



DCS/CSCI 2350:
Social & Economic Networks

*Information Networks:
WWW and PageRank
Chapters 13, 14*

Mohammad T. Irfan

1

Questions

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



3

Web

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



5

Web's organization of information

- Different from library system
- Different from folders in a computer
- Different from indexing

Hypertext

6

Inception of hypertext

Vannevar Bush (1945)

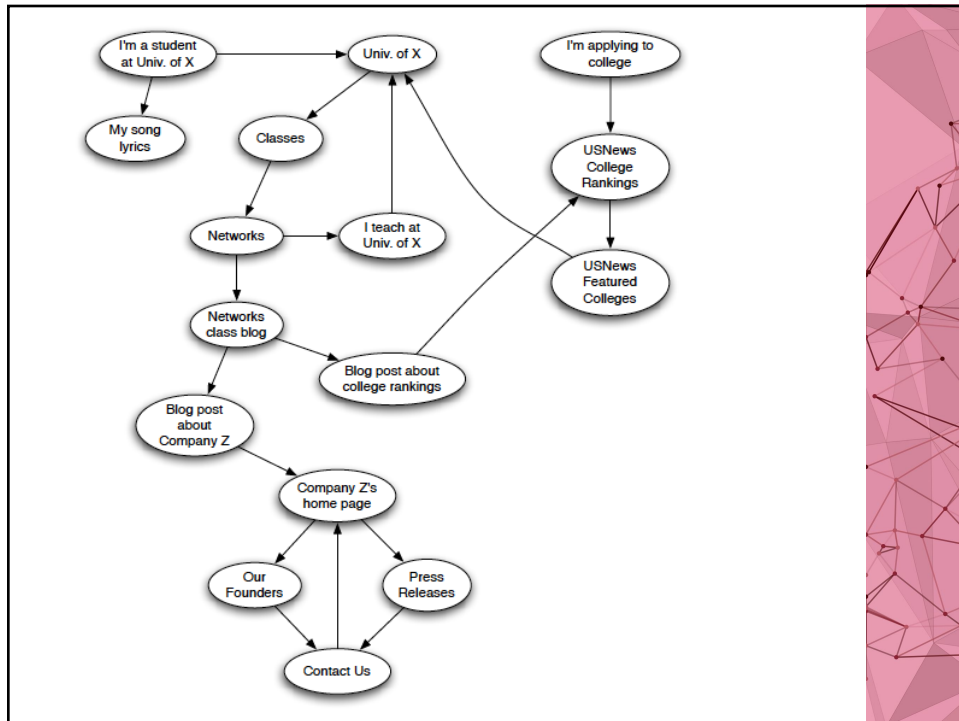
- “As we may think” – check out Canvas
- Associative memory in “Memex”
- Cited by Tim Berners-Lee

10

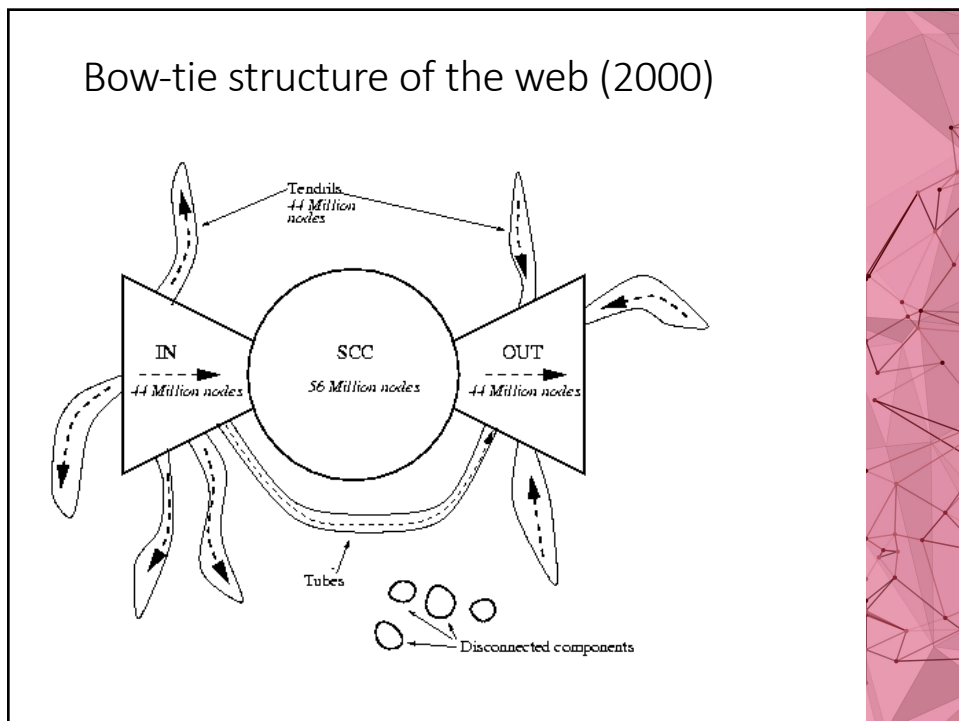
Web as a directed graph

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

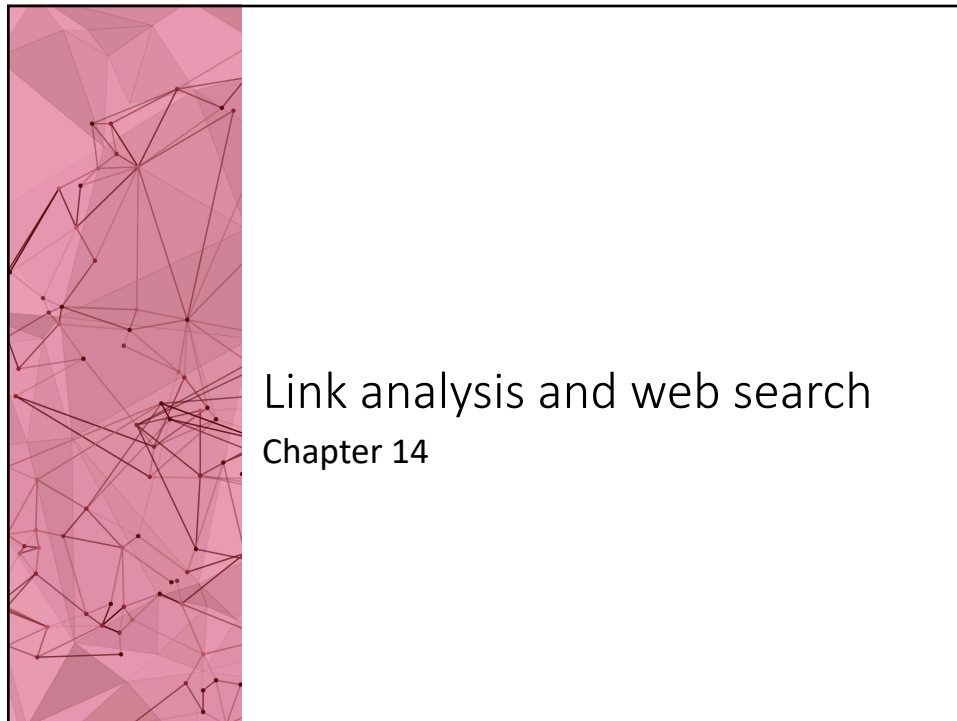
12



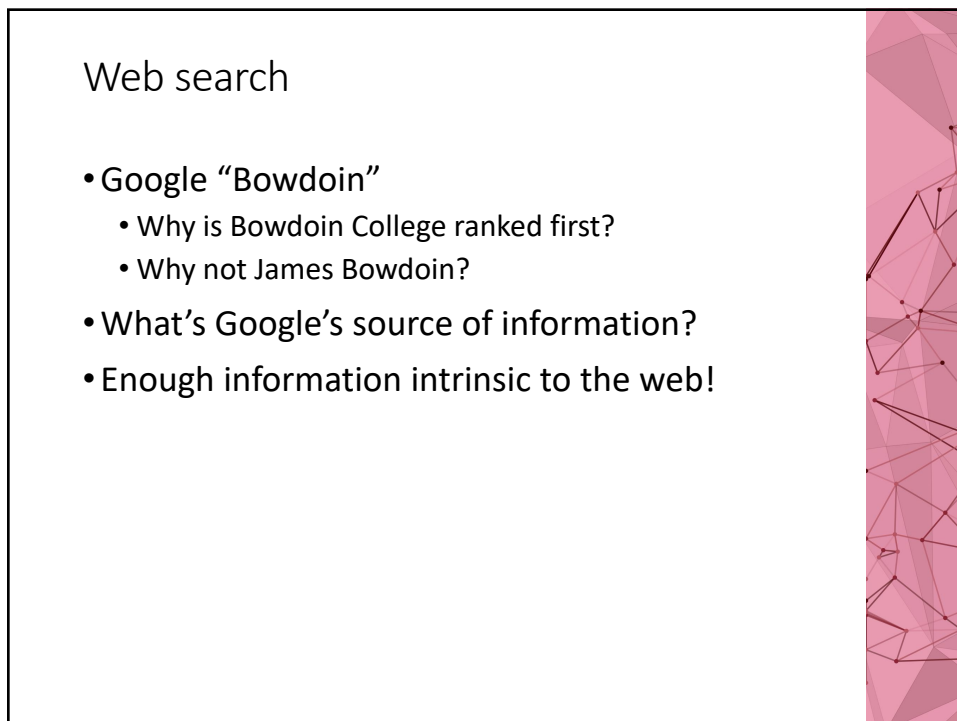
13



16

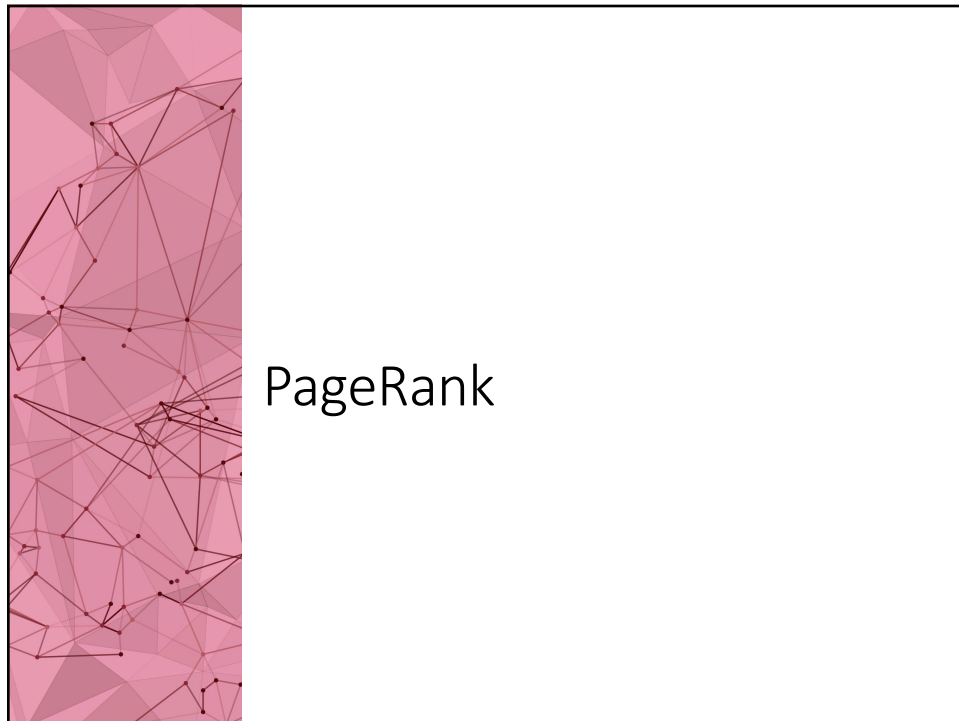


17

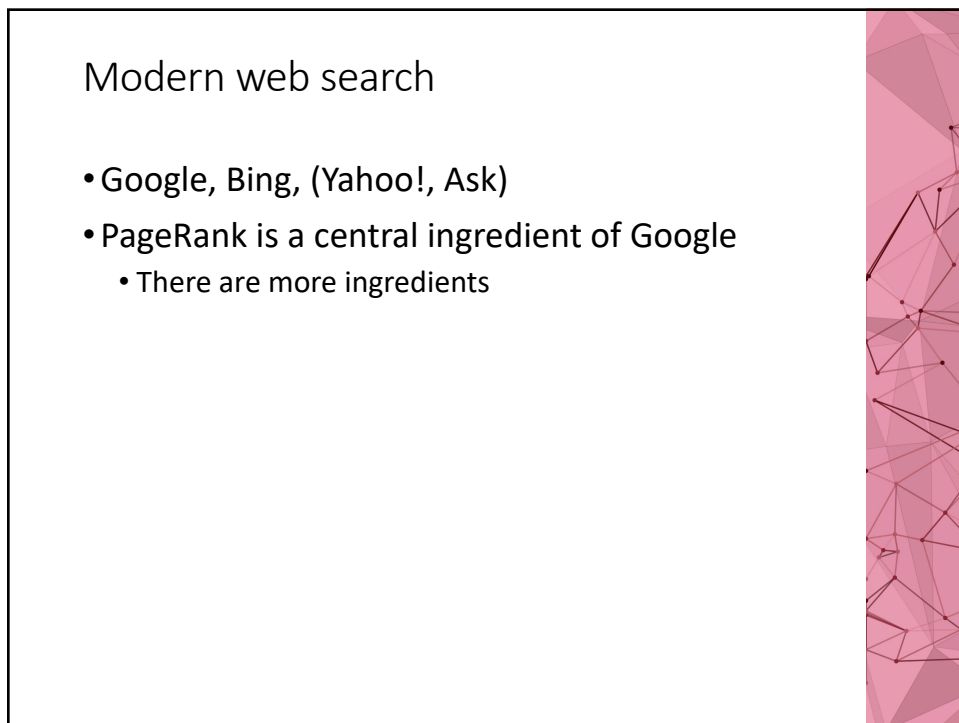


- Google "Bowdoin"
 - Why is Bowdoin College ranked first?
 - Why not James Bowdoin?
- What's Google's source of information?
- Enough information intrinsic to the web!

18



23



24

Modern web search

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

PageRank (PR) algorithm, 1998

- Idea
- NetLogo demo
- Update rule

26

Idea

- A webpage is important if it is cited by other important webpages
 - Bonacich's idea on centrality (1987)
- Iteratively refine the PR of each webpage

27

Demo

Netlogo -> Models Library -> Computer Science -> PageRank

29

PageRank (PR) algorithm

- Input: directed network with **n nodes** and desired **number of rounds k**
- Steps
 1. Assign each node initial $PR = 1/n$
 2. Repeat for k rounds:
 - **Out-Phase:** Each node divides its current PR equally across its outgoing links and passes these equal shares to the nodes it points to.
 - **In-Phase:** Each node replaces its PR with the sum of the shares it receives.
- Q: What if PR values do not change between two consecutive rounds?

30

Sufficient conditions for convergence

- Network is strongly connected
 - There's a directed path from any node to any other node
- Network is aperiodic
 - GCD of all cycle lengths = 1

31

Equilibrium interpretation of convergence

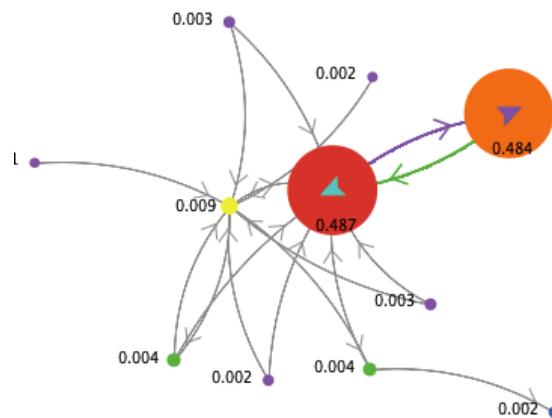
Convergence means doing another round of Out-Phase and In-Phase will not change PR

- Stable outcome or equilibrium

32

Slow leak problem

PR getting trapped into a few nodes due to lack of outlets



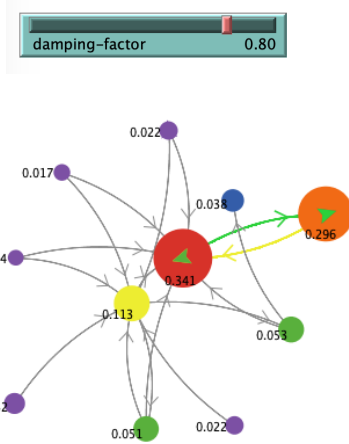
33

Solution to slow leak

- Intuition: Why doesn't all the water on earth get trapped into the lowest point on earth?
- **Scaled update rule**
 - Scaling parameter (or damping factor) s ($0 < s < 1$)
 - Scale all the PR by s (sum of PR = s)
 - $(1 - s)$ evaporates
 - Rain down $(1 - s)$: equally distribute $(1 - s)$ to all nodes

34

Solution in NetLogo



35

JCPenney scandal (2011)



36

How JCPenney did it

- Hired SearchDex
- Black hat optimization

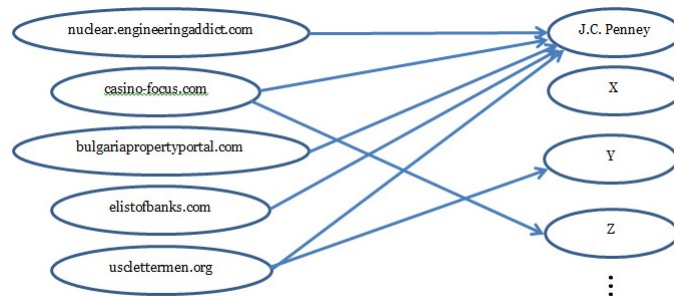


Image source: <http://blogs.cornell.edu/info2040/2011/11/03/j-c-penney%E2%80%99s-pagerank/>

37

How they got caught

- NY Times + Blue Fountain Media
- Punishment (Feb 9, 2011)
 - 7 pm: J. C. Penney was still the No. 1 result for "Samsonite carry on luggage."
 - 9 pm: It was at No. 71.
 - Similar with other keywords
- Another case: BMW in Germany (2006)



Google's spam cop
Matt Cutts

(Image source: NY Times)

38