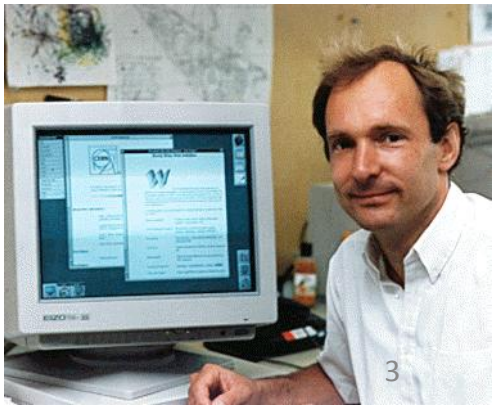
The Structure of the Web

Objectives

- So far: networks with connected people or other social entities
- Next: information networks with connected are pieces of information
- Similarities and differences between the two different types of networks
- WWW as information network
  - ideas, history, structure

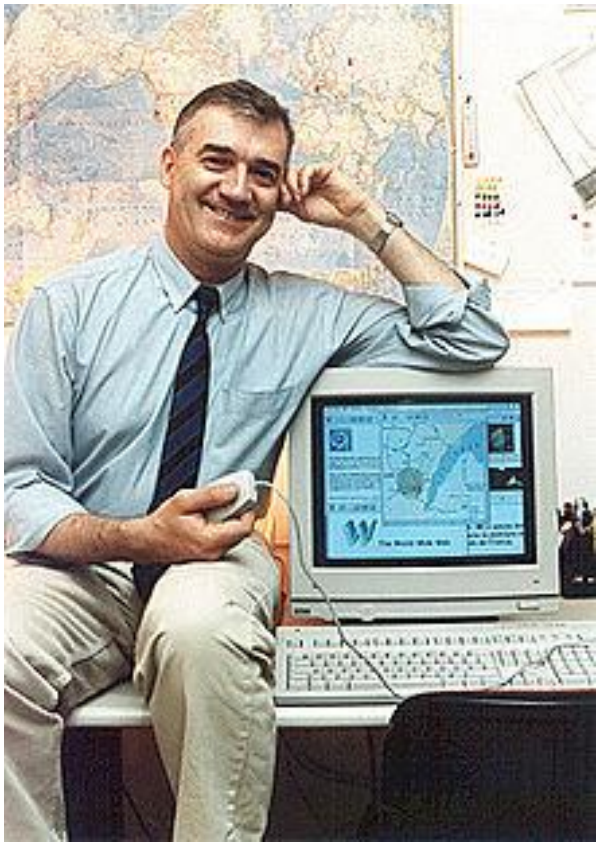
Emergence of the World Wide Web

- A collection of information stored on the networked computers over the world
- The WWW was proposed in 1989 by Tim Berners-Lee at CERN
  - code for a hypertext server program
    - Hypertext server:
      - Stores files written in hypertext markup language
      - Lets other computers connect to it and read files
    - Hypertext Markup Language (HTML)
      - Includes a set of codes (or tags) attached to text



Co-inventor of WWW

- Robert Cailliau, born 26 January 1947, is a Belgian informatics engineer and computer scientist who, together with Sir Tim Berners-Lee, developed the World Wide Web



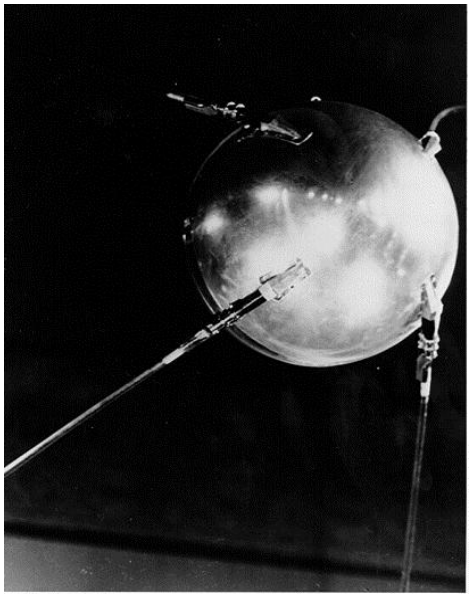
From ARPANET to Internet and to WWW

<HISTORY>

Creation of ARPANET

- 1957 – USSR launched Sputnik I

  United States were shocked
- Advanced Research Projects Agency
  - Thechnological think-tank
  - Space, ballistic missiles and nuclear test monitoring
  - Communication between operational base and subcontracters

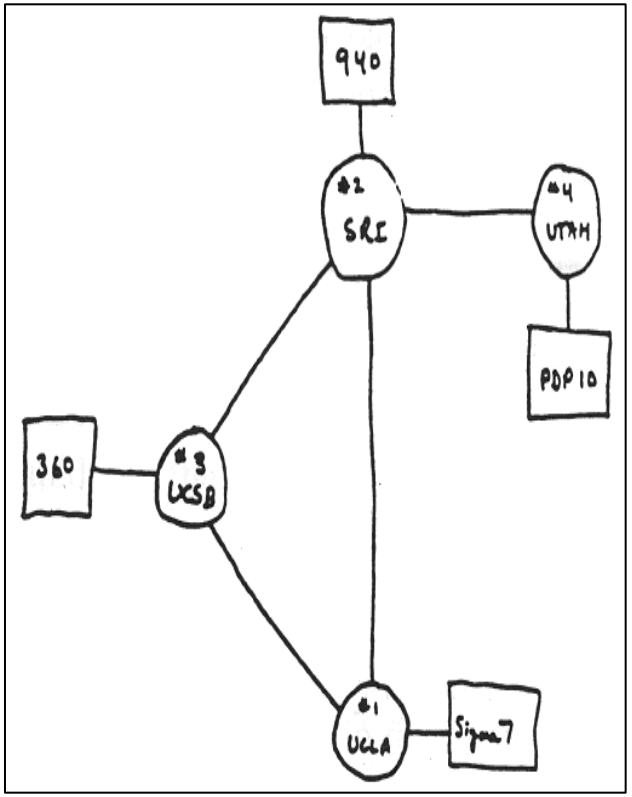


Creation of ARPANET

- 1962 – computer research program
  - Leaded by John Licklider (MIT)
  - Leonard Kleinrock published his first paper on packet-switching theory
- 1965 – first "wide area network" created
  - Connection between Berkeley and MIT

Creation of ARPANET

- 1967 – plans for ARPANET were published
  - MIT – NPL (UK) – RAND
- 1969 – Interface Message Processor (IMP)
  - 4 computers (UCLA, SRI, UCSB and UTAH)
- 1971 – 23 host computers (15 nodes)



From ARPANET to Internet

- 1972 – ARPANET went ‘public’
  - ICCC
  - First program for person-to-person communication (e-mail)
- 1973
  - 75% of all ARPANET traffic is e-mail
  - First international connection (University College of London)

From ARPANET to Internet

- 1974 – TCP/IP
  - Each network should work on its own
  - Within each network there would be a ‘gateway’
  - Packages would be routed through the fastest available route

  - Large mainframe computers
  - Several years of modification and redesign

From ARPANET to Internet

- 1974/1982 – Networks launched
  - Telenet – first commercial version of ARPANET
  - MFENet – researchers into Magnetic Fusion Energy
  - HEPNet – researchers into High Energy Physics
  - SPAN – space physicists
  - Usenet – open system focusing on e-mail and newsgroups
  - Bitnet – university scientists using IBM computers
  - CSNet – Computer Scientists in universities, industry and government
  - Eunet – European version of the Unix network
  - EARN – European version of Bitnet

From ARPANET to Internet

- 1974/1982
  - Very chaotic
  - Different competing techniques and protocols
  - ARPANET is still the backbone

- 1982 – The internet is born using the TCP/IP standard

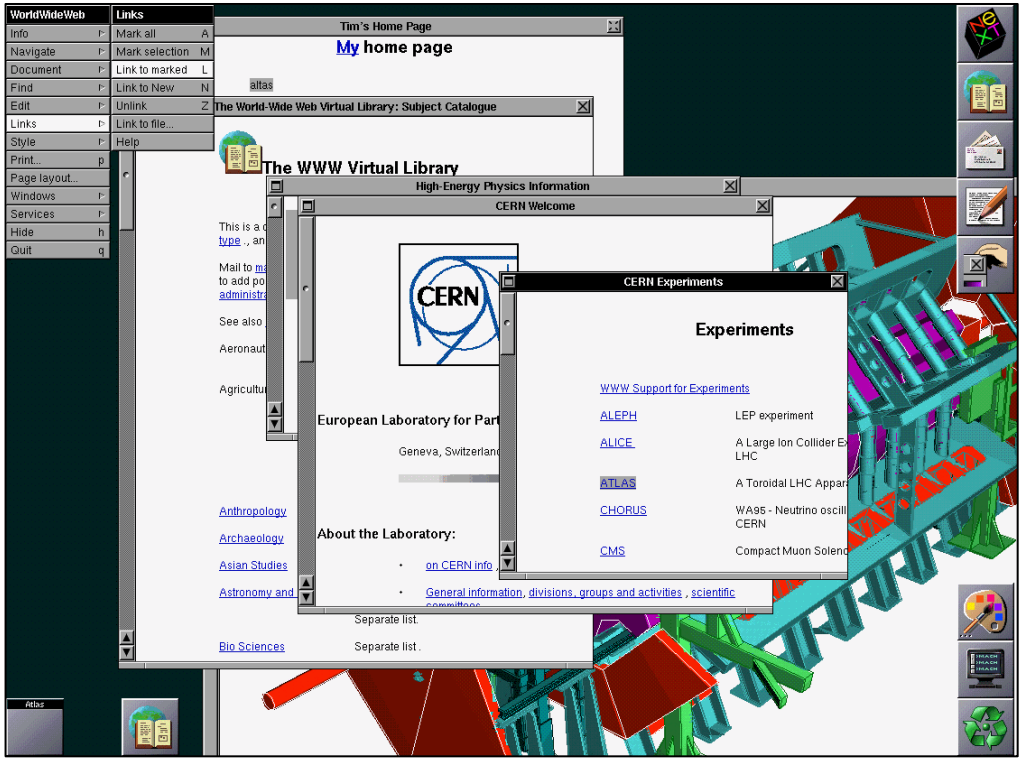
From Internet to WWW

- System expands
  - Advances in computer capacities and speeds
  - Introduction of glass-fibre cables
- Problems created by its own success
  - More computers are linked (1984 – 1000 hosts)
  - Large volume of traffic (success of e-mail)

- 1984 – Introduction DNS

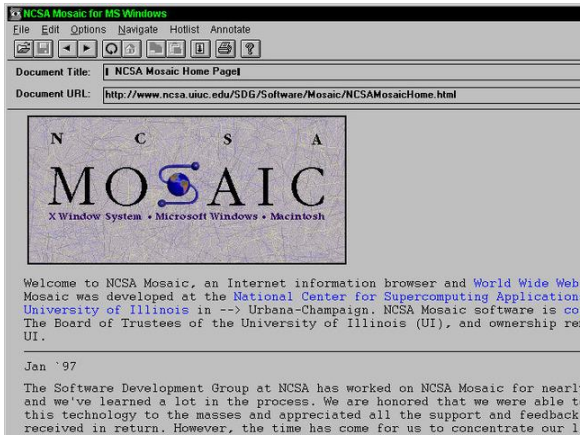
From Internet to WWW

- 1989 – WWW concept by Tim Berners-Lee
- 1990 – first browser/editor program



From Internet to WWW

- National Center for SuperComputing Applications launched Mosaic X
- Commercial websites began their proliferation
- Followed by local school/club/family sites
- The web exploded
  - 1994 – 3,2 million hosts and 3,000 websites
  - 1995 – 6,4 million hosts and 25,000 websites
  - 1997 – 19,5 million hosts and 1,2 million websites
  - January 2001 – 110 million hosts and 30 million websites
  - December 2010 – 200 million hosts 255 million websites



http://royal.pingdom.com/2011/01/12/internet-2010-in-numbers/

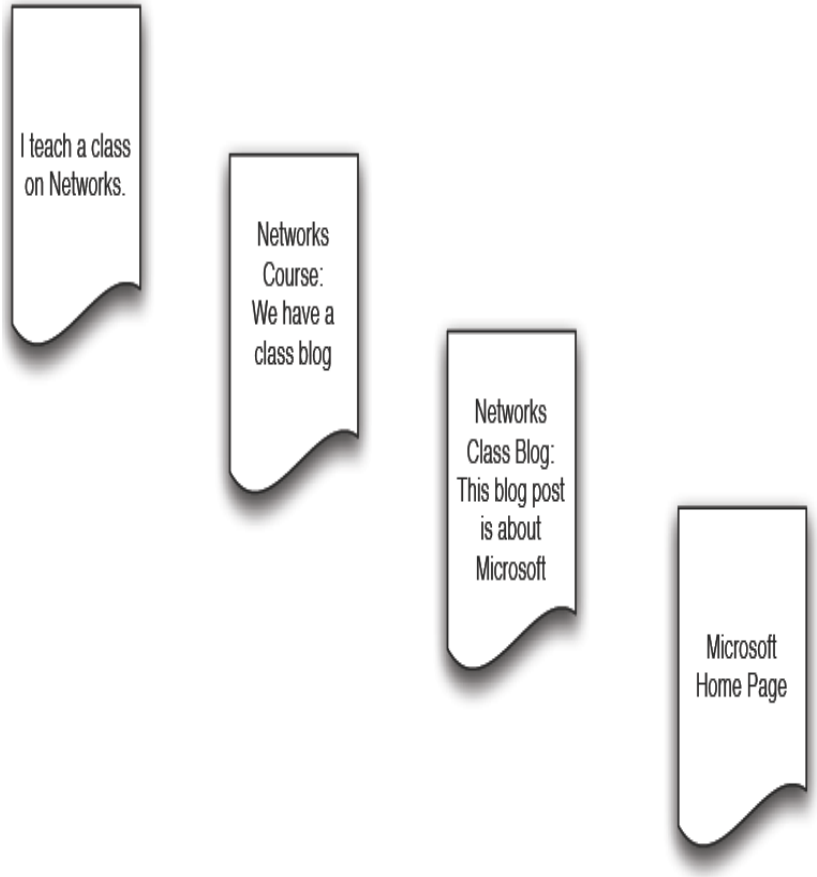
From Internet to WWW

- Who defines the Web standards?
  - The Web standards are defined by the World Wide Web Consortium (W3C)
- The specifications form the Web standards.
  - HTML, CSS, XML, XHTML, …

</HISTORY>

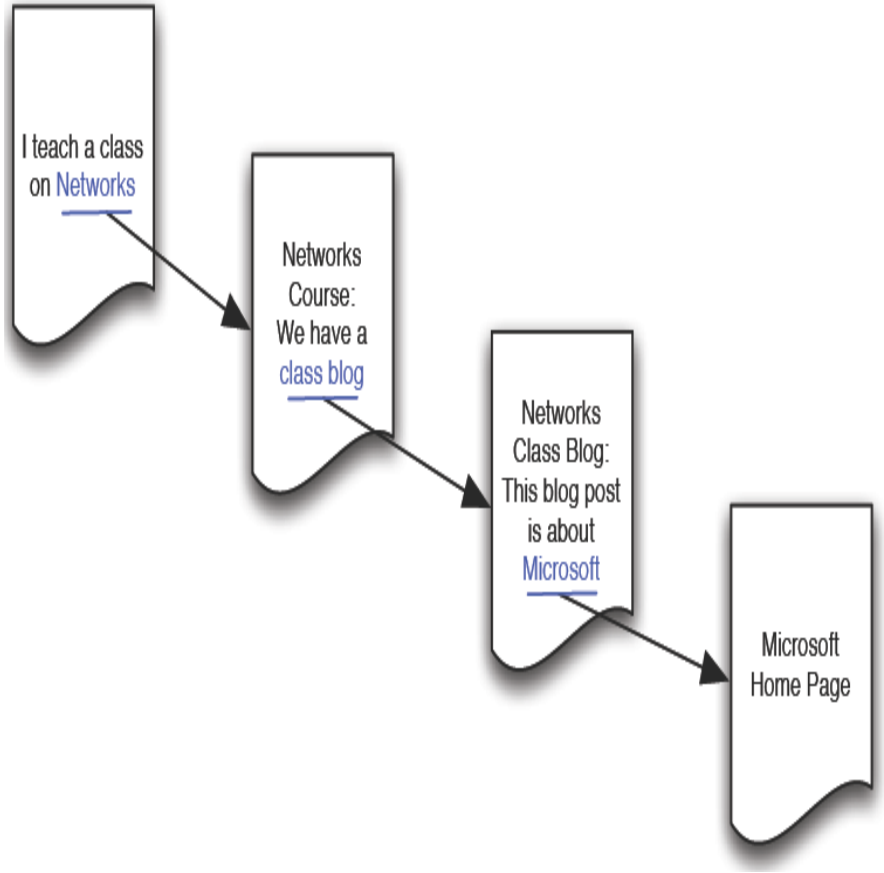
WWW as pages and browsers

- Approximation, but this is still how we experience the Web today
- Ex:
  - home page of a college instructor who teaches a class on networks; the home page of the networks class he teaches; the blog for the class, with a post about Microsoft listed at the top; and the corporate home page for Microsoft
  - pages as part of a single coherent system (WWW)
  - pages files on four separate computers, controlled by several organizations, and publically accessible through Web browsers



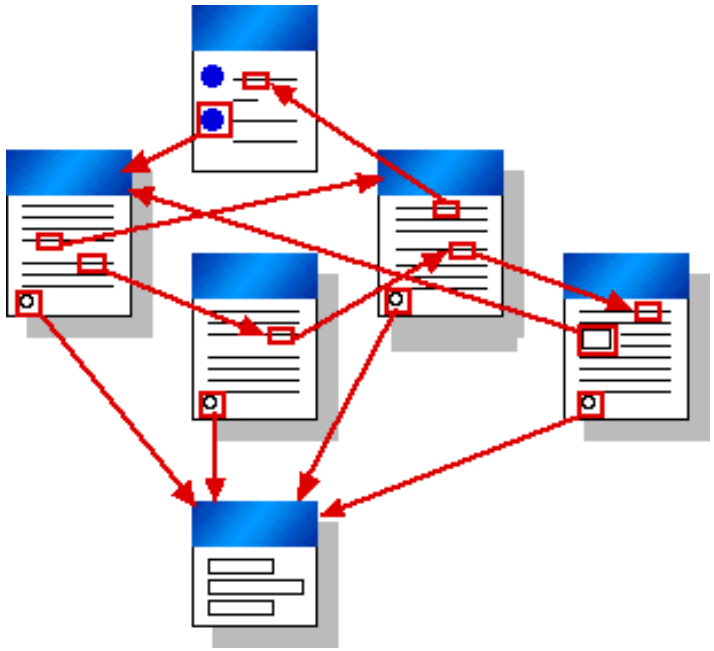
WWW as Information Network

- Hypertext: annotate any portion of a Web page with a virtual link to another Web page
- Network structure of WWW
  - inspired and non-obvious idea
    - alternatives: hierarchy of folders (PC); alphabetically or indexed (phone directory, libraries)
  - globalizing power WWW
    - highlight relationship with any other page, anywhere in world
- *How did we get the idea for hypertext?*



Hypertext

- Replace traditional linear structure of text with a network structure
  - any portion of a text linking to any other part
- Web brought hypertext to a global audience
- Web is the largest information network today

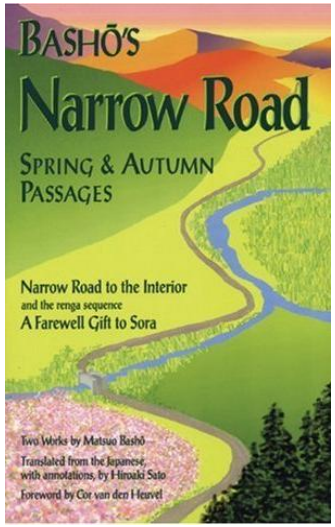


Hypertext

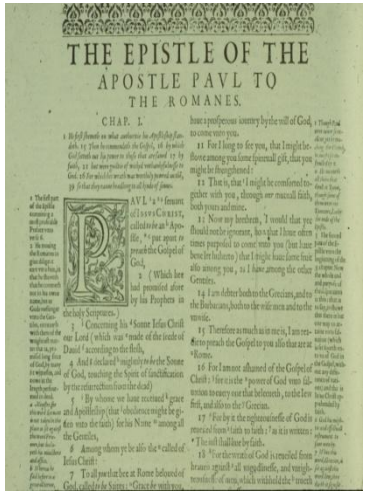
- Ted Nelson coined the term hypertext
  - the concept behind WWW links
  - Nelson influenced from film-making and media
- Xanadu project:
  - robust two-way hyperlinks, version management, controversy management, annotation and copyright management
- Nelson considers WWW an over-simplification
  - "*HTML is precisely what we were trying to PREVENT—ever-breaking links, links going outward only, quotes you can't follow to their origins, no version management, no rights management*" -Nelson

Intellectual Precursors of Hypertext

Japanese linked poetry
Renga and Basho



Western religious commentaries

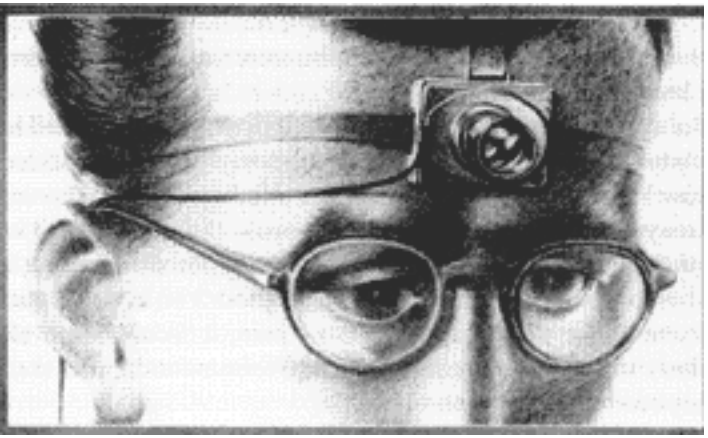

THE EPISTLE OF THE
APOSTLE PAVL TO
THE ROMANES.
CHAP. I.

"The Garden of Forking Paths"

Author	Jorge Luis Borges
Original title	"El jardín de senderos que se bifurcan"
Translator	Anthony Boucher
Country	Argentina

Intellectual Precursors of Hypertext

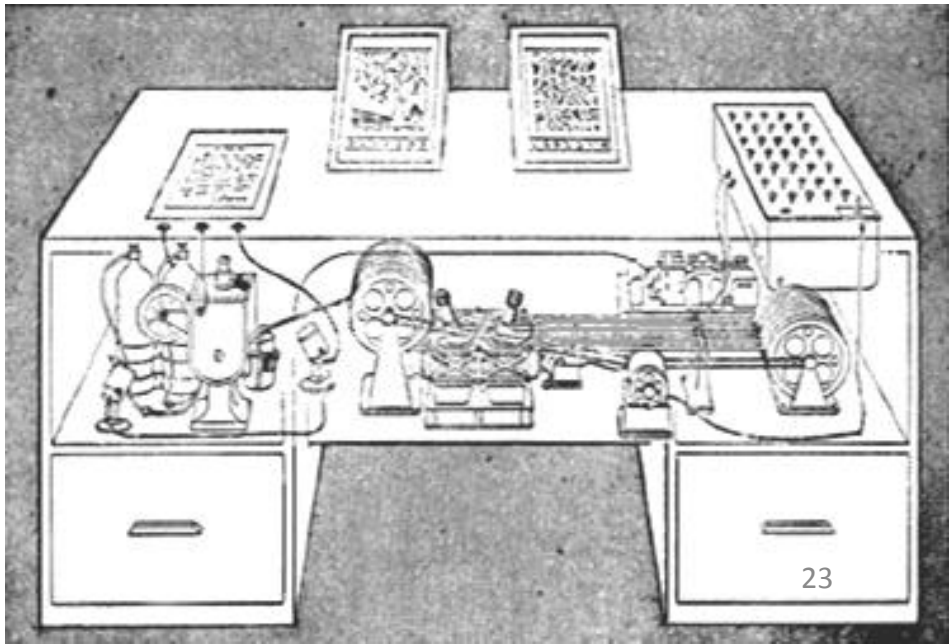
Vannevar Bush – As We May Think



Memex (1945)

"As We May Think", Vannevar Bush in The Atlantic Monthly, 1945

- Hypothetical proto-hypertext system
- A device to stores books, records, and communications
  - A desk (operated also from a distance), with screen, keyboard, and microfilms
- Indexed repository of knowledge any section of which could be called up with a few keystrokes



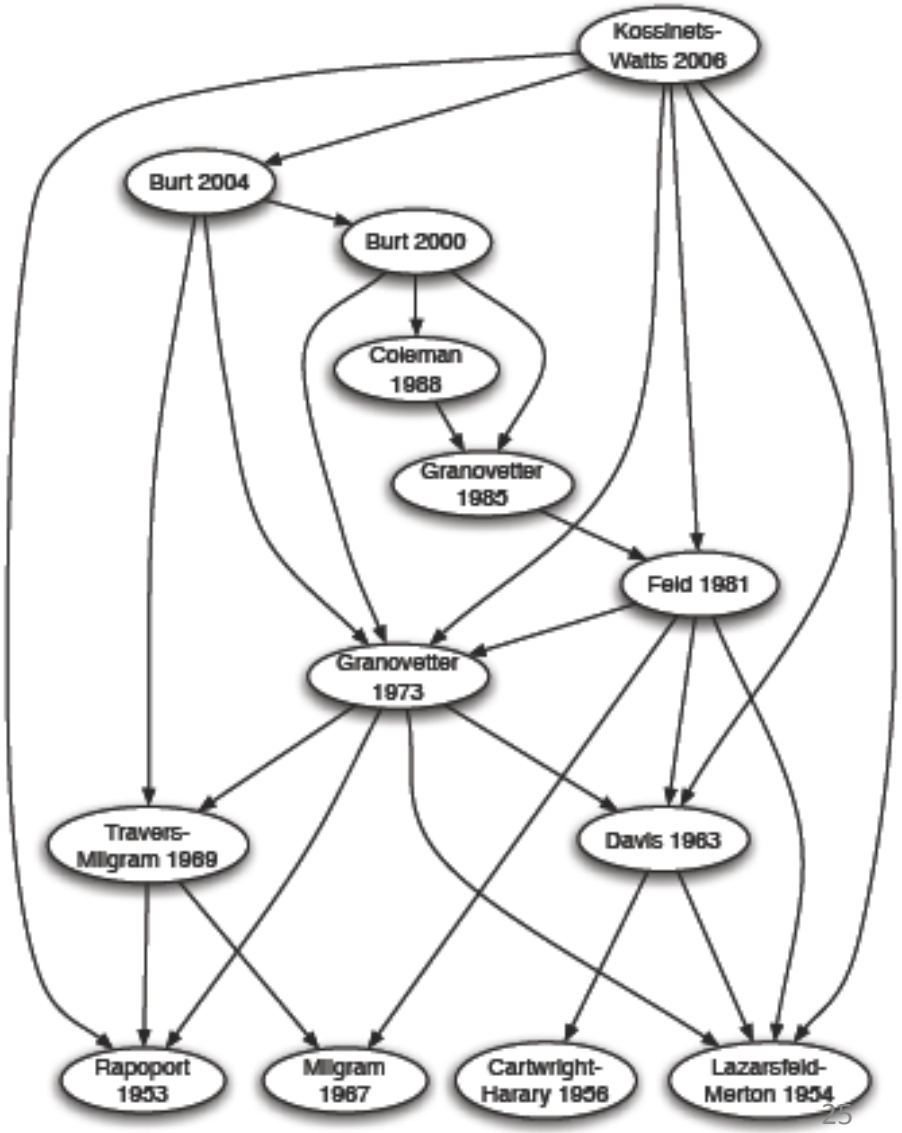
Intellectual Precursors of Hypertext

Vannevar Bush – As We May Think

- An associative trail creates a new *linear* sequence of microfilm frames across any arbitrary sequence of microfilm frames by creating a chained sequence of links in the way just described, along with personal comments and *side trails*
- Store information that was analogous to the mental association of the human brain
- Memex functioned very much like the Web
  - the Web as universal encyclopedia
  - the Web as giant socio-economic system
  - the Web as global brain
  - the Web as human collaborative platform (Web 2.0)

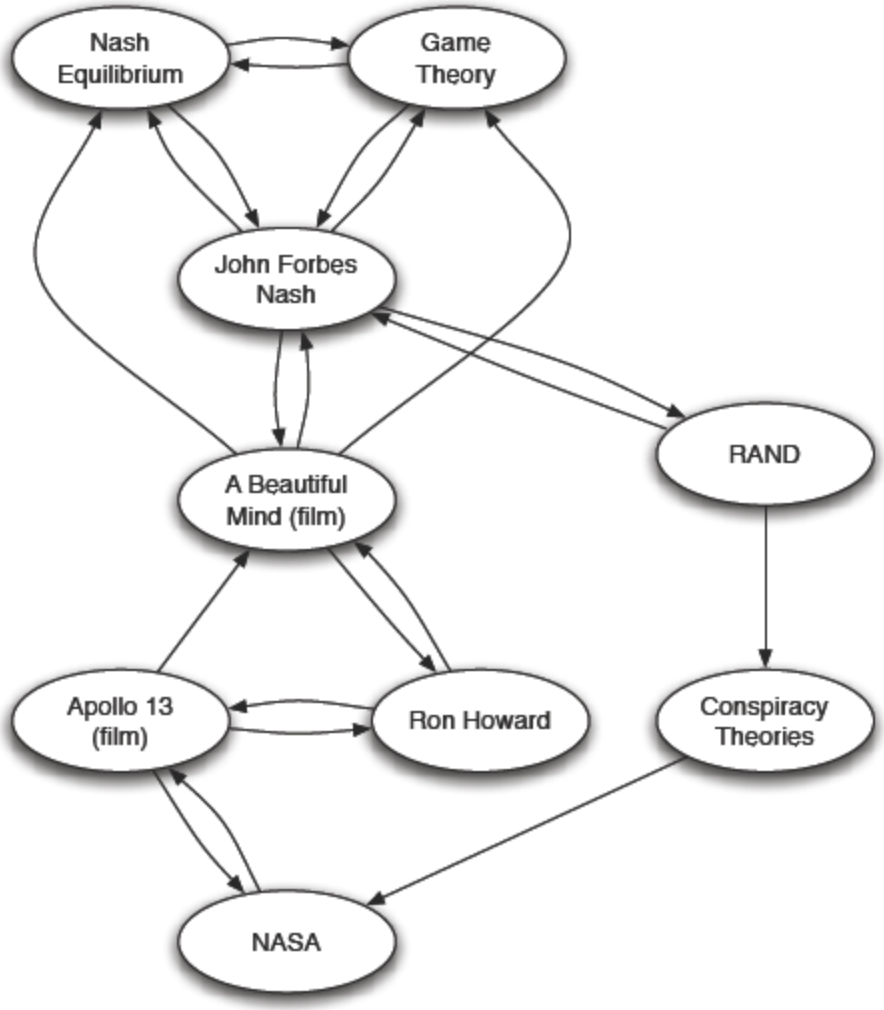
Intellectual Precursors of Hypertext

- Citations among scholarly books and articles
- Citation graphs
  - directed
  - "arrow of time" (Web doesn't have it)



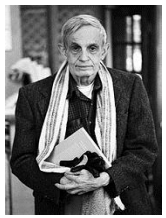
Intellectual Precursors of Hypertext

- Cross references within an encyclopedia
  - printed encyclopedias
  - Wikipedia (independent of the fact that it exists on the Web)
- Ex (figure):
  - articles on game theory + referenced articles
  - supports serendipitous browsing

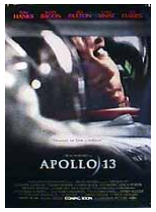


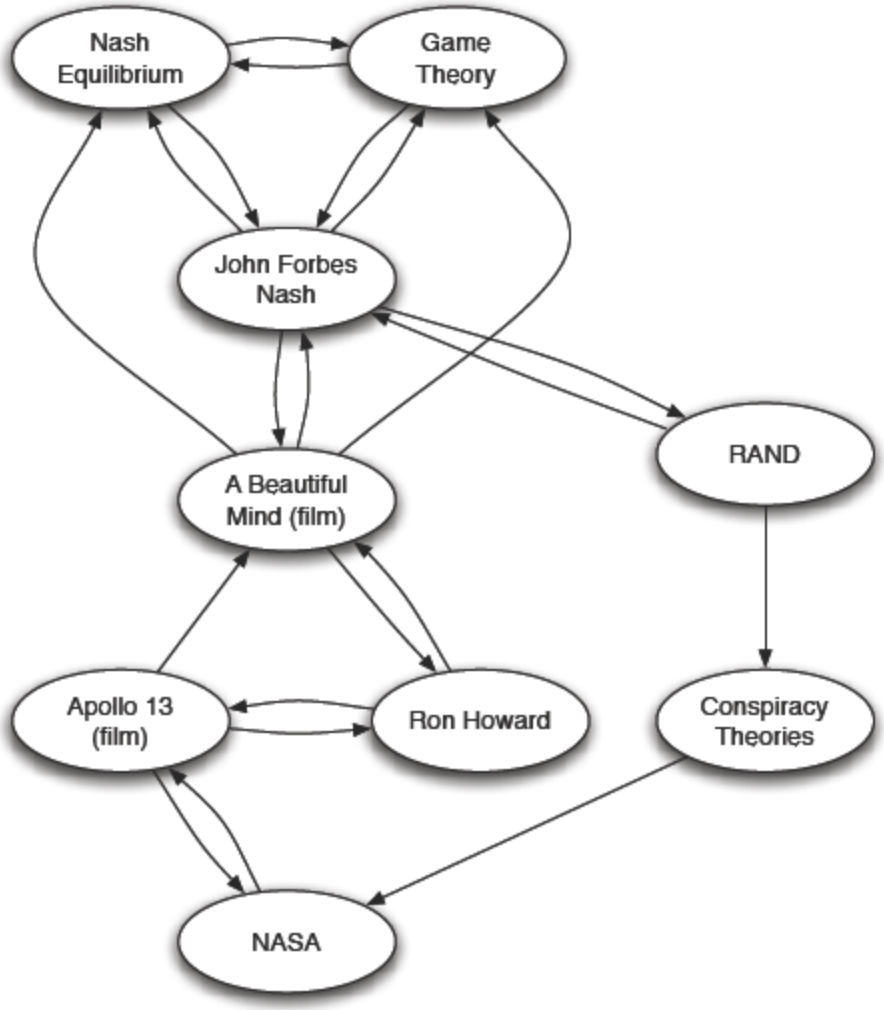
Information Networks and Serendipity

- From Nash Equilibrium to NASA:
  - Nash equilibrium was created by Nash whose life was the subject of a movie (“Beautiful Mind”) by a director (Ron Howard) who also made a movie about NASA (“Apollo 13”)



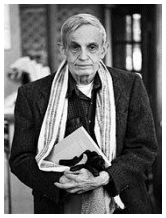


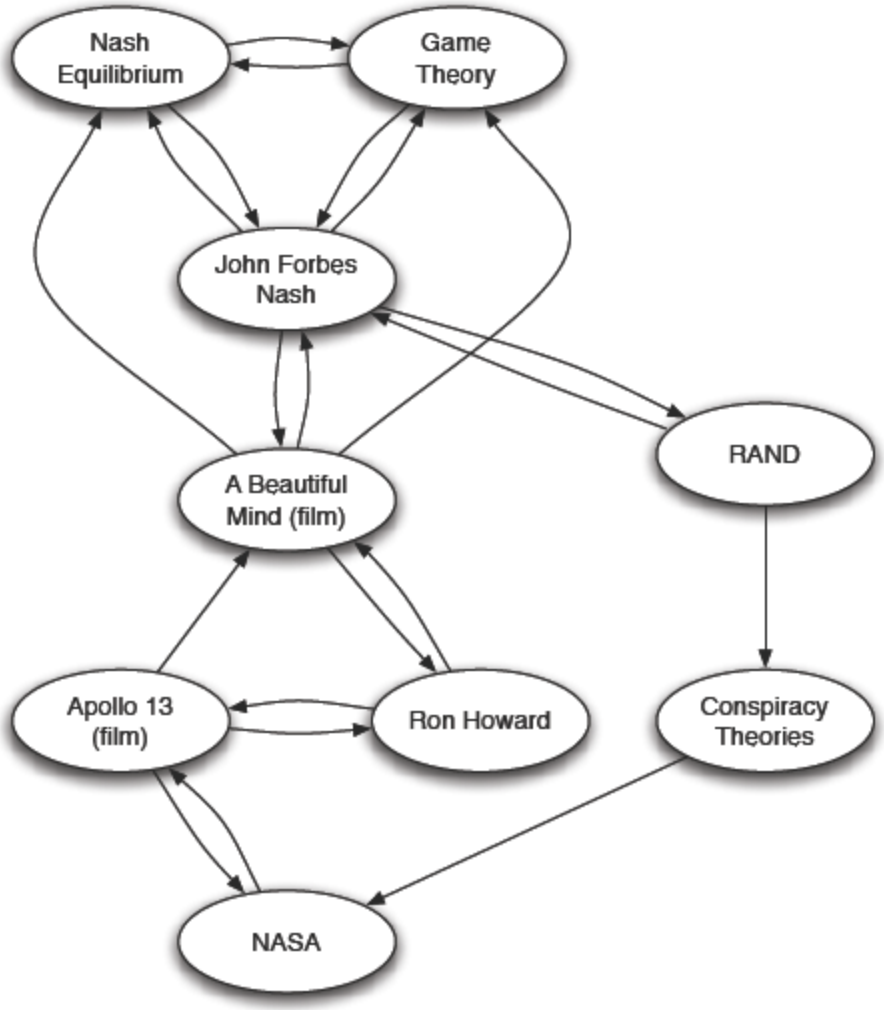




Information Networks and Serendipity

- From Nash Equilibrium to NASA (second path):
  - John Nash worked for a period of time at RAND (policy think tank of USAF for research and analysis), and RAND is the subject of several conspiracy theories, as is NASA





Information Networks and Serendipity

Information Networks

- Short paths between seemingly distant concepts
- Closely related to the stream-of consciousness way in which we mentally free-associates between different ideas
- Word association games ("Tell me what you think of when I say the word 'cold' ")

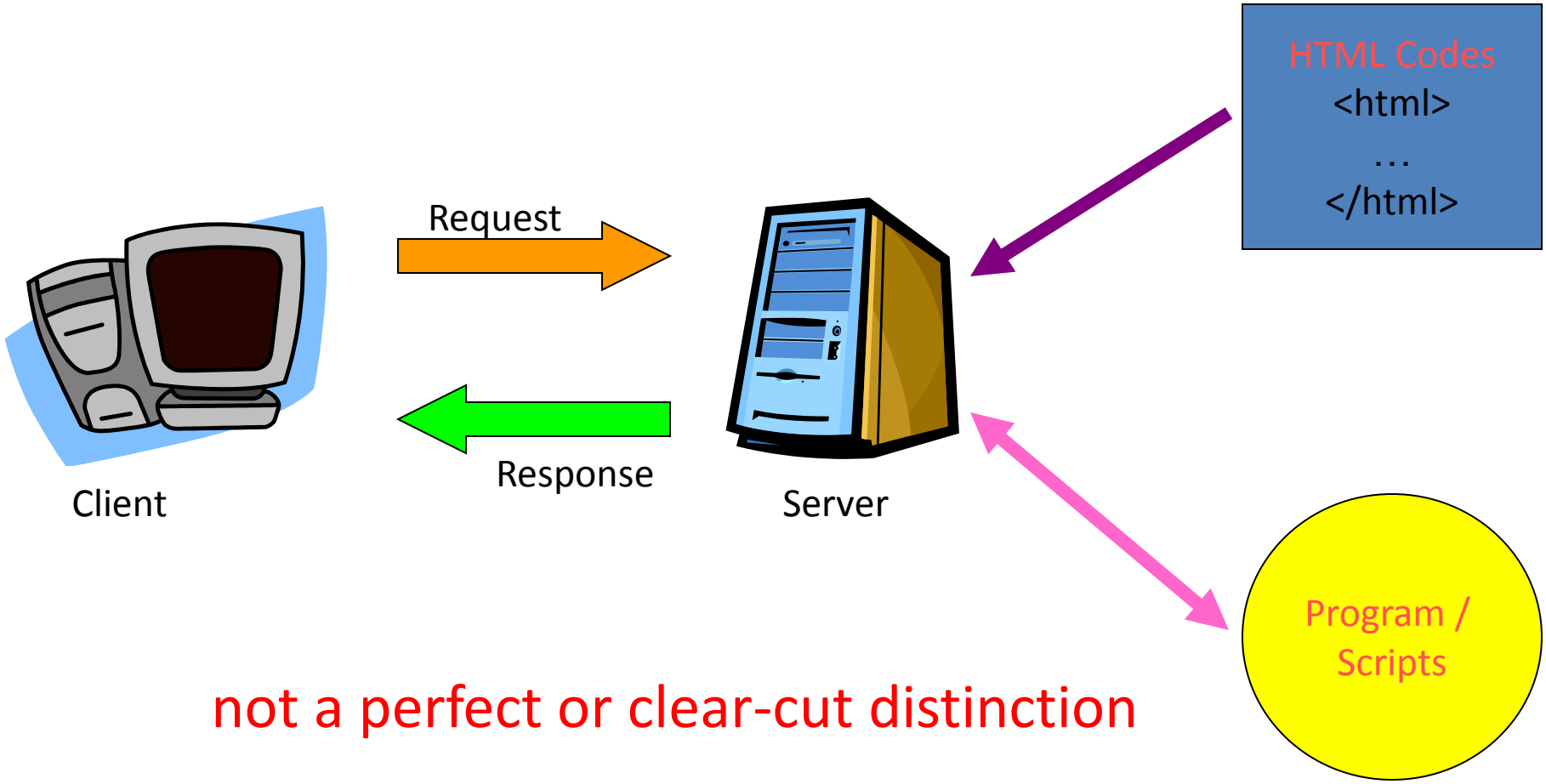
Social Networks

- Similarly short paths link apparently distant pairs of people ("six degrees of separation")

The Web and its Evolution

- In 1990's:
  - most pages relatively static and most links served navigational functions (hypertext)
  - Web servers passive hosting and responding to requests
- Today:
  - links often trigger complex programs on Web servers and activate computational transactions
    - "Add to Shopping Cart", "Submit my Query", "Update my Calendar", or "Upload my Image"
- Ex:
  - "Buy Now" -> receipt page
  - the purpose of "Buy Now" link was not to transport (hypertextually) to "receipt page"; rather to perform the transaction

Navigational vs. Transactional Links

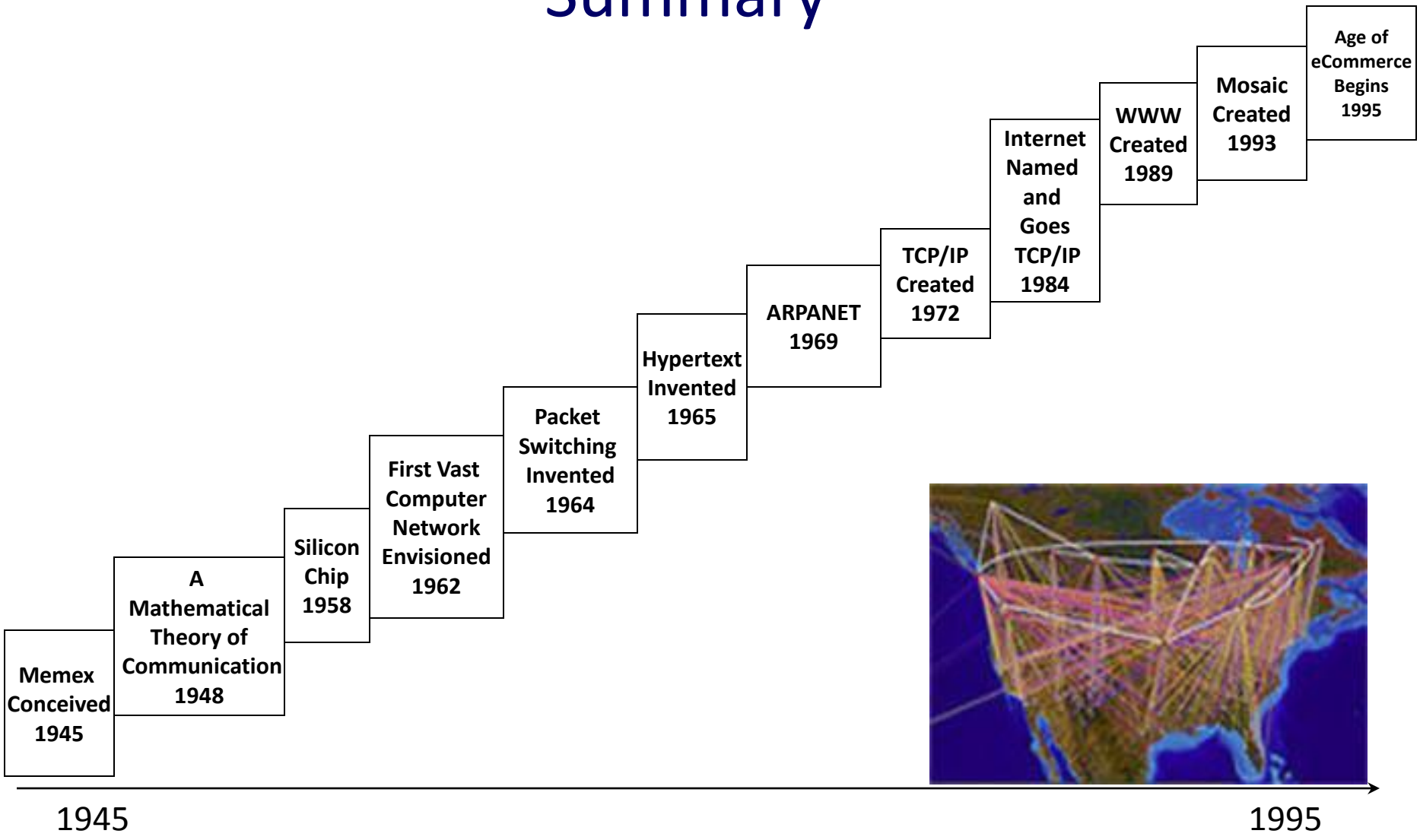


not a perfect or clear-cut distinction

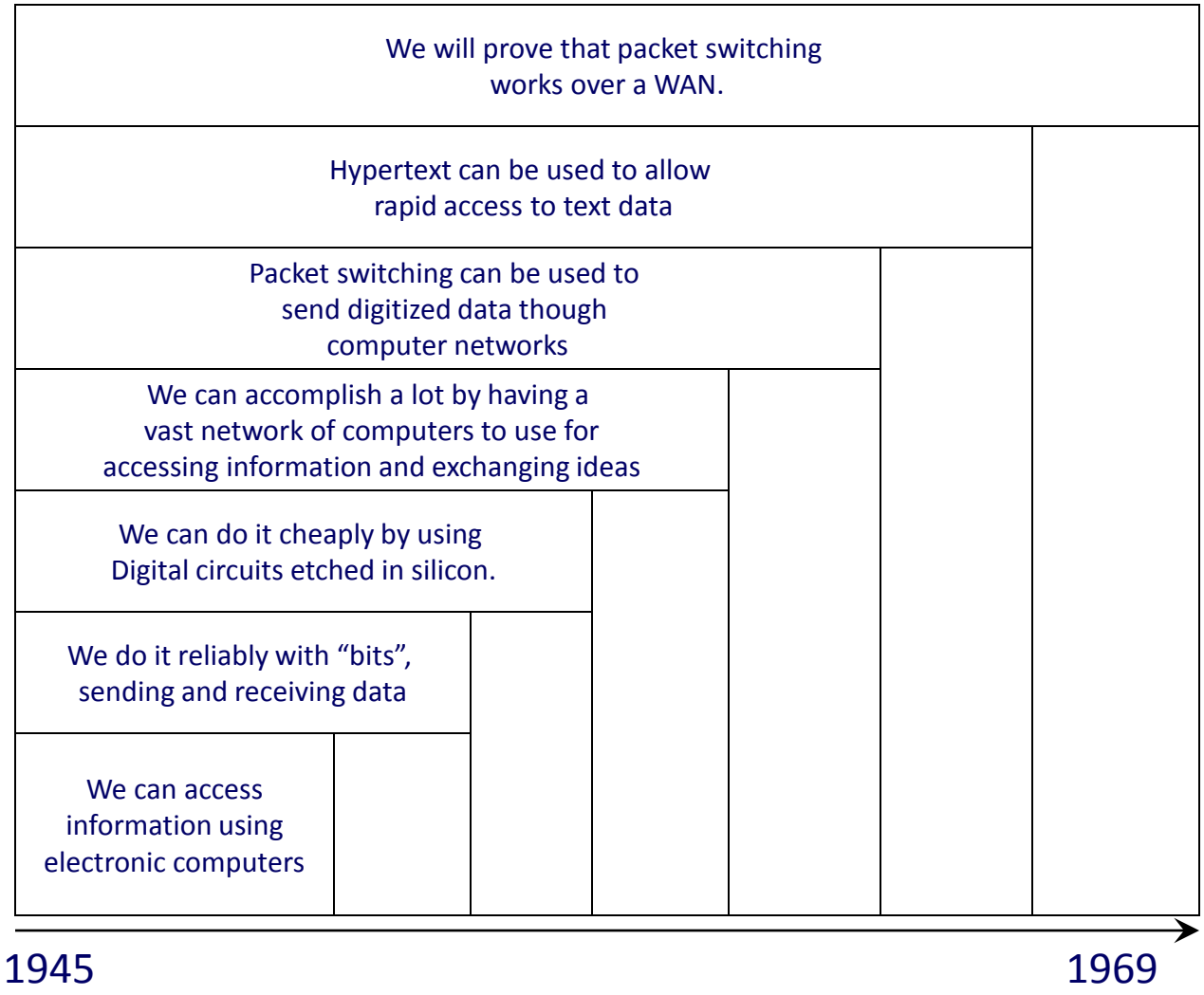
Navigational vs. Transactional Links

- A lot of content on the Web has transactional nature
- This content is linked together by a navigational "backbone"
  - reachable via relatively stable Web pages connected to each other by more traditional navigational links
- For the analysis of Web's global structure we focus on navigational "backbone"
  - search engines index content reachable via navigational links

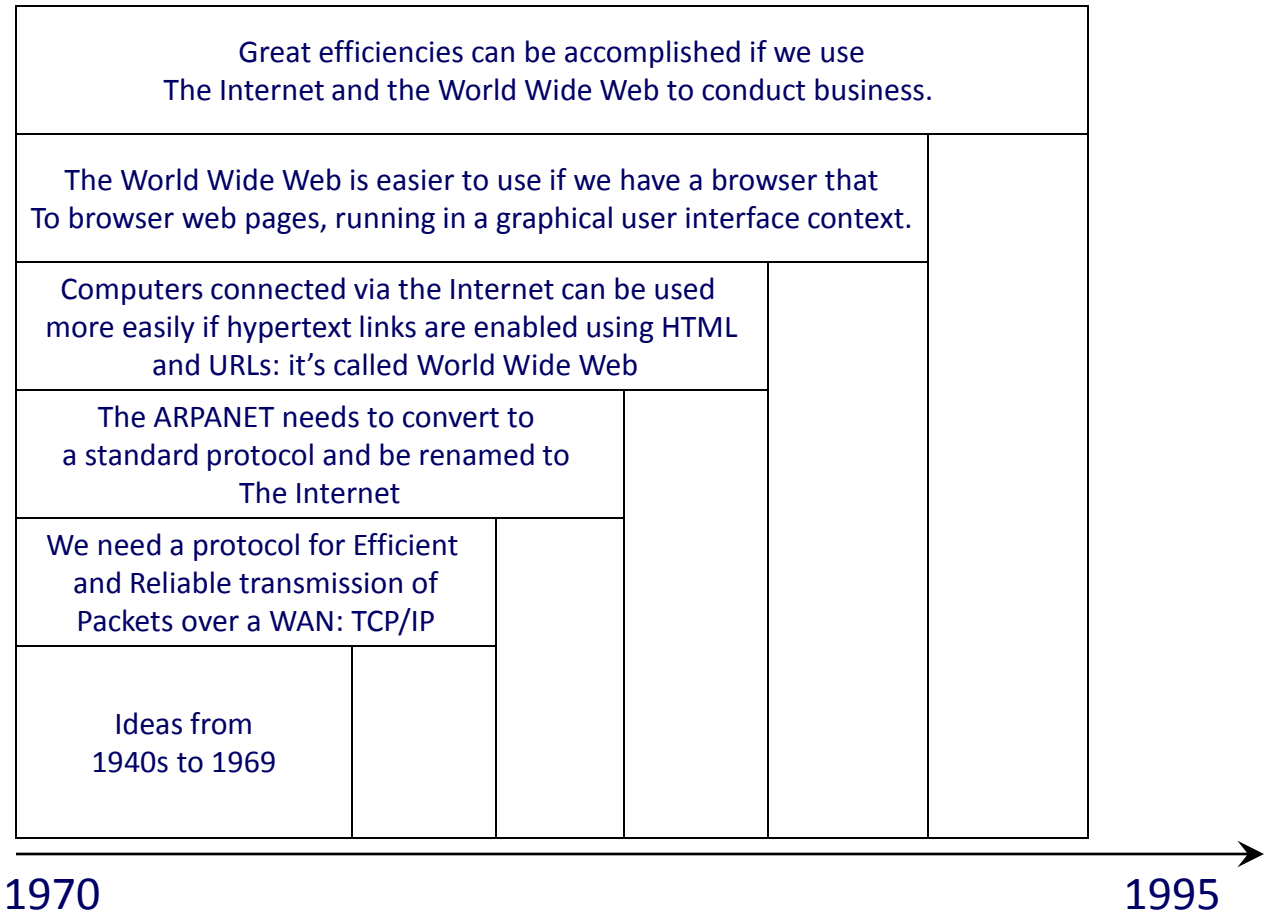
Summary


Memex Conceived 1945
A Mathematical Theory of Communication 1948
Silicon Chip 1958
First Vast Computer Network Envisioned 1962
Packet Switching Invented 1964
Hypertext Invented 1965
ARPANET 1969
TCP/IP Created 1972
Internet Named and Goes TCP/IP 1984
WWW Created 1989
Mosaic Created 1993
Age of eCommerce Begins 1995
1945
1995

Summary

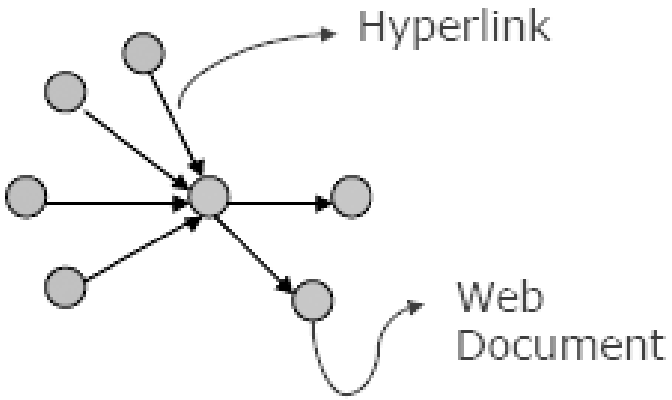

We will prove that packet switching works over a WAN.
Hypertext can be used to allow rapid access to text data
Packet switching can be used to send digitized data though computer networks
We can accomplish a lot by having a vast network of computers to use for accessing information and exchanging ideas
We can do it cheaply by using Digital circuits etched in silicon.
We do it reliably with “bits”, sending and receiving data
We can access information using electronic computers
1945
1969

Summary


Great efficiencies can be accomplished if we use
The Internet and the World Wide Web to conduct business.
The World Wide Web is easier to use if we have a browser that
To browser web pages, running in a graphical user interface context.
Computers connected via the Internet can be used
more easily if hypertext links are enabled using HTML
and URLs: it's called World Wide Web
The ARPANET needs to convert to
a standard protocol and be renamed to
The Internet
We need a protocol for Efficient
and Reliable transmission of
Packets over a WAN: TCP/IP
Ideas from
1940s to 1969
1970
1995

The Web as a Directed Graph

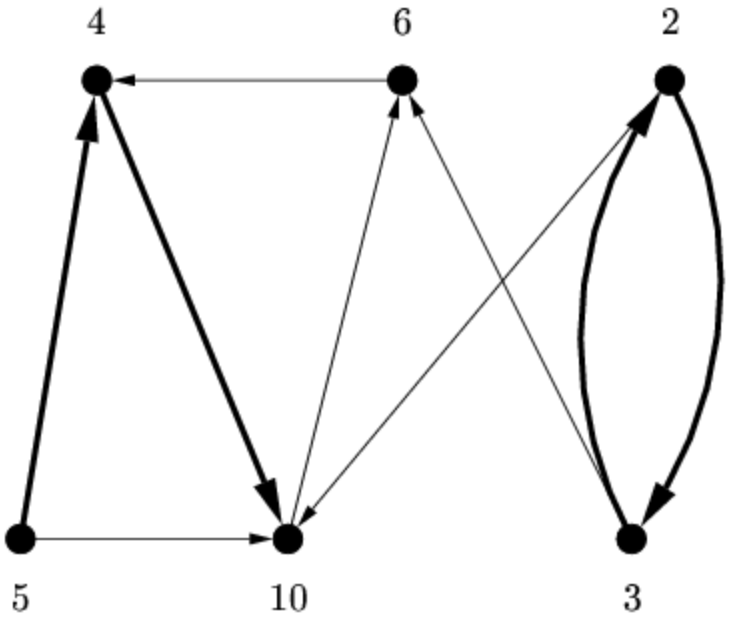
- Nodes: web pages
- Directed edges: navigational links
  - no reciprocal relationship
  - different from (most) social networks



Web Graph Structure

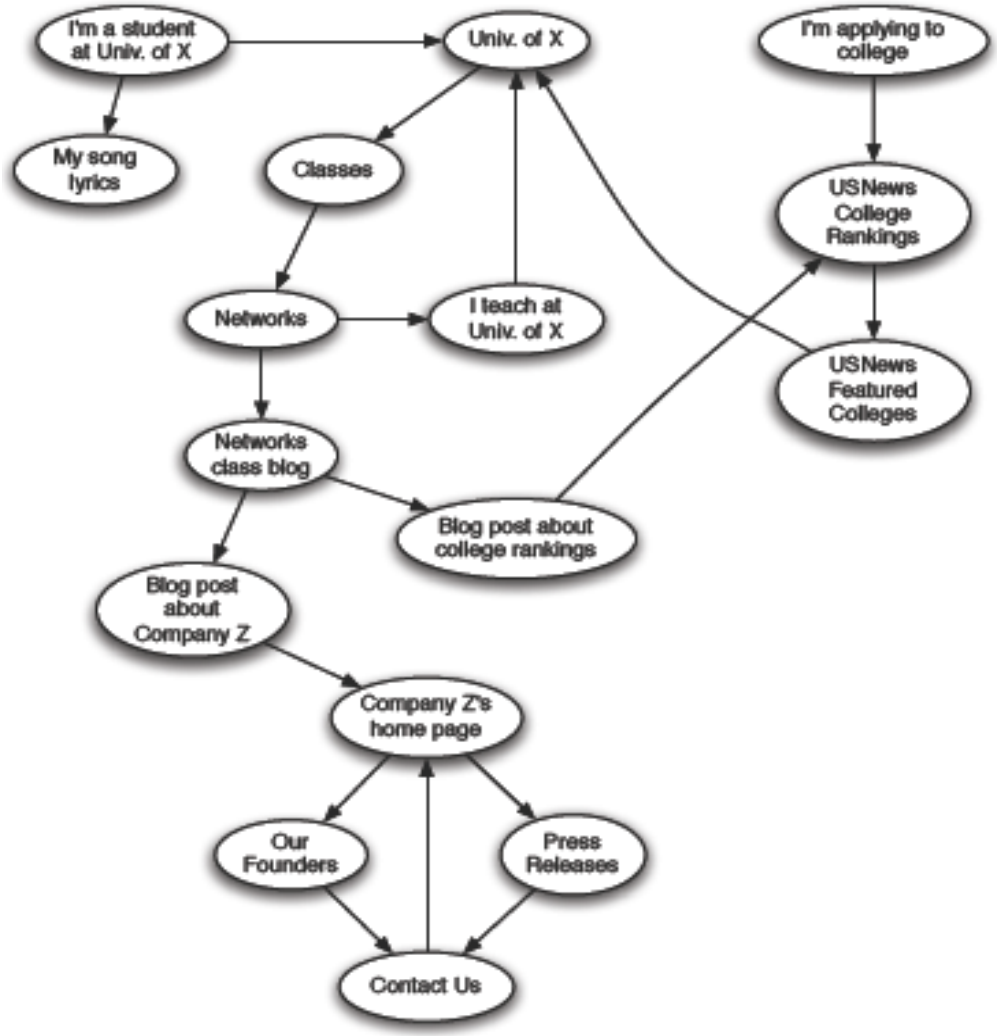
Paths and Strong Connectivity

- Path from node A to node B: sequence of nodes beginning with A and ending with B, where each consecutive pair of nodes is connected by a directed edge (forward direction)



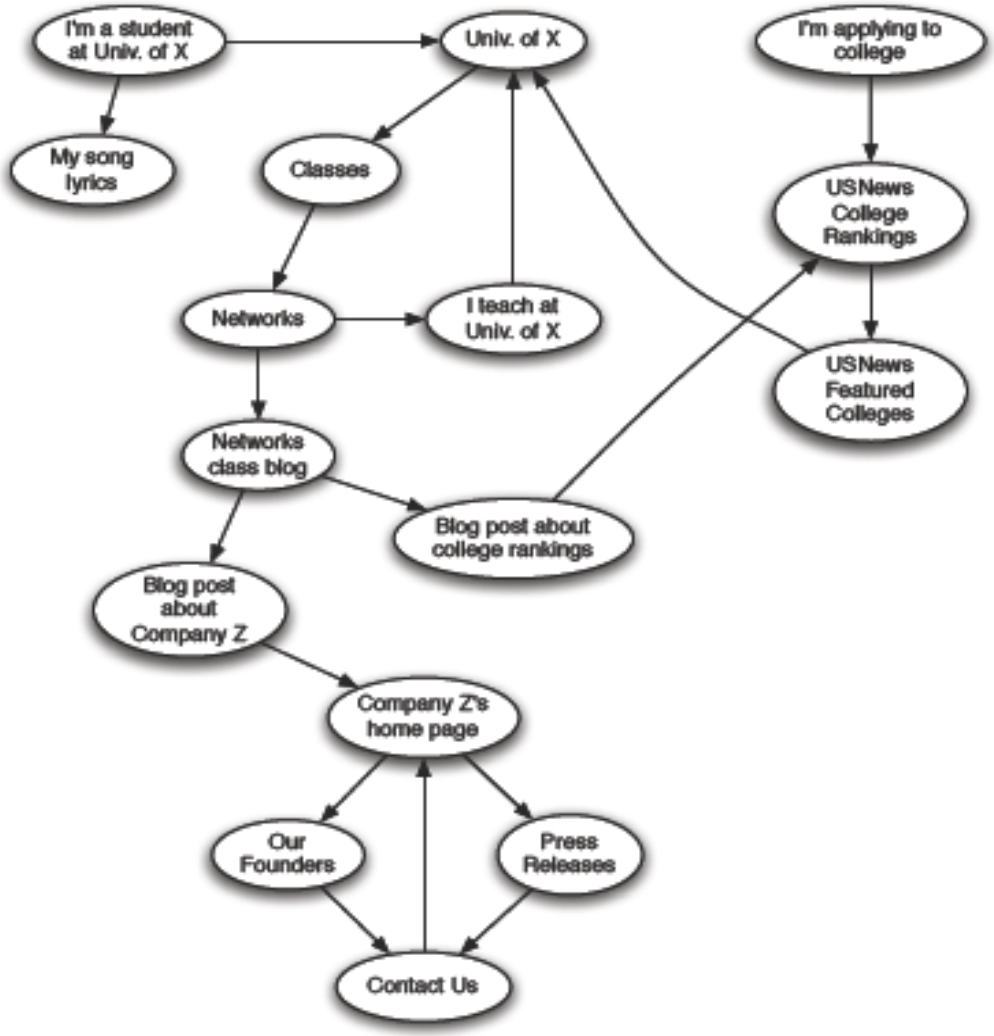
Example

- Path from node “Univ. of X” to “US News College Rankings”
  - “Univ. of X” -> “Classes” -> “Networks” -> “Networks class blog” -> “Blog post about college rankings” -> “US News College Rankings”
- No path from node “Company Z’s home page” to “US News College Rankings”



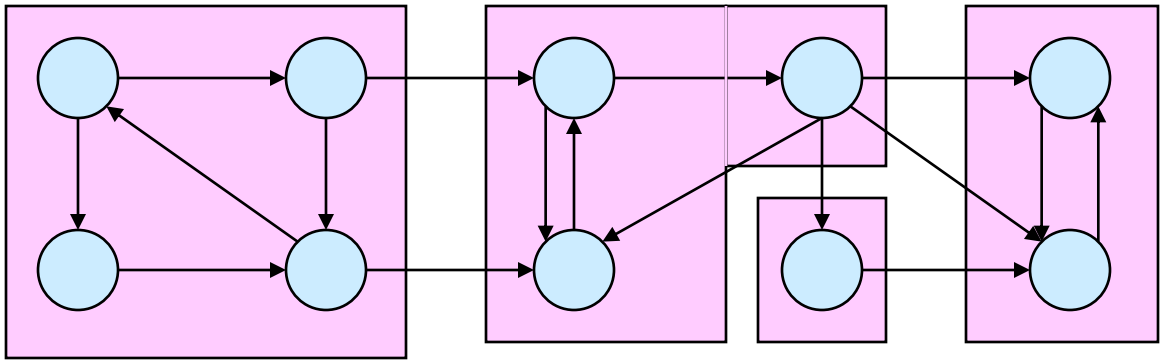
Strongly Connected Directed Graphs

- A directed graph is strongly connected, if there is a path from every node to every other node
- 3 options:
  - pairs of nodes for which each can reach the other ("Univ. of X" and "US News College Rankings")
  - pairs for which one can reach the other but not vice versa ("US News College Rankings" and "Company Z's home page")
  - pairs for which neither can reach the other ("I'm a student at Univ. of X" and "I'm applying to college")



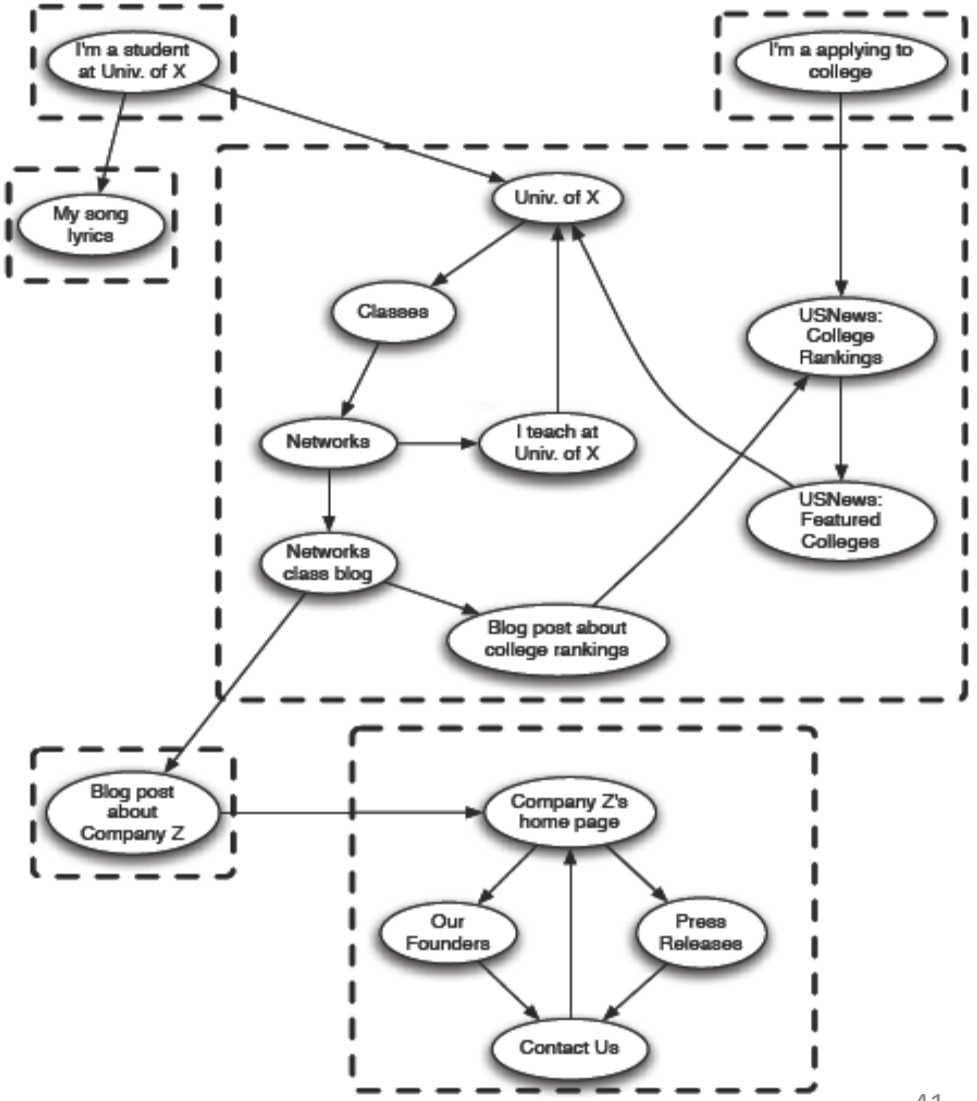
Strongly Connected Components

- A **strongly connected component** (***SCC***) of G is a maximal set of vertices $C \subseteq V$ such that for all $u, v \in C$, both $u \rightsquigarrow v$ and $v \rightsquigarrow u$ exist.



SCC: Example

- SCC in a directed graph is a subset of the nodes such that:
  - (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



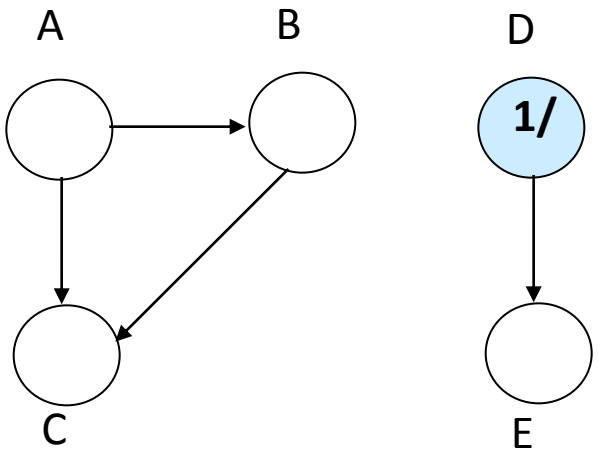
Method for Finding SCCs

- DFS (label nodes with starting/finishing times)
- Transpose of a directed Graph
- Algorithm for finding SCCs
- Deriving the Component Graph

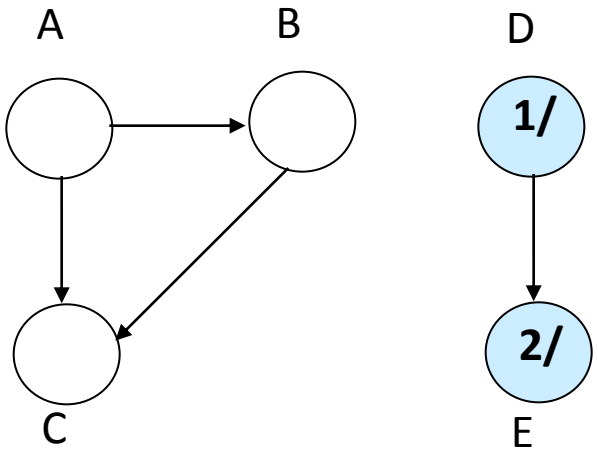
http://www.personal.kent.edu/~rmuhamma/Algorithms/MyAlgorithms/GraphAlgor/strongComponent.htm

DFS and Starting/Finishing Times: Example

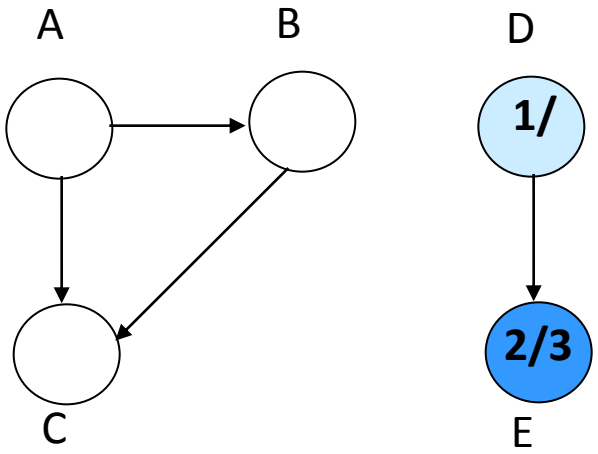
X/Y
X = Starting time
Y = Finishing time



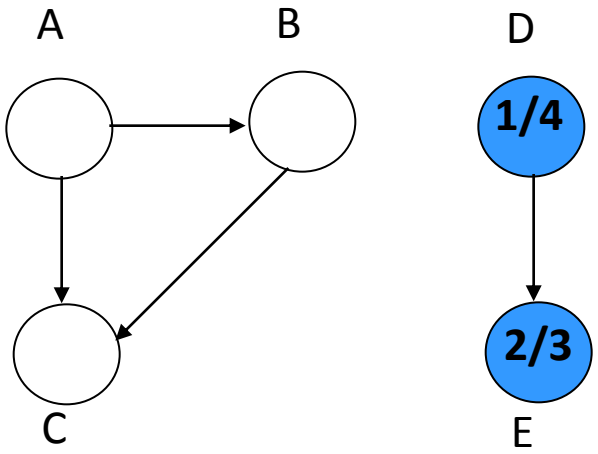
DFS and Starting/Finishing Times: Example



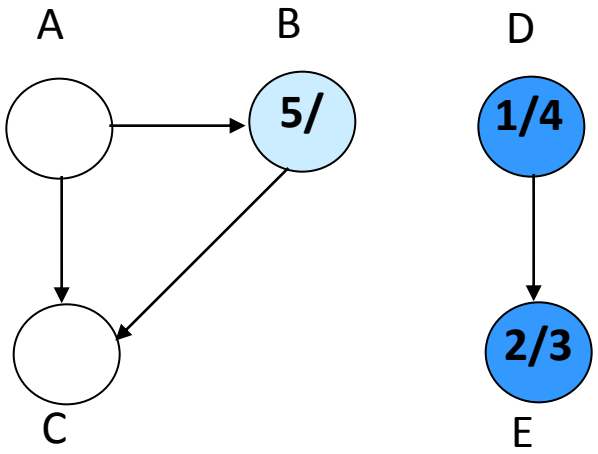
DFS and Starting/Finishing Times: Example



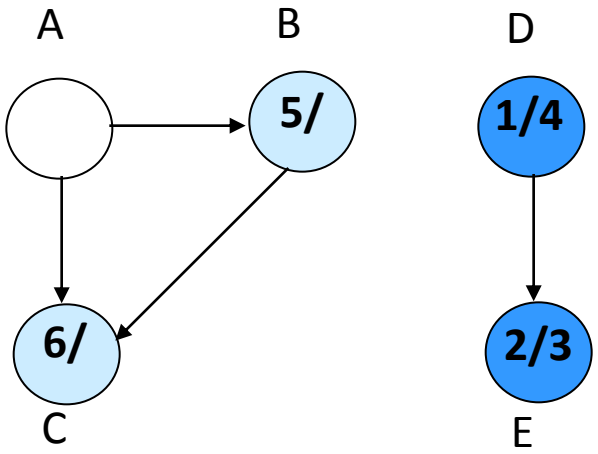
DFS and Starting/Finishing Times: Example



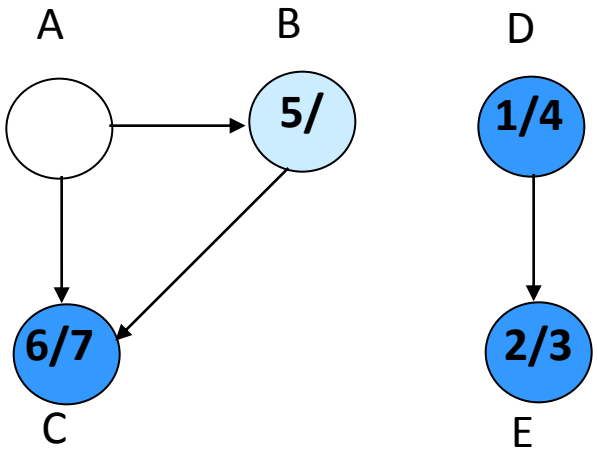
DFS and Starting/Finishing Times: Example



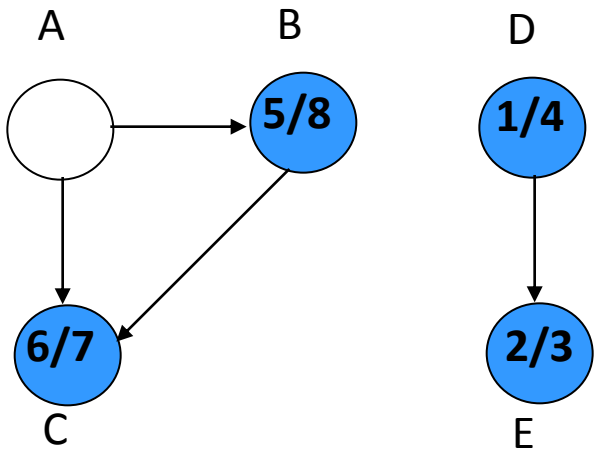
DFS and Starting/Finishing Times: Example



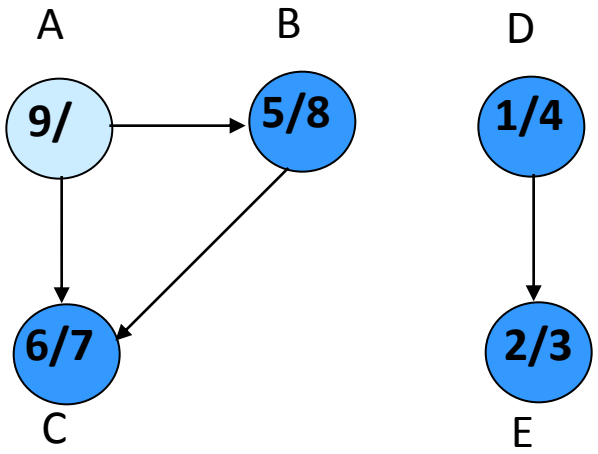
DFS and Starting/Finishing Times: Example



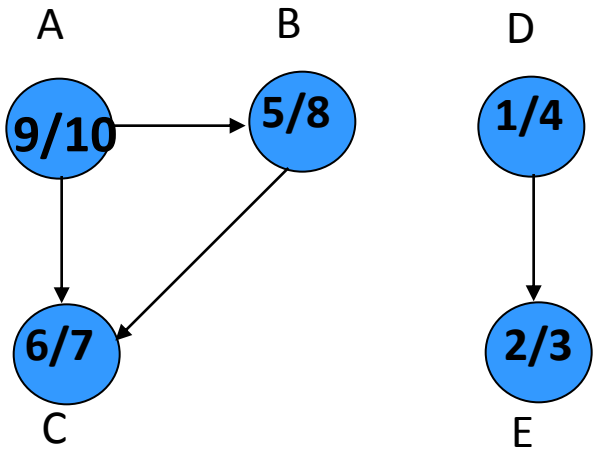
DFS and Starting/Finishing Times: Example



DFS and Starting/Finishing Times: Example



DFS and Starting/Finishing Times: Example



Transpose of a Directed Graph

- G^T = **transpose** of directed G.
  - $G^T = (V, E^T)$, $E^T = \{(u, v) : (v, u) \in E\}$.
  - G^T is G with all edges reversed
- Can create G^T in $\Theta(V + E)$ time if using adjacency lists.
- G and G^T have the *same* SCC's (u and v are reachable from each other in G if and only if reachable from each other in G^T)

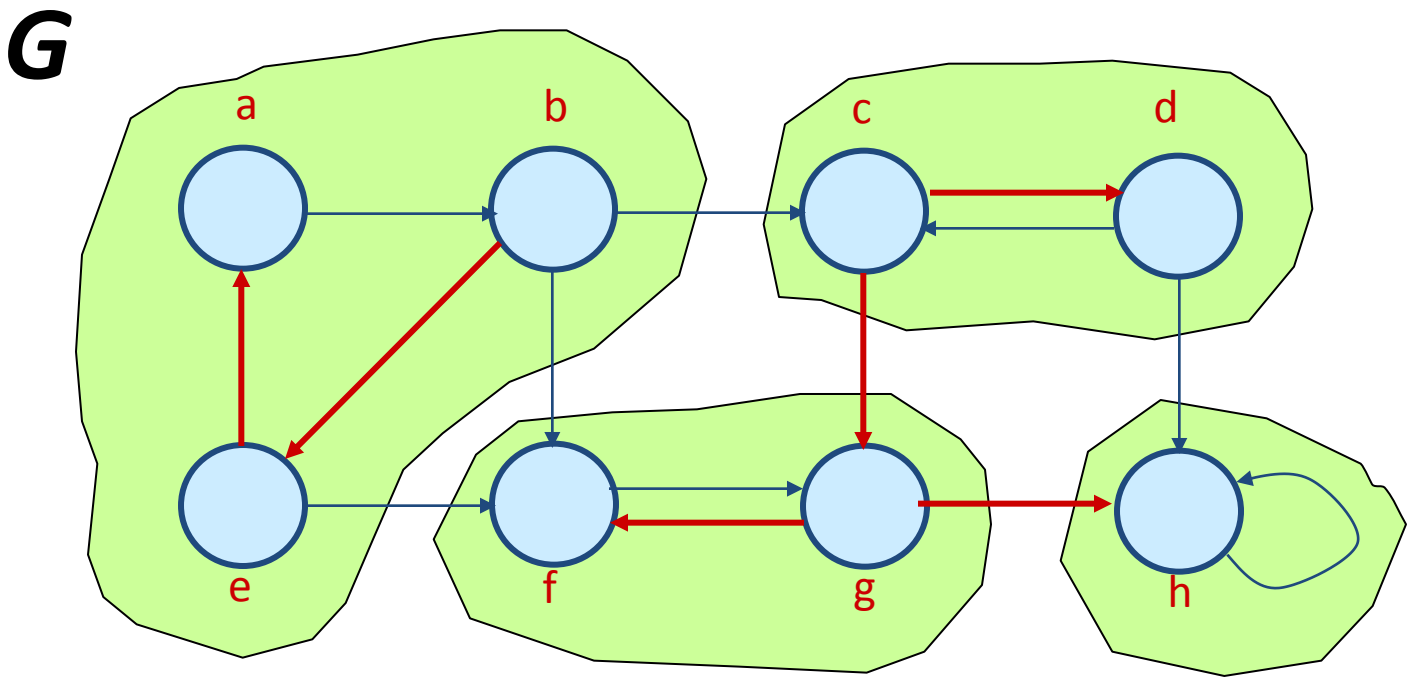
Algorithm to find SCCs

SCC(G)

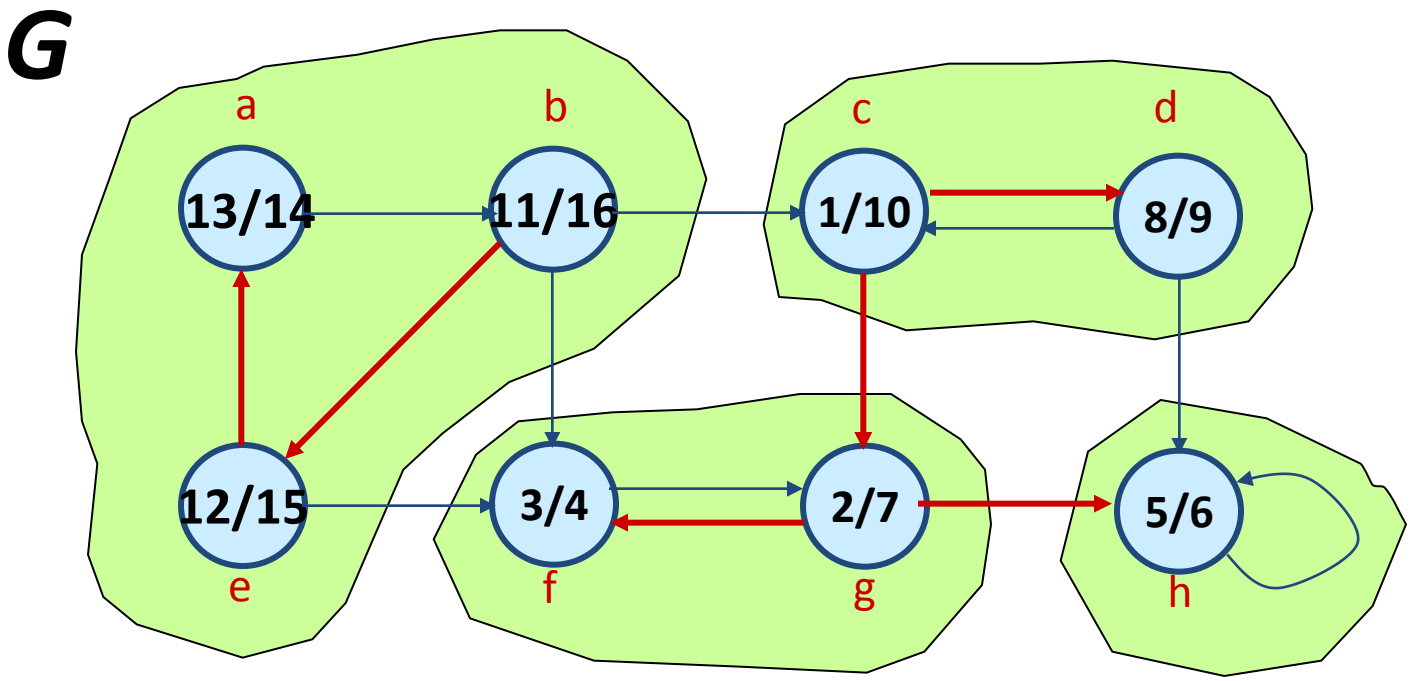
1. call DFS(G) to compute finishing times $f[u]$ for all u
2. compute G^T
3. call DFS(G^T), but in the main loop, consider vertices in order of decreasing $f[u]$ (as computed in first DFS)
4. output the vertices in each tree of the depth-first forest formed in second DFS as a separate SCC

Time: $\Theta(V + E)$.

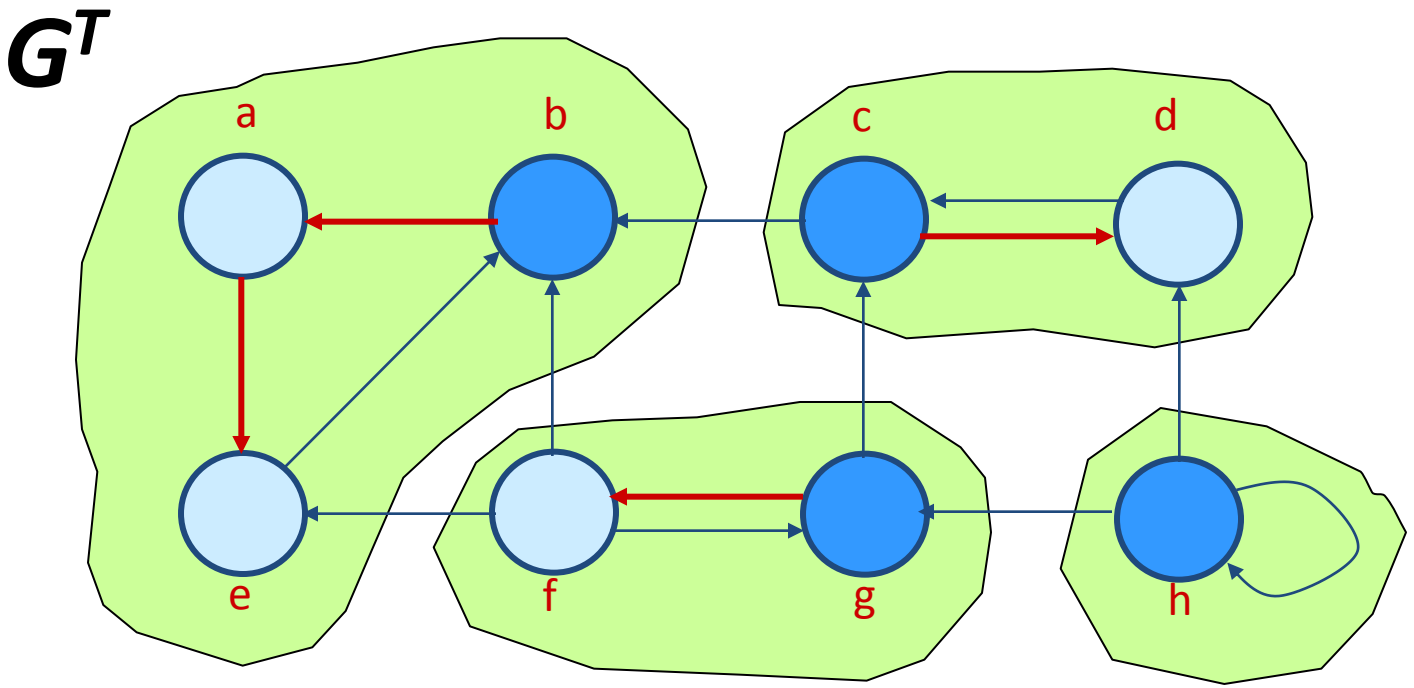
Example

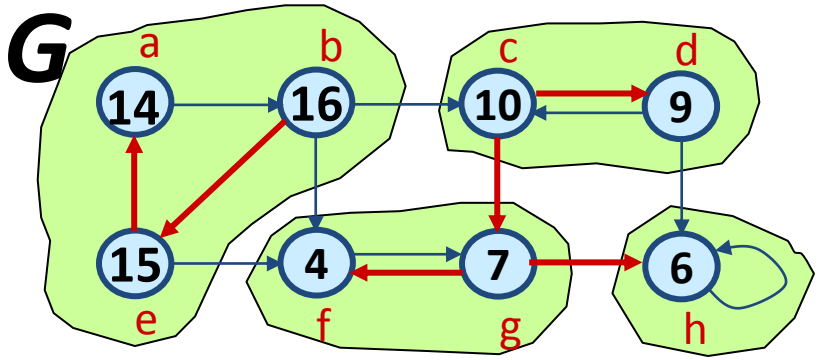


Example



Example

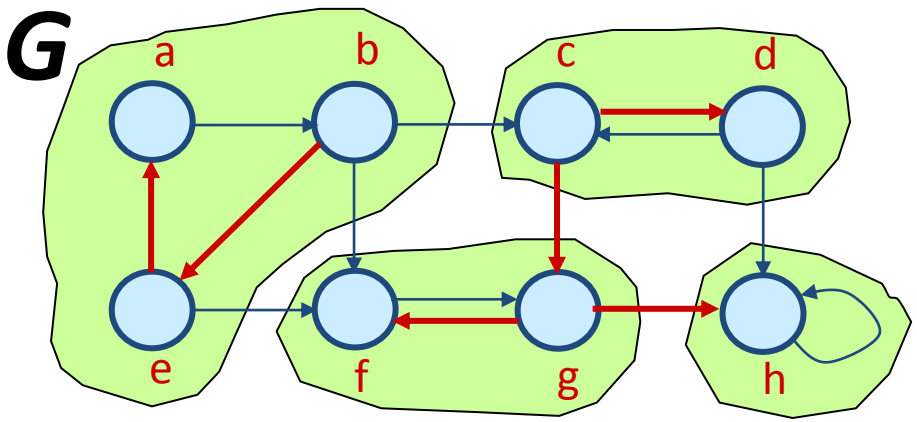


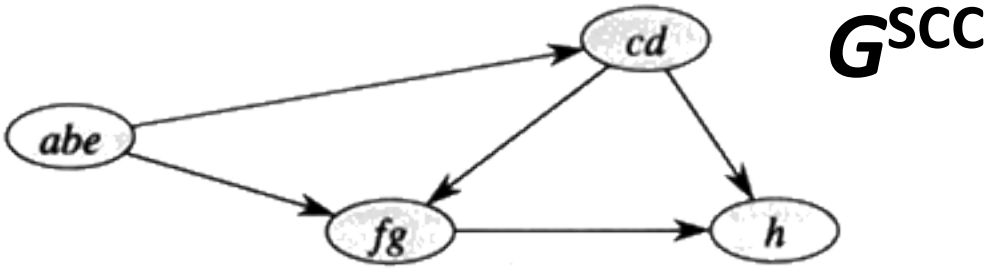


Component Graph

- $G^{SCC} = (V^{SCC}, E^{SCC})$
- V^{SCC} has one vertex for each SCC in G
- E^{SCC} has an edge if there's an edge between the corresponding SCC's in G

Example





The Bow-Tie Structure of the Web

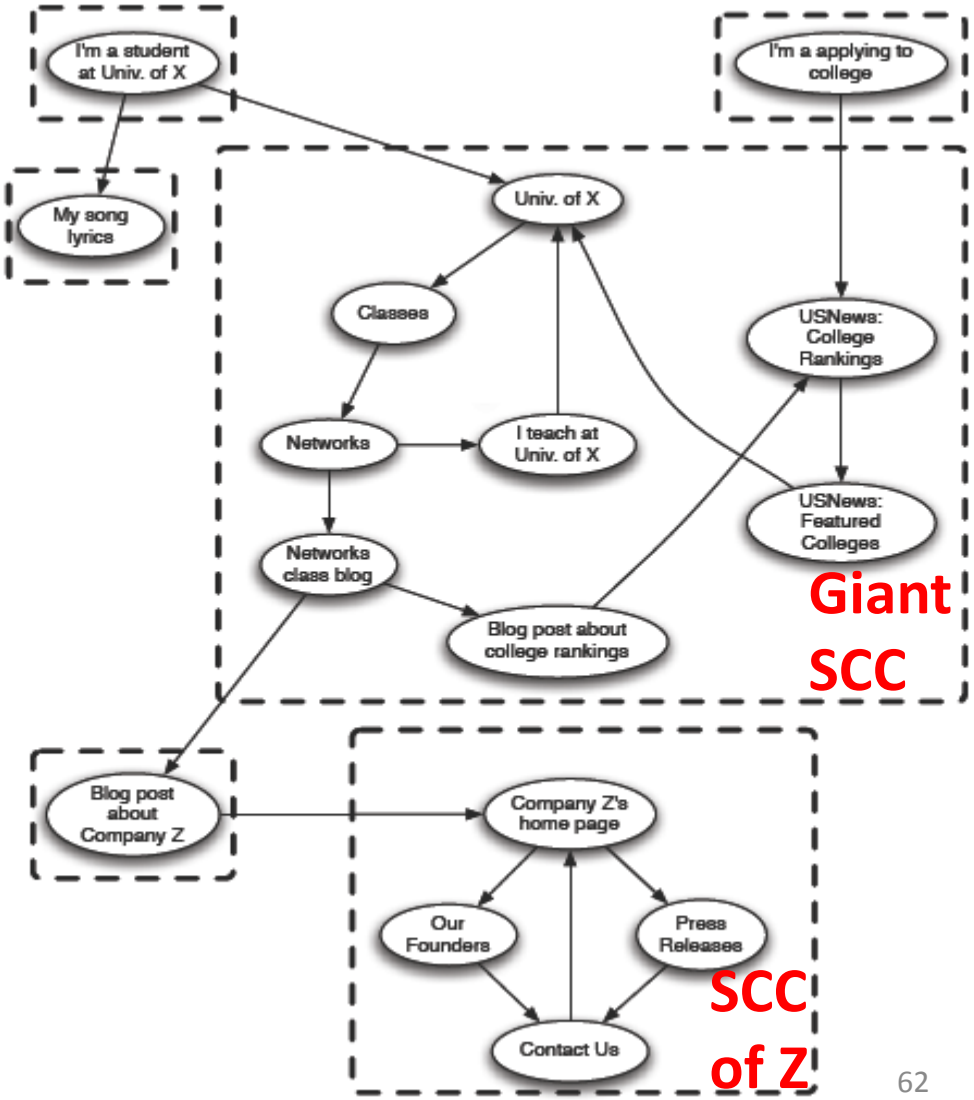
- Andrei Broder et al., 1999
  - A global map of the Web, using SCCs as the basic building blocks (component graph)
  - Dividing Web into a few large pieces and show how they fit together
- Data
  - navigational “backbone” indexed by AltaVista (1999)
  - Pioneering research verified by others (newer studies with navigational “backbone” indexed by Google, Wikipedia, etc.)

The Bow-Tie Structure of the Web

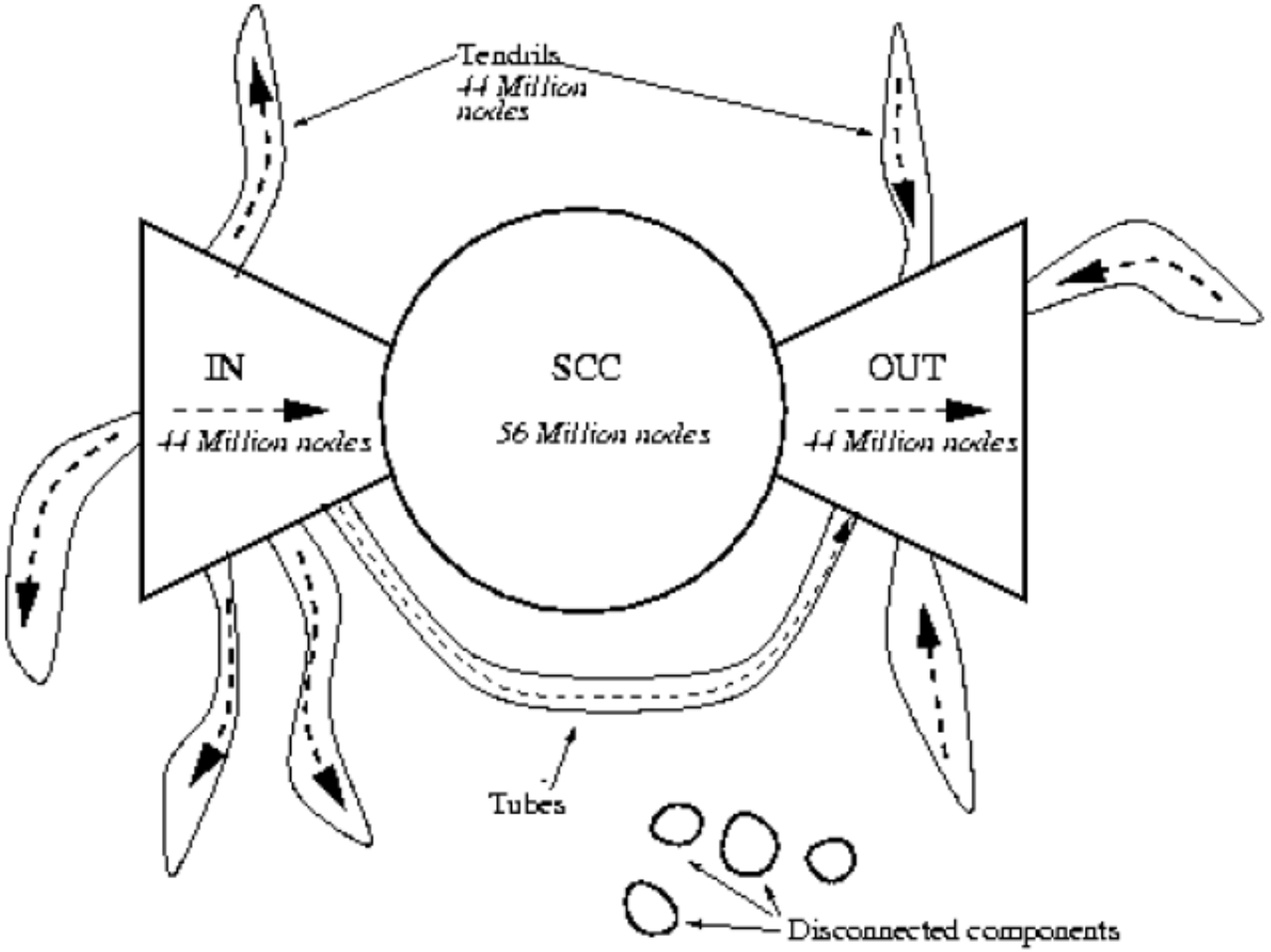
- The Web contains a single giant SCC
  - giant SCC contains a significant fraction of all pages (also most important pages: major commercial, governmental, and non-profit organizations)
- Position all the remaining SCCs in relation to the giant SCC, by classifying nodes by their ability to reach and be reached from the giant SCC
  - IN: nodes that can reach the giant SCC but cannot be reached from it
    - Pages not “discovered” by members of the giant SCC
  - OUT: nodes that can be reached from the giant SCC but cannot reach it
    - Pages receiving links from the giant SCC, but not linking back

Example

- IN:
  - "I'm a student at Univ. of X"
  - "I'm applying to college constitute"
- OUT:
  - "Blog post about Company Z"
  - SCC involving Company Z



The Bow-Tie Structure of the Web

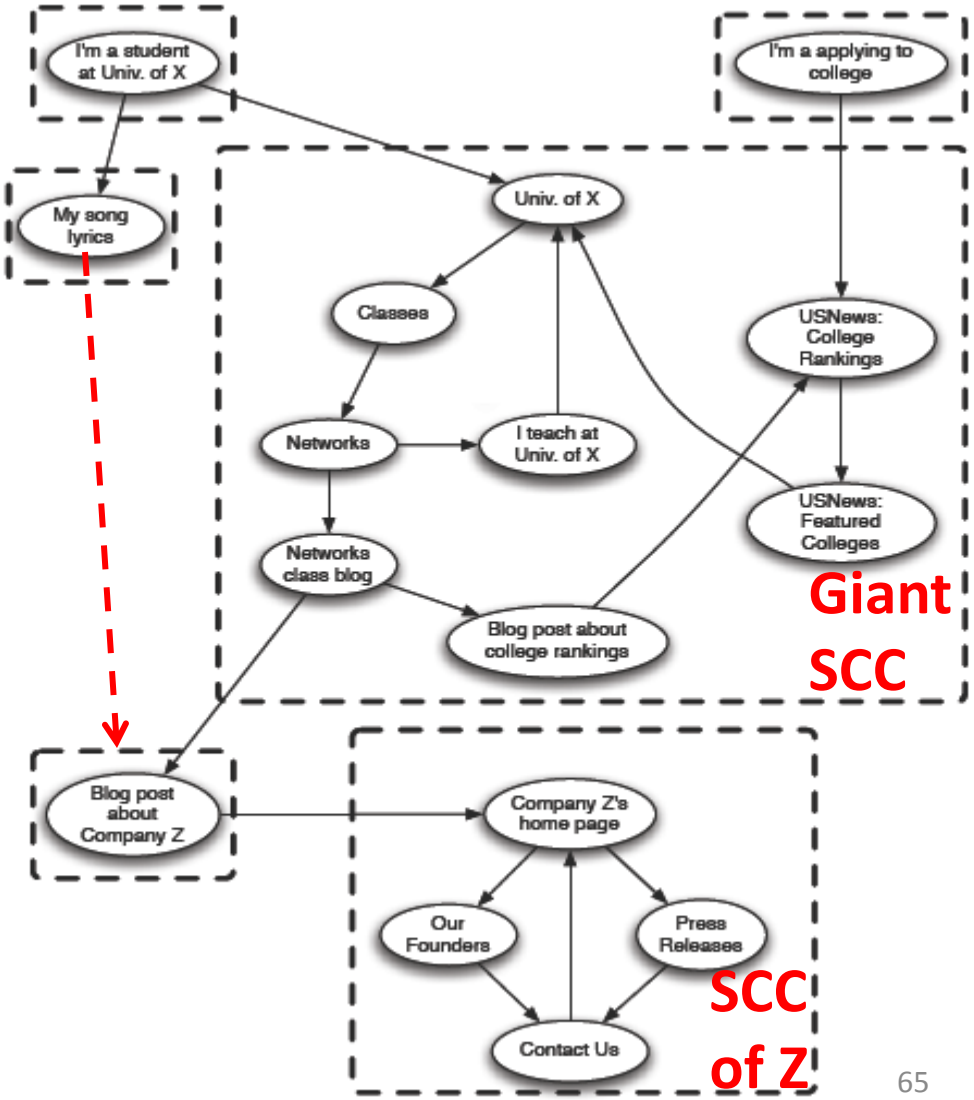


The Bow-Tie Structure of the Web

- There are pages that belong to none of IN, OUT, or the giant SCC
  - neither reach the giant SCC nor be reached from it
- Categories of such pages:
  - Tendrils: (a) nodes reachable from IN that cannot reach the giant SCC, or (b) nodes that can reach OUT but cannot be reached from the giant SCC
  - Tube: nodes satisfying both (a) and (b) above
  - Disconnected: otherwise (nodes that would not have a path to the giant SCC even if we ignored directions of the edges)

Example

- Tendril:
  - "My song lyrics" (reachable from IN but has no path to the giant SCC)
- Tube:
  - "My song lyrics", if linked to "Blog post about Company Z" (dashed red link)



Final notes

- Dynamic structure
  - changes as new pages and links are created
  - nodes entering (and also leaving) the giant SCC over time
  - Newer studies show that the aggregate picture remains relatively stable over time
- Limitations
  - Bow-tie picture gives a global view of the Web
  - Doesn’t give us insight into the more fine-grained patterns of connections within the parts
  - detailed network analysis can highlight important Web pages (next lectures)