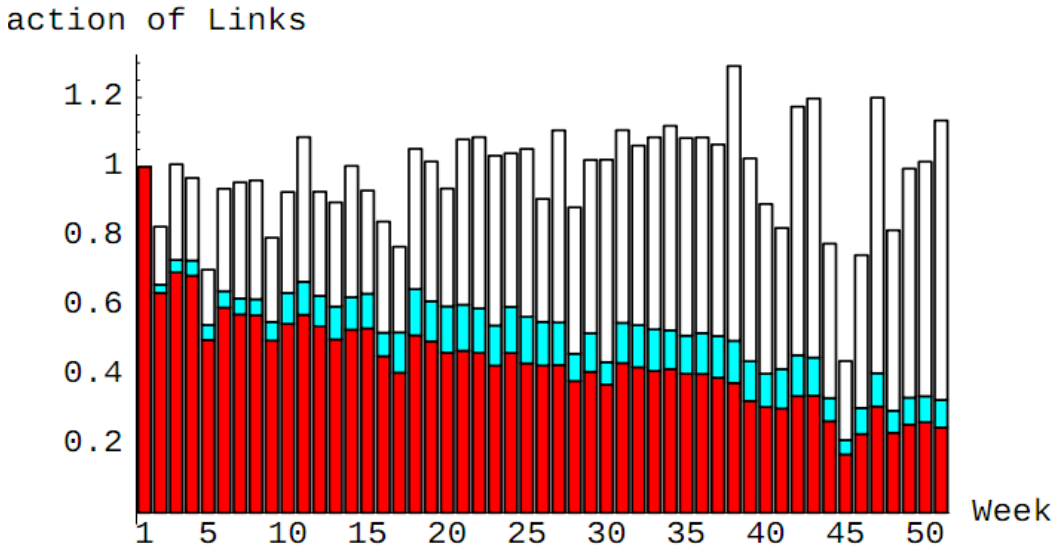
- Sin embargo, las “nuevas” paginas (que reemplazan a otras) toman “prestado” el contenido de existentes.

Dinámica [Ntoulas et al.]

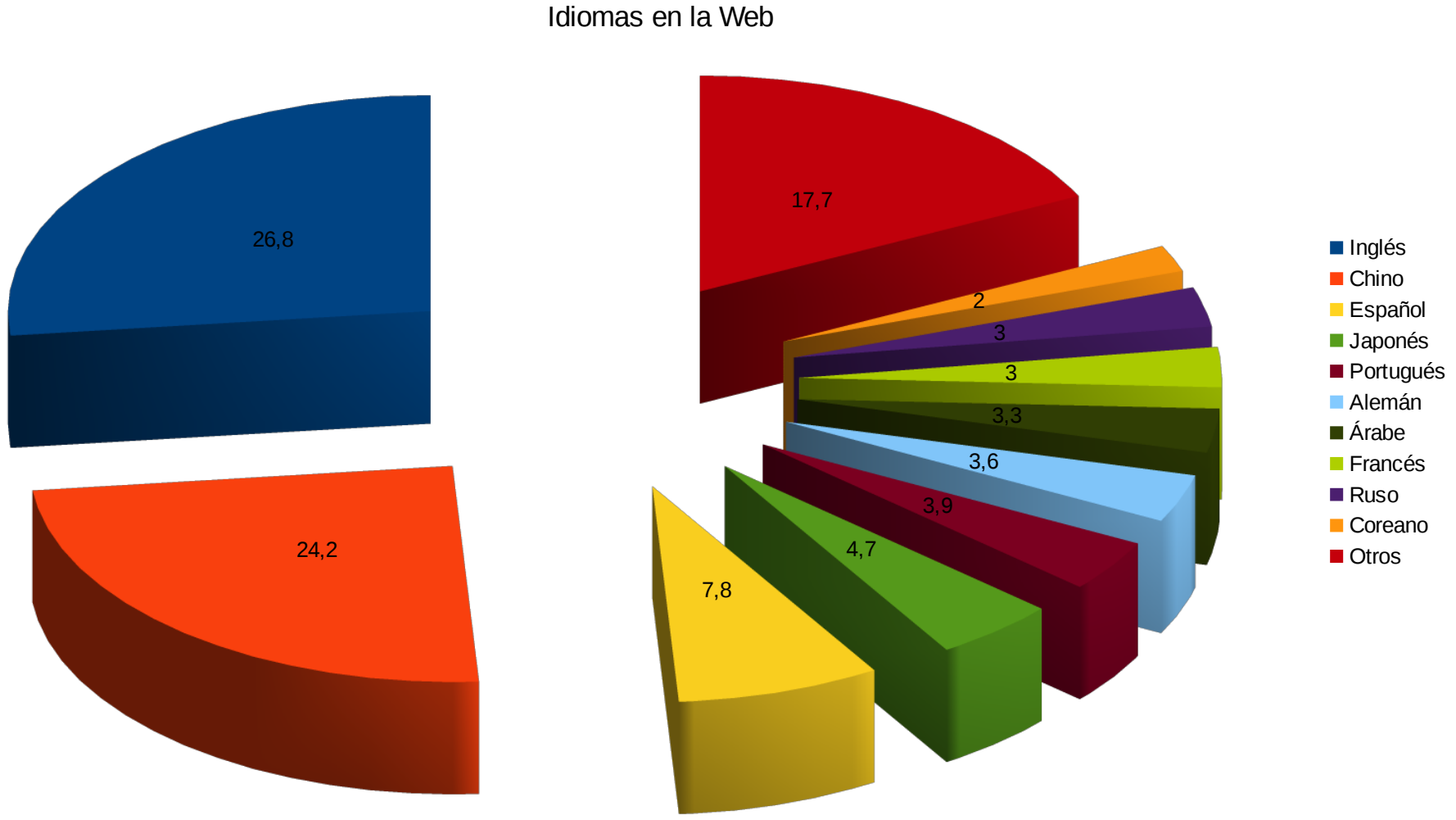
Fracción de páginas desde el primer crawl que permanecen después de *n* semanas (rojo) y nuevas (celeste)



Fracción de links desde la primera muestra aún después de *n* semanas (rojo), nuevos links en páginas existentes (celeste) y nuevos links en páginas nuevas (blanco)

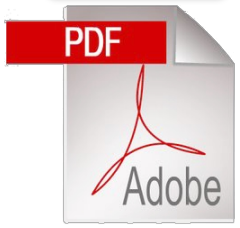


Heterogeneidad: Idiomas



Heterogeneidad

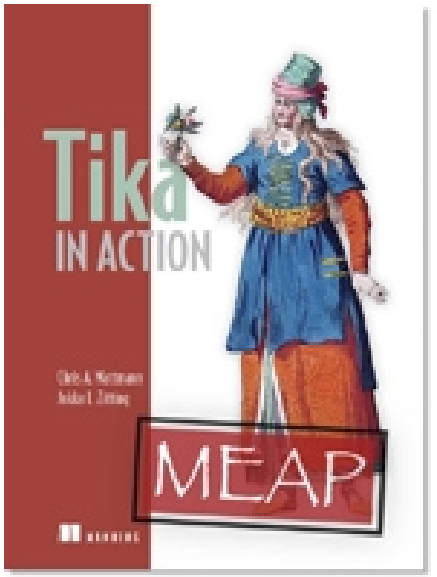
- Páginas estáticas
  - HTML → [90-95%]
    - Resto: PDF y texto plano → [70-85%]
    - Luego, .doc y .ppt
    - Código fuente
    - Archivos comprimidos
- Problema?
  - Parsing (extraer texto y estructura)
  - Identificar idioma. ¿Para qué?


(Parsing)

- Apache TIKA [http://tika.apache.org/]
  - Soporta varios formatos: HTML, XML, Office, OpenDocument, iWorks, PDF, RTF, Texto, Comprimidos, Audio, Imagen, Video, Java, Mail, Autocad, y mas...

- Usos:
  - Motores de búsqueda
  - Machine learning
  - Análisis estadístico
  - Otros (texto)



Finalizando...

- **“Characterization of National Web Domains.”** Ricardo Baeza-yates, Carlos Castillo, Efthimis N. Efthimiadis. ACM Transactions on Internet Technology. 2006.
- **“Characterization of the Argentinian Web.”** Gabriel Tolosa, Fernando Bordignon, Ricardo Baeza-Yates, Carlos Castillo. Cybermetrics 11(1), 2007.

- **Estudios sobre contenido, enlaces y tecnologías en:**
  - Africa, Austria, Brasil, Chile, Grecia, Indochina, Italia, Portugal, Corea del Sur, España, Tailanda, Reino Unido
  - Y **Argentina!**