Web Characterization
Web Design

Week 3
LBSC 690
Information Technology
Why is there a Web?

• Affordable storage
  – 300,000 words/$ in 1995

• Adequate backbone capacity
  – 25,000 simultaneous transfers in 1995

• Adequate “last mile” bandwidth
  – 1 second/screen in 1995

• Display capability
  – 10% of US population in 1995

• Effective search capabilities
  – Lycos and Yahoo were started in 1995
What is the Web?

• Protocols
  – HTTP, HTML, or URL?

• Perspective
  – Content or behavior?

• Content
  – Static, dynamic or streaming?

• Access
  – Public, protected, or internal?
Some Perspectives

• Web “sites”
  – In 2002, OCLC counted any server at port 80
  – Total was 3 million, an undercount
    • Misses many servers at other ports
    • Some servers host unrelated content (e.g., TerpConnect)
    • Some content requires specialized servers (e.g., rtsp)

• Web “pages”
  – In 2012, Google counted any URL it has seen
  – Total was 30 trillion, an overcount
    • Includes dead links, spam, …

• Web “use”
  – Google users pose 3 billion queries a day
Crawling the Web
Robots Exclusion Protocol

• Requires voluntary compliance by crawlers

• Exclusion by site
  – Create a robots.txt file at the server’s top level
  – Indicate which directories not to crawl

• Exclusion by document (in HTML head)
  – Not implemented by all crawlers
    <meta name="robots" content="noindex,nofollow">
Link Structure of the Web

Nature 405, 113 (11 May 2000) | doi:10.1038/35012155

Central core
56 million pages

IN
44 million pages

OUT
44 million pages

Tendrils and tubes
44 million pages

Disconnected components
17 million pages
Web Crawl Challenges

• Discovering “islands” and “peninsulas”

• Duplicate and near-duplicate content
  – 30-40% of total content

• Link rot
  – Changes at ~1% per week

• Network instability
  – Temporary server interruptions
  – Server and network loads

• Dynamic content generation
Duplicate Detection

• Structural
  – Identical directory structure (e.g., mirrors, aliases)

• Syntactic
  – Identical bytes
  – Identical markup (HTML, XML, …)

• Semantic
  – Identical content
  – Similar content (e.g., with a different banner ad)
  – Related content (e.g., translated)
Hands on: The Internet Archive

• alexa.com Web crawls since 1997

• Check out the iSchool’s Web site from 1998!
Most Widely-Spoken Languages

Source: Ethnologue (SIL), 1999
Global Trade

Leading economies of merchandise trade, 2009

The “Deep Web”

- Dynamic pages, generated from databases
- Much larger than surface Web
- Not easily discovered using crawling