Fundamental Internet and WWW Technologies

Chapter 1

Lecture 1

April 2008

What Is the Internet?

The international network of networks with communication formats defined by International Standards

- set of open software standards
- overseer bodies are coordinated by the -- -
International Organization for Standardization (ISO)
- Internet Engineering Task Force (IETF)
- Internet Society (ISOC)
- RFC Editors publish working notes of the Internet research and development community

What Is the World Wide Web?

The evolving network of information resources.
WWW relies principally on three mechanisms to make these resources readily available to the widest possible audience:

- uniform naming scheme for locating resources on the web (e.g., URIs).
- protocols, for access to named resources over the web (e.g., HTTP).
- hypertext, for easy navigation among resources (e.g., HTML).

History: The original 1990 proposal for the WWW http://www.w3.org/Proposal

WWW standards
The World Wide Web Consortium (W3C) develops interoperable technologies (specifications, guidelines, software, and tools). Today focus is on XML and “Semantic” technologies

Terminology: Internet vs. Web

- Internet is general term and includes physical aspect of underlying networks and software layer that connects
  - Protocols of TCP/IP, DHCP, DNS, SMTP (email), FTP, HTTP…
- Web is term associated with information stored on the Internet
  - Markup languages HTML, XML, RDF, WSDL
  - Protocols of applications HTTPS, XQuery, SOAP,…
- Theoretic studies of both Internet and web use dynamic network models with interesting analytical properties.

DNS System: From Internet to Web

The domain name system (DNS) provide name resolution. Thousands of DNS servers run continually on the Internet, and form what is probably the largest and most successful distributed application in history.
Steps involved in doing a DNS lookup:

- 1. Browser checks for www.xyz.com in its cache
- 2. Query goes to OS cache
- 3. Local name server is contacted
- 4. Local name server makes recursive call, eventually terminating at a root nameserver
- 5. Root nameserver returns IP address for xyz.com DNS server
- 6. Local name server contacts xyz.com DNS server
- 7. xyz.com DNS server returns IP address for www.xyz.com
- 8. IP address returned to operating system of user machine
- 9. IP address returned to browser, which updates its cache
- 10. HTML request begins using HTTP and the IP address

Top Level Domains (TLD)

- Top level domain names, .com, .edu, .gov and ISO 3166 country codes (e.g., .us, .uk, .de)
- There are three types of top-level domains:
- Generic domains were created for use by the Internet public
- Country code domains were created to be used by individual country
- The .arpa domain **A**ddress and **R**outing **P**arameter **A**rea domain is designated to be used exclusively for Internet-infrastructure purposes

Registrars

- Domain names ending with .aero, .biz, .com, .coop, .info, .museum, .name, .net, .org, or .pro can be registered through many different companies (known as "registrars") that compete with one another
- InterNIC (internic.net) provides the central public information regarding DNS registration services (service of the U.S. Department of Commerce).
- InterNIC is now licensed to ICANN (Internet Corporation for Assigned Names and Numbers) which oversees all naming issues and public registrars.

Functions of HTML

- HTML gives authors the means to publish online documents with headings, text, tables, lists, photos, etc
  - Include spread-sheets, video clips, sound clips, and other applications directly in their documents
- Link information via hypertext links
- Design forms for conducting transactions with remote services, for use in searching for information, making reservations, ordering products, etc

Sample Webpage HTML Structure

```
<HTML>
<Head>
<Title>Sample Web Page</Title>
</Head>
<Body>
<P ALIGN=Center><IMG SRC="../gifs/writer.jpg">
<H1><Center>Creating the HTML File</Center></H1>
<P Align=Left>
For tips on creating a web page, see the
<A HREF="http://www.cs.uc.edu/computer_services/web"> Computer
Services Web Kit</A> available on the web.<P>
Learning how HTML works just takes some practice. It's best to start by
creating your own simple page.
</Body>
</HTML>
```

HTML Hyperlink

- ***web kit***
- A link is a connection from one Web resource to another
- It has two ends, called anchors, and a direction
- Starts at the "source" anchor and points to the "destination" anchor, which may be any Web resource (e.g., an image, a video clip, a sound bite, a program, an HTML document)

Introduction to URIs

- Every resource available on the Web has an address that may be encoded by a URI
- URIs typically consist of three pieces:
- The naming scheme of the mechanism used to access the resource. (HTTP, FTP)
- The name of the machine hosting the resource
- The name of the resource itself, given as a path

URI Example

- *http://www.w3.org/TR*
- There is a resource or hypertext document available via the HTTP protocol
- Residing on the machines hosting the domain name www.w3.org
- Accessible via the path "/TR"

Resources URIs, URNs, and XRIs

- Current Internet infrastructure is based primarily on two layers of identifiers
  - machine-friendly IP addresses
  - human-friendly DNS names.
- URIs (Uniform Resource Identifiers) helps people identify resource on the Internet. When a URI points to a specific resource at a specific location we call it a URL (Uniform Resource Locator), and when it points to a specific resource at a nonspecific location we call it a URN (Uniform Resource Name).
- URIs are based on DNS (and therefore IP) and is the linking syntax for the World Wide Web.

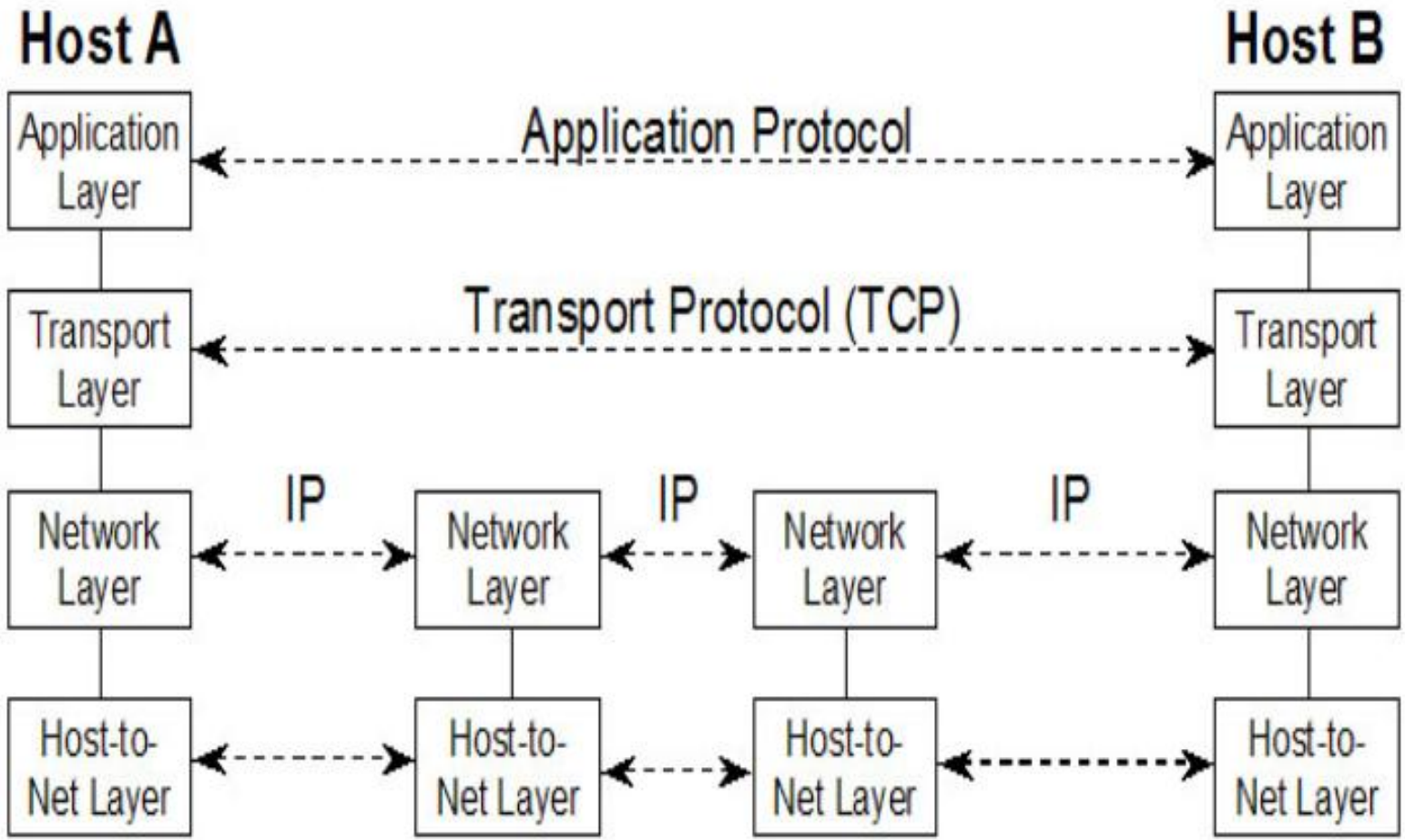
Alternatives to URIs

- There are a number of new identity management technologies, such as DOI, XRI, OpenID and Microsoft CardSpace.
- A DOI, or digital object identifier, in contrast to a URL, is not dependent upon the electronic document's location
- A DOI consists of a unique alphanumeric character string divided into two parts: a prefix and a suffix. Example of a complete DOI is: doi:10.1000/182
- An XRI, or eXtensible Resource Identifier, is a proposed syntax for abstract identifiers and a unified Internet-based resolution protocol, including support for secure distributed resolution, and support for reassignable names that are more human-friendly.
- Example of XRIs with mixes of persistent and reassignable segments (XRI allows any combination of the two):
  - !!1002!A745/(+phone.number)
  - @Jones.and.Company/!D90F.88/(+area.code)

HTTP and Data Transport

- HTTP is client-server protocol that transfers HTML as an application between a browser and a server
- HTTP is an application built on top of TCP – transport control protocol that lives in most every operating system and supported by the Internet standards
- HTTP communication is established via a TCP connection and server at port 80

TCP/IP Layering Architecture



TCP/IP Layering Architecture

- A simplified model, provides the end-to-end reliable connection
- The network layer
  - Hosts drop packages into this layer, layer routes towards destination
  - Only promise of layer: “Try my best”
- The transport layer
  - Reliable byte-oriented stream

GET Method in HTTP

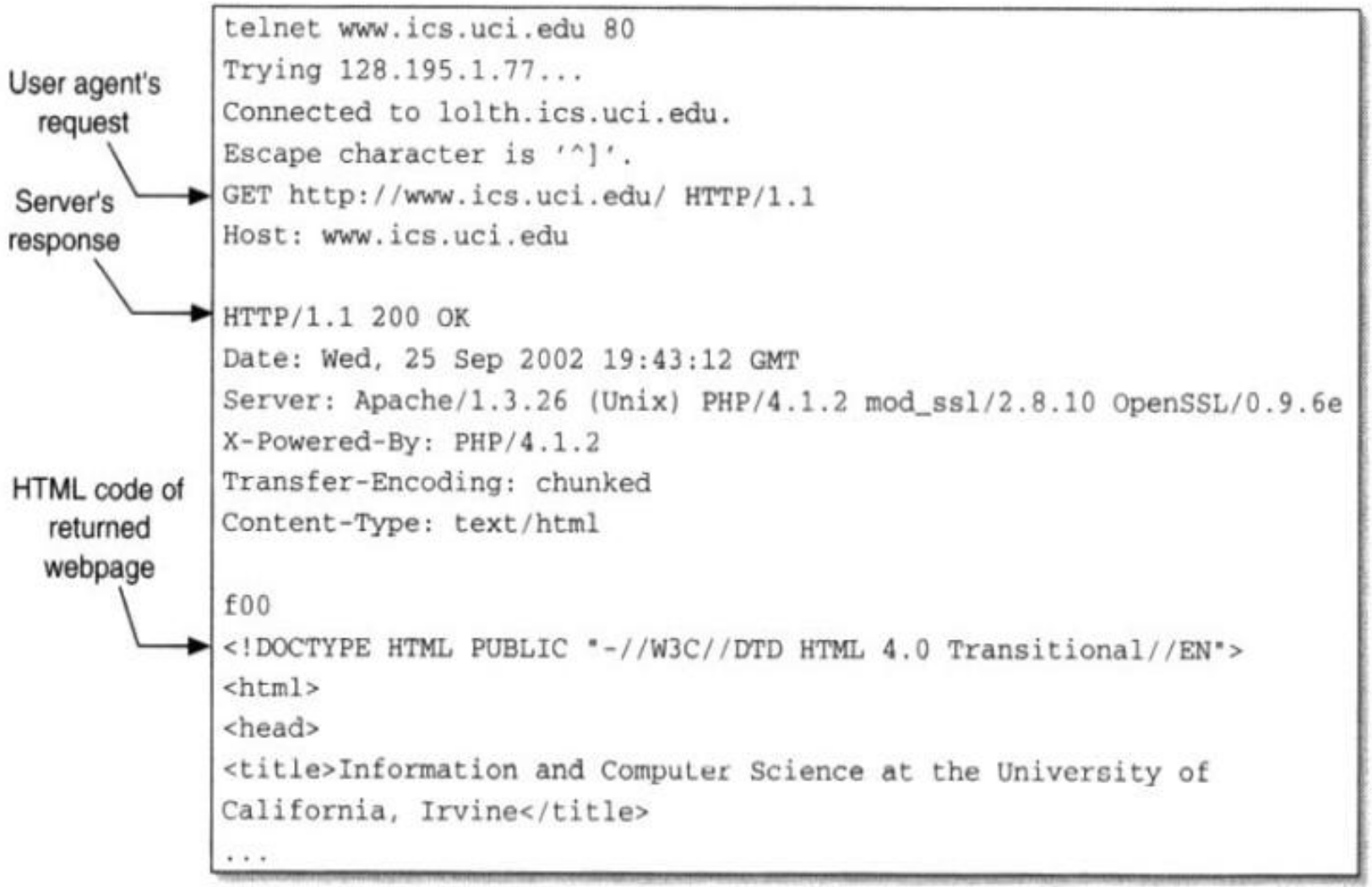


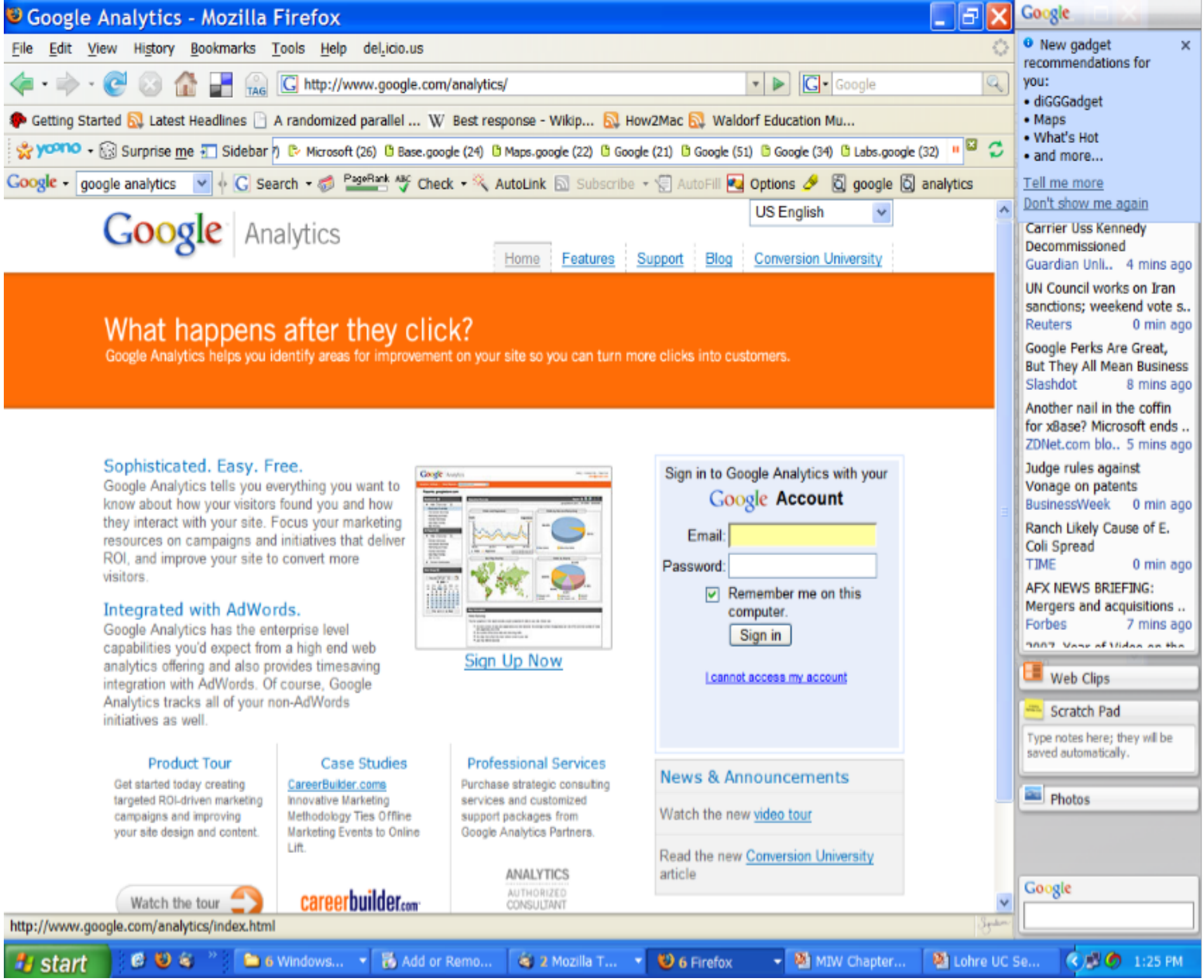
Figure 2.4 Example of the use of the GET method in an HTTP 1.1 session.

Server Log Files

- Server Transfer Log: transactions between a browser and server are logged
- IP address, the time of the request
- HTTP Method of the request (GET, HEAD, POST…)
- Status code, a response from the server
- Size in byte of the transaction
- Referrer Log: where the request originated
- Agent Log: browser software making the request (spider)
- Error Log: request resulted in errors (404)

Server Log Analysis

- Most and least visited web pages
- Entry and exit pages
- Referrals from other sites or search engines
- What are the searched keywords
- How many clicks/page views a page received
- Error reports, like broken links


Google Analytics - Mozilla Firefox
File Edit View History Bookmarks Tools Help del.icio.us
http://www.google.com/analytics/
Getting Started Latest Headlines A randomized parallel ... Best response - Wikip... How2Mac Waldorf Education Mu...
Surprise me Sidebar Microsoft (26) Base.google (24) Maps.google (22) Google (21) Google (51) Google (34) Labs.google (32)
google analytics Search PageRank Check AutoLink Subscribe AutoFill Options google analytics
Google Analytics
US English
Home Features Support Blog Conversion University
What happens after they click?
Google Analytics helps you identify areas for improvement on your site so you can turn more clicks into customers.
Sophisticated. Easy. Free.
Google Analytics tells you everything you want to know about how your visitors found you and how they interact with your site. Focus your marketing resources on campaigns and initiatives that deliver ROI, and improve your site to convert more visitors.
Integrated with AdWords.
Google Analytics has the enterprise level capabilities you'd expect from a high end web analytics offering and also provides timesaving integration with AdWords. Of course, Google Analytics tracks all of your non-AdWords initiatives as well.
Sign Up Now
Sign in to Google Analytics with your Google Account
Email:
Password:
Remember me on this computer.
Sign in
I cannot access my account
Product Tour
Get started today creating targeted ROI-driven marketing campaigns and improving your site design and content.
Watch the tour
Case Studies
CareerBuilder.com: Innovative Marketing Methodology Ties Offline Marketing Events to Online Lift.
careerbuilder.com
Professional Services
Purchase strategic consulting services and customized support packages from Google Analytics Partners.
ANALYTICS AUTHORIZED CONSULTANT
News & Announcements
Watch the new video tour
Read the new Conversion University article
http://www.google.com/analytics/index.html
Google
New gadget recommendations for you:
• diGGGadget
• Maps
• What's Hot
• and more...
Tell me more
Don't show me again
Carrier Uss Kennedy Decommissioned
Guardian Unli.. 4 mins ago
UN Council works on Iran sanctions; weekend vote s..
Reuters 0 min ago
Google Perks Are Great, But They All Mean Business
Slashdot 8 mins ago
Another nail in the coffin for xBase? Microsoft ends ..
ZDNet.com blo.. 5 mins ago
Judge rules against Vonage on patents
BusinessWeek 0 min ago
Ranch Likely Cause of E. Coli Spread
TIME 0 min ago
AFX NEWS BRIEFING: Mergers and acquisitions ..
Forbes 7 mins ago
Web Clips
Scratch Pad
Type notes here; they will be saved automatically.
Photos
Google
start
6 Windows... Add or Remo... 2 Mozilla T... 6 Firefox MIW Chapter... Lohre UC Se... 1:25 PM

- Life on Internet- Pew Research (pewinternet.org)

QuickTime™ and a
decompressor
are needed to see this picture.

Search Engines

- According to Pew, search engines are the most popular way to locate information.
- About 75% or 150 million adults in U.S. are Internet users, and majority query search engines daily.
- Search Engines are measured by coverage of the web, recency, perceived ability to find most relevant information (page rank)
- Google is reported to carry 80% of all searches

Statistical Experiments in Coverage

We'll use *Overlap analysis* to estimate the size of the web

Capture/Recapture Method of Estimating Animal Population Size

- Need to make some independence assumptions
- Let W be set of all (indexable) webpages (unknown size)
- One web crawling engine is not sufficient to estimate W

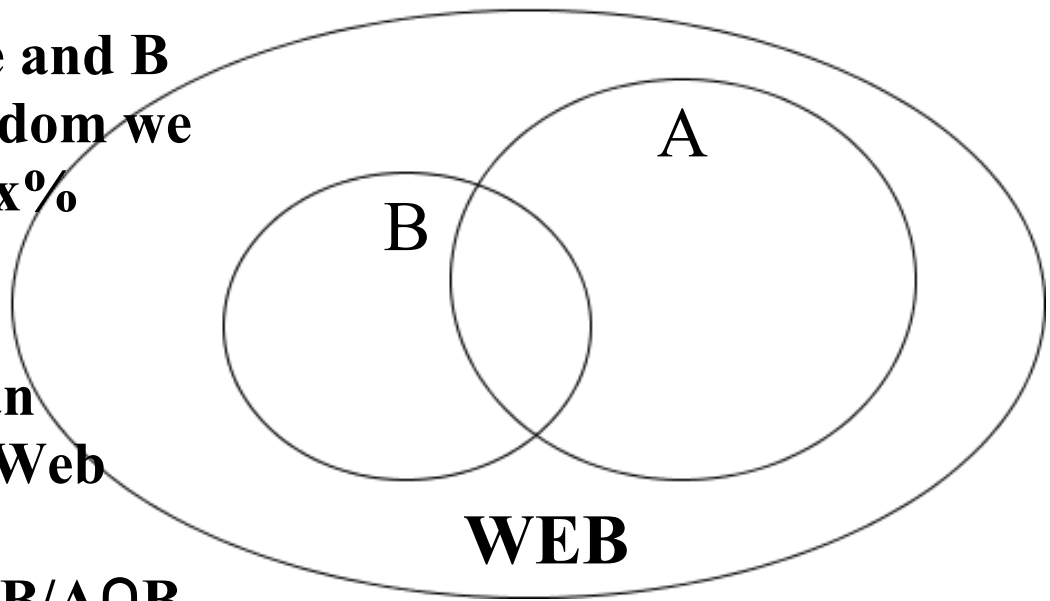
Capture – Recapture technique

- Assumes there is method of generating a statistically independent random subset.

Suppose B is a random sample and B covers x% of A. Since B is random we can assume, B covers approx. x% of the entire Web as well

Hence knowing size of B we can compute estimated size of the Web

Size of the Web = 100*B/x = AB/A∩B



Bharat & Broder : 200 M (Nov 97), 275 M (Mar 98)
Lawrence & Giles : 320 M (Dec 97)

Overlap Analysis – Independent Crawls

We now present a slightly more formal analysis using probability.
Let Wa and Wb be set of pages crawled by two independent engines. By independent we mean that the event that a page is crawled by one engine has no impact on the event that the same page is crawled by the other engine.
Let P(Wa) and P(Wb) be probabilities that a random page p was crawled by a and b, respectively
P(Wa)=|*Wa*| / |*W*|*, and P(Wb)*=|*Wb*| / |*W*|
Let us look at conditional probabilities
By definition,
P(Wa|*Wb) = P(Wa* $\cap$ *Wb)/ P(Wb)*
= |*Wa* $\cap$ *Wb*| / |*Wb*|

Now using the assumption that a and b are independent crawlers
We have that (by definition), P(Wa| *Wb) = P(Wa)*
= |*Wa*| / |*W*|
So, |*W*| = |*Wa*| * |*Wb*| / |*Wa* $\cap$ *Wb*|

Homework Assignment: Statistical Estimation of UC Web

- Download and learn to use web crawler software (see http://en.wikipedia.org/wiki/Web_crawler)
- I have had success with java websuck http://www.ake.nu/software/websuck/
- Use crawler to try to estimate the size of the number of webpages in uc.edu domain
- Run a series of experiments with different starting sites
- Count the number of distinct page links (e.g., ignore image links) found for each experiment
- To apply overlap analysis we need a simple set intersection algorithm (any ideas)

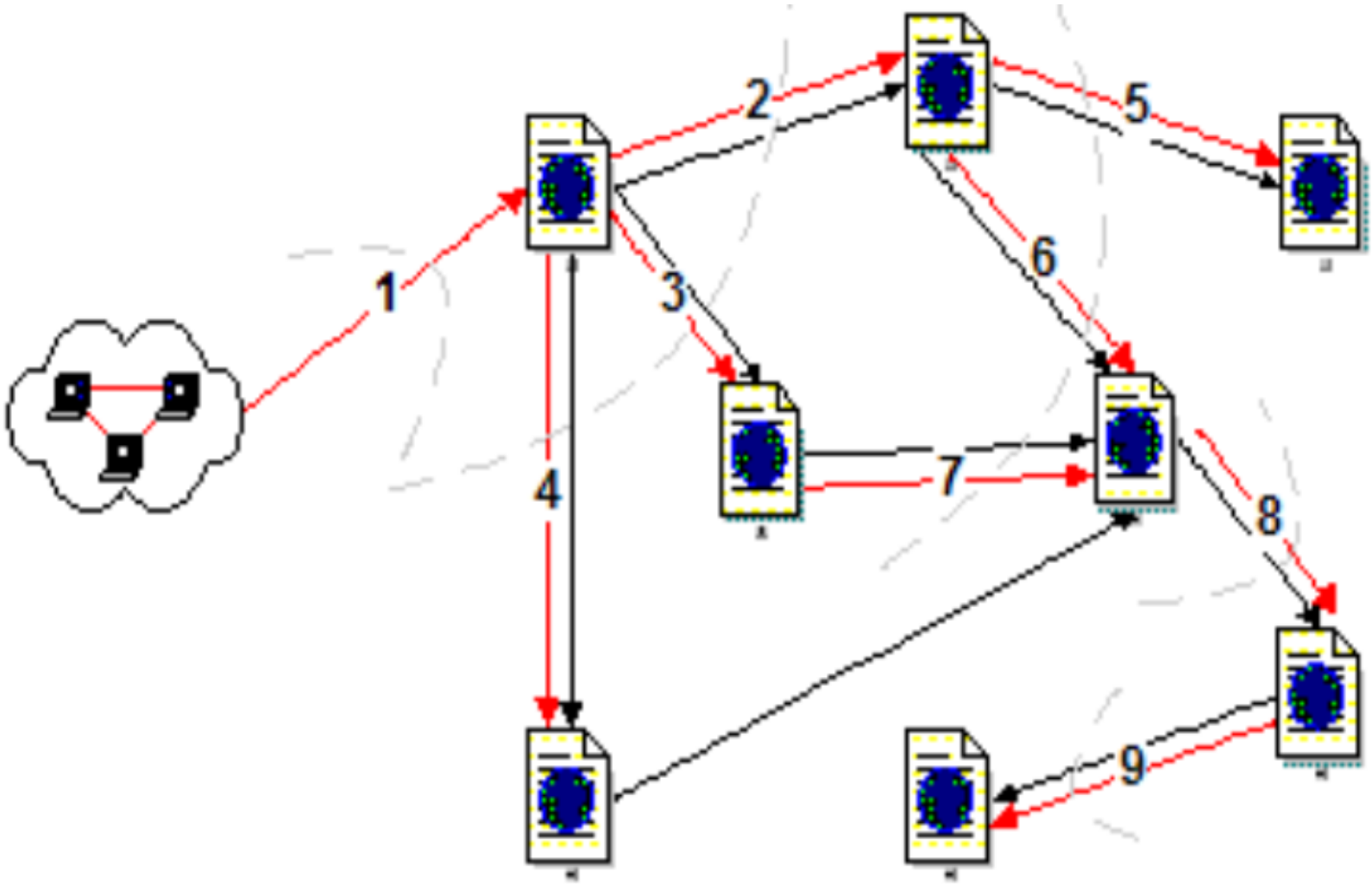
Web Crawler

- A crawler is a program that picks up a page and follows all the links on that page
- Crawler = Spider
- Types of crawler:
  - Breadth First
  - Depth First

Breadth First Crawlers

- Use breadth-first search (BFS) algorithm
- Get all links from the starting page, and add them to a queue
- Pick the 1st link from the queue, get all links on the page and add to the queue
- Repeat above step till queue is empty

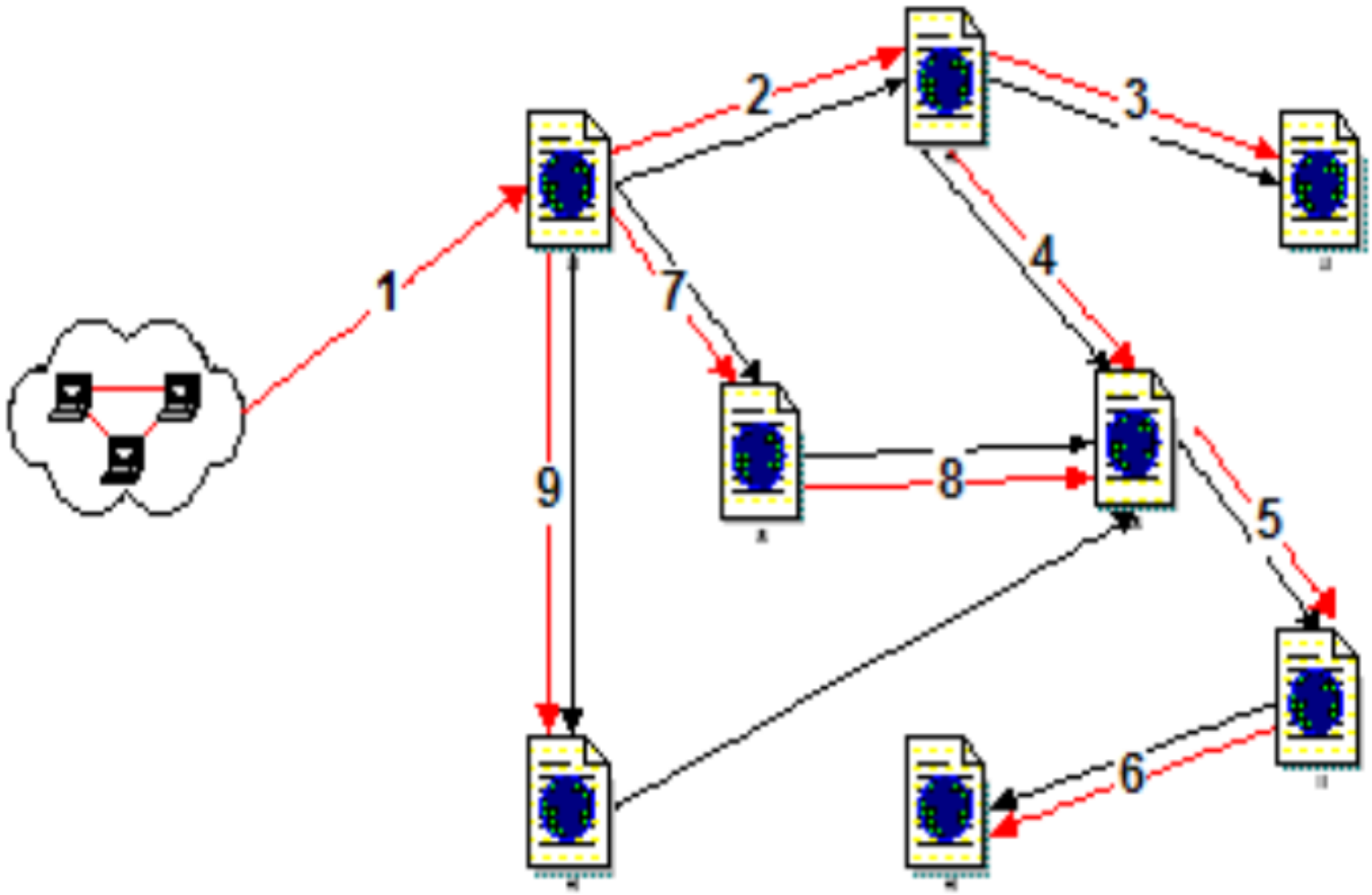
Breadth First Crawlers



Depth First Crawlers

- Use depth first search (DFS) algorithm
- Get the 1st link not visited from the start page
- Visit link and get 1st non-visited link
- Repeat above step till no no-visited links
- Go to next non-visited link in the previous level and repeat 2nd step

Depth First Crawlers



Appendix

Manipulating DNS to solve Locality Problems

The DNS system can be manipulated to solve some important name resolution problems
Example: where is the “best” web server for me to access.
Private companies (e.g., Akamai) provide services to high traffic websites that direct users to optimal resource servers (potentially managed by 3rd parties).

- Requires changes to the usual DNS lookup system. Here is an example:
- the address www.xyz.com must not translate directly to an IP address such as 18.7.21.70, but rather must be aliased to an intermediate address
- known in DNS as a “CNAME”, for example a CNAME for xyz.com could be a212.g.akamai.net
- CNAME address is then resolved by Akamai to an optimally located server