CS 59000-ENS: The Structure of the Web

Assefaw Gebremedhin
Purdue University
agebreme@purdue.edu

http://www.cs.purdue.edu/homes/agebreme/Networks

Spring 2013
Overview

- Web basics
- Information networks and Web precursors
- Web as a directed graph
- Structure of the Web
- Web 2.0
WWW basics

- Created around 1990 (Tim Berners-Lee)
- Two central features
  - Web pages
  - Browsers

We can think of these pages both as part of a single coherent system (the Web) and also as files that reside on four different computers.
Hypertext

- **Crucial design decision:**
  Organize the information using a network metaphor => turns webpages into the “web”

- **Hypertext:**
  replace traditional linear structure of text with a network structure
Intellectual precursors of Hypertext: Other examples of information networks

**Citation network**

Cross reference in encyclopedia:
In this case wikipedia article on topics in game theory

**Semantic networks**
(in cognitive science)

Any analogies with “six degrees of separation”?
Web: some history and evolution

- Memex
  - Imagined by Vannevar Bush in 1945 in an article in the Atlantic Monthly titled “As We May Think”

- Evolution of the Web
  - Navigational
  - Transactional
The Web as a directed graph

- **Basic units:**
  - Nodes: webpages
  - Edges: hyperlinks
- **Path**
  - usual defn + direction
- **Strongly connected**
  - There is a path from every node to every other node
Strongly connected components

- SCC of a graph is a subset of nodes s.t.
  - (i) every node in the subset has a path to every other, and
  - (ii) the subset is not part of some larger set with the property that every node can reach every other
Directed graphs

- **Fact:** every directed graph is a DAG (directed acyclic graph) on its SCCs.
  - (1) SCCs partition the nodes of G
    - Each node is in exactly one SCC
  - (2) If we build a graph $G'$, where the nodes are the SCCs of G and there is an edge in $G'$ whenever there is an edge between the SCCs in G, then $G'$ is a DAG
Proofs

1. **SCCs partition nodes**
   - Assume the contrary (a node v belongs to two SCCs, S and S’).
   - Then, S U S’ is a SCC, a contradiction!

2. **G’ is DAG**
   - Assume the contrary (there is a cycle).
   - Then, original SCC was not maximal, a contradiction!
Structure of the Web

- Brodier et al (2000) set out to build global map of the Web using SCC as building blocks

- Computational matters:
  - How to find SCC containing node \( v \)?
  - Observation:
    - \( \text{Out}(v) = \text{nodes that can be reached from } v \)
    - \( \text{In}(v) = \text{nodes that can reach } v \)

  - SCC containing \( v \) is \( \text{Out}(v) \) intersection \( \text{In}(v) \)

  - Also \( \text{Out}(v, G) = \text{In}(v, G^R) \), where \( G^R \) is \( G \) with its edges reversed in direction
Out(A) Int. In(A) = SCC

Example

- Out (A) = \{A, B, C, D, E, F, G\}
- In(A) = \{A, B, D, E, F, G\}
- SCC(A) = Out(A) int. In(A)
  = \{A, B, D, E, F, G\}
Broder et al experiment

- Altavista Crawl from October 1999
  - 203 million URLs
  - 1.5 billion links
  - Computer: server with 12GB memory
Broder et al finding

Giant SCC

Bow-tie structure
 IN
 OUT
 Tendrils
 Disconnected

IN
 SCC
 OUT

Tendrils
 IN
 OUT

Connected components

Disconnected components
Web 2.0

Three major drivers

- Growth of web authoring styles that enabled many people to collectively create and maintain shared content
- Movement of personal online data from personal computers to services offered by large companies
- Growth of linking styles that emphasize connection between people, not just documents