Web Mining

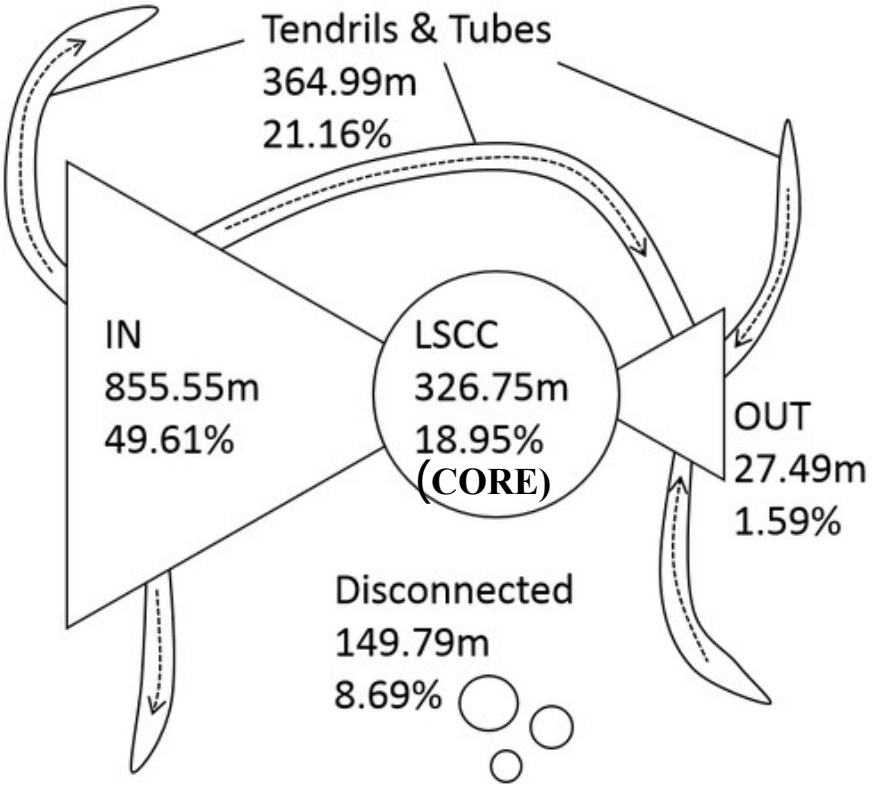
¿Qué es web mining?

- Web Mining es descubrir información y patrones a partir de las fuentes de datos disponibles en la web:
  - Contenido de las páginas (minería de texto, de imágenes)
  - Estructura de links de las páginas (minería de grafos)
  - Comentarios de los usuarios en las redes sociales
  - Estadísticas de uso de las páginas por parte de los usuarios (minería de logs)

- Como el uso de la web es interactivo, muchas veces la información y patrones deben generarse en tiempo real para ser presentados a los usuarios.

La estructura de la web

- Estimados: 4780 millones de páginas (técnicamente infinita, muchas son generadas dinámicamente).
- Se estima que el 25-30% del contenido en la web es duplicado (o casi).



La estructura no ha cambiado, pero si las proporciones:

	Broder (2000)	WDC (2014)
CORE	28%	18,95%
IN	21%	49,61%
OUT	21%	1,59%
TENDRILS	22%	21,16%
ISLANDS	9%	8,69%
Total	203,5 M	1540 M

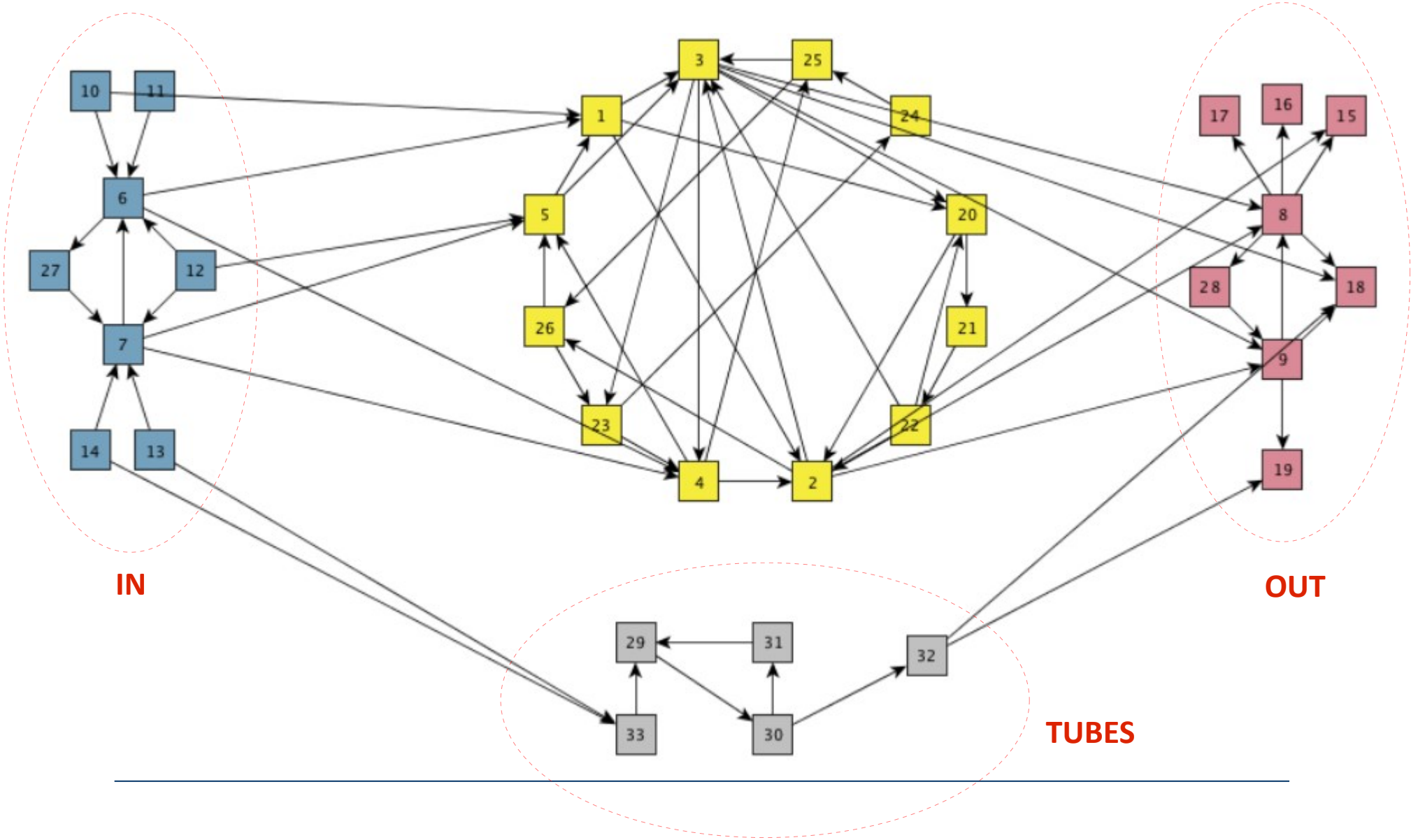
"Topology of the 2014 WDC Hyperlink Graph", WWW 2014
http://webdatacommons.org/hyperlinkgraph/2014-04/topology.html

La estructura de la web

- **CORE:** Un grupo de páginas en donde siempre hay un camino de una página a otra dentro del grupo.
- **IN:** Nodos que forman parte de un camino dirigido que llega a CORE, pero no forman parte del SCC
- **OUT:** Nodos que forman parte de un camino dirigido que parte de CORE, pero no forman parte del SCC
- **TENDRLIS-IN:** Nodos que forman parte de un camino dirigido que parte desde IN, y sin conexión a los otros componentes.
- **TENDRLIS-OUT:** Nodos que forman parte de un camino dirigido que llega a OUT, y sin conexión a los otros componentes.
- **ISLANDS:** Pequeños grupos fuertemente conexos, pero sin conexiones a otros componentes. Las islas replican la estructura de IN, CORE y OUT en pequeña escala.
- **TUBES:** Nodos que forman parte de un camino que va de un nodo en IN a un nodo en OUT, pero sin ningún camino que los conecte a CORE. Si se “partiera” un tubo, sería uno o mas TENDRILS.

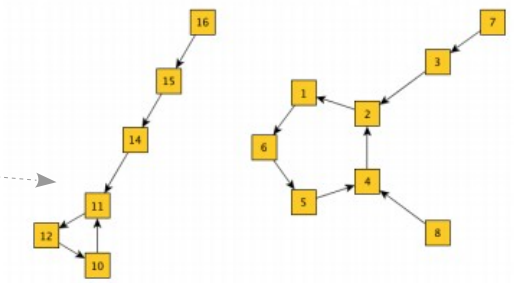
Metaxas, P. Why Is the Shape of the Web a Bowtie?, WWW 2012

Ejemplo de un moño



¿Por qué tiene forma de moño?

La estructura comienza a aparece con 2 links por nodo y “preferential attachment”= La probabilidad de conección es proporcional a lo conectado que esté el destino.



Con 1 link por nodo:

- Al agregar un link a cada nodo, aparecen naturalmente “pseudo árboles” conteniendo ciclos.
- Cada nodo puede formar parte de un componente o de IN de un componente. Con 1 solo link por nodo, OUT o TENDRILs no pueden aparecer: o el link se usa para formar parte del componente, o lleva al componente (es parte de IN). Si apunta "para afuera" no puede formar parte del componente (porque tiene el nodo puede tener 1 solo link).

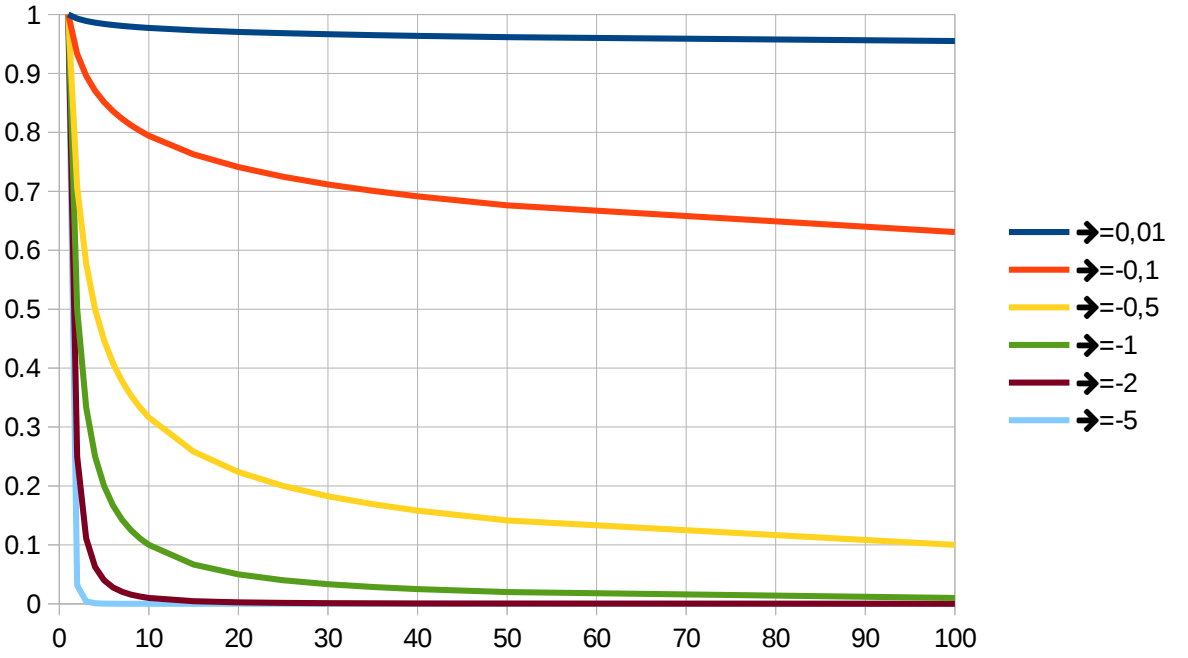
Con 2 links por nodo:

- Ahora es posible crear OUT mediante nodos con un link hacia el CORE y otro hacia afuera.
  - CORE (y otros componentes) crecen de tamaño. Esto ocurre p.ej.:
  - cuando un nodo en CORE agrega un nuevo link a un nodo en IN (arma un ciclo),
  - cuando una pagina en OUT agrega un nuevo link a un nodo en CORE (arma un ciclo),
  - cuando una pagina en OUT agrega un nuevo link a un nodo en IN (arma un ciclo),
- El numero de ISLANDS decrece: Cuando se agreuga un link desde o hacia un nodo entre una ISLAND y IN, SCC o OUT, la isla desaparece. Si el nuevo link.,.
  - viene desde una ISLAND, entonces esta se incorpora a IN
  - viende de CORE o OUT, la ISLAND se agrega a OUT
  - viene de IN entonces la ISLAND se agrega a un TENDRIL.

Fuente:
Why Is the Shape of the Web a Bowtie?, WWW 2012

Ley de Potencias

$y=kx^{\alpha}$, con $\alpha \leq 1$



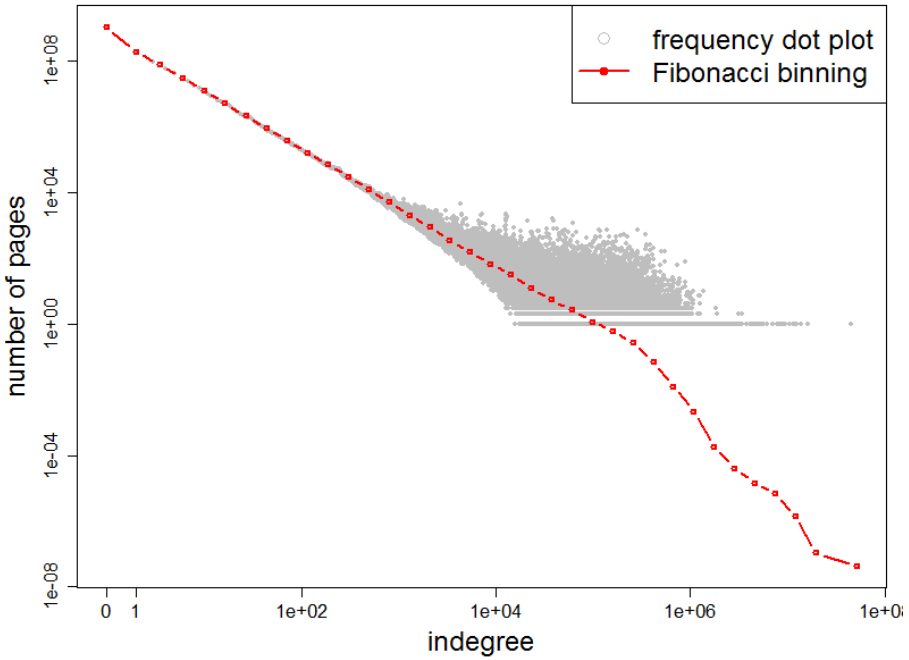
Ejemplos:

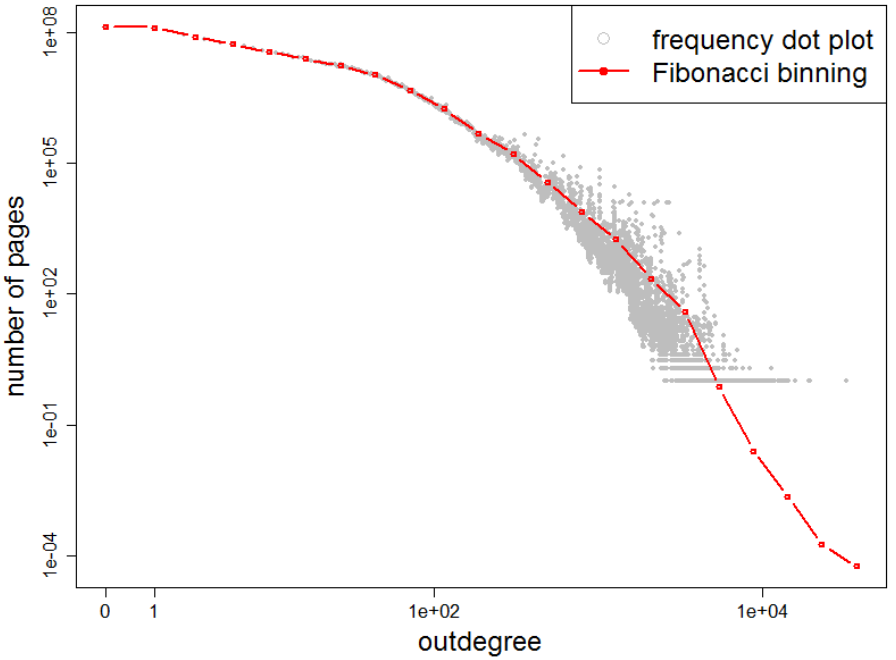
- Frecuencia de palabras (Zipf)
- Numero de citas de papers
- Cantidad de accesos por sitio web

- Una muy larga cola con caída lenta en la cola (el valor promedio no indica nada).
- La mayoría del área debajo de la curva esta en los primero valores (regla 80/20).
- El 50% del área de la curva está en $x_{min}2^{(-\alpha-1)}$

Newman, E., "Power laws, Pareto distributions and Zipf's law"
http://arxiv.org/pdf/cond-mat/0412004.pdf

In versus out-degree (2014)

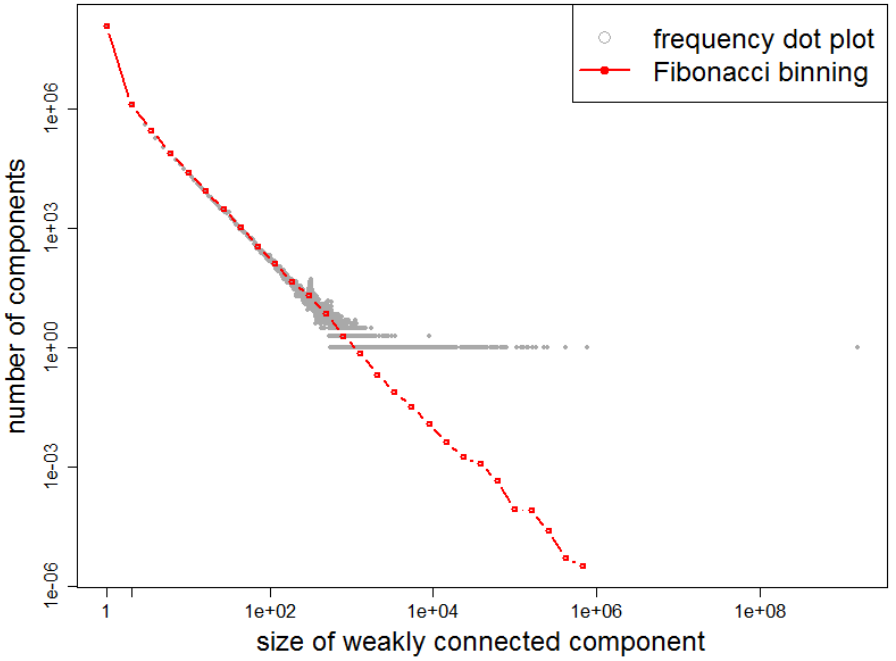


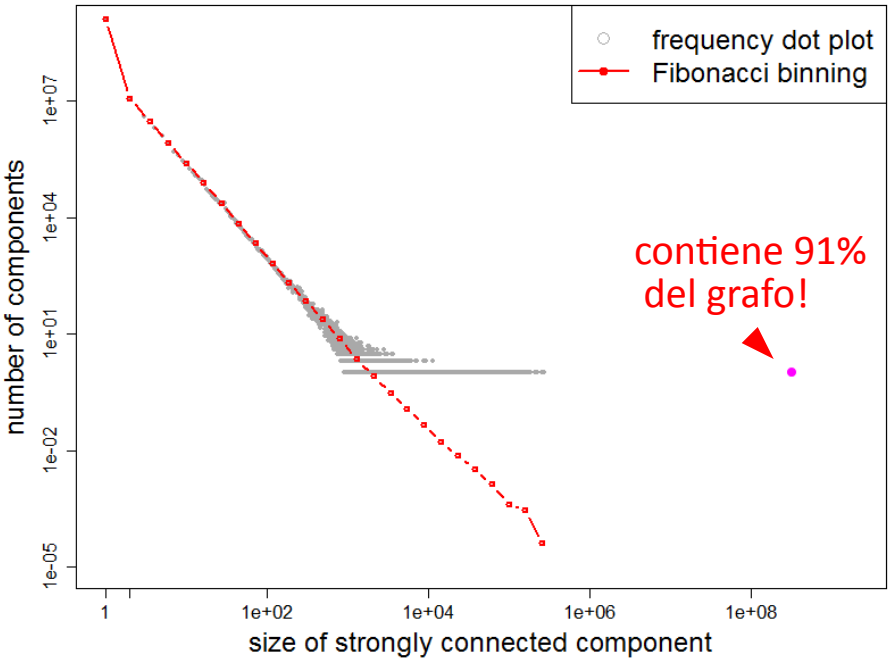


- In degree NO sigue una ley de potencia pero solo dentro de cierto rango.
- La grado promedio ha aumentado significativamente desde el 2000 (~ 5 veces).

Meusel, R. "Graph Structure of the Web", WWW 2014

Conectividad (2014)





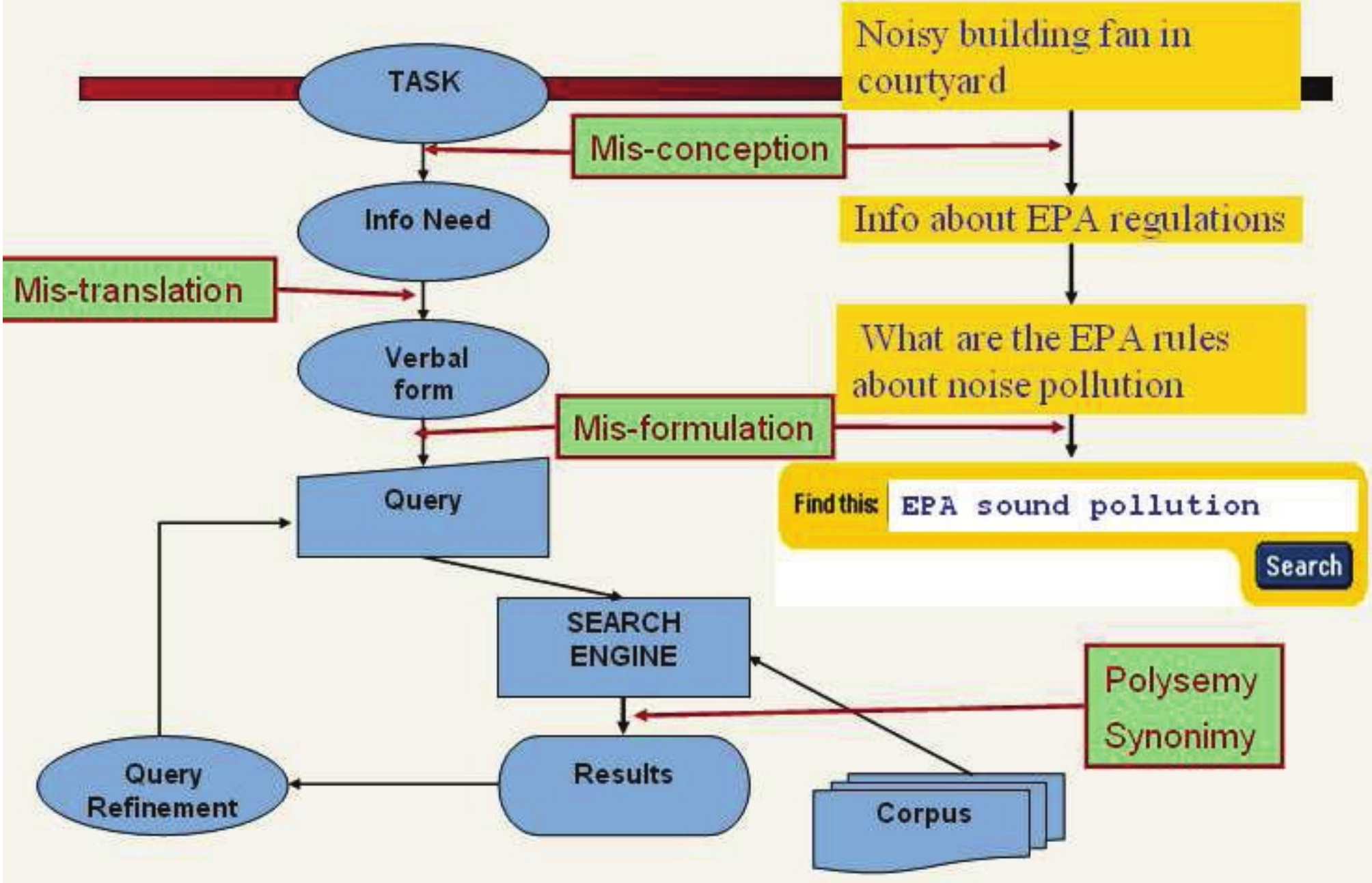
- Un solo componente conexo contiene al 91% del grafo; CORE ha crecido con el tiempo.
- La conectividad entre pares de páginas es el doble que en el 2000.

Meusel, R. "Graph Structure of the Web", WWW 2014

Buscar en la web

- La búsqueda es la única manera en que la web es explotable por personas.
- La publicidad online hace a los buscadores rentables.
  - Es un negocio de volumen; el CTR promedio de Google es 2%
  - Es un problema de optimización: basado en la búsqueda y patrones de visitas a otros sitios, que avisos mostrar en un espacio limitado de manera de maximizar la probabilidad de click del usuario en uno de esos avisos.
  - Es un problema en donde existe la publicidad engañosa, el fraude de clicks...

True example*

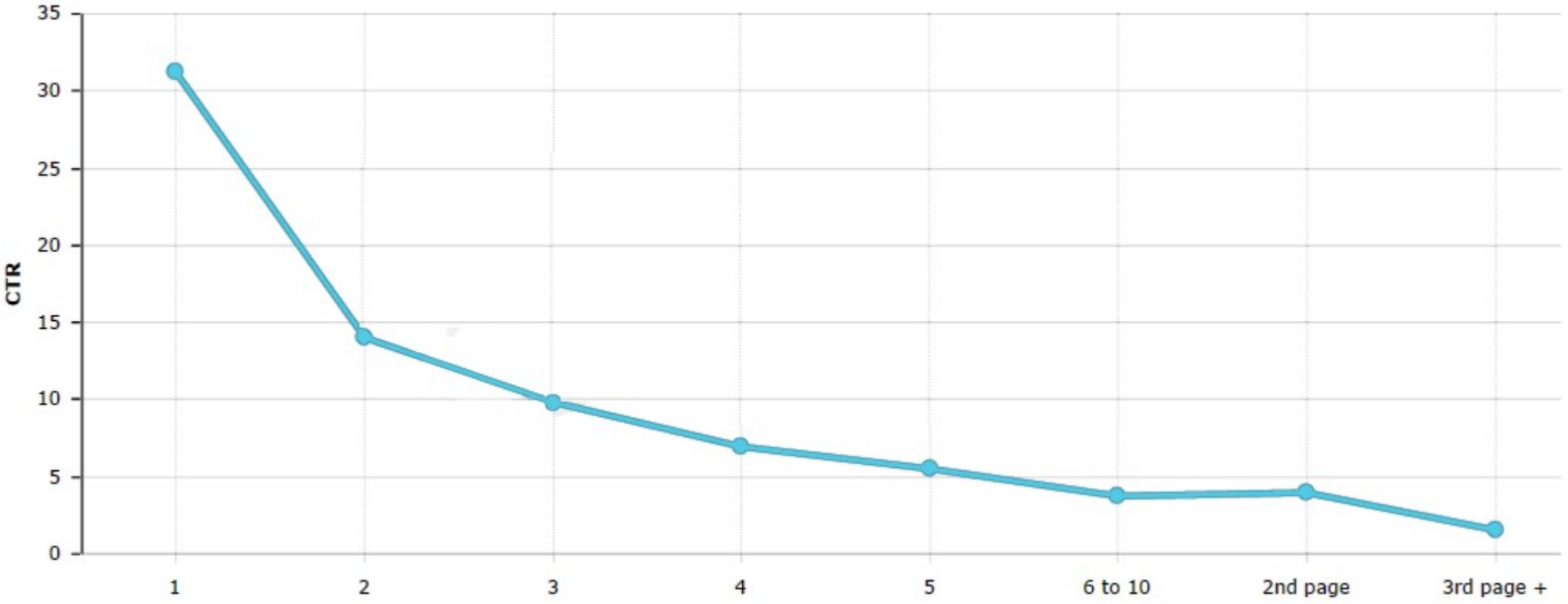


* To Google or to GOTO, Business Week Online, September 28, 2001

Características de las búsquedas

- Frases cortas, imprecisas, sin operadores, poco esfuerzo.
  - Promedio de longitud (en EEUU): (www.keywordiscovery.com/keyword-stats.html)
    - En 2008: 2,29, el 80% de las consultas con < 3 palabras.
    - En 2013: 2,16, el 72% con < 3 palabras.
    - En 2014: 3,62, el 62% con < 3 palabras.
    - En 2015: 3,65, el 60% con < 3 palabras.
- Comportamiento bastante definidos:
  - 85% mira sólo a la primera pantalla de resultados (y sólo “above the fold”)
  - 78% de los queries no son reformulaciones (1 query: 1 sesión)
- Muchísima variación en las necesidades, espectativas y conocimiento de los que buscan.
- La distribución de queries sigue una ley de potencias: algunos queries muy frecuentes, muchos poco frecuentes.
- Crece la búsqueda mediante dispositivos móviles. En algunos países (EEE, Japón), hay más búsquedas por teléfonos y tablets que por PCs.

¿Qué tan lejos la gente busca resultados?



Position	1	2	3	4	5	6 to 10	2nd page	3rd page+
CTR	31.24	14.04	9.85	6.97	5.50	3.73	3.99	1.60

Fuente: http://www.advancedwebranking.com/google-ctr-study-2014.html

Una búsqueda tiene atrás una necesidad

Necesidades detrás de una búsqueda:

- **Informacion**: quiere aprender acerca de algo (~40% a 65%)
  - P.ej. “baja hemoglobina”
- **Navegación:** quiere ir a una página o sitio en particular (15% a 25%)
  - P.ej. “American Airlines”
- **Transaccional:** quiere hacer algo a través de la web (20% a 35%)
  - Acceder a un servicio (“pronóstico meteorológico de la plata”)
  - Descargar (“50 shades of gray descargar gratis”)
  - Comprar (“alquiler de autos brasil”)

Andrei Broder, “A Taxonomy of web search”, 2002

¿Cómo identificar la intención de un query?

Por el texto del query y la acción de los usuarios:

- **Navigational Searching:** queries que contienen...
  - nombres de compañías / negocios / gente
  - Sufijos de dominios (.ar, .com, etc)
  - Queries con menos de 3 tokens
  - Queries donde la mayoría de la gente clickeó en 1 resultado.
- **Transactional Searching**: queries que contienen...
  - Términos relacionados a películas, canciones, letras de canciones y porno
  - extensiones de imágenes, video, o pdf
  - Términos como “comprar”, “chatear”,”descargar”
- **Informational Searching**: queries que contienen:
  - Frases de pregunta: “como hago...” “como llegar”, “que es”, “quien es”, “cuanto sale”...
  - Queries donde la mayoría de la gente clickeó en varios resultados.
  - Queries con 3 o más palabras

Jansen, J. “Determining the User Intent of Web Search Engine Queries”, WWW 2007

¿Cómo identificar la intención de un query?

Por el comportamiento de los usuarios y de los autores de páginas:

- **Comportamiento de Usuarios anteriores**: Si el objetivo de un query es navegación, entonces la mayoría de los usuarios anteriores deben haber hecho click en un mismo resultado (la autoridad).
- **Número de clicks promedio por query:** Para un query navegacional esperamos muy pocos clicks entre los resultados porque está buscando 1 en particular; para un query de información, esperamos que mire varios resultados.
- **Distribución del texto de los links:** Para un query navegacional, esperamos que el mismo texto de un links lleve casi siempre a la misma página, porque el resultado es casi único y tienen un nombre concreto. Para queries de información, esperamos que el mimos texto de link lleva a diferentes páginas.
- Usando regresión logísticas sobre estas 3 features, consiguen casi 90% de precisión para distinguir entre queries de nagevacion y de información.

Lee, U. "Automatic Identification of User Goals in Web Search", WWW 2005

Respondiendo a la necesidad del query

- Un query es comúnmente un indicador impreciso de lo que el usuario realmente quiere.
- Podemos:
  - Restringir resultados a un idioma que el usuario entienda
  - Restringir resultados geográficamente (p.ej. “como jubilarme”)
  - Reordenar resultados según la intención de la búsqueda, y el tipo de recurso (web, imágenes, videos)
  - Corregir la consulta / expandir la consulta / sugerir consultas similares.

Distribución de queries

- Top mundial en 2014

Robin Williams

World Cup

Ebola

Malaysia Airlines

ALS Ice Bucket Challenge

Flappy Bird

Conchita Wurst

ISIS

Frozen

Sochi Olympics

- Top 10 en Argentina en 2014

Mundial 2014

Camus Hacker

Robin Williams

Cyber Monday

Gran DT

Google Street View

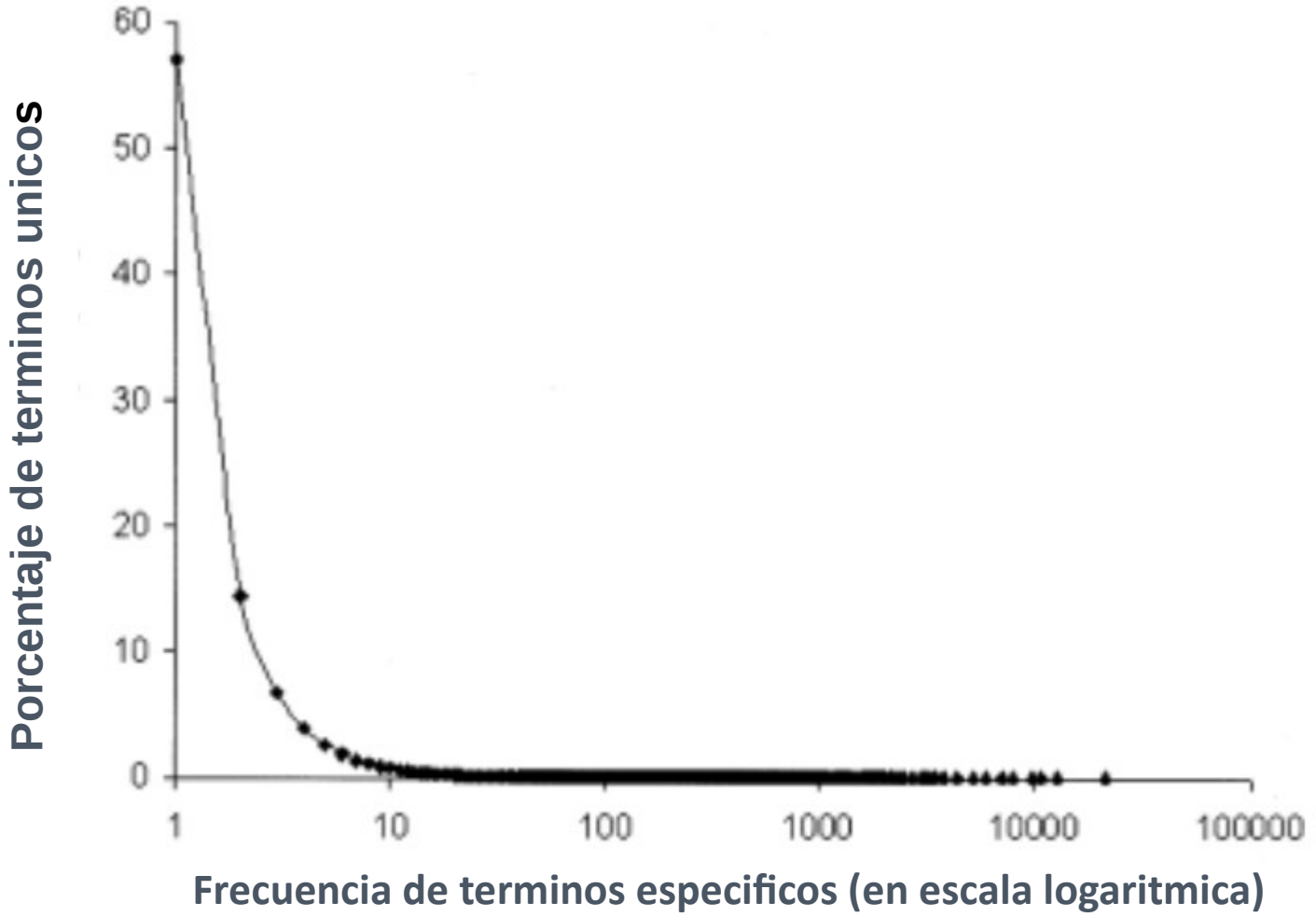
Jorge Ibáñez

Ébola

Dakar

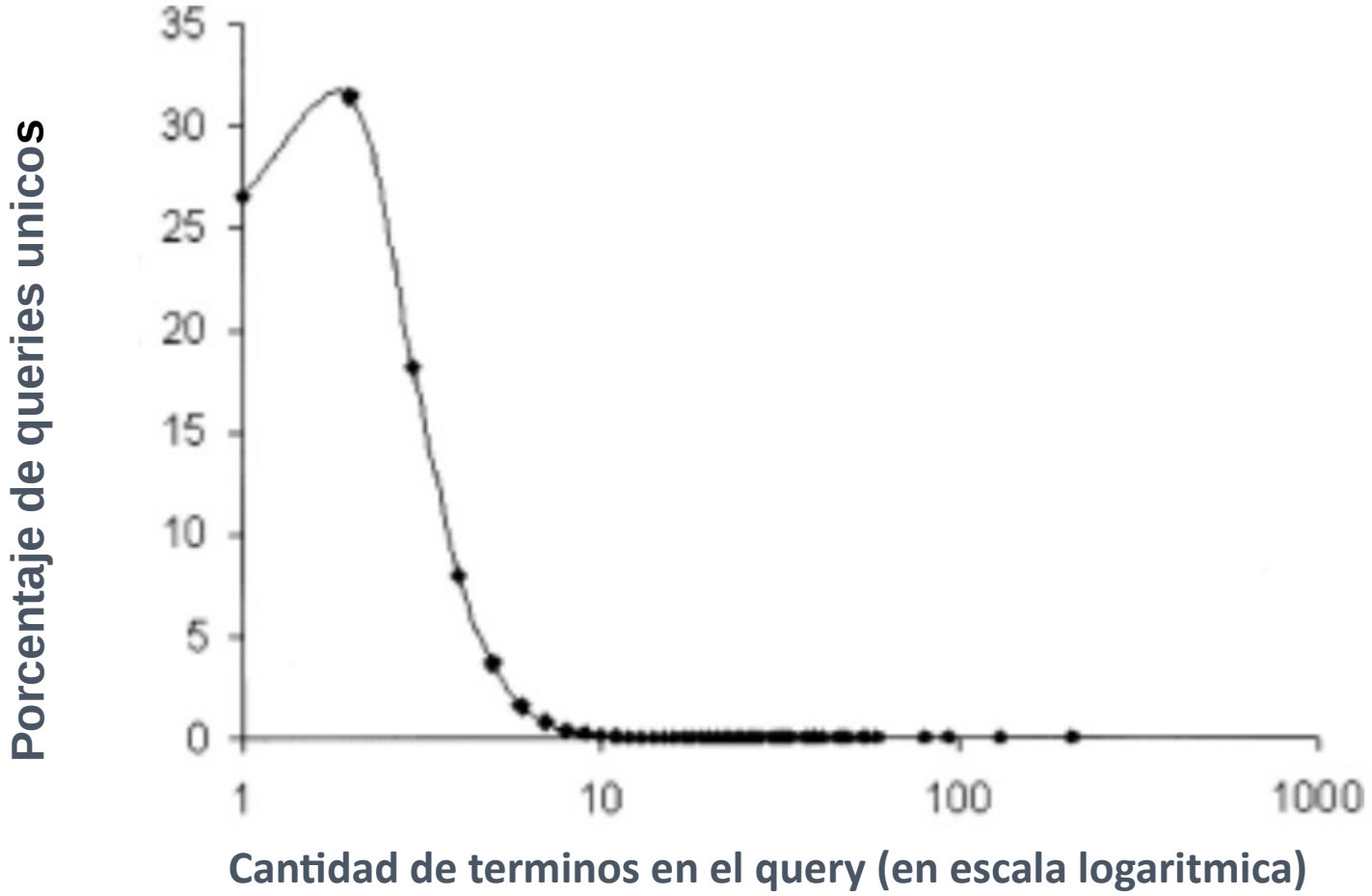
Ezequiel Lavezzi

Un numero pequeño de términos es usado en muchísimos queries



Spink, A. "Searching the web: The public and their queries", 2001

La mayoría de los queries tienen menos de 3 palabras



Spink, A. "Searching the web: The public and their queries", 2001

¿Cómo la gente evalúa resultados?

- Qué tan relevantes son los resultados (precision) cuantos resultados son relevantes (recall).
- Precision es más importante que recall, pero en algunos queries recall importa mucho: Recall importa al buscar antecedentes (p.ej. de un doctor), o al buscar algo que sabemos que existe pero que es difícil de encontrar.
- Confianza en los resultados (no tienen malware), duplicados, legibilidad, velocidad de carga de la página.
- La percepción de los resultados puede no ser objetiva, pero es significativa considerada en promedio.