Questions

1. What does the web look like? [Ch 13]
2. How does Google search it? [Ch 14]
Information network

- Common elements of social and economic network
  - Graphs, paths, giant components
  - Connections to matching markets and auctions

Web

- Application for sharing info over the Internet
- Created by Tim Berners-Lee (1989)

- 2 perspectives
  - Web pages: Make documents easily available to anyone on the Internet
  - Browser: Retrieve and display documents
Web

- Web organizes information in a unique fashion
- Different from library system
- Different from folders in a computer
- Different from indexing
- Hypertext

Hypertext

- Replaces linear structure of text by pointers
- Concept dates back to 1950s
Precursor to hypertext

- Citation network

- Semantic network
Precursor to hypertext

- Vannevar Bush (1945)
  - Associative memory in “Memex”
  - Cited by Tim Berners-Lee

Evolution of the web

- Navigational functions (1990s)
  - Static web pages
- Transactional functions
  - Dynamic, real-time operations
- Web 2.0
  - New attitude to technology, not new technology
    1. Collective creation and maintenance of shared content (Wikipedia)
    2. Move personal data to corporate servers (Gmail)
    3. Network among individuals, not just web pages (Facebook)
Web as a directed graph

- Nodes: Web pages
- Directed edges: Links
- `bowdoin.edu` → Arts → Museum of Art → Exhibitions → … → `bowdoin.edu`
- A directed cycle

Example
Strongly Connected Component (SCC)

Bow-tie structure of the web
Link analysis and web search
Chapter 14

Web search

- Google “Bowdoin”
  - What do you see?
  - Why is Bowdoin College ranked first? (Why not James Bowdoin?)
- Google's source of information is the web itself
  - No expert intervention
- There must be enough information intrinsic to the web!
Information retrieval

- 1960s: Search repositories of newspapers, patents, etc. by keywords
  - Done by specialized people
- Challenges in web search
  - Synonymy: scallion vs. onion
  - Polysemy: jaguar (you mean the animal or the car or the football team?)
  - Search results must be dynamic
  - Abundance of information (opposite of needle-in-haystack)

Ranking algorithms

- Voting by in-links
- PageRank
- Hubs and authorities
Voting by in-links

- Highest in-degree node is ranked first, and so on...

PageRank
Modern web search

- Google, Bing, (Yahoo!, Ask)
- PageRank is a central ingredient of Google
  - There are more ingredients

Modern web search

- Combination of links, text, and clicks
  - Anchor text: “I’m a student of Bowdoin College.”
- Moving target
  - Google’s changes in algorithm causes millions of dollars of damage to many companies
  - Companies seek help from SEOs to climb up the ranking
    - “white hat” vs. “black hat” optimization (later)
PageRank (PR) (1998)

- Intuition
- Update rule
- Demo
  - NetLogo