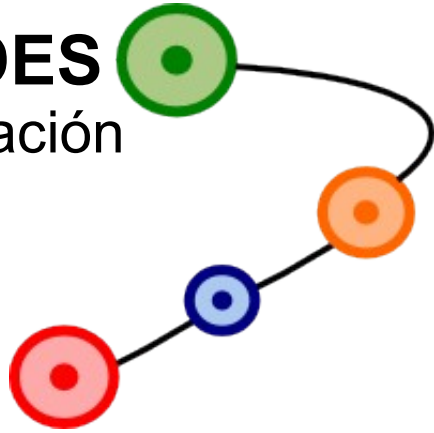
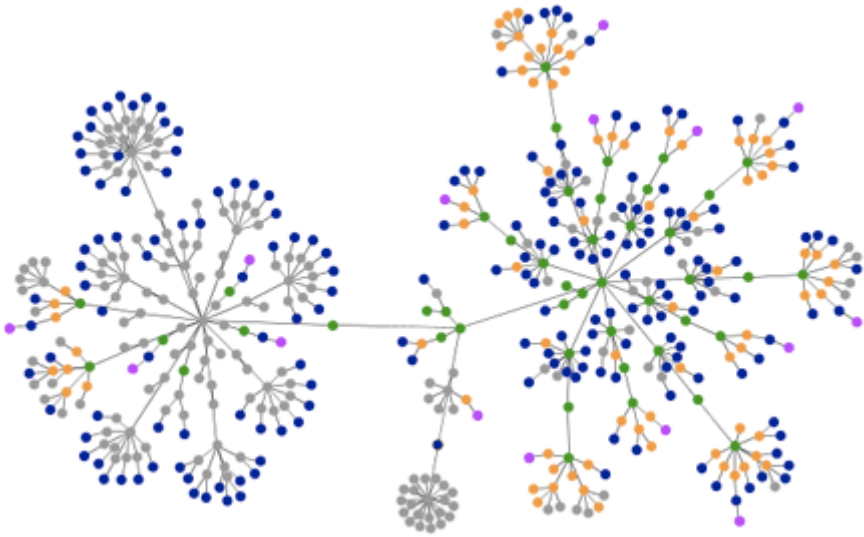
Laboratorio de REDES
Recuperación de Información
y Estudios de la Web



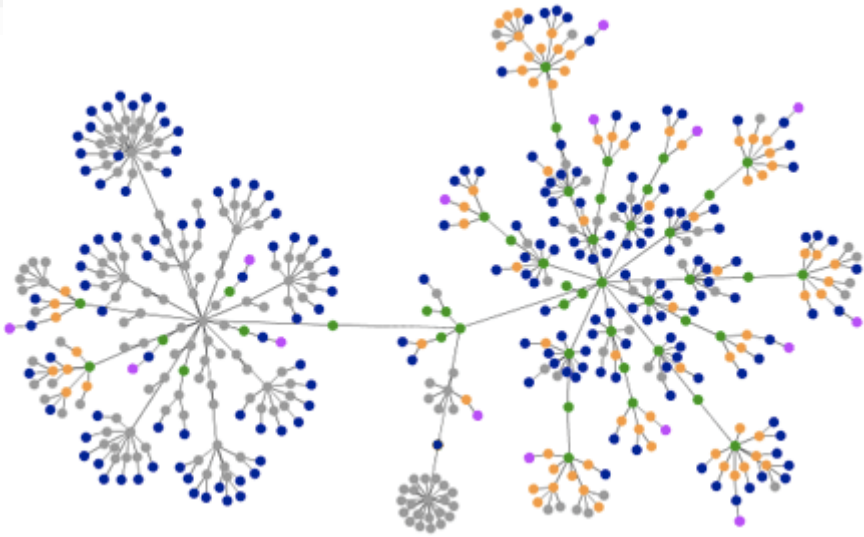
Recuperación de Información en la Web y Motores de Búsqueda

Dr.Gabriel H. Tolosa
tolosoft@unlu.edu.ar



Estructura y Características de la Web

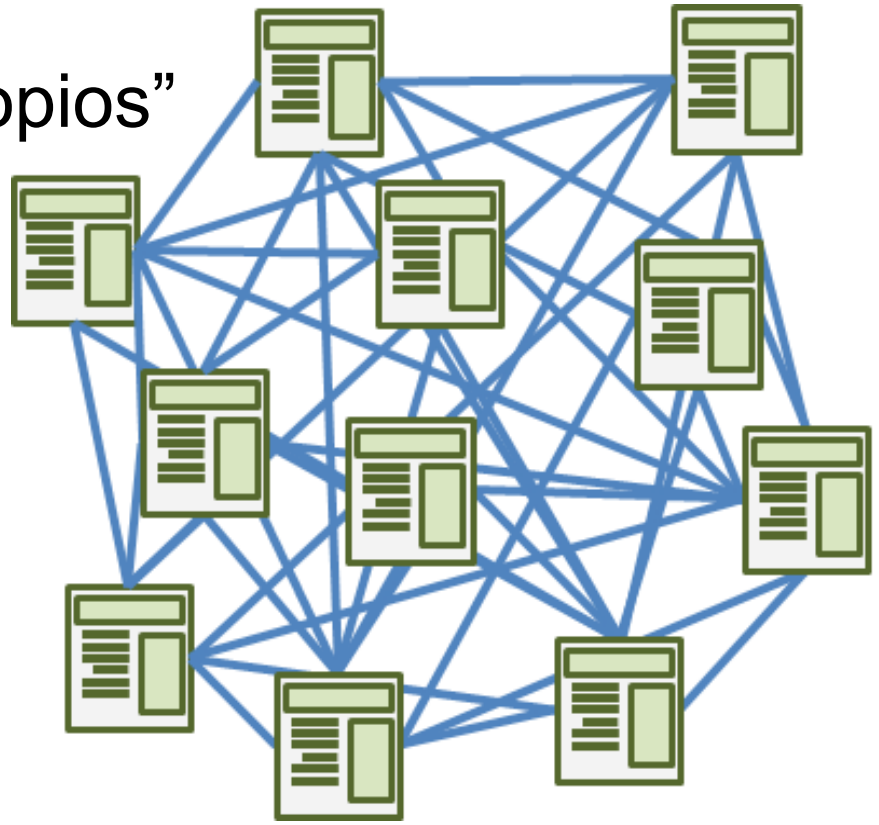
WWW



- 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!!!

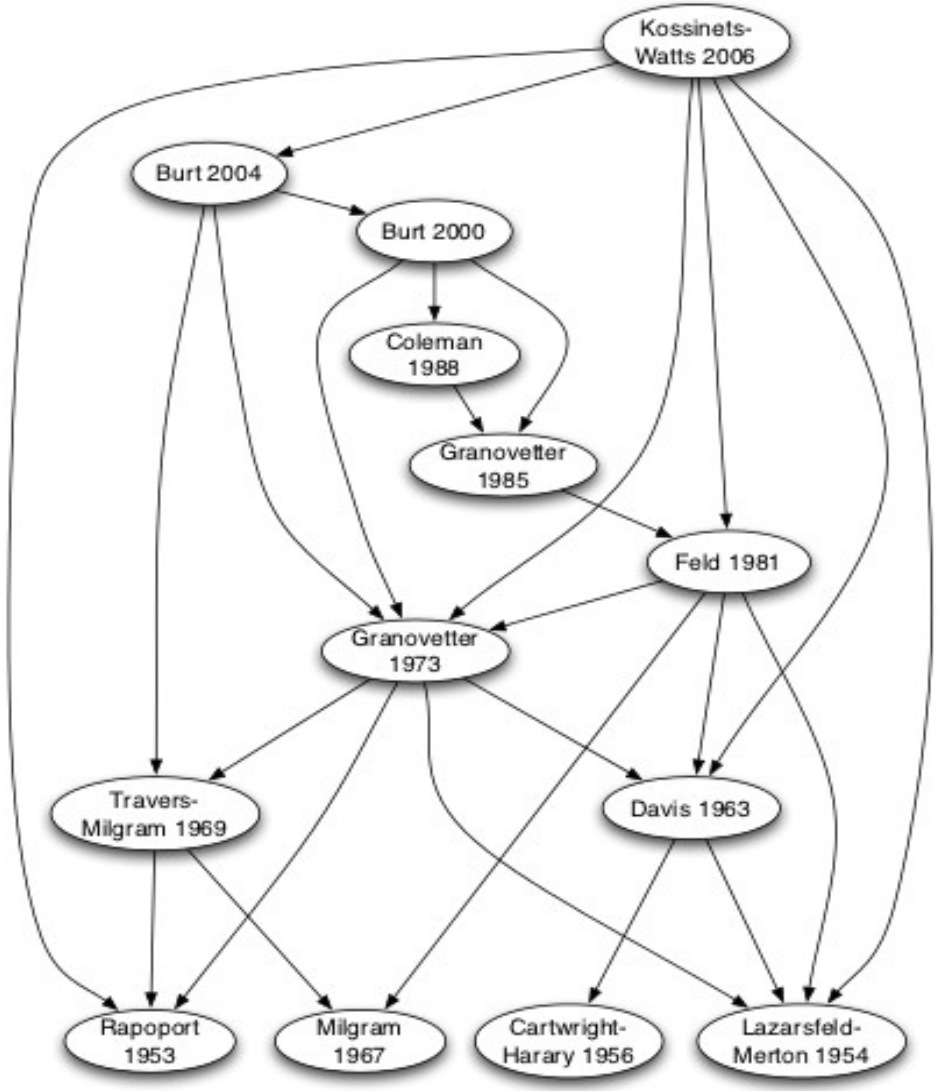
Qué es? (a los efectos de RI)

- 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)

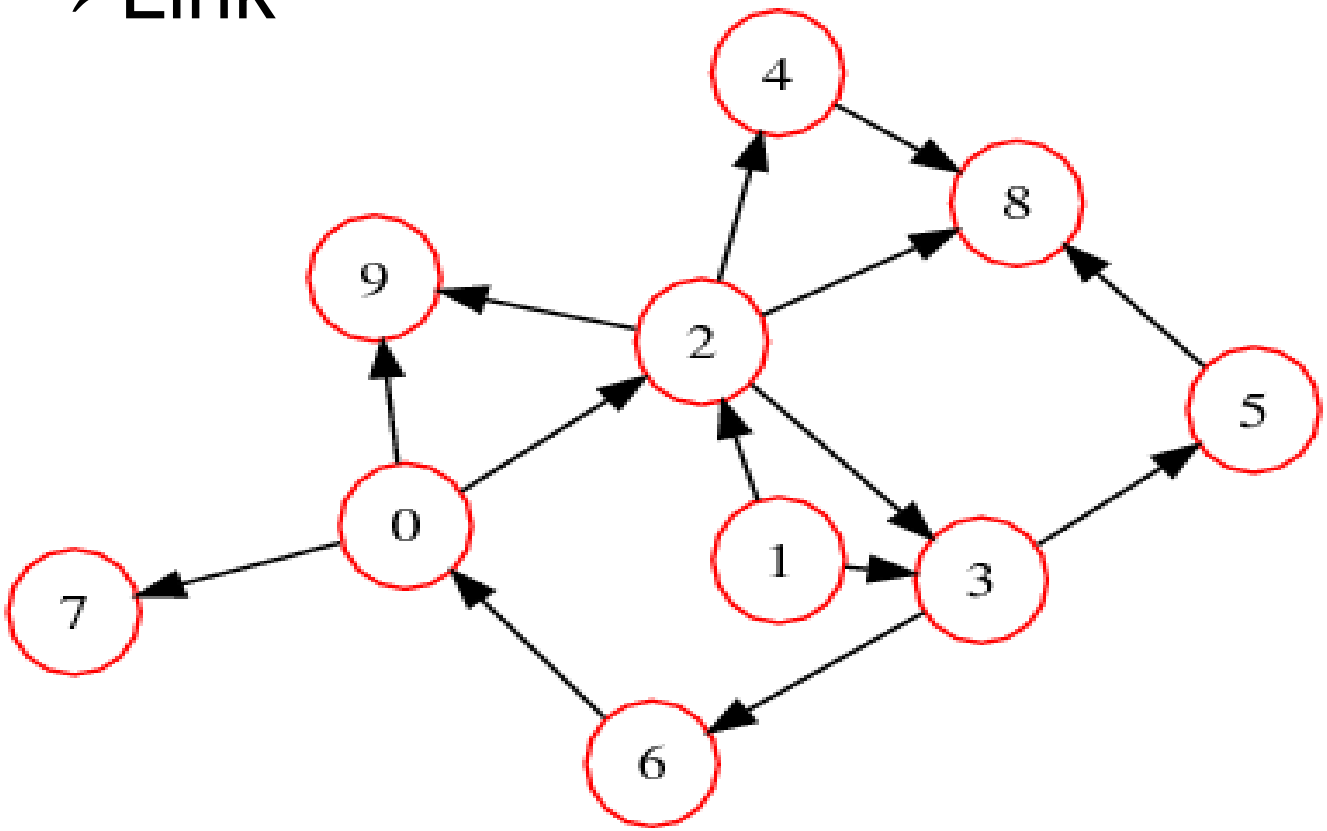
...



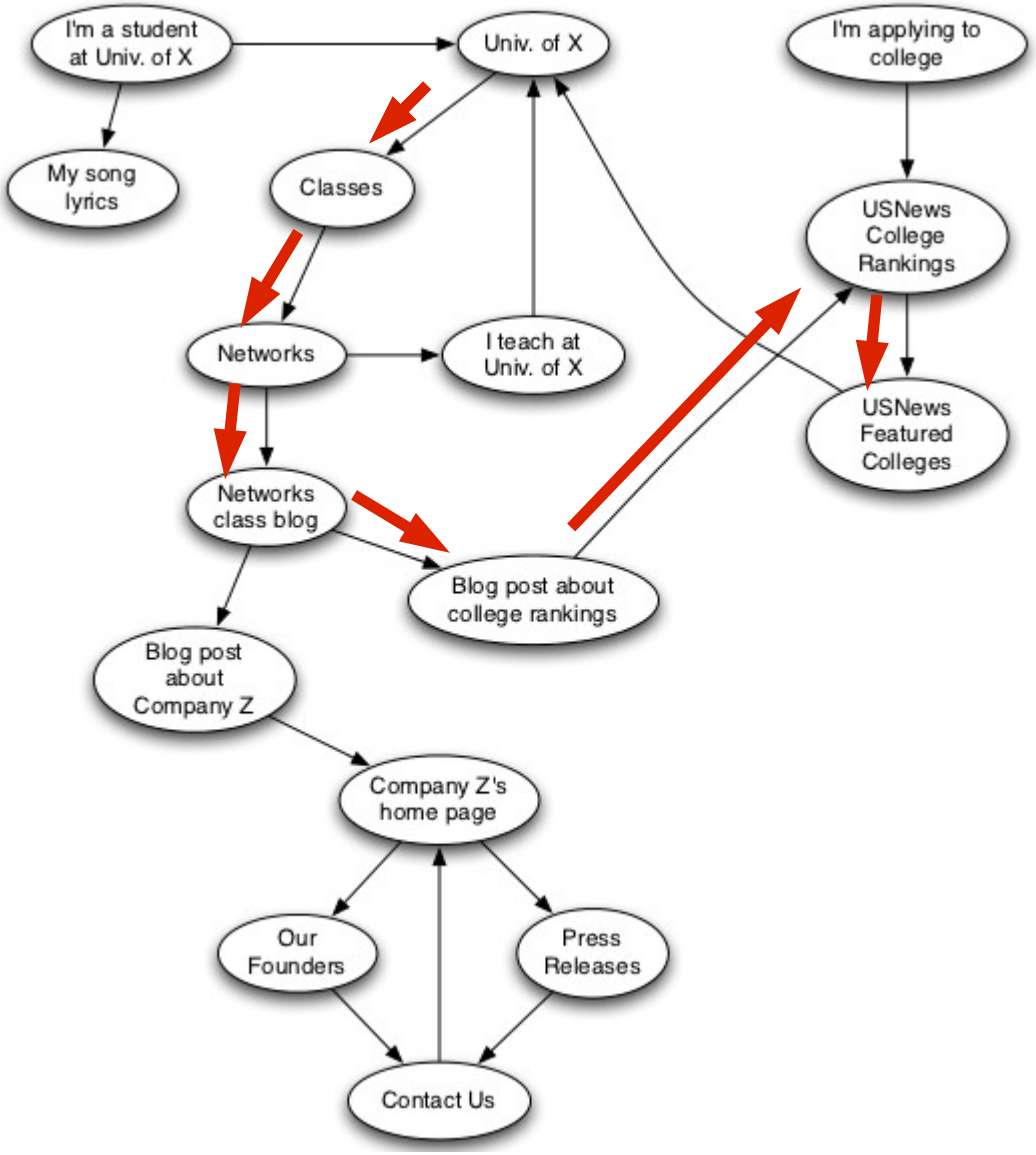
Networks, Crowds, and Markets: Reasoning About a Highly Connected World. D. Easley y J.Kleinberg http://www.cs.cornell.edu/home/kleinber/networks-book/

Estructura de grafo

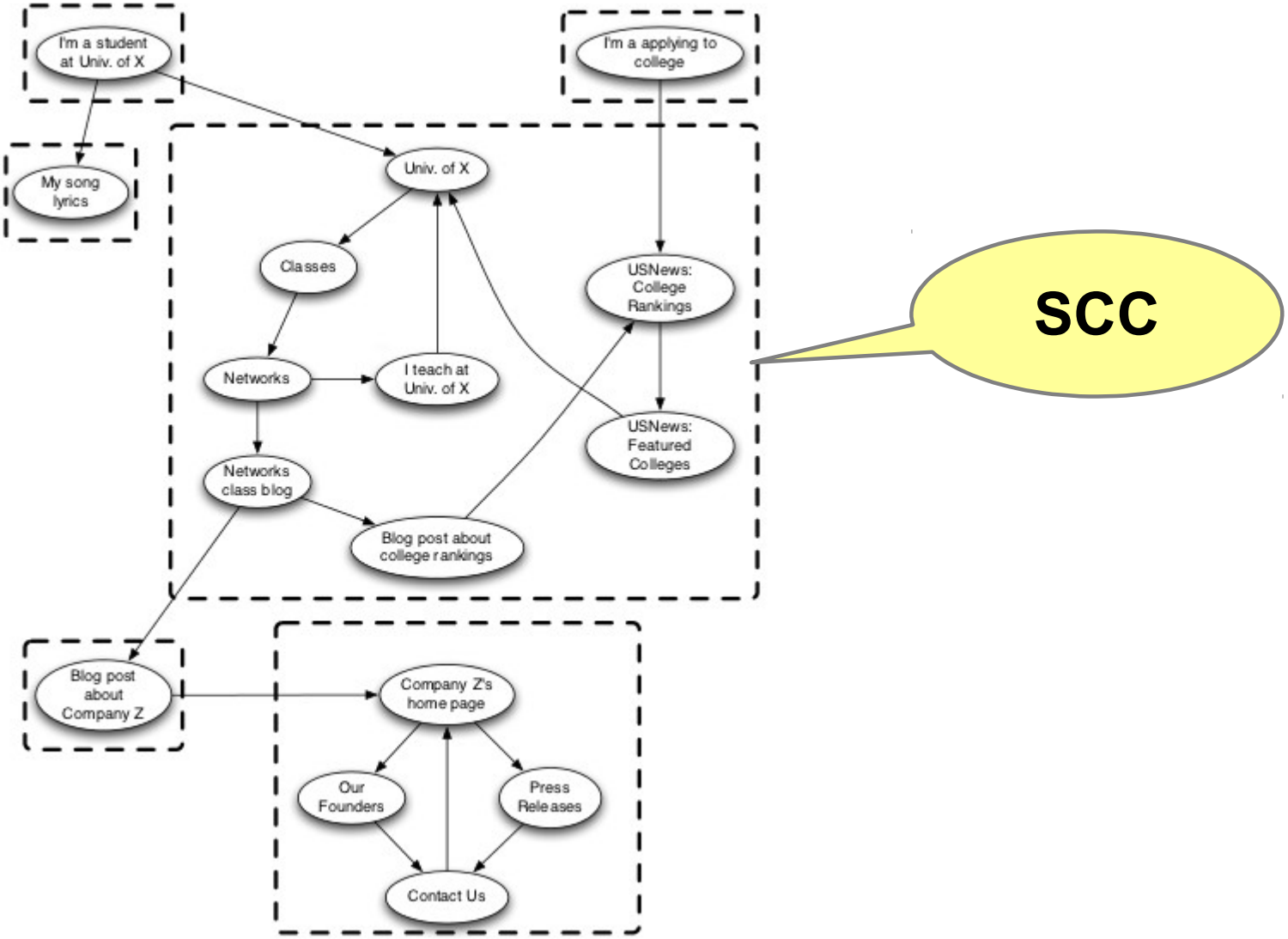
- Nodo → Página web
- Arco → Link



Estructura de grafo

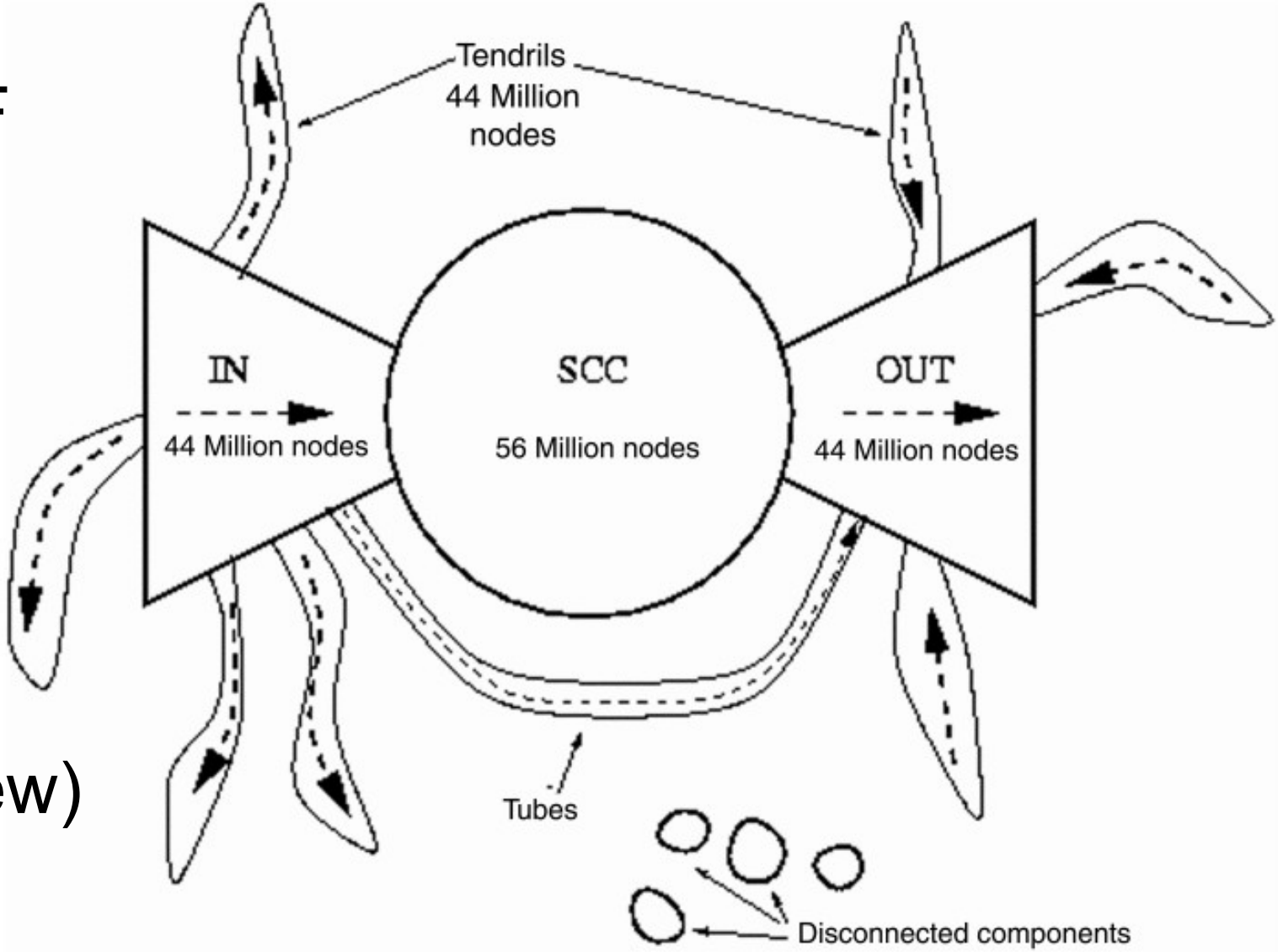


Estructura de grafo



Estructura de grafo

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

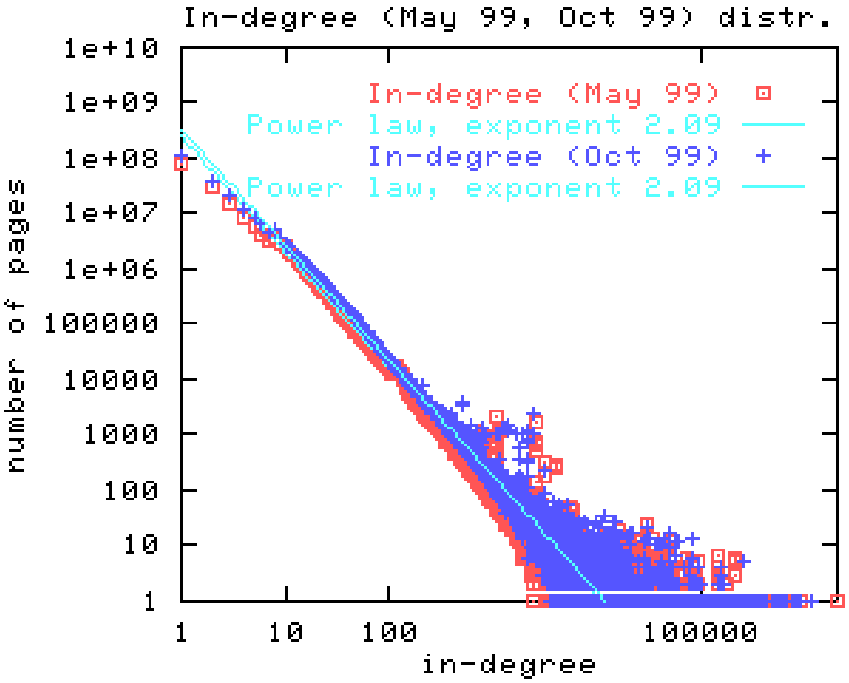


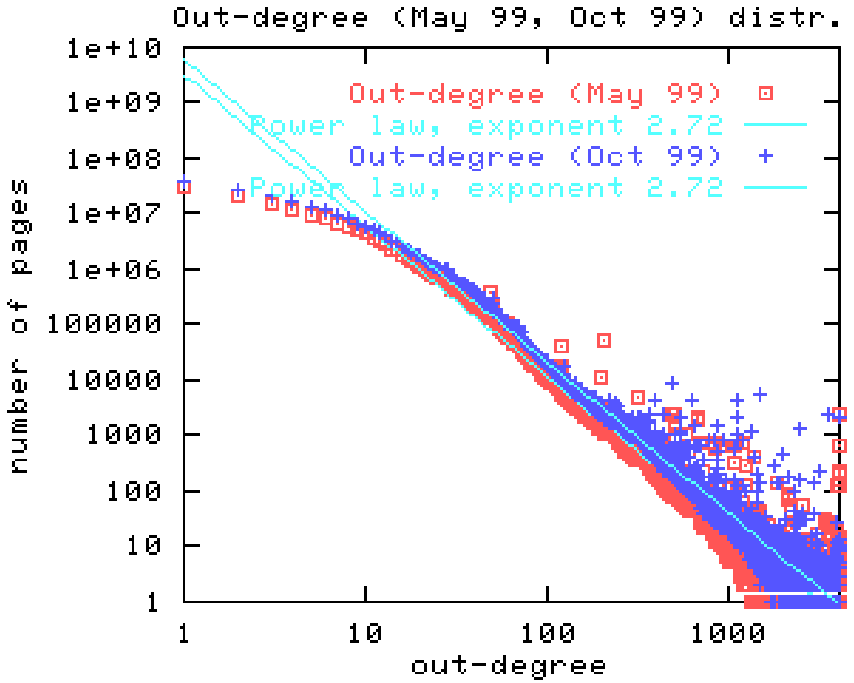
"*Graph Structure on the Web*" [Broder et al., 1999]

Estructura de grafo

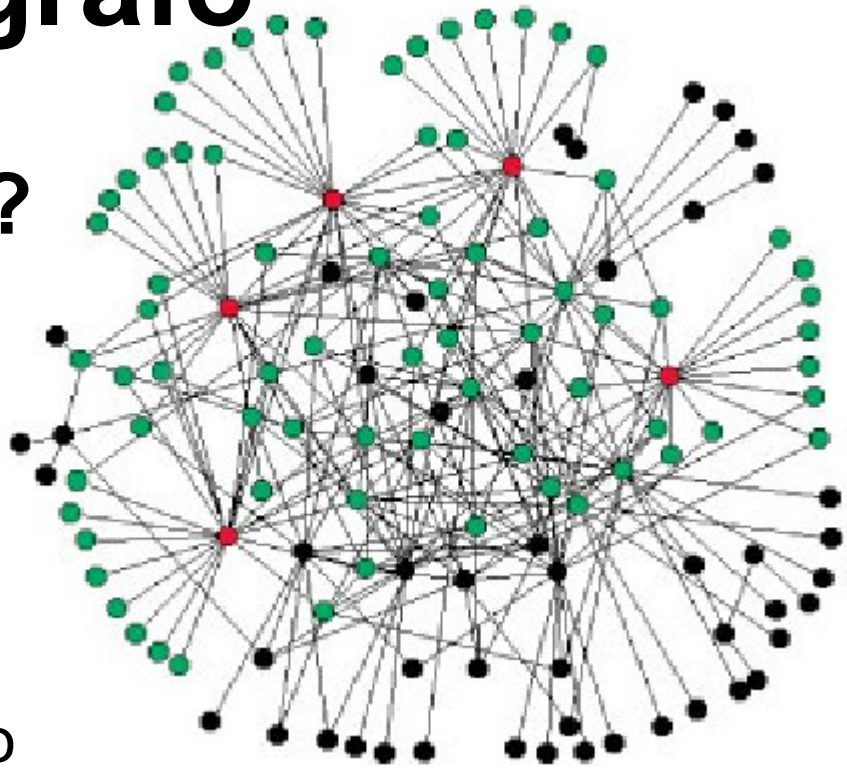
- Grado entrante/saliente → Distribuciones: Power-Law

indegree : $\frac{1}{n^{2.1}}$ outdegree : $\frac{1}{n^{2.72}}$





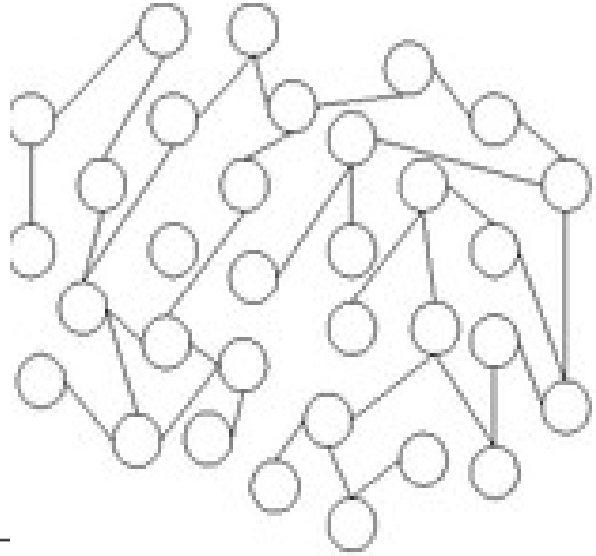
Estructura de grafo

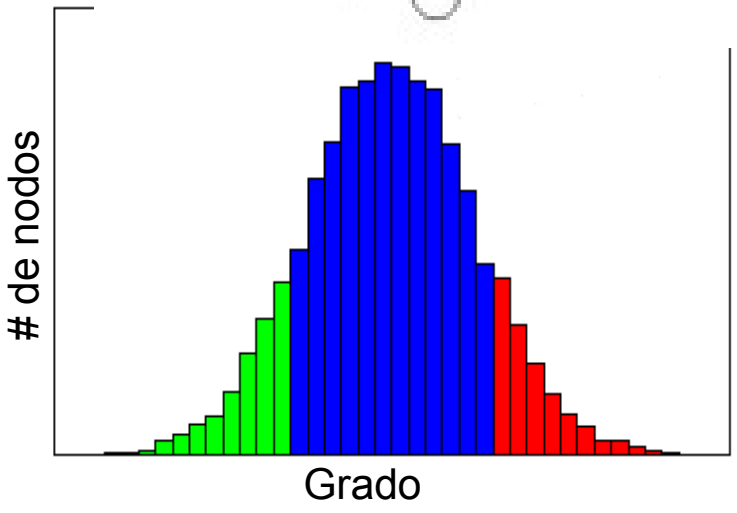


- **Por qué una “Power-Law”?**
- Efecto: **Richer-Get-Richer**
  - Un nuevo nodo se une a la red
  - Establece links con L de los existentes
    - El nodo ***X*** se conecta a un nodo ***Y*** con probabilidad proporcional al grado de ***Y***.
  - Entonces, los nodos con más enlaces tienden a “atraer”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

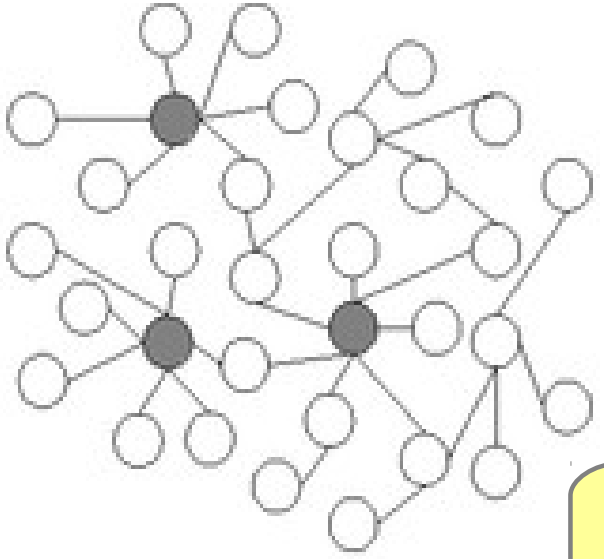
Estructura de grafo

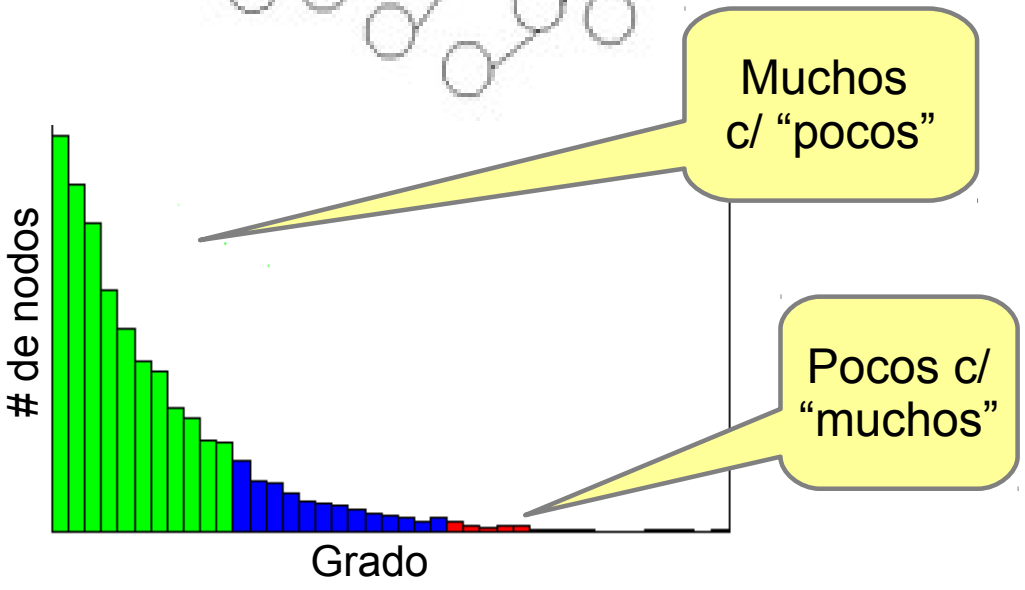
Random





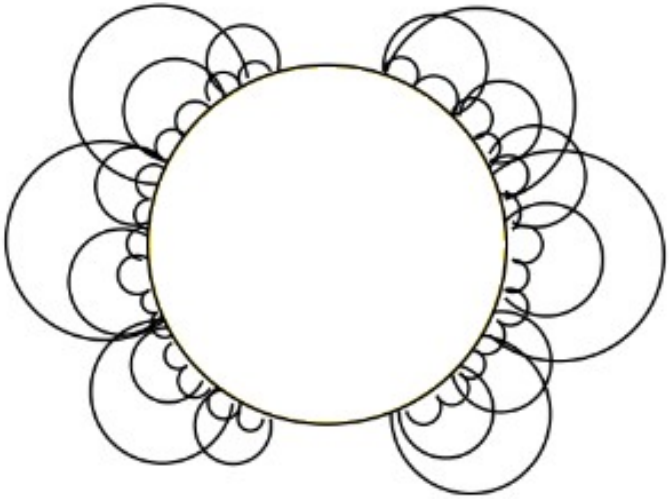
Scale-free





Estructura de grafo

Estructura de grafo

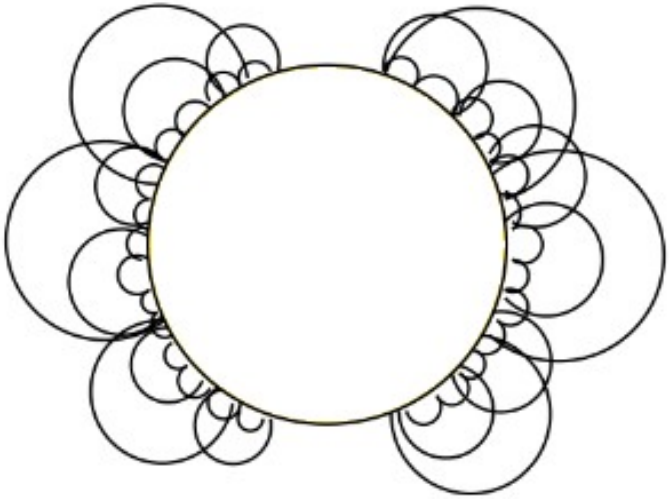


D. Donato, S. Leonardi, S. Millozzi, and P. Tsaparas,

“Mining the inner structure of the web graph”

In Eigth international workshop on the Weband databases WebDB, June 2005

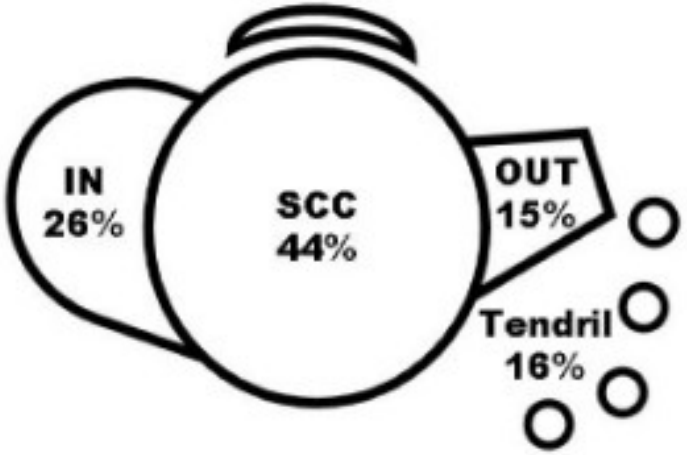
Estructura de grafo



D. Donato, S. Leonardi, S. Millozzi, and P. Tsaparas,

“Mining the inner structure of the web graph”

In Eigth international workshop on the Weband databases WebDB, June 2005

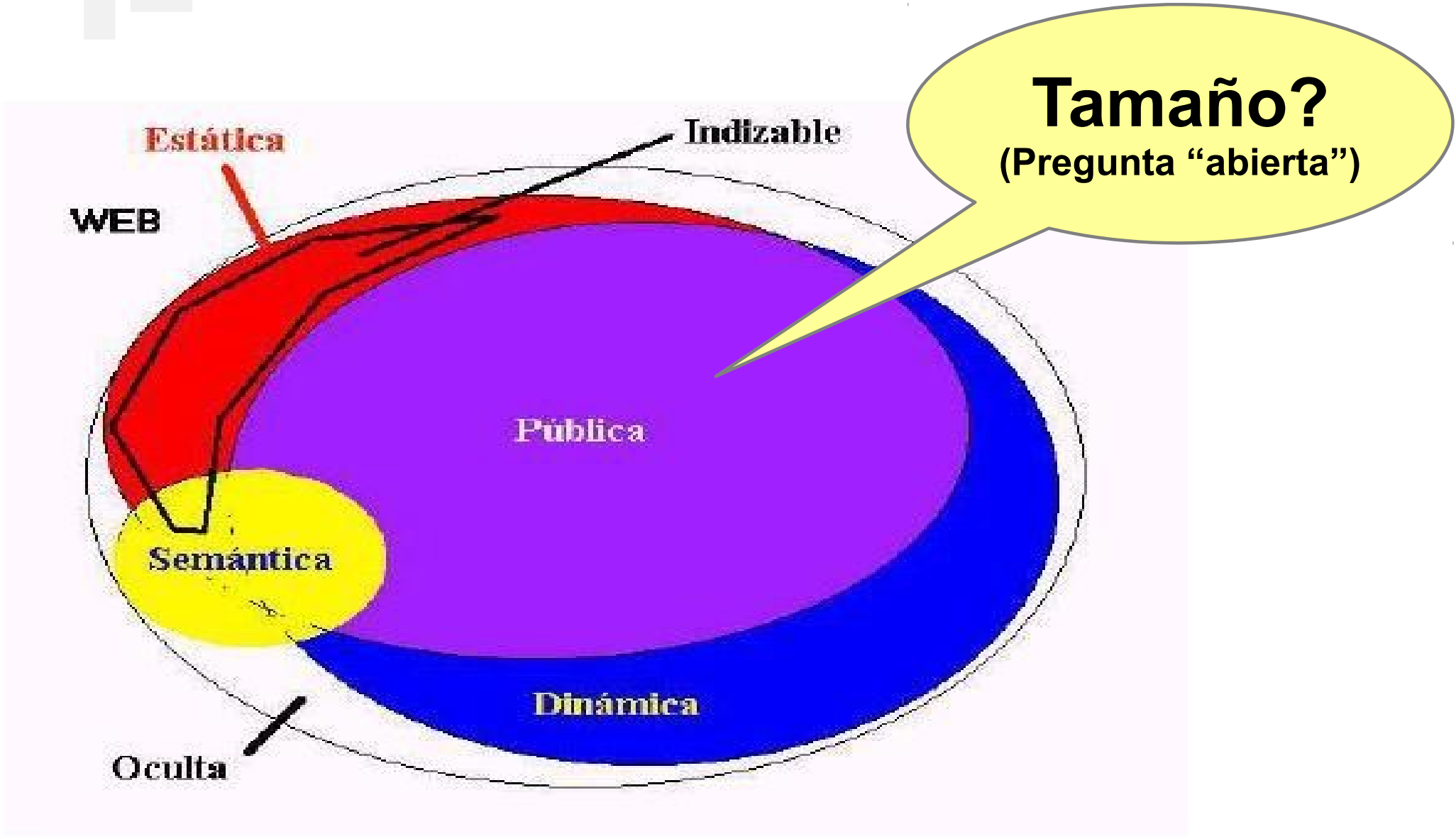


J. J. H. Zhu, T. Meng, Z. Xie, G. Li, and X. Li,

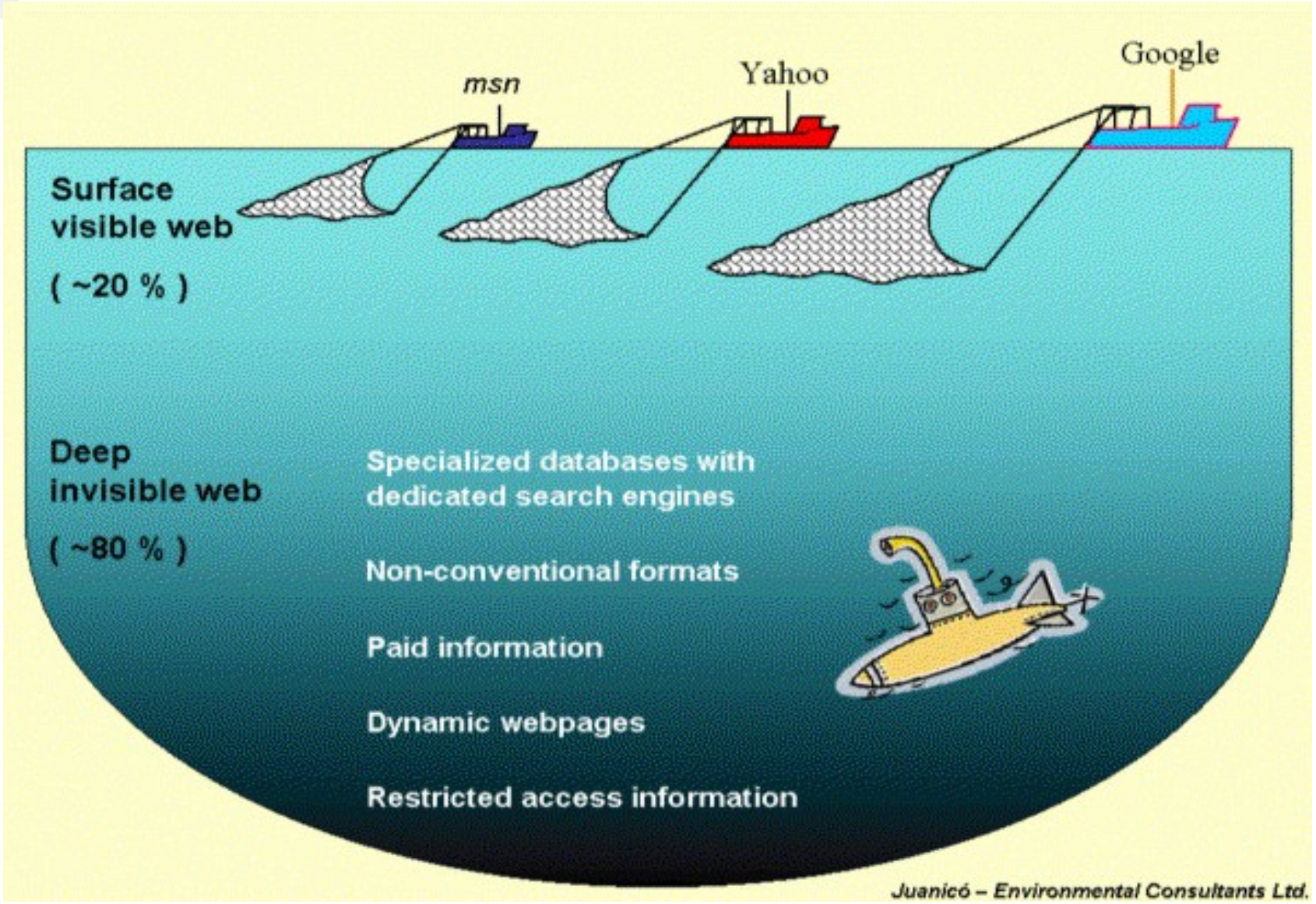
“A teapot graph andits hierarchical structure of the chinese web.”

In WWW.ACM, 2008,pp. 1133–1134

Otra vista [Baeza-Yates, 2003]



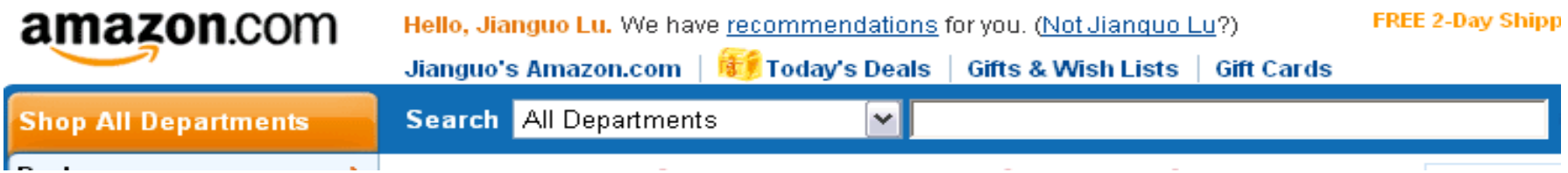
Web “profunda”



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

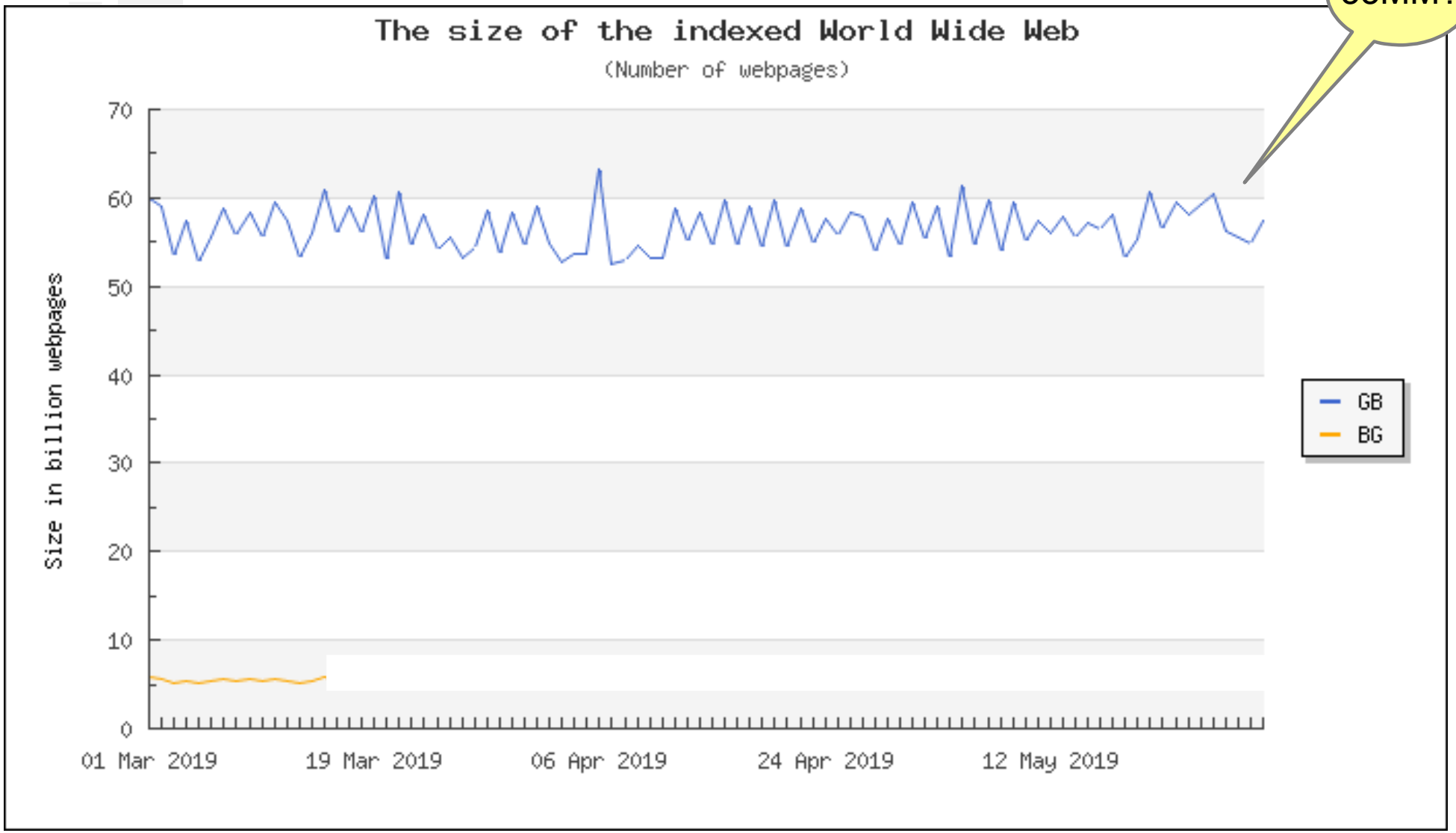
Madhavan, Jayant; David Ko, Łucja Kot, Vignesh Ganapathy, Alex Rasmussen, Alon Halevy. **Google's Deep-Web Crawl**. VLDB, 2008.



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



http://www.worldwidewebsize.com/

Todo crece!!!

3,915,201,961
Internet Users in the world

1,875,258,768
Total number of Websites

156,497,474,320
Emails sent today

2018

3,743,427,816
Google searches today

3,527,790
Blog posts written today

449,448,548
Tweets sent today

+8%

http://www.internetlivestats.com/

Todo crece!!!

3,915,201,961
Internet Users in the world

1,875,258,768
Total number of Websites

156,497,474,320
Emails sent today

2018

3,743,427,816
Google searches today

3,527,790
Blog posts written today

449,448,548
Tweets sent today

2019

+8%

4,240,847,746
Internet Users in the world

1,690,111,028
Total number of Websites

156,175,656,829
Emails sent today

4,007,970,685
Google searches today

3,808,355
Blog posts written today

456,013,936
Tweets sent today

http://www.internetlivestats.com/

Todo crece!!!

3,915,201,961
Internet Users in the world

1,875,258,768
Total number of Websites

156,497,474,320
Emails sent today

2018

3,743,427,816
Google searches today

3,527,790
Blog posts written today

449,448,548
Tweets sent today

2019

+8%

4,240,847,746
Internet Users in the world

-10%

1,690,111,028
Total number of Websites

-1%

156,175,656,829
Emails sent today

+7%

4,007,970,685
Google searches today

+7%

3,808,355
Blog posts written today

+1%

456,013,936
Tweets sent today

http://www.internetlivestats.com/

Todo crece!!!

3,915,201,961
Internet Users in the world

1,875,258,768
Total number of Websites

156,497,474,320
Emails sent today

2018

3,743,427,816
Google searches today

3,527,790
Blog posts written today

449,448,548
Tweets sent today

2019

4,240,847,746
Internet Users in the world
+8%

1,690,111,028
Total number of Websites
-10%

156,175,656,829
Emails sent today
-1%

4,007,970,685
Google searches today
+7%

3,808,355
Blog posts written today
+7%

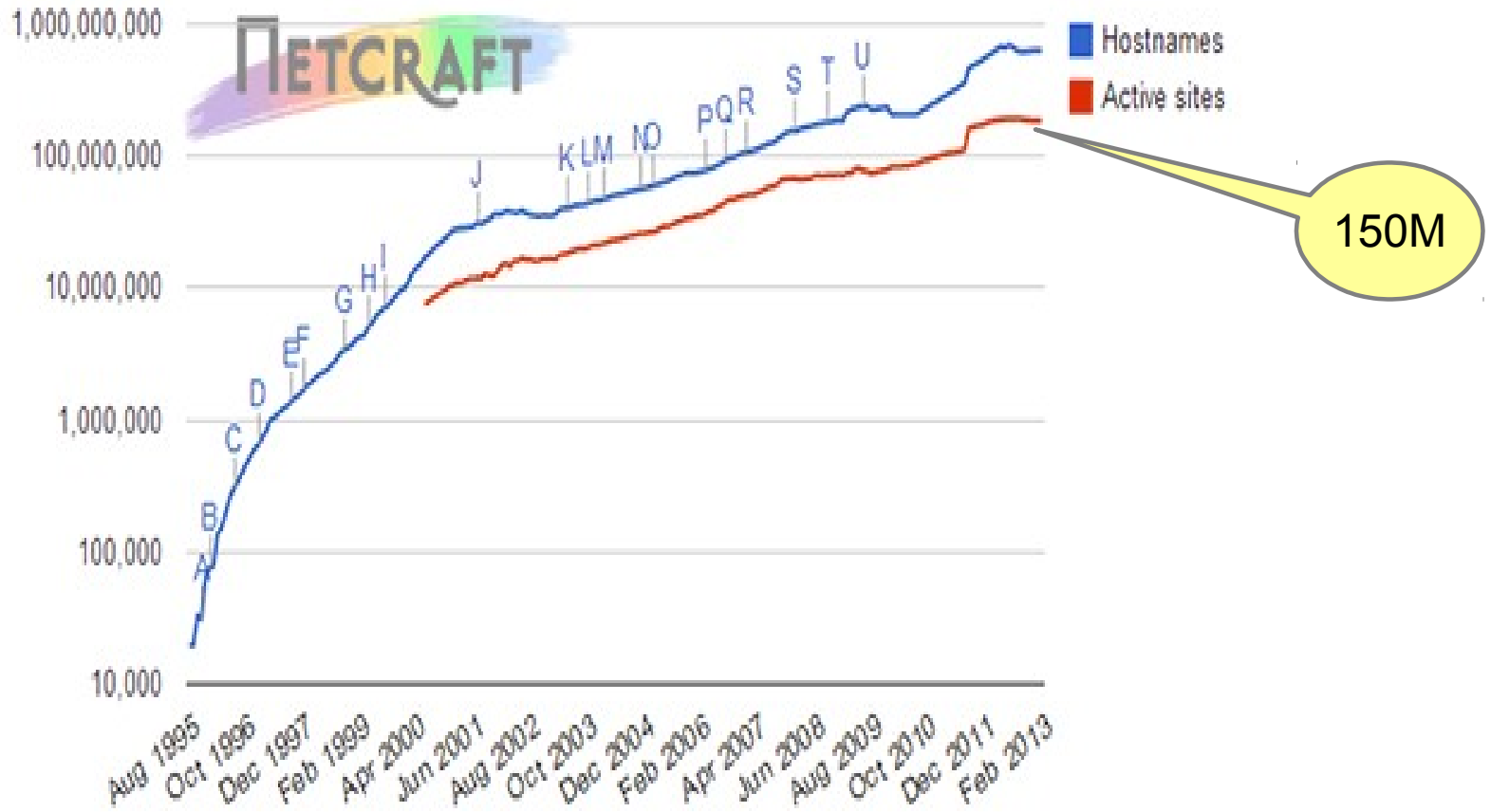
456,013,936
Tweets sent today
+1%

- **Preguntas abiertas**
  - Nodos temporales
  - Dinámica
  - Duplicados
  - Profunda: 95%?

http://www.internetlivestats.com/

Tamaño

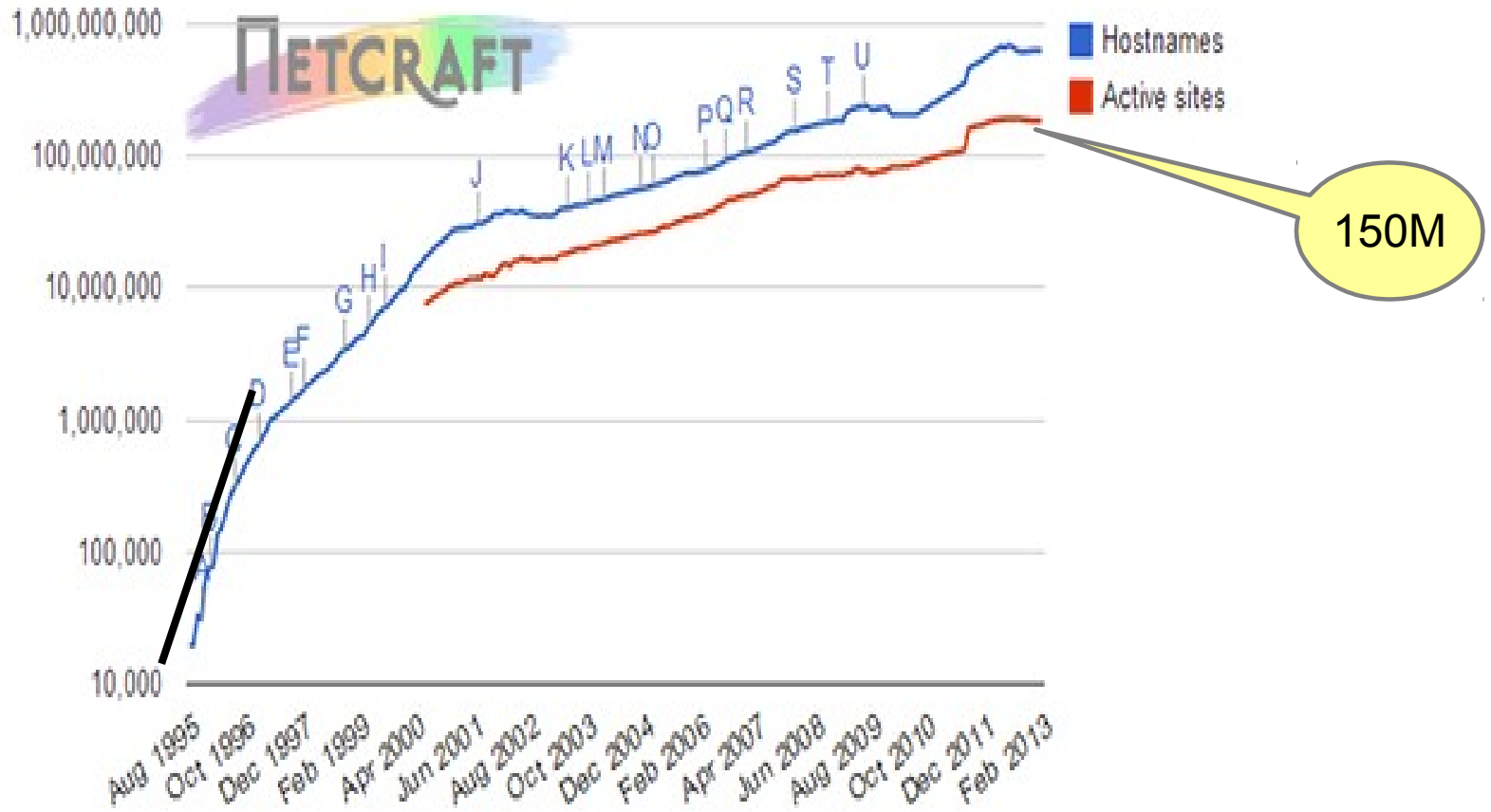
Cantidad de Sitios [1996-2013]



http://www.netcraft.com/

Tamaño

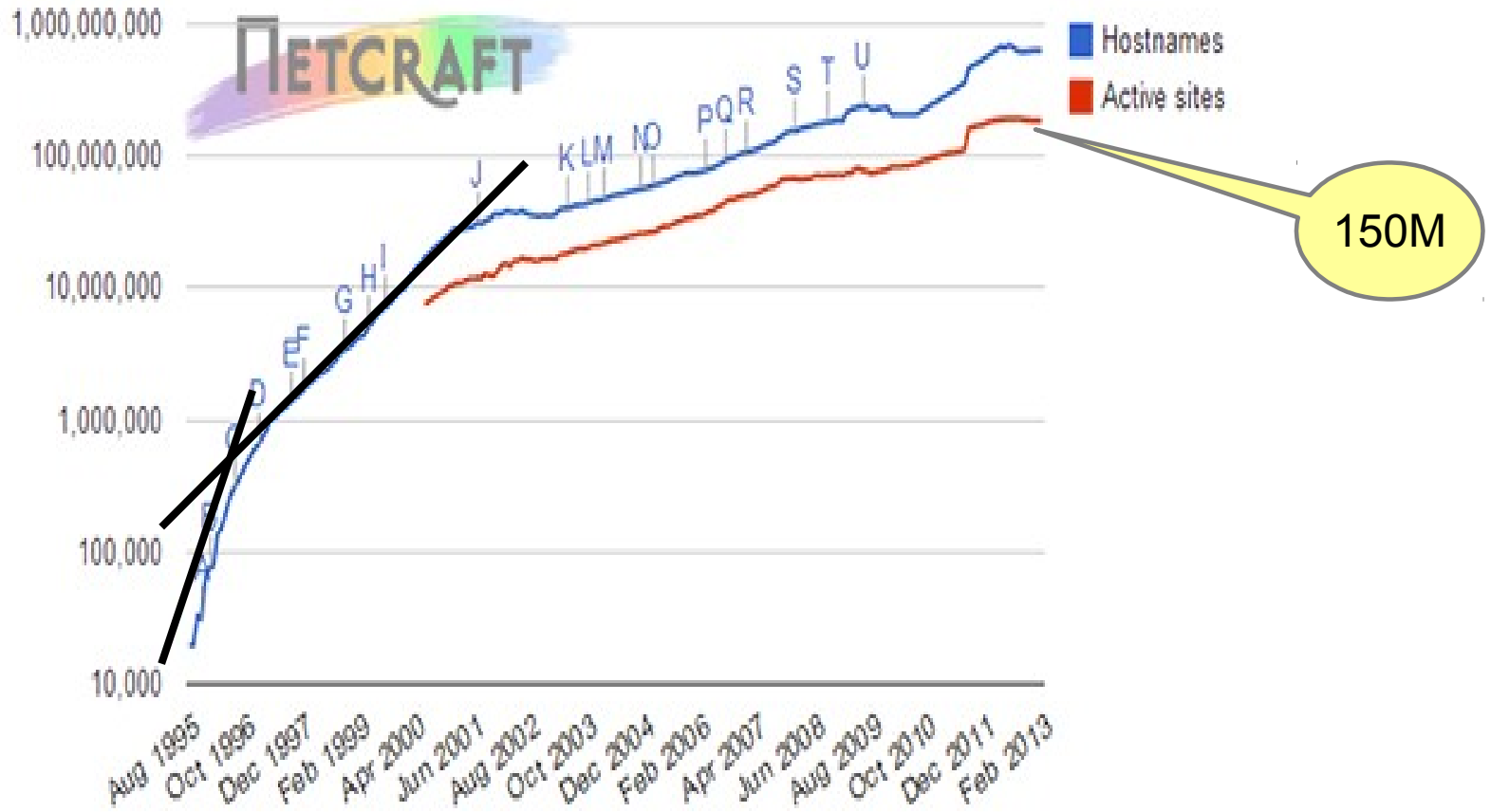
Cantidad de Sitios [1996-2013]



http://www.netcraft.com/

Tamaño

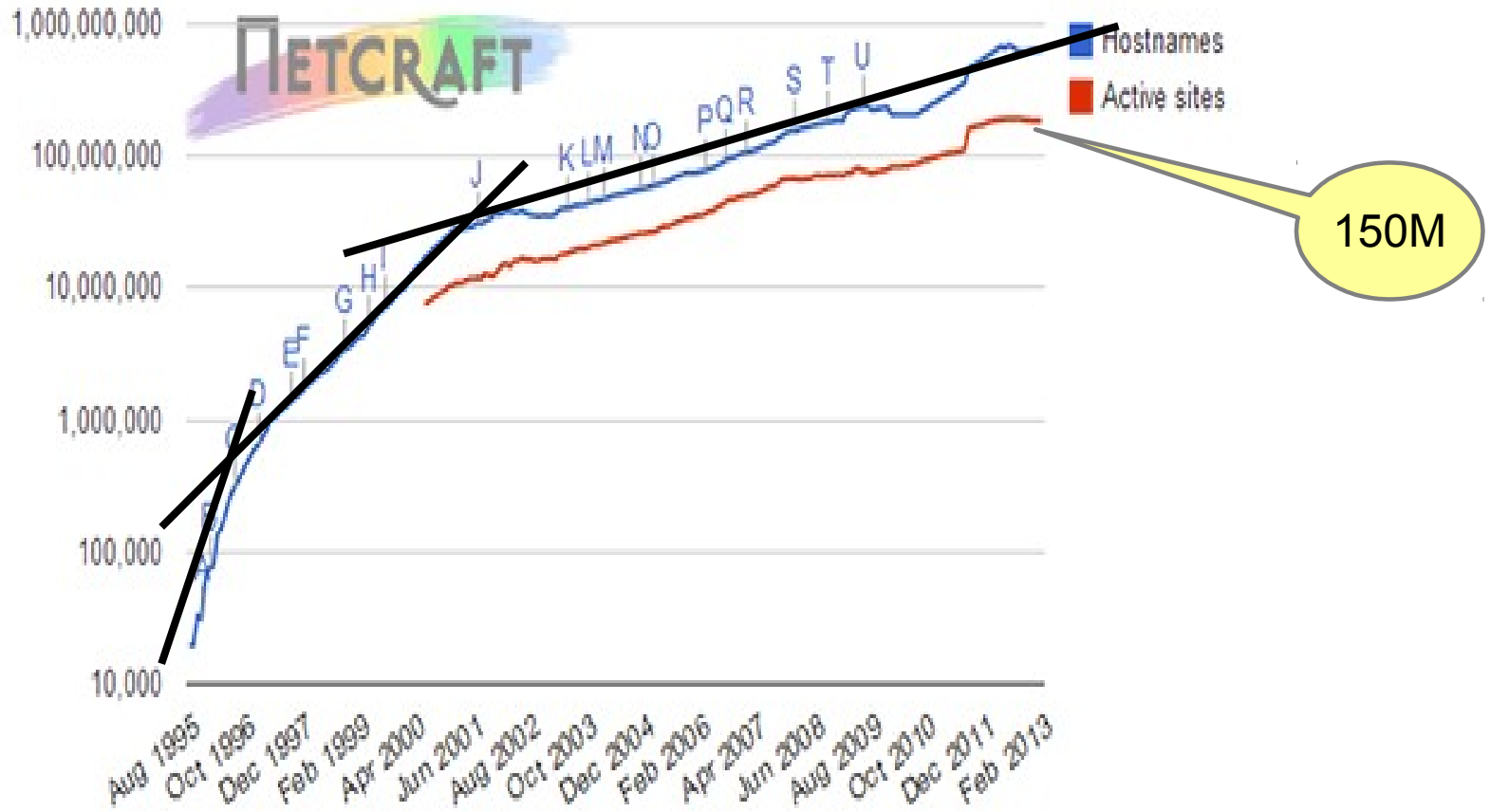
Cantidad de Sitios [1996-2013]



http://www.netcraft.com/

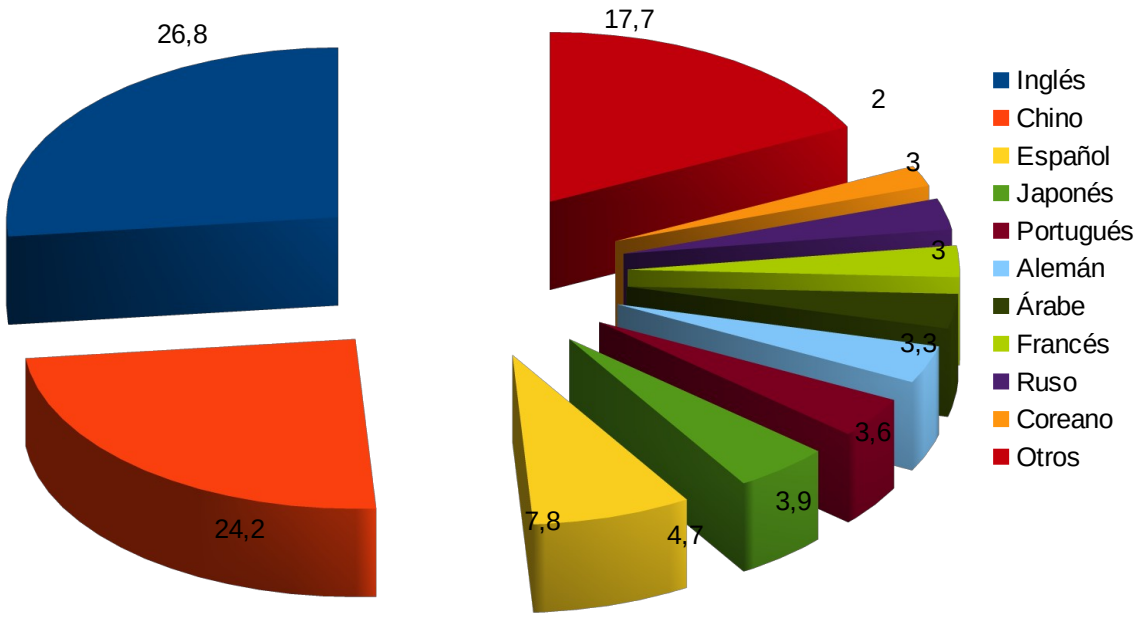
Tamaño

Cantidad de Sitios [1996-2013]



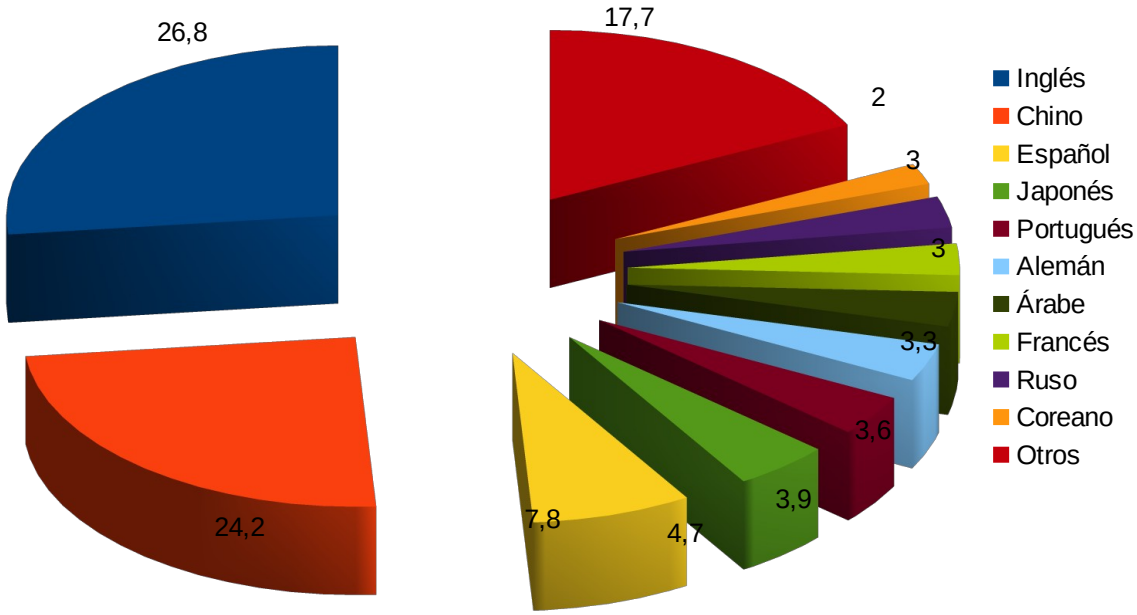
http://www.netcraft.com/

Heterogeneidad: Idiomas



Heterogeneidad: Idiomas

Idiomas en la Web

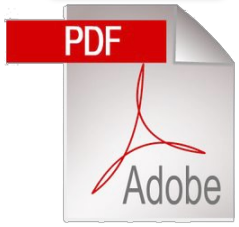


Rank	Language	Internet users	Percentage
1	English	1,105,919,154	25.2%
2	Chinese	863,230,794	19.3%
3	Spanish	344,448,932	7.9%
4	Arabic	226,595,470	5.2%
5	Portuguese	171,583,004	3.9%
6	Indonesian / Malaysian	169,685,798	3.9%
7	French	144,695,288	3.3%
8	Japanese	118,626,672	2.7%
9	Russian	109,552,842	2.5%
10	German	92,304,792	2.1%
1-10	Top 10 languages	3,346,642,747	76.3%
-	Others	1,039,842,794	23.7%
Total		4,386,485,541	100%

https://en.wikipedia.org/wiki/Langu

Heterogeneidad

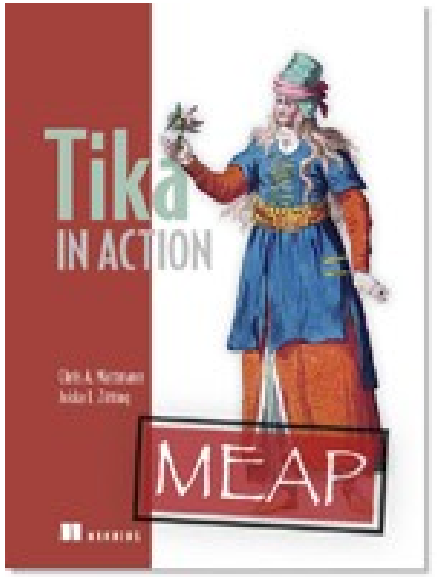
- Páginas estáticas
  - HTML → [70-80%] (estáticas y dinámicas)
    - Resto: PDF y texto plano → [70-85%]
    - Luego, .doc y .ppt
    - Código fuente
    - Archivos comprimidos
- Problema?
  - Parsing (extraer texto y estructura)
  - Identificar idioma. ¿Para qué?

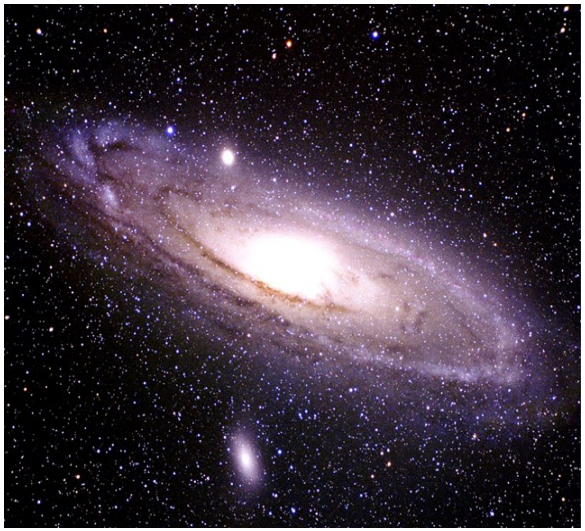

(Parsing)

- 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)



Qué es lo que dificulta la tarea de búsqueda?

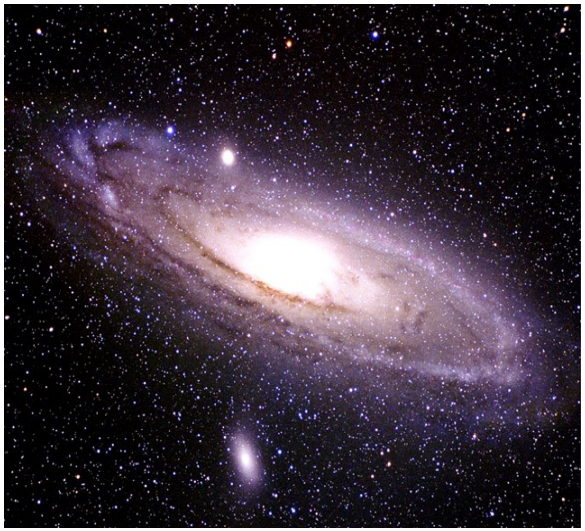


Tamaño

Diversidad

Dinamismo

Qué es lo que dificulta la tarea de búsqueda?



Tamaño **Diversidad** **Dinamismo**

Estas tres características también se observan en los **usuarios!!!!**

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!**