IN4522-Web Mining
Auxiliar 3: Estructura de la Web

Juan Velásquez
Felipe Bravo Márquez
Gabriel Oberreuter

18 de octubre de 2010

La Web

Internet

Es un conjunto descentralizado de redes de comunicación interconectadas que utilizan la familia de protocolos **TCP/IP**.

¿Qué es la Web?

World Wide Web, es un sistema de documentos de hipertexto enlazados y accesibles a través de **Internet**.

¿Cómo funciona la Web?

- Un cliente (generalmente un browser) realiza una petición a un recurso (URL) vía protocolo HTTP.
- Un servidor DNS identifica el host asociado y así el cliente realiza la petición a un servidor web.
- El servidor retorna un documento de hipertexto (HTML).
- El cliente despliega el contenido al usuario (texto, imágenes, hipervínculos, multimedia,..etc)[Velasquez and Palade, 2008].

Anatomía de la Web [1]

Tipos de páginas Web

- ***Estáticas***: Existen como archivo en un servidor Web, ejemplo (/bio.html).
- ***Dinámicas***: Se crean dinámicamente mediante la interacción de un usuario con un servidor Web. Generalmente el servidor Web interactúa con una capa de datos para crear contenido dinámico (formularios), ejemplo (/bio.php?nombre=pepe&ciudad=santiago).
- ***Semántica***: Contienen *metadatos* en lenguajes como *RDF* (Resource Domain Framework) que le entregan significado a los datos contenidos en la Web (precios, autores, lugares, licencias). Los metadatos ayudan agentes de software a “entender” la información contenida.
- ***Públicas***: Accesibles por cualquier usuario.
- ***Privadas***: Protegidas por clave o pertenecientes a una Intranet.

Anatomía de la Web [2]

Figura: Sitio generado dinámicamente

```
<header>
<identifier>oai:brill:1234567</identifier>
<datestamp>2003-06-09</datestamp>
</header>
<metadata>
<oai_cc:cc
      xmlns:cc="http://creativecommons.org/metadata/schema/"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemalocation="http://creativecommons.org/metadata/schema/
      http://creativecommons.org/metadata/schema/cclicences.xsd">
      <cc:license rdf:about="http://creativecommons.org/licenses/by-nd-nc/1.0">
            <cc:permits rdf:resource="http://web.resource.org/cc/Reproduction" />
            <cc:permits rdf:resource="http://web.resource.org/cc/Distribution" />
            <cc:requires rdf:resource="http://web.resource.org/cc/Notice" />
            <cc:requires rdf:resource="http://web.resource.org/cc/Attribution" />
            <cc:prohibits rdf:resource="http://web.resource.org/cc/CommercialUse"
      />
      </cc:license>
</oai_cc:cc>
</metadata>
```

Figura: Página con Metadatos

Anatomía de la Web [3]

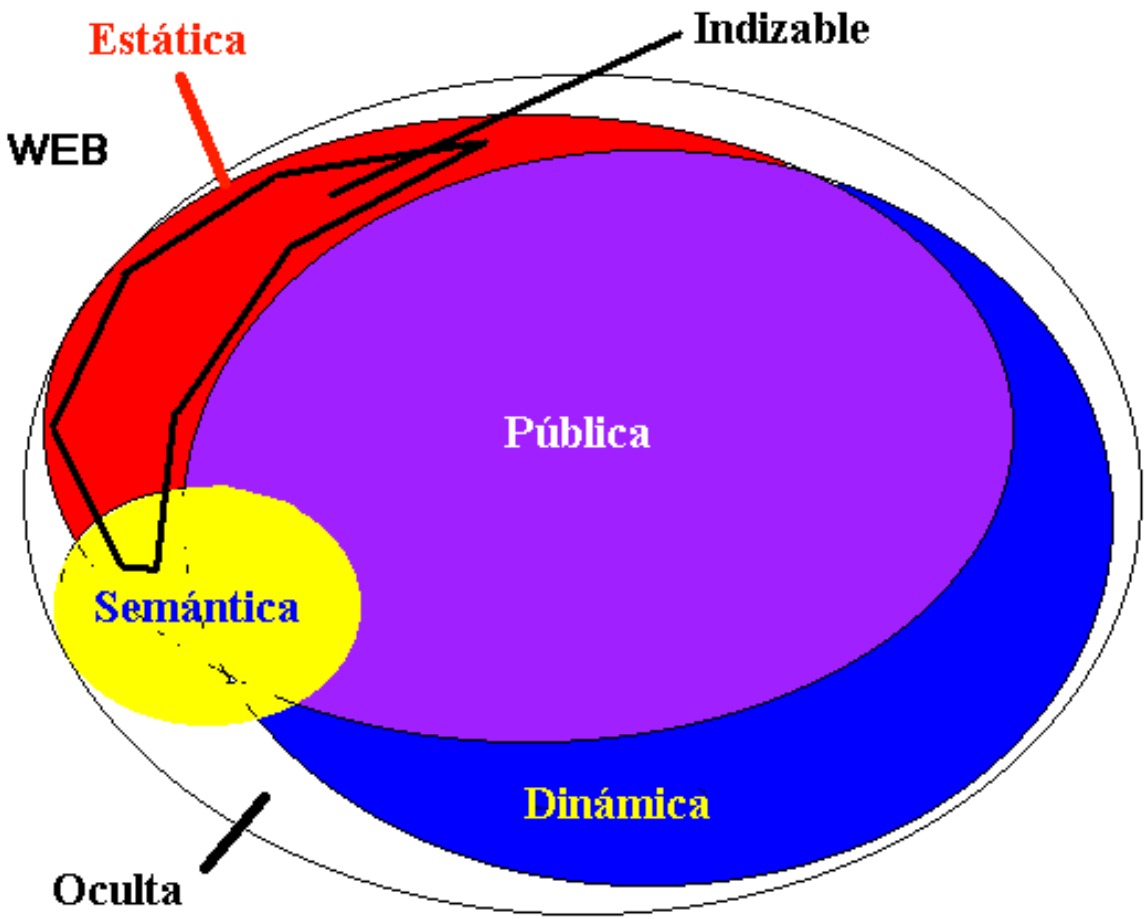


Figura: Características de la Web[Gutiérrez et al., 2008]

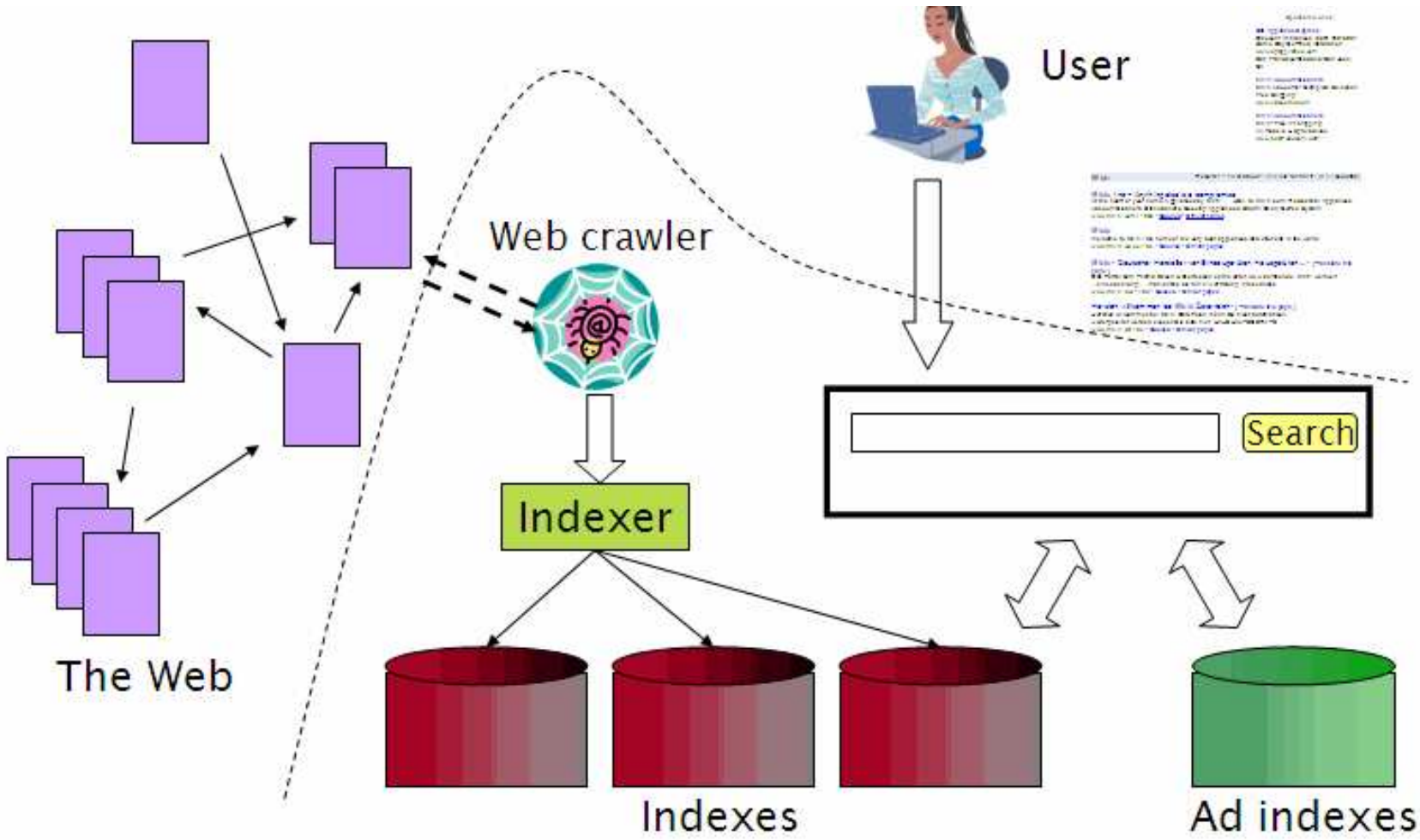
Anatomía de la Web [5]

Tamaño de la Web

- Internet sobrepasó el 2006 los 430 millones de computadores conectados.
- Los servidores web crecen de manera exponencial.
- Las páginas dinámicas hacen a la Web un conjunto infinito.
- La cantidad de páginas indexadas por los motores de búsqueda se estima como al menos 15.35 billones (27-09-2010) [a]

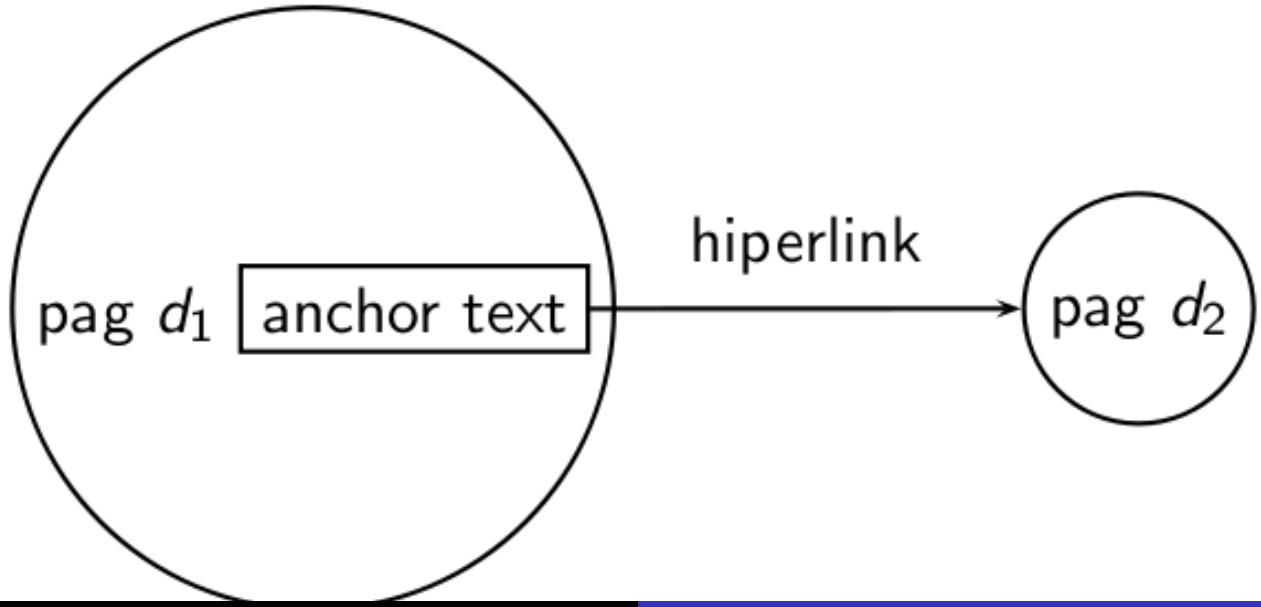
[a] http://www.worldwidewebsize.com/

Arquitectura de un Motor de Búsqueda



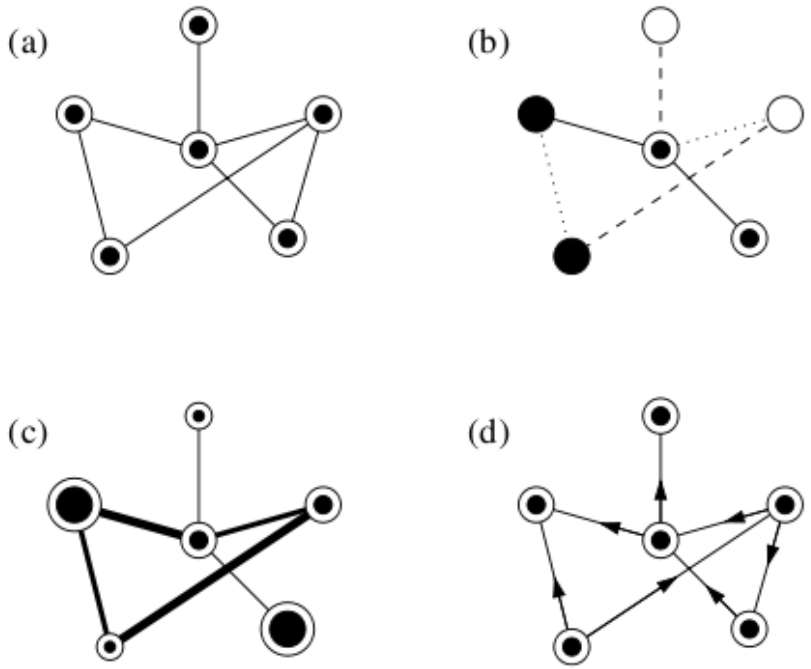
La Web como un Grafo

- La Web es un grafo dirigido donde los vértices son páginas y arcos son hipervínculos entre ellos.
- ` anchor text`
- El anchor text entrega información sobre el recurso apuntado y puede ser utilizado por los crawlers al describir el recurso.
- Definimos a los links que apuntan a una página como *in-links* y a los que apunta la página como *out-links*



Grafos [1]

- Se define un grafo $G(V, E)$ como un estructura compuesta por vértices V (ítemes) y arcos (aristas) E que representan relaciones entre vértices.
- El tipo de grafo varía por los tipos de vértices que tenemos, tipos de arcos, dirección de arcos y pesos de arcos.[Newman, 2003]



- (a) grafo no dirigido simple. (b) distintos tipos de vértices y arcos. (c) distintos pesos de vértices y arcos. (d) grafo dirigido.

Tipos de redes modeladas con Grafos

- **Redes sociales**: Grupos de personas con patrones de contacto o interacción entre ellos. Ejemplo: Transferencia de trabajo en oficina, Facebook.
- **Redes de información**: Unidades de información inter-referenciadas. Ejemplo: Citas de publicaciones (acíclica), la Web.
- **Redes tecnológicas**: Redes diseñadas por humanos para la distribución de un recurso o commodity(redes de trasmisión de voltaje, Internet, ferrocarriles).

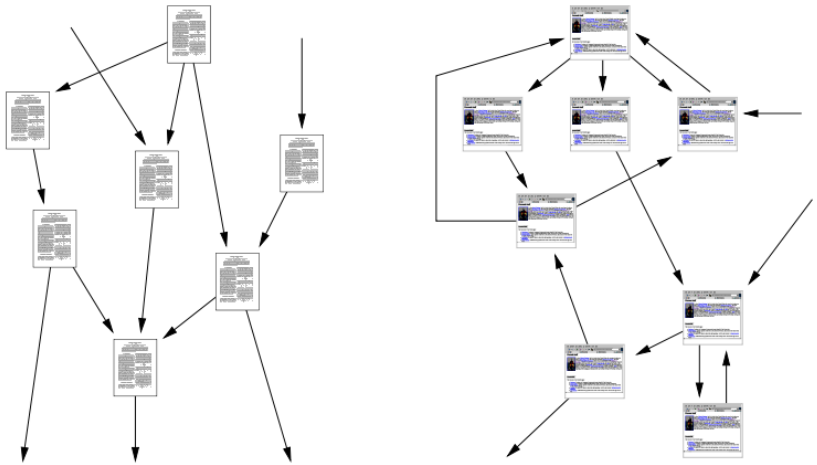
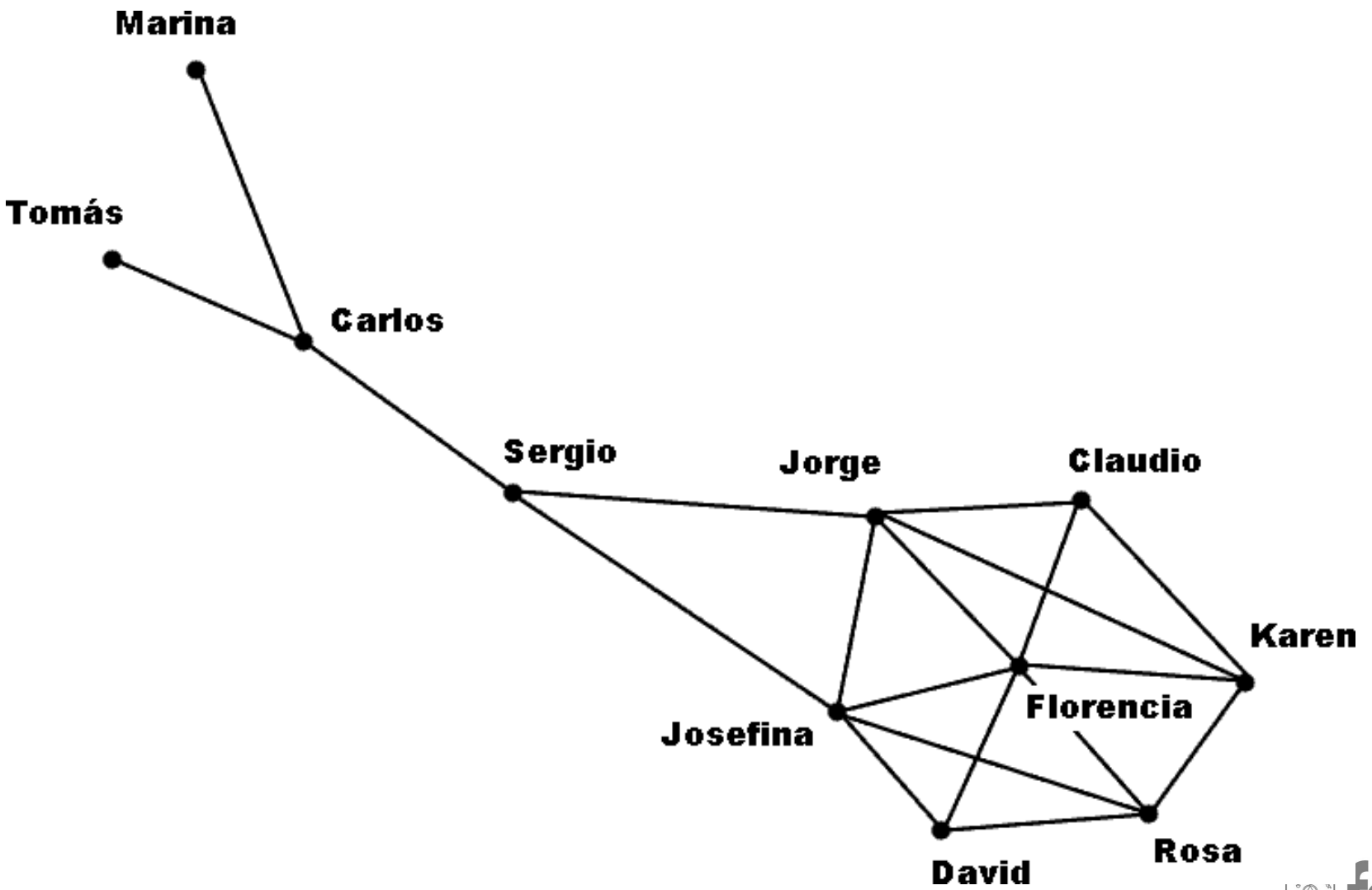


Figura: izquierda:red de citas (acíclica) , derecha: la Web (permite ciclos)

Ejemplo de Red Social no Dirigida



Propiedades de Grafos [1]

- **Grado**: Se define el grado(degree) de un vértice v, $deg(v)$ como el número de arcos conectado al vértice. Para un grafo dirigido se tiene un in-degree y un out-degree. El grado de *Sergio* es 3.
- **Componente**: El componente de un vértice, es el conjunto de vértices alcanzables por medio de caminos.
- **Camino geodésico**: Es el camino más corto entre dos vértices , $d(v_i, v_j)$.
- **El camino geodésico promedio** l de un grafo, indica cuando pasos se requieren en promedio para llegar desde un vértice a otro. Se define para un grafo no dirigido como:

$$l = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d(v_i, v_j)$$

Propiedades de Grafos[2]

Small world effect

- En 1960 Stanley Miligram postuló experimentalmente que bastaban 6 pasos para conectar a todas las personas de Estados Unidos en el envío de correspondencia.
- Estos resultados fueron las primeras demostraciones del ***small world effect*** que dice que muchos tipos de redes de elevadas dimensiones poseen caminos geodésicos promedio de bajo valor.
- Básicamente cada persona está conectada con todas las personas cercanas (geográficamente) y sólo con algunas lejanas de manera **aleatoria** con distribución **uniforme**.

El diámetro de la Web es de 19 según un experimento de Albert, Jeong y Barabási a fines de los 90 con una muestra de 330 mil páginas.

Propiedades de Grafos [3]

- **Diámetro**: Es el camino geodésico más largo en el grafo. Un grafo de diámetro pequeño tiene muchas conexiones.
- **Centralidad**: La centralidad de un vértice busca medir la importancia de éste en el grafo. Un vértice con alto nivel de centralidad es relevante en la **conectividad** del grafo, entonces si eliminamos un vértice con alta centralidad se elevarían los caminos geodésicos de muchos vértices. En la Web, se usa la centralidad para medir la relevancia de una página.

Algunas formas de medir centralidad

- **Centralidad de grado**: (Degree Centrality) Se usa el grado de un vértice normalizado por la cantidad de vértices del grafo para medir centralidad. En el ejemplo, **Florencia** tendría una alta centralidad de grado.
- **Centralidad de intermediación**: (Betweenness Centrality) Es la cantidad de veces en que el vértice se encuentra en el camino geodésico de dos vértices. Define la importancia del vértice en la conectividad de la red. En el ejemplo, *Sergio* tendría un alto *betweenness*.

Grafos y Probabilidades

La forma en que distribuyen los grados de los vértices de un grafo, ayudan a entender los fenómenos detrás de éstos (Web, redes sociales, Internet).

Distribuciones de Grados

- Sea p_k la fracción de vértices del grafo con grado k.
- Entonces p_k es la probabilidad que un vértice v tomado uniformemente del grafo tenga grado k. Algunos casos de distribuciones de grados son:
  - **Binomial**: Cada vértice v está conectado a v' con probabilidad p y no conectado con probabilidad $1-p$. A las redes que distribuyen binomial se les conoce como **redes aleatorias**
  - **Ley de Potencias** (power-law) : $p_k \sim \frac{1}{k^\alpha}$. La ley de Zipf de frecuencia de términos distribute bajo ley de potencias con $\alpha = 1$.
  - **Exponencial**: $p_k \sim e^{\frac{-k}{\kappa}}$

Propiedades del Grafo Web

La Web como una red libre de escala

- La Web no se comporta como una red aleatoria.
- Existen páginas mejores que otras, que reciben más links (autoridades), y otras que apuntan a más links que otras (hubs).
- El *in-degree* y el *out-degree* de la Web se asemejan más a distribuciones **power-law**. Con $\alpha = 2{,}1$ para in-degree y $\alpha = 2{,}7$ para out-degree según estudio de Altavista en Mayo y Octubre de 1999. El número promedio de *in-links* de una página vale entre 8 y 15 [Manning et al., 2008].
- Unas pocas páginas reciben muchos links, mientras que la mayoría de las páginas son muy poco apuntadas.
- A las redes de este tipo, se les conoce como ***red libre de escala***.
- Identificar los vértices centrales es muy importante en la Web.

Propiedades del Grafo Web[2]

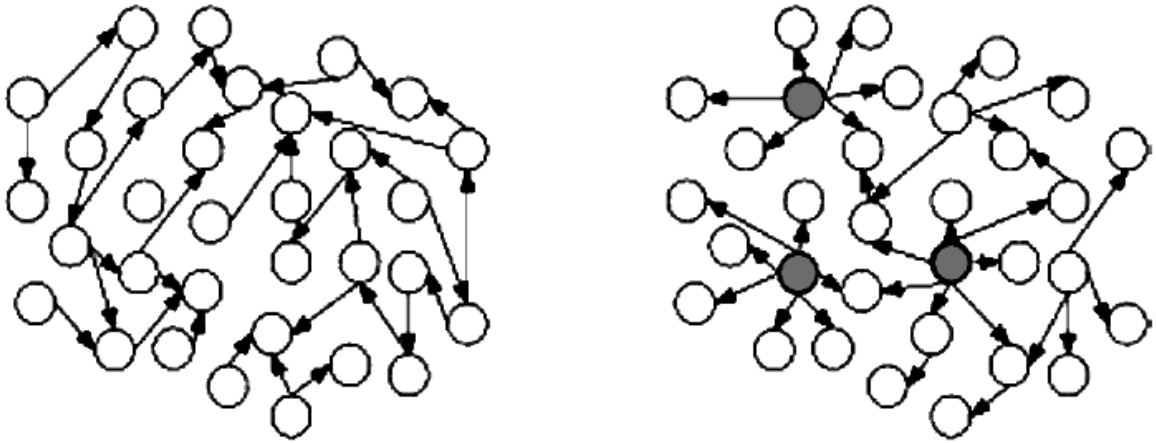


Figura: izq: red aleatoria, der: red libre escala

Propiedades del Grafo Web [3]

La Web tiene forma bowtie

Diversos estudios muestran que la Web tiene forma de humita (bowtie) [Manning et al., 2008, Gutiérrez et al., 2008].

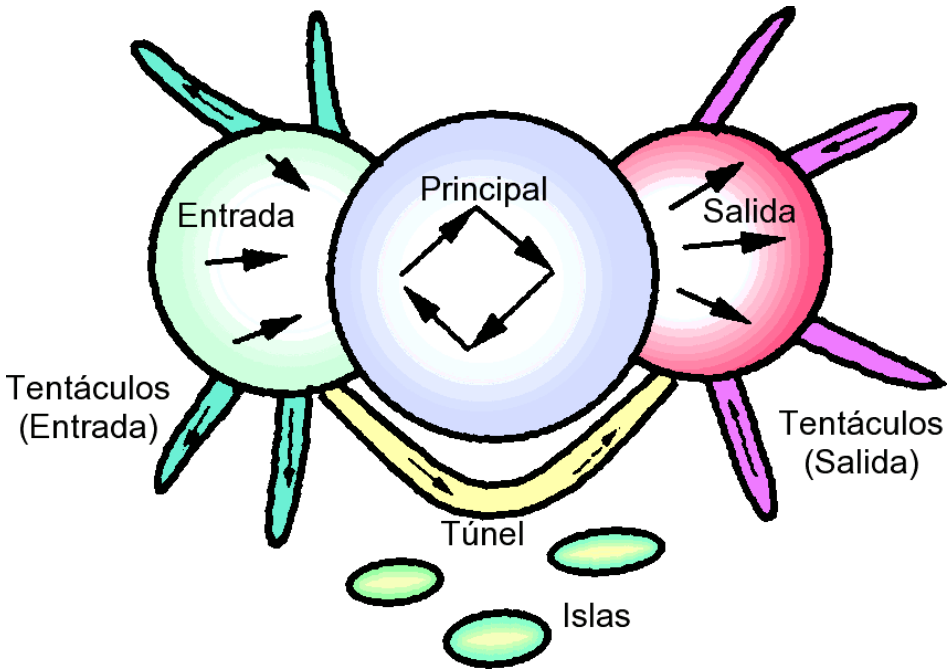


Figura: bowtie Web

Propiedades del Grafo Web [4]

Según el estudio de Altavista el año 1999 sobre alrededor de 200 millones de páginas se tuvo la siguiente distribución:

- IN (entrada) (43 millones): Páginas que apuntan al núcleo pero no se puede llegar a ellos desde él (páginas nuevas).
- SCC (principal) (56 millones): Núcleo de la Web (páginas consolidadas).
- OUT (salida) (43 millones): Páginas viejas que no apuntan al núcleo.
- Tendrils (tentáculos) (44 millones) : Links a sitios inexistentes.
- Tube (túnel): Enlaces de páginas nuevas a páginas viejas.
- Islas (17 millones): Páginas aisladas.

El diámetro obtenido fue de 28.

Dinámica de la Web

- Un sitio nuevo comienza como ISLA o IN. Si llega a ser conocido pasa al centro de la Web.
- Si luego no decide apuntar a un sitio importante o no se actualiza para a OUT o puede dejar de ser apuntada y pasar a ser una ISLA.
- El tiempo promedio de vida de una página es de alrededor de tres meses.
- A veces los sitios cambian de dominio.
- La tasa de nacimiento y muerte de páginas es alto. Muchos de los sitios nuevos son copias de sitios anteriores.
- Los sitios nuevos tienen bajo in-degree, por ende bajo puntaje para los motores de búsqueda.

Crawling

- Se comienza desde un conjunto de URLs **semillas** muy populares o enviadas por administradores de Web sites. Se siguen los enlaces recursivamente manteniendo control sobre las URLS ya visitadas.
- El recorrido puede ser *breadth-first* (cobertura amplia no profunda) o *depth-first* (cobertura vertical).
- Se distribuye en distintas máquinas, lo que dificulta la coordinación. Una alternativa simple es repartir los dominios entre los crawlers.
- El crawler tiene un **scheduler**(agendador) que le da prioridad de crawling a sitios más relevantes, y define que sitios reindexar con más periocidad (emol vs mi página personal).
- Existen protocolos de buen comportamiento para que los crawlers no saturen los servidores. (robots.txt, metatags)

Evolución de los Motores de Búsqueda

- 1995–1997: Basados en texto (Tf-Idf), Altavista, Excite, Infoseek, Inktomi
- Comenzando en 1998/1999: Google Algoritmo de ranking basado en links. Considera además del texto a las páginas con mayor centralidad mayor relevancia (PageRank es una variación de una medida de centralidad basada en vectores propios).
- 2001: Se separan los resultados y los adds. Nuevos ingresos para los motores de búsqueda: add words.
- Hoy en día: links+ texto + clicks. Web Query Mining se basa en analizar las consultas de los usuarios y los clicks realizadas para entrenar modelos predictivos en base a la relevancia indicada.
- La evolución de los buscadores sigue las áreas de **Web Mining**: Content Mining, Structure Mining, Usage Mining.

Search Engine Optimization

Técnicas para hacer subir un sitio en los rankings de motores de búsqueda comerciales: consideran texto, links y clicks.

- Buenas prácticas de diseño pueden mejorar los resultados (no poder todo como páginas dinámicas, difíciles de indexar), poseer buena estructura interna de links, ingresar a comunidades de sitios relevantes, etc..
- Algunas técnicas hacen daño a la calidad de los buscadores. Como **spamdexing** (términos falsos en los metatags), link farms, etc..
- Los motores de búsqueda usan técnicas de machine learning para evitar técnicas sucias de SEO. **Adversarial information retrieval**

References I


Gutiérrez, C., Navarro, G., Baeza-Yates, R., Hurtado, C., Arenas, M., Marín, M., Piquer, J. M., Rodríguez, M., del Solar, J. R., and Velasco, J. (2008).
Cómo funciona la Web.
Autoeditada.


Manning, C. D., Raghavan, P., and Schütze, H. (2008).
Introduction to Information Retrieval.
Cambridge University Press, New York, NY, USA.


Newman, M. E. J. (2003).
The structure and function of complex networks.
SIAM Review, 45(2):167–256.


Velasquez, J. D. and Palade, V. (2008).
Adaptive Web Sites: A Knowledge Extraction from Web Data Approach.