LECTURE 3:BASIC NETWORK PROPERTIES AND WEB GRAPH

COMP4641: Social Information Network Analysis and Engineering

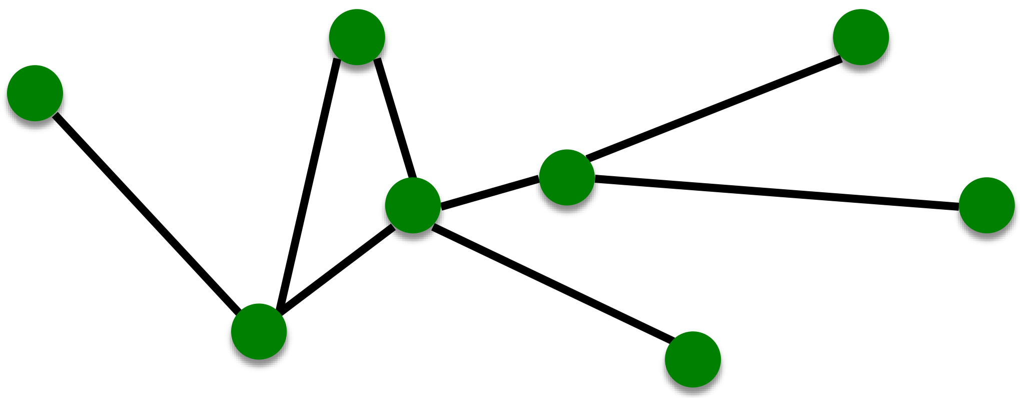
Wednesday February 11th 2015

Structure of Networks?

Network is a collection of objects where some pairs of objects are connected by links

What is the structure of the network?

Components of a Network



- **Objects:** nodes, vertices N
- **Interactions:** links, edges E
- **System:** network, graph $G(N,E)$

Networks or Graphs?

- **Network** often refers to real systems
  - Web, Social network, Metabolic network

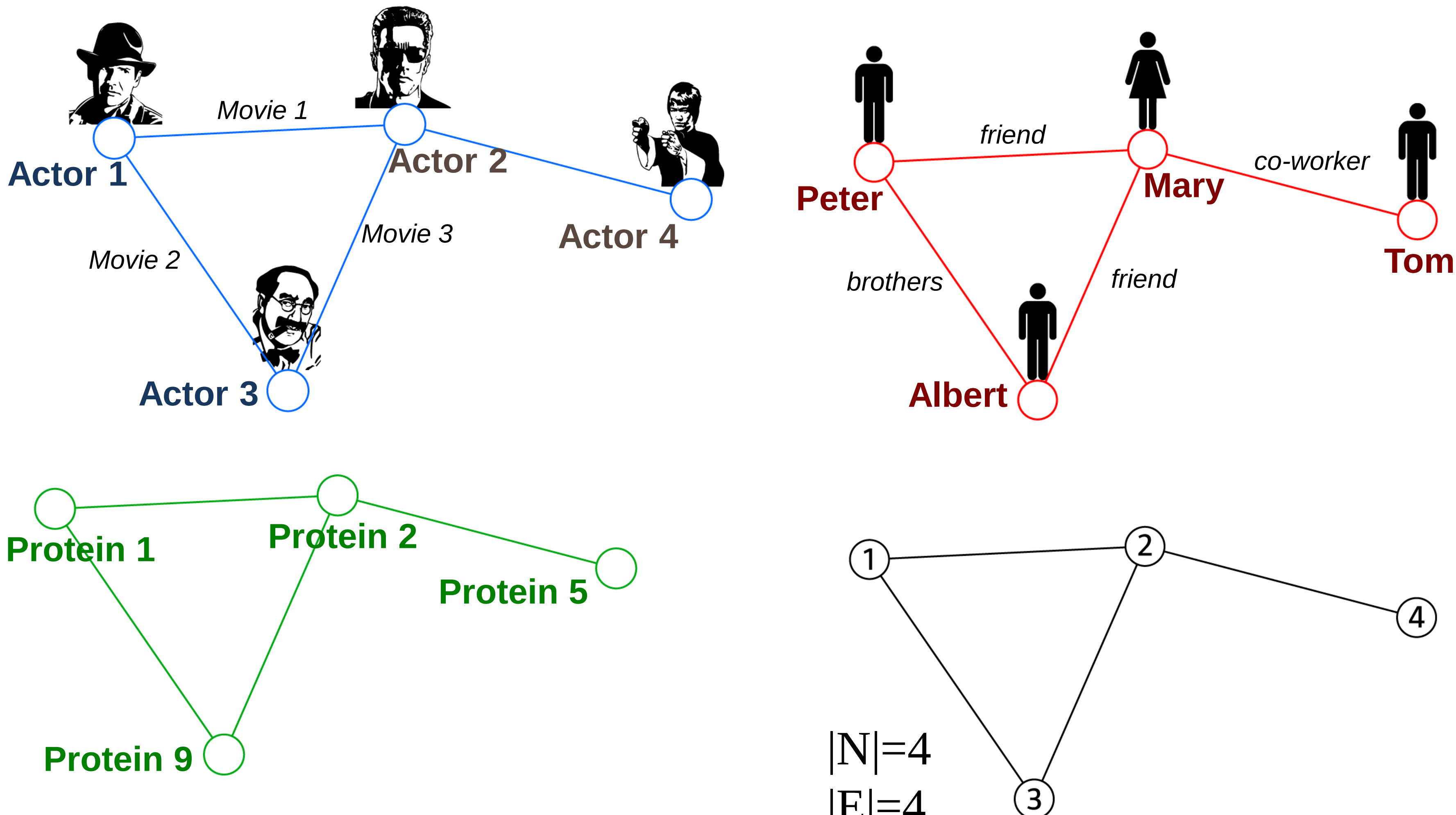
  Language: Network, node, link

- **Graph:** mathematical representation of a network
  - Web graph, Social graph (a Facebook term)

  Language: Graph, vertex, edge

We will try to make this distinction whenever it is appropriate, but in most cases we will use the two terms interchangeably

Networks: Common Language



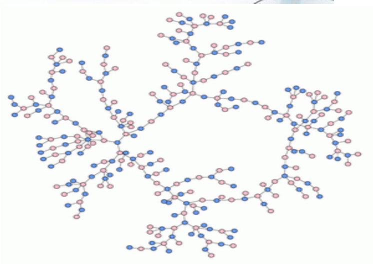
$|N|=4$
$|E|=4$

Choosing Proper Representation

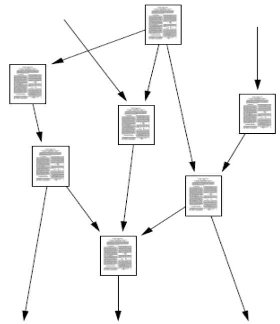
- **Choice of the proper network representation determines our ability to use networks successfully:**
  - In some cases there is a unique, unambiguous representation
  - In other cases, the representation is by no means unique
  - The way you assign links will determine the nature of the question you can study

Choosing Proper Representation

- If you connect individuals that work with each other, you will explore a **professional network**



- If you connect those that have a sexual relationship, you will be exploring **sexual networks**



- If you connect scientific papers that cite each other, you will be studying the **citation network**

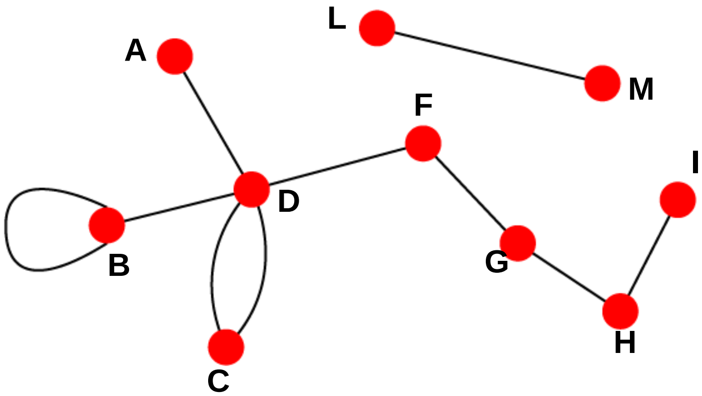
- **If you connect all papers with the same word in the title, you will be exploring what?** It is a network, nevertheless

NETWORK PROPERTIES: HOW TO CHARACTERIZE A NETWORK?

Undirected vs. Directed Networks

Undirected

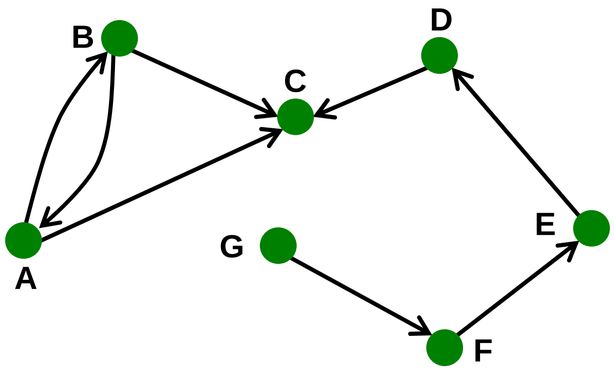
- Links: undirected (symmetrical)



- Examples:
  - Collaborations
  - Friendship on Facebook

Directed

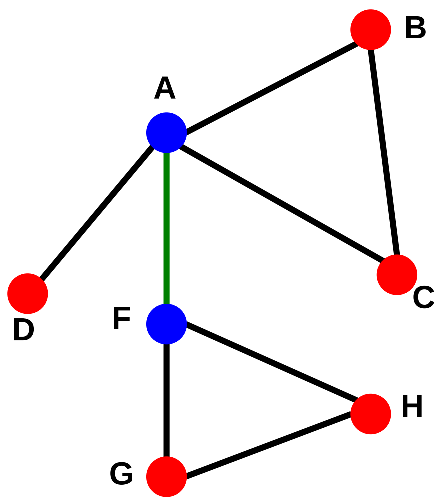
- Links: directed (arcs)

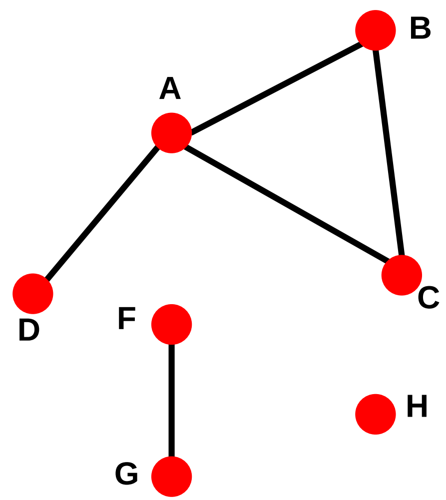


- Examples:
  - Phone calls
  - Following on Twitter

Connectivity of Graphs

- **Connected (undirected) graph:**
  - Any two vertices can be joined by a path.
- A disconnected graph is made up by two or more connected components





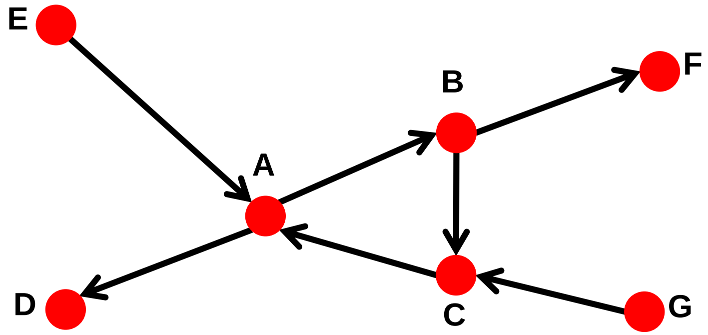
Largest Component:
Giant Component

Isolated node (node H)

Bridge edge: If we erase it, the graph becomes disconnected.
Articulation point: If we erase it, the graph becomes disconnected.

Connectivity of Directed Graphs

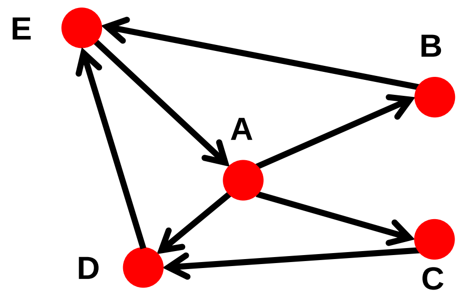
- **Strongly connected directed graph**
  - has a path from each node to every other node and vice versa (e.g., A-B path and B-A path)
- **Weakly connected directed graph**
  - is connected if we disregard the edge directions

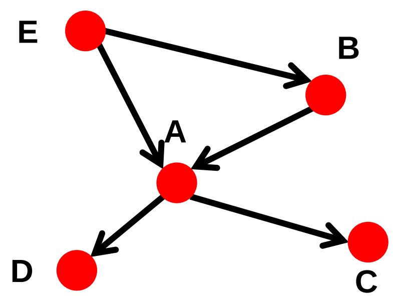


Graph on the left is connected but not strongly connected (e.g., there is no way to get from F to G by following the edge directions).

Directed Graphs

- **Two types of directed graphs:**
  - **Strongly connected:**
    - Any node can reach any node via a directed path
  - **DAG – Directed Acyclic Graph:**
    - Has no cycles: if u can reach v, then v can not reach u

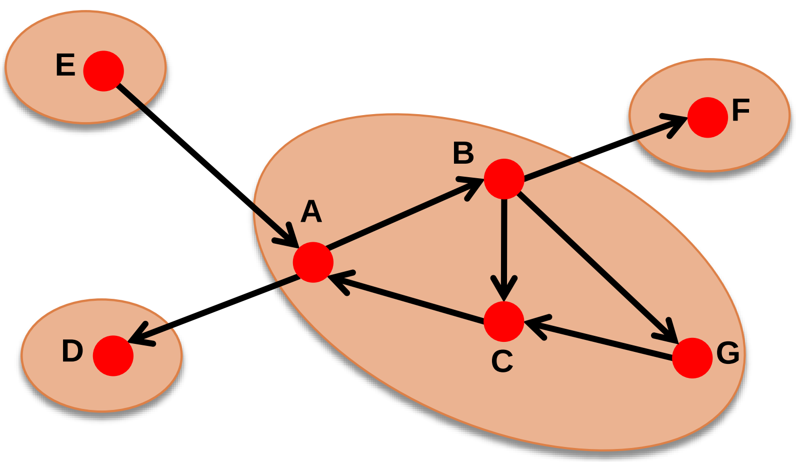




- **Any directed graph can be expressed in terms of these two types!**

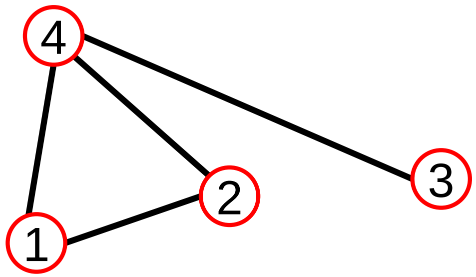
Strongly Connected Component

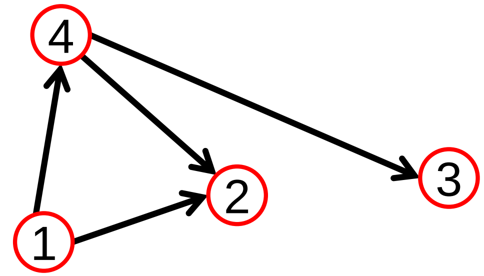
- **Strongly connected component (SCC)** is a set of nodes S so that:
  - Every pair of nodes in S can reach each other
  - There is no larger set containing S with this property



Strongly connected components of the graph: {A,B,C,G}, {D}, {E}, {F}

Adjacency Matrix





$\boldsymbol{A_{ij} = 1}$ if there is a link from node *i* to node *j*

$\boldsymbol{A_{ij} = 0}$ otherwise

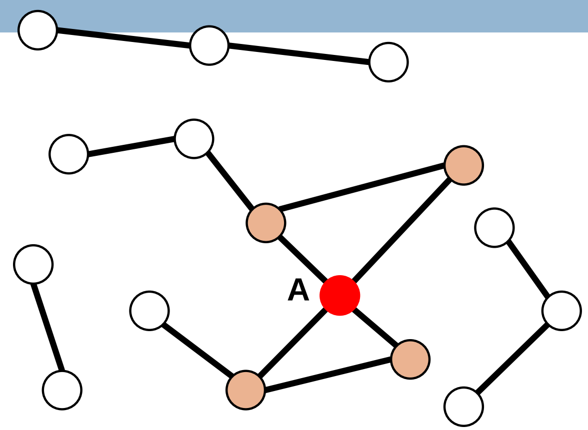
$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

Note that for a directed graph (right) the matrix is not symmetric.

Node Degrees

Undirected

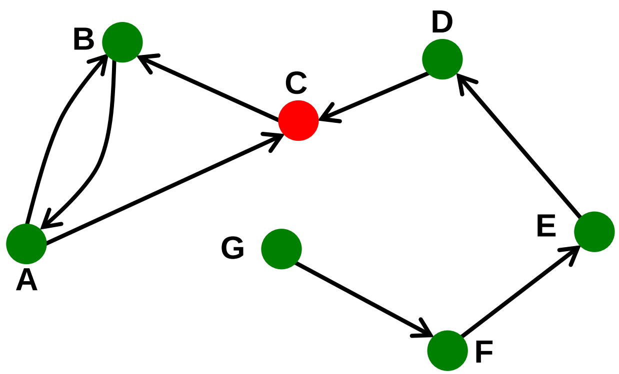


Node degree, k_i: the number of edges adjacent to node i

$$k_A = 4$$

Avg. degree: $\bar{k} = \langle k \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{2E}{N}$

Directed



Source: node with $k^{in} = 0$

Sink: node with $k^{out} = 0$

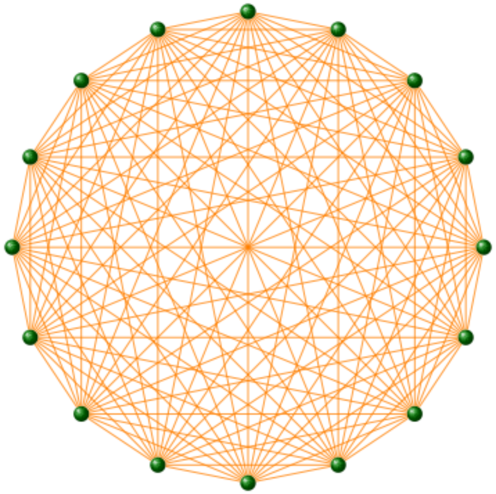
In directed networks we define an **in-degree** and **out-degree.**

The (total) degree of a node is the sum of in- and out-degrees.

$$k_C^{in} = 2 \qquad k_C^{out} = 1 \qquad k_C = 3$$

$$\bar{k} = \frac{E}{N} \qquad \overline{k^{in}} = \overline{k^{out}}$$

Complete Graph



The **maximum number of edges** in an undirected graph on N nodes is

$$E_{\max} = \binom{N}{2} = \frac{N(N-1)}{2}$$

A graph with the number of edges $E = E_{max}$ is a **complete graph**, and its average degree is N-1

Networks are Sparse Graphs

Most real-world networks are sparse

$$E \ll E_{max} \quad (\text{or } \bar{k} \ll N-1)$$

WWW (Stanford-Berkeley):	N=319,717	$\langle k \rangle$=9.65
Social networks (LinkedIn):	N=6,946,668	$\langle k \rangle$=8.87
Communication (MSN IM):	N=242,720,596	$\langle k \rangle$=11.1
Coauthorships (DBLP):	N=317,080	$\langle k \rangle$=6.62
Internet (AS-Skitter):	N=1,719,037	$\langle k \rangle$=14.91
Roads (California):	N=1,957,027	$\langle k \rangle$=2.82
Protein (S. Cerevisiae):	N=1,870	$\langle k \rangle$=2.39

(Source: *Leskovec et al., Internet Mathematics, 2009*)

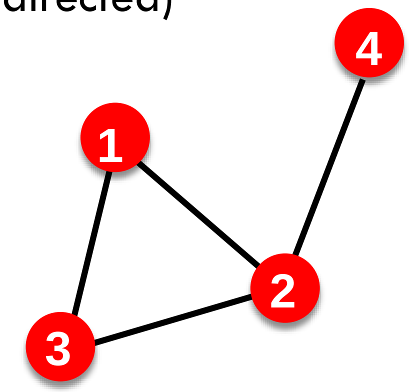
Consequence: Adjacency matrix is filled with zeros!

(**Density (E/N^2):** WWW=1.51×10^{-5}, MSN IM = 2.27×10^{-8})

More Types of Graphs:

- **Unweighted**

(undirected)



$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

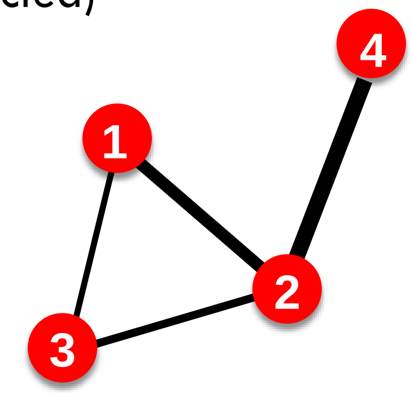
$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

$$E = \frac{1}{2}\sum_{i,j=1}^{N} A_{ij} \qquad \bar{k} = \frac{2E}{N}$$

Examples: Friendship, Sex

- **Weighted**

(undirected)



$$A_{ij} = \begin{pmatrix} 0 & 2 & 0.5 & 0 \\ 2 & 0 & 1 & 4 \\ 0.5 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$$

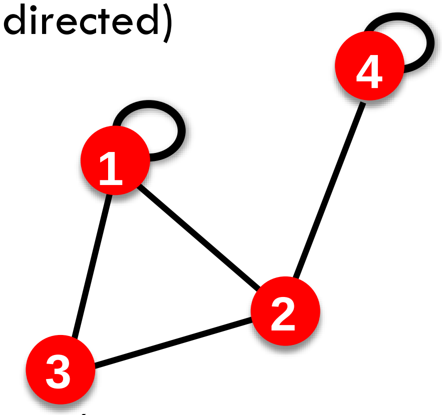
$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

$$E = \frac{1}{2}\sum_{i,j=1}^{N} nonzero(A_{ij}) \qquad \bar{k} = \frac{2E}{N}$$

Examples: Collaboration, Internet, Roads

More Types of Graphs:

- **Self-edges (self-loops)**
  (undirected)



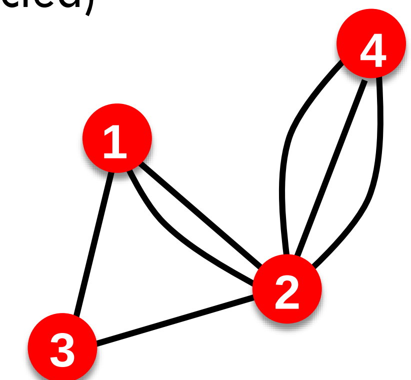
$$A_{ij} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

$$A_{ii} \neq 0 \qquad A_{ij} = A_{ji}$$

$$E = \frac{1}{2}\sum_{i,j=1, i \neq j}^{N} A_{ij} + \sum_{i=1}^{N} A_{ii}$$

Examples: Proteins, Hyperlinks

- **Multigraph**
  (undirected)



$$A_{ij} = \begin{pmatrix} 0 & 2 & 1 & 0 \\ 2 & 0 & 1 & 3 \\ 1 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

$$E = \frac{1}{2}\sum_{i,j=1}^{N} nonzero(A_{ij}) \qquad \bar{k} = \frac{2E}{N}$$

Examples: Communication, Collaboration

Network Representations

WWW >> directed multigraph with self-interactions

Facebook friendships >> undirected, unweighted

Citation networks >> unweighted, directed, acyclic

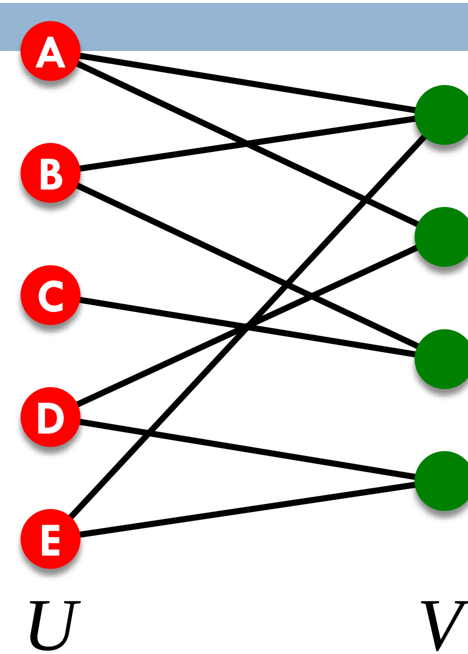
Collaboration networks >> undirected multigraph or weighted graph

Mobile phone calls >> directed, (weighted?) multigraph

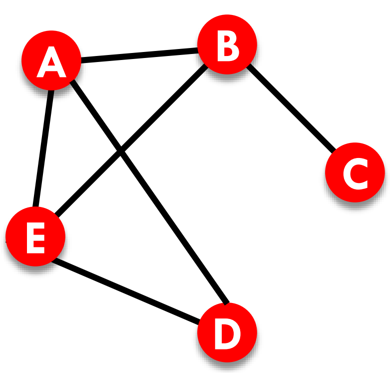
Protein Interactions >> undirected, unweighted with self-interactions

Bipartite Graph

- **Bipartite graph** is a graph whose nodes can be divided into two disjoint sets U and V such that every link connects a node in U to one in V; that is, U and V are independent sets.



- **Examples:**
  - Authors-to-papers (they authored)
  - Actors-to-Movies (they appeared in)
  - Users-to-Movies (they rated)
- **"Folded" networks:**
  - Author collaboration networks
  - Movie co-rating networks



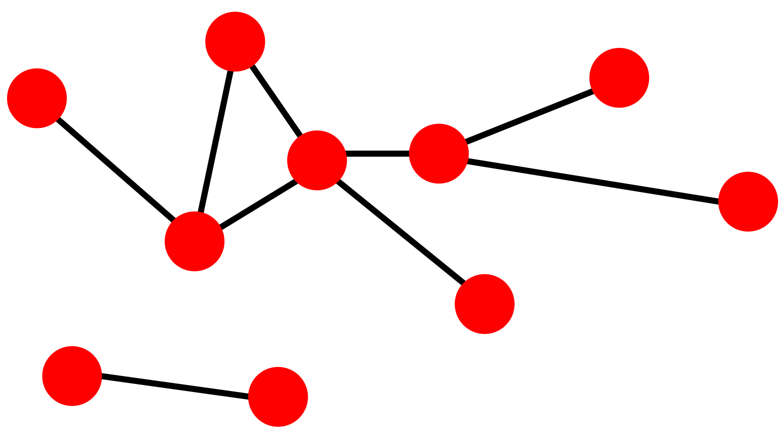
Degree Distribution

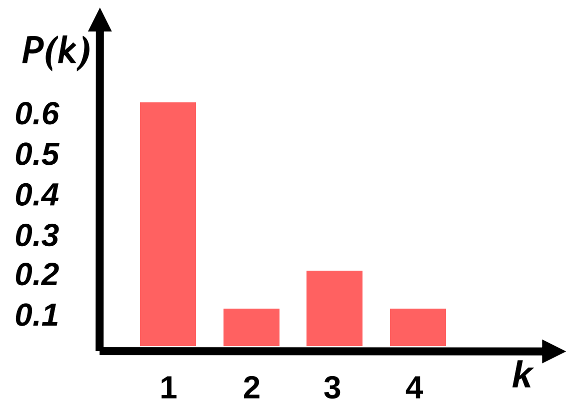
- **Degree distribution $P(k)$:** Probability that a randomly chosen node has degree k

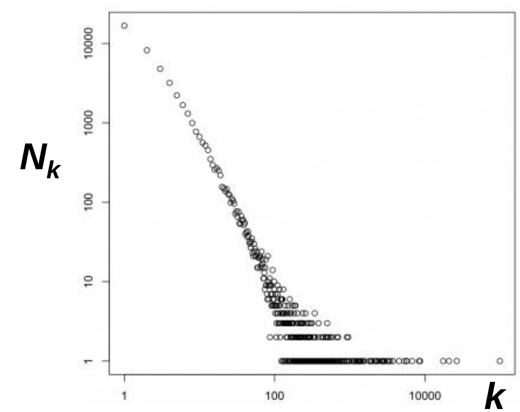
  N_k = # nodes with degree k

- Normalized histogram:

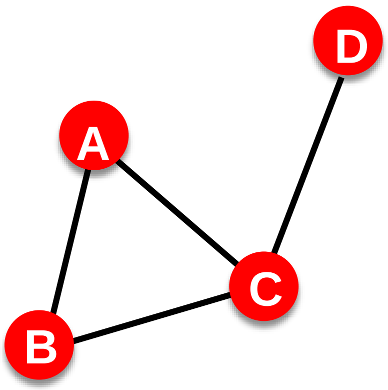
  $P(k) = N_k / N$ ➔ **plot**



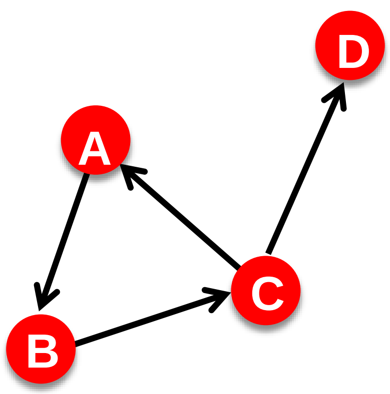




Distance in a Graph



$h_{B,D} = 2$



$h_{B,C} = 1$, $h_{C,B} = 2$

- **Distance (shortest path, geodesic)** between a pair of nodes is defined as the number of edges along the shortest path connecting the nodes
  - *If the two nodes are disconnected, the distance is usually defined as infinite
- In **directed graphs** paths need to follow the direction of the arrows
  - Consequence: Distance is not symmetric: $h_{A,C} \neq h_{C,A}$

Network Diameter

- **Diameter:** the maximum (shortest path) distance between any pair of nodes in a graph

- **Average path length** for a connected graph (component) or a strongly connected (component of a) directed graph

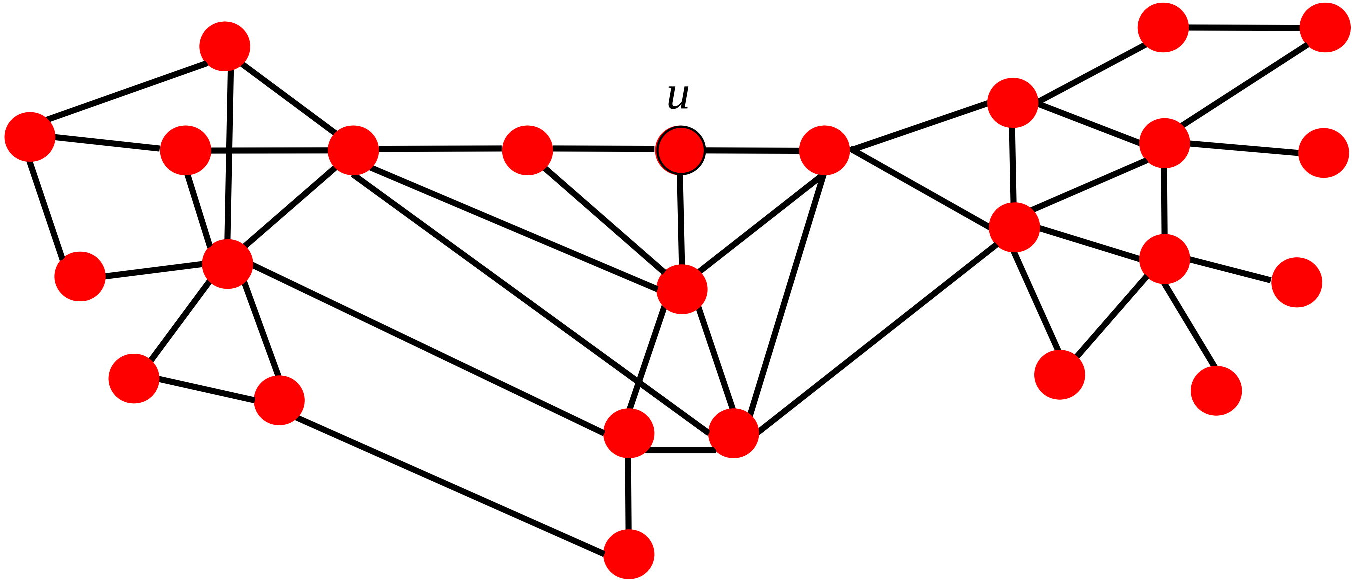
$$\bar{h} = \frac{1}{2E_{\max}} \sum_{i, j \neq i} h_{ij}$$

where h_{ij} is the distance from node i to node j

  - Many times we compute the average only over the connected pairs of nodes (we ignore “infinite” length paths)

Finding Shortest Paths

- **Breath-First Search:**
  - Start with node u, mark it to be at distance $h_u(u)=0$, add u to the queue
  - While the queue not empty:
    - Take node v off the queue, put its unmarked neighbors w into the queue and mark $h_u(w)=h_u(v)+1$

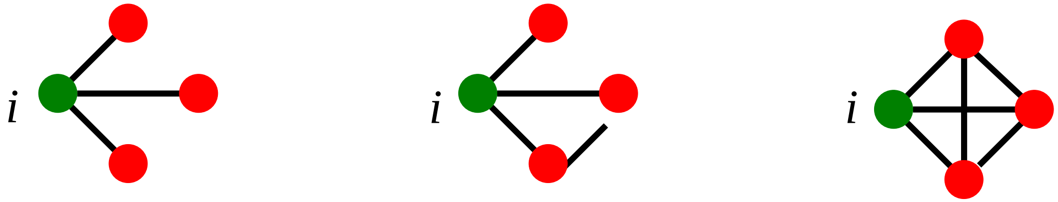


Clustering Coefficient

- **Clustering coefficient:**
  - What portion of i's neighbors are connected?
  - Node i with degree k_i
  - $C_i \in [0,1]$
  -

$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$

where e_i is the number of edges between the neighbors of node i



C_i=0 C_i=1/3 C_i=1

- **Average Clustering Coefficient:**

$$C = \frac{1}{N}\sum_{i}^{N} C_i$$

Key Network Properties

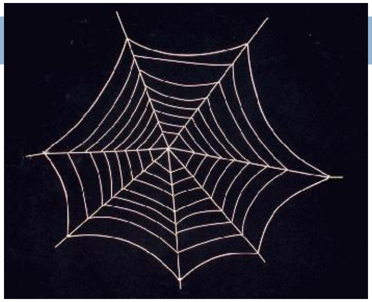
Degree distribution: $P(k)$

Path length: h

Clustering coefficient: C

STRUCTURE OF THE WEB GRAPH

Web as a Graph

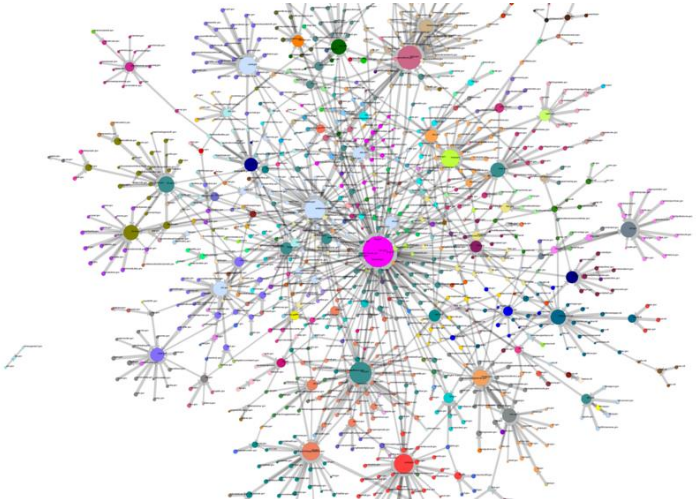


- **Q: What does the Web "look like"?**

- **Here is what we will do next:**
  - We will take a real system (i.e., the Web)
  - We will collect lots of Web data
  - We will represent the Web as a graph
  - We will use language of graph theory to reason about the structure of the graph
  - Do a computational experiment on the Web graph
  - **Learn something about the structure of the Web!**

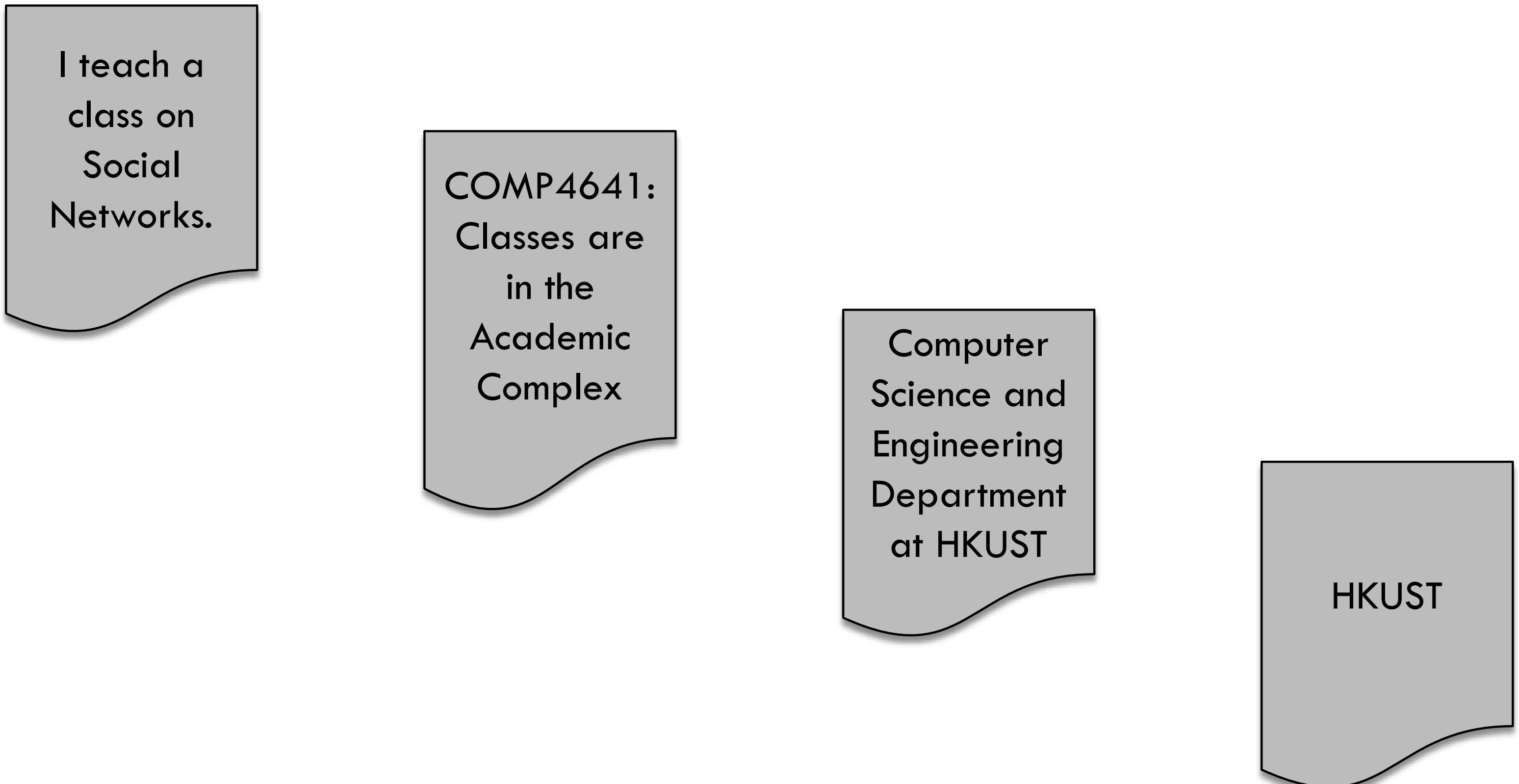
Web as a Graph

Q: What does the Web "look like" at a global level?

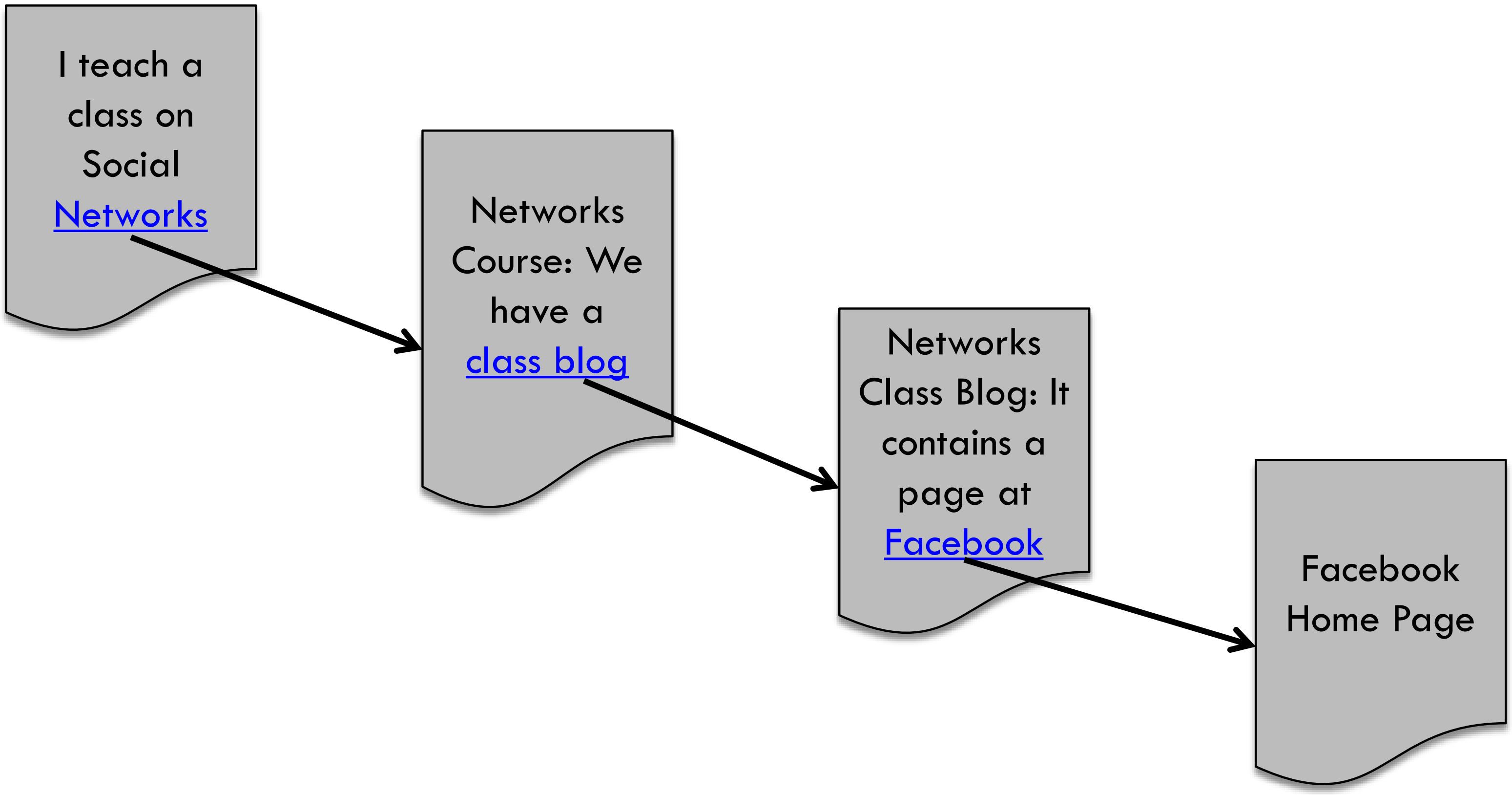


- **Web as a graph:**
  - Nodes = web pages
  - Edges = hyperlinks
  - Side issue: What is a node?
    - Dynamic pages created on the fly
    - "dark matter" – inaccessible database generated pages

The Web as a Graph

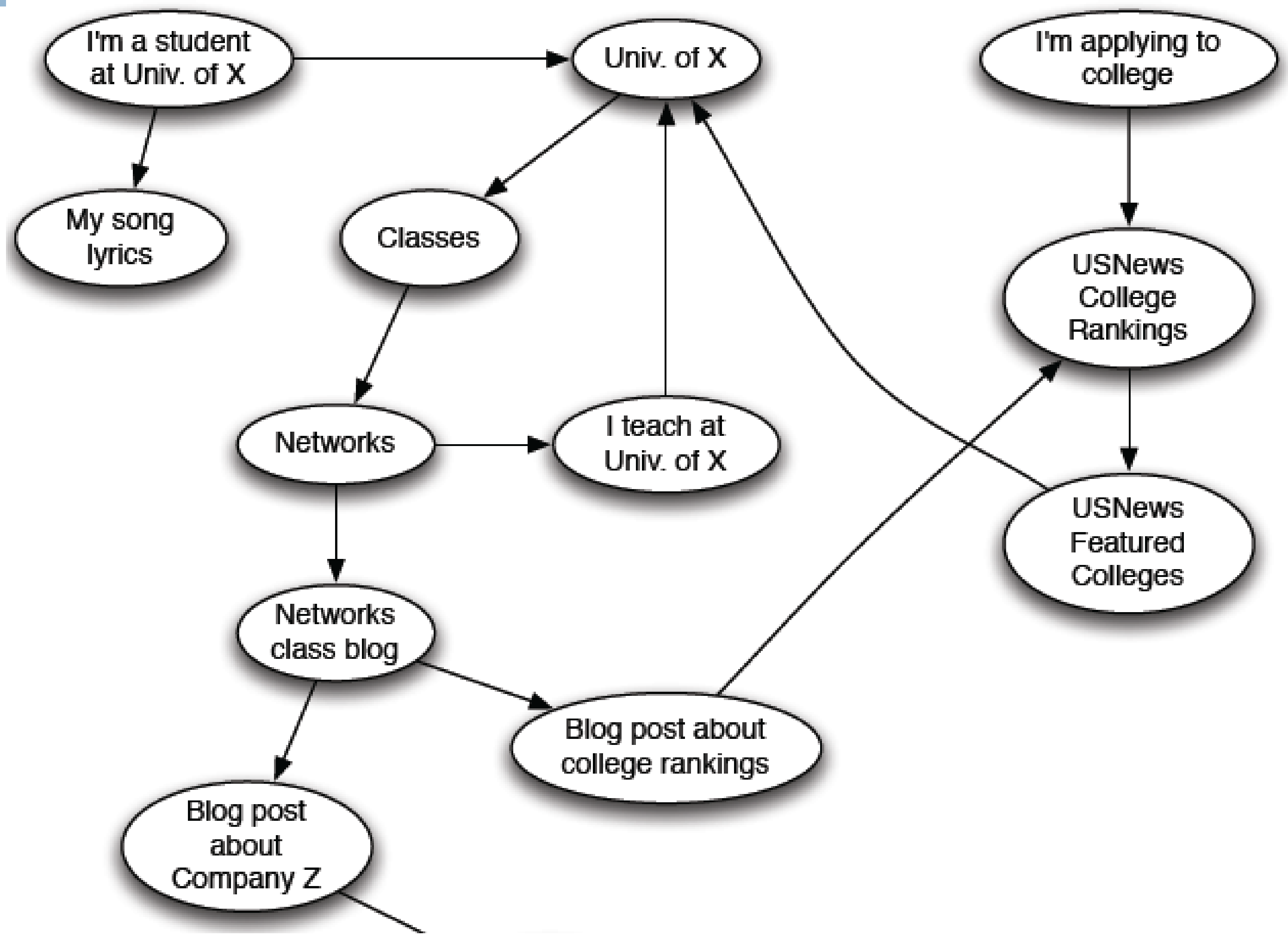


The Web as a Graph

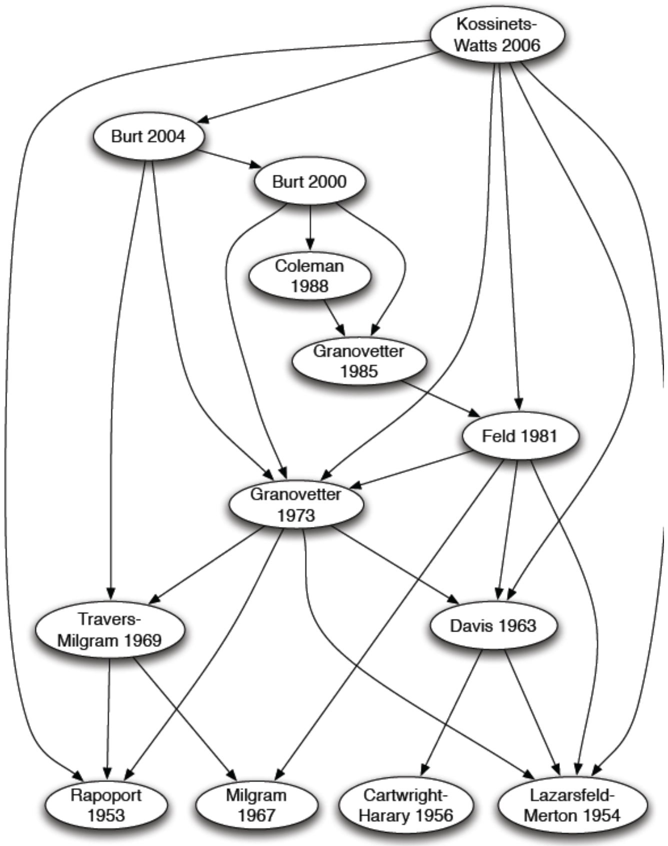


- In early days of the Web links were **navigational**
- Today many links are **transactional**

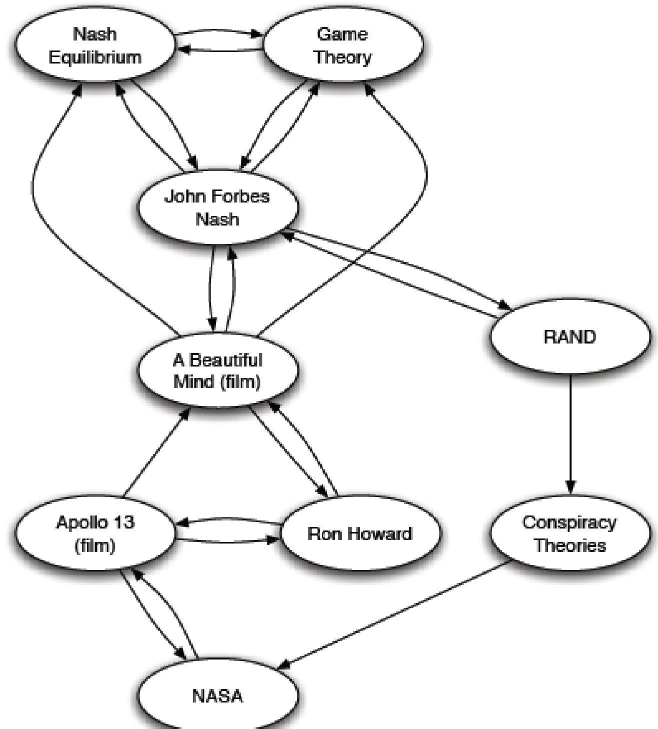
The Web as a Directed Graph



Other Information Networks



Citations



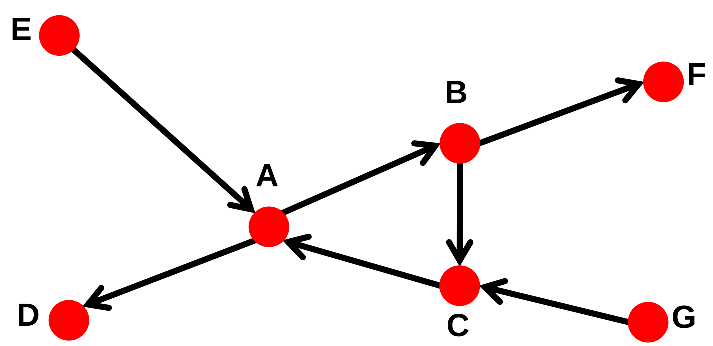
References in an Encyclopedia

What Does the Web Look Like?

- How is the Web linked?
- What is the "map" of the Web?

Web as a directed graph [Broder et al. 2000]:

- Given node v, what can v reach?
- What other nodes can reach v?

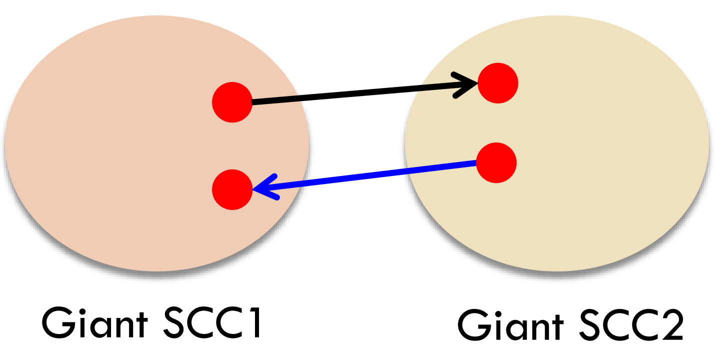


$$In(v) = \{w \mid w \text{ can reach } v\}$$
$$Out(v) = \{w \mid v \text{ can reach } w\}$$

For example:
In(A) = {A,B,C,E,G}
Out(A)={A,B,C,D,F}

Graph Structure of the Web

- **There is a giant SCC**
- **There won't be 2 giant SCCs**
- Heuristic argument:
  - It just takes 1 page from one SCC to link to the other SCC
  - If the 2 SCCs have millions of pages the likelihood of this not happening is very very small



Structure of the Web

- **Broder et al., 2000:**
  - Altavista crawl from October 1999
    - 203 million URLS
    - 1.5 billion links
  - Computer: Server with 12GB of memory
- **Undirected version of the Web graph:**
  - 91% nodes in the largest weakly conn. component
  - Are hubs making the web graph connected?
    - Even if they deleted links to pages with in-degree >10 WCC was still ≈50% of the graph

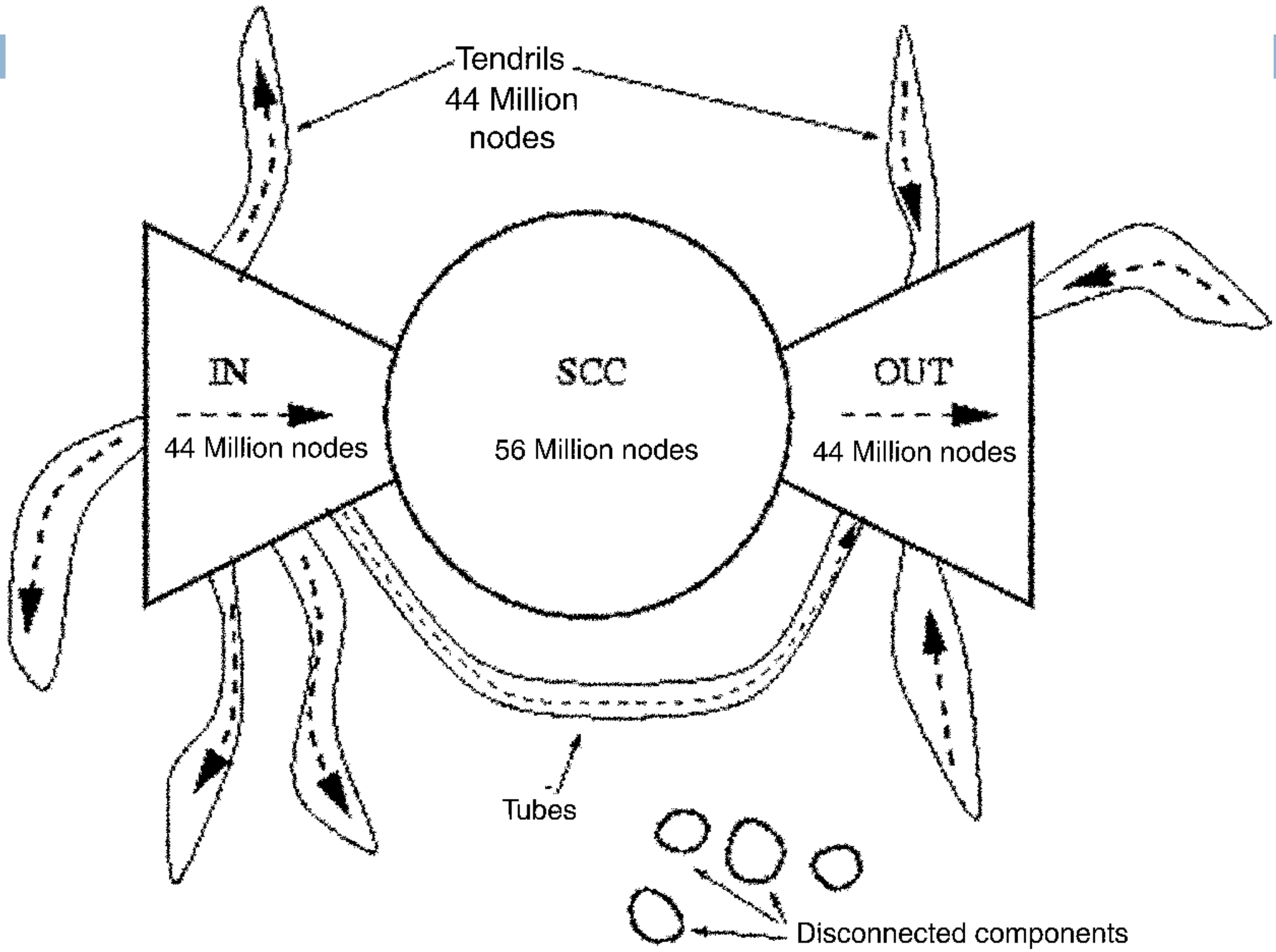
Question about the bias coming from the BFS nature of crawling the graph.

Structure of the Web

- **Directed version of the Web graph:**
  - **Largest SCC:** 28% of the nodes (56 million)
  - Taking a random node v
    - Out(v) ≈ 50% (100 million)
    - In(v) ≈ 50% (100 million)

- **What does this tell us about the conceptual picture of the Web graph?**

Bow-tie Structure of the Web



203 million pages, 1.5 billion links [Broder et al. 2000]

What did We Learn/Not Learn ?

- Learn:
  - Some conceptual organization of the Web (i.e., the bowtie)
- Not learn:
  - **Treats all pages as equal**
    - Google's homepage == my homepage
  - **What are the most important pages**
    - How many pages have k in-links as a function of k?
      The degree distribution: $\sim 1 / k^2$
    - Link analysis ranking -- as done by search engines (PageRank)
  - **Internal structure inside giant SCC**
    - Clusters, implicit communities?
  - **How far apart are nodes in the giant SCC:**
    - Distance = # of edges in shortest path
    - Avg = 16 [Broder et al.]