The Structure of the Web

Getting to knowing the Web

- How big is the web and how do you measure it?
- How many people use the web?
- How many use search engines?
- What is the shape of the web?

- How do people search for information?
- Can we categorize web searchers?

The Web Is a Graph

Examples of Graphs

“Map of the Internet” (1998)
Graph G: Formal Definition

- A graph G consists of two sets G = (V, E)
  - A set V of vertices, or nodes (entities)
  - A set E ⊆ V × V of edges (relationships between entities)

Nodes are students in a large HS, edges join two who had a romantic relationship at some point during the 18-month period in which the study was conducted.

Examples of Relationships (Graphs)

- Connected, Paths and Distances
  - A Connected graph
    - has a path between each pair of distinct vertices
  - A Path A → B is
    - A sequence of edges from A to B
  - Minimum Distance
    - The shortest path
    - How do you measure it?

Computing Minimum Distances

- Algorithm:
Directed Graphs and DAGs

- Directed graph
  - Each edge is a directed edge, or an arc, or a link
  - Can have two arcs between a pair of vertices, one in each direction
  - Vertex $y$ is adjacent to vertex $x$ iff there is a directed edge from $x$ to $y$

- Directed Path
  - A sequence of directed edges between two vertices

- Directed Acyclic Graph (DAG)
  - Directed graph that has no cycles

Structure of The Web

- A Web Page corresponds to a node
- A Hyperlink corresponds to a directed edge

Structure of The Web

- The Web is a directed graph with one large
  Strongly Connected Component
- Is there a directed path from the Univ. of X to Company Z’s Home?
- How about to USNews College Rankings?
- The other way(s) around?
How big is the web?

- **Number of accessible** web pages –
  
  May 2005 estimate: 11.5 Billion pages
  
  Most recent estimates? _______

- **The deep (or hidden or invisible) web**
  
  “contains 400-550 times more information”
  
  (Are they serious?)

- **Coverage** (i.e., the proportion of the web indexed)
  
  is crucial for search engines.
  
  Today, ________________ pages are indexed

How do you measure the size of web?

- **Capture-recapture** method

  - **SE1** = # of pages indexed by search engine 1
  
  - **QSE2** = # of pages returned by search engine 2 for “typical” queries.
  
  - **OVR** = # of pages returned by both search engines for typical queries.

  **Estimate**:
  
  \[
  \text{SE1} / \text{WWW} = \text{OVR} / \text{QSE2} \implies \\
  \text{WWW} = (\text{SE1} \times \text{QSE2}) / \text{OVR}
  \]

How hard is it to go from one page to another?

- Over 75% of the time there is no directed path from one random web page to another.

- When a directed path exists
  
  its average length is 16 clicks.

- When an undirected path exists
  
  its average length is 7 clicks.

- Short average path between pairs of nodes is characteristic of a **small-world network**.

  Kleinberg: The small-world phenomenon (we will study later)