

Class 3 - Paths, Small World and Connectivity

Course: Computational Network Analysis

Prof. Dr. Claudia Müller-Birn

Institute of Computer Science, «Human-Centered Computing»

Feb 24, 2016

Recap

- We learnt about the difference between a network and a graph.
- We never forget again how an adjacency list is defined.
- We received a general overview on possible network visualizations.
- We understood that collecting “good” data and defining an appropriate network might be challenging.

Today's outline

- Overview on measuring network properties
- Basic properties of networks
 - Degree
 - Mean Degree
 - Degree distribution
 - The power law degree distribution
 - Density
 - Path
 - Distance
 - The Small-World Phenomenon
 - Eccentricity
 - Connectivity
 - Components
 - The structure of the Web

Dimensions of Analyzing Networks

Semantic Dimension of Network Analysis

Presentation of qualitative data (e.g. tabular with frequency, bar chart)

Presentation of quantitative data (e.g. histogram)

Measures of central tendency and variability (e.g. mean, range)

Syntactical Dimension of Network Analysis

Local structure

- Degree
- Degree Centrality
- Closeness Centrality
- Betweenness Centrality
- Local CC

Global structure

- Mean degree
- Degree distribution
- Density
- Network Centralization
- Components

Partitions

- Local definition, such as clique, k-core, k-plex
- Global definition with null model (modularity with modularity optimization and edge betweenness)

Dimensions of Analyzing Networks

Semantic Dimension of Network Analysis

Presentation of qualitative data (e.g. tabular with frequency, bar chart)

Presentation of quantitative data (e.g. histogram)

Measures of central tendency and variability (e.g. mean, range)

Syntactical Dimension of Network Analysis

Local structure

- **Degree**
- Degree Centrality
- Closeness Centrality
- Betweenness Centrality
- Local CC

Global structure

- **Mean degree**
- **Degree distribution**
- **Density**
- Network Centralization
- Components

Partitions

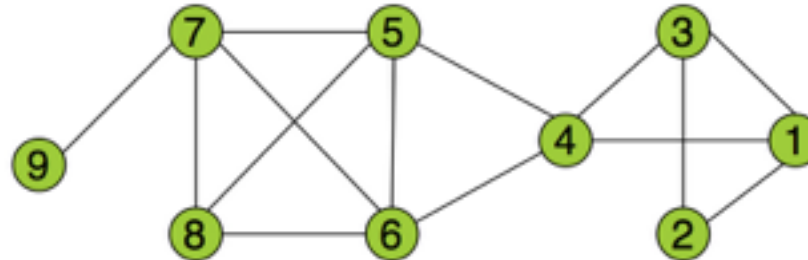
- Local definition, such as clique, k-core, k-plex
- Global definition with null model (modularity with modularity optimization and edge betweenness)

Properties of Large Networks

- Large network comparison is computationally hard due to NP-completeness of the underlying subgraph isomorphism problem
- Given two graphs G and H as input, determine whether G contains a subgraph that is isomorphic to H .
- Thus, network comparisons rely on easily computable **heuristics** (approximate solutions), called “**network properties**”

Basic properties of networks

Basic properties



- N_i : Adjacent nodes of node i
- k_i : Degree of node i ($d_i = |N_i|$)
- g_{ij} : Geodesic distance between nodes j and i
- D : Maximum of the distances

Degree

- The **degree** of a vertex v in an undirected graph $G = (V, E)$, denoted by $d(v)$, is the number of edges in E that have v as an endvertex.
- If G is a multigraph, parallel edges are counted according to their multiplicity in E .
- The **degree sequence** of an undirected graph is the non-increasing sequence of its vertex degrees
- The **rank** r_v of a vertex v is its index in the degree sequence

In vs. out-degree

- In directed networks each vertex has two degrees
 - In-degree: number of ingoing edges connected to the vertex
 - Out-degree: number of outgoing edges connected to the vertex
- Number of edges m in a directed network equals to the total number of ingoing ends of edges at all vertices or equals to the number of outgoing edges

$$m = \sum_{i=1}^n k_i^{in} = \sum_{j=1}^n k_j^{out}$$

- Thus, the mean in-degree and the mean out-degree are equal

How can the in-/out-degree be interpreted?

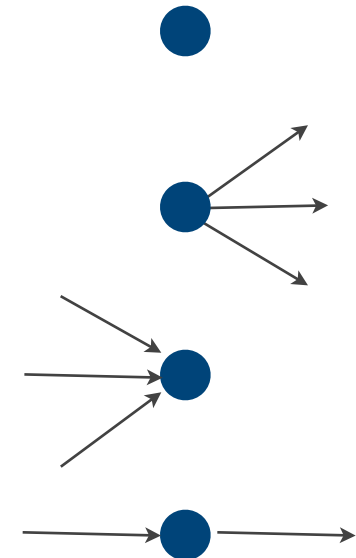
- Types of vertices in a directed (communication) network

- Isolate if $k_i^{in} = k_i^{out} = 0$

- Transmitter if $k_i^{in} = 0$ and $k_i^{out} > 0$

- Receiver if $k_i^{in} > 0$ and $k_i^{out} = 0$


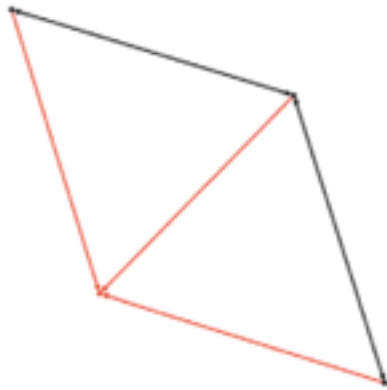
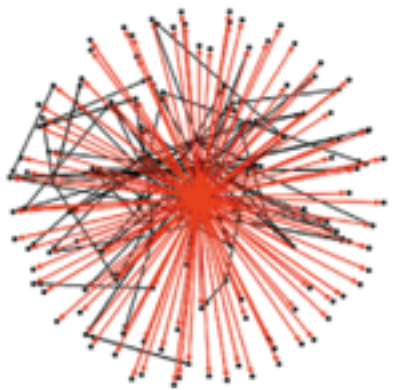
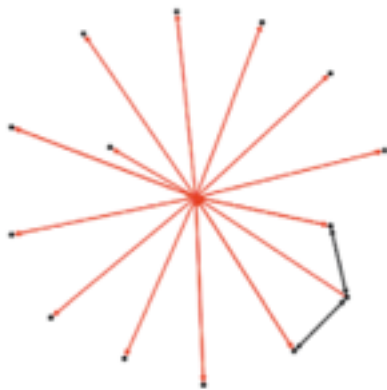
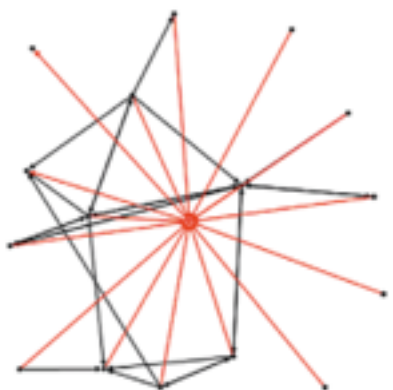
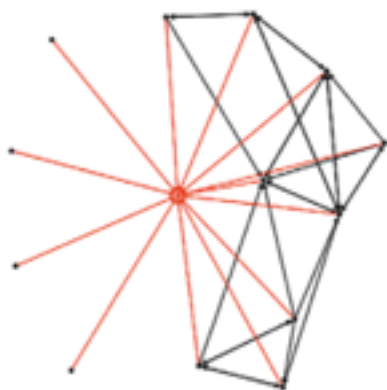
- Carrier if $k_i^{in} > 0$ and $k_i^{out} > 0$



Communication network of a technical support email list

- Network building
 - Threaded conversations is represented in a reply network
 - Each time someone replies to another person's message, she creates a directed link to that other person
 - If she replies to the same person multiple times, a stronger weighted edge is created
- Goal
 - Users contribute in different patterns and styles (i.e. social roles)
 - Understanding the composition of social roles within your community can provide many insights that can help you be a more effective community manager or provide new functionalities

Using syntactical data for social roles

Question People <ul style="list-style-type: none"> • Low In- and Out-Degree • High Avg Degree of Neighbors 		
Answer People <ul style="list-style-type: none"> • High % Out-Degree • Low Clustering Coefficient • Low Avg Degree of Neighbors 		
Discussion Starters <ul style="list-style-type: none"> • Low % Out-Degree • High Clustering Coefficient • High Avg Degree of Neighbors 		

Including semantic data for social roles

Metric	Description
$(\text{User's Thread Count}) \div (\text{User's Post Count})$	Brevity is preferred. Larger values = fewer messages per thread.
$(\text{User's Reply Posts}) \div (\text{User's Total Posts})$	Initiation is avoided. Larger values = avoids starting threads.
$(\text{User's Degree}) \div (\text{Total Users})$	Talks to many people. Larger values = replies to a significant fraction of community members.
$(1 - \text{Clustering Coefficient})$	Talks to people who aren't well connected to each other. Larger values = lower clustering coefficient (i.e., less well-connected neighbors)
$1 \div \text{Avg of Neighbor's Degree}$	Talks to people who connect to few others. Larger values = talks to more isolates
$(\text{User's Days Active}) \div (\text{User's Possible Active Days})$	Posts on most days. Larger values = posts on multiple days more often.
$(\text{User's Out-Degree}) \div (\text{User's Out-Degree} + \text{User's In-Degree})$	Percent out-degree. Larger values = is connected to more people because of replying to them than because of receiving from them.

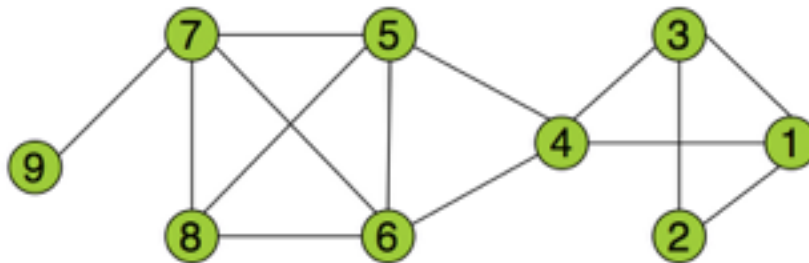
Mean Degree

- The mean degree of an undirected graph is $c = \frac{1}{n} \sum_{i=1}^n k_i$
- An undirected graph $G = (V, E)$ is called **regular** if all its vertices have the same degree
- In an undirected network there are m edges in total and therefore there are $2m$ ends of edges, therefore, the number of all ends of edges is equal to the degree of all vertices

$$2m = \sum_{i=1}^n k_i$$

Degree distribution

- For a graph $G = (V, E)$ and for $k = 1, 2, \dots$, the degree distribution of G is P_k fraction of vertices with degree k .
- If there are N nodes in total in a network and N_k of them have degree k , we have $P(k) = N_k/N$. If we pick a vertex $v \in V$ uniformly at random, $\text{Prob}[\text{deg}(v) = k] = P_k$.



- The degree of distribution of a graph is the sequence $P(k=0), P(k=1), \dots$
- $P(k=0) = 0/9$
- $P(k=1) = 1/9$
- $P(k=2) = 2/9 \dots$

Famous example of a degree distribution

On Power-Law Relationships of the Internet Topology

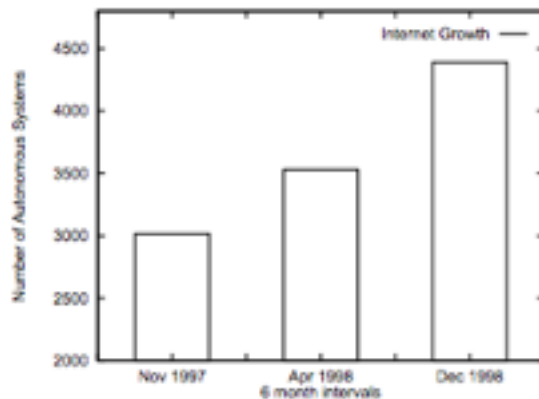
Michalis Faloutsos
U.C. Riverside
Dept. of Comp. Science
`michalis@cs.ucr.edu`

Petros Faloutsos
U. of Toronto
Dept. of Comp. Science
`pfal@cs.toronto.edu`

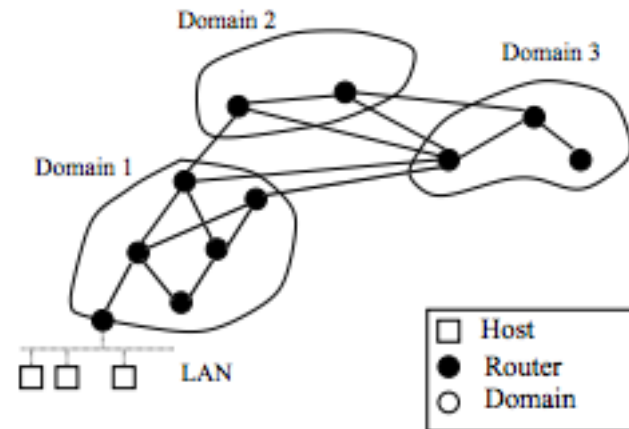
*Christos Faloutsos **
Carnegie Mellon Univ.
Dept. of Comp. Science
`christos@cs.cmu.edu`

<http://www.cs.cmu.edu/~christos/PUBLICATIONS/sigcomm99.pdf>

Famous example of a degree distribution is power law degree distribution



Growth of the Internet

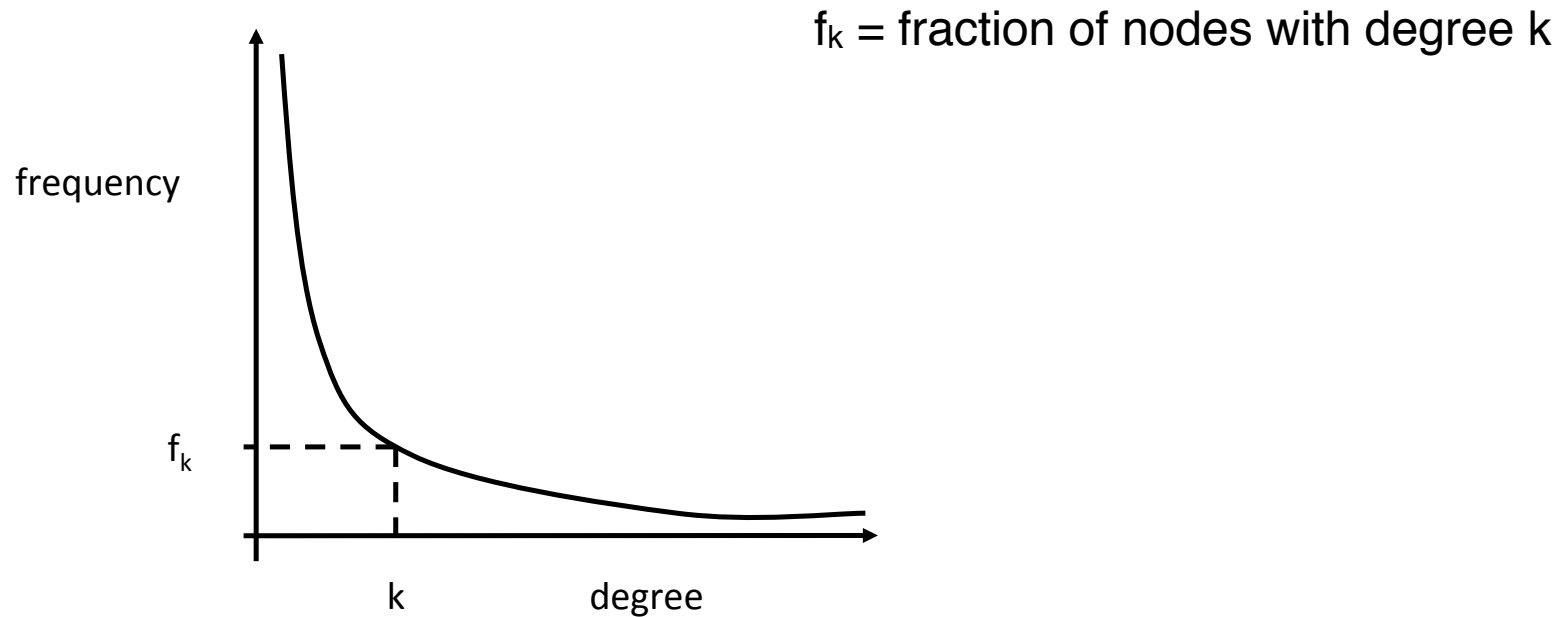


Structure of the Internet at the router level

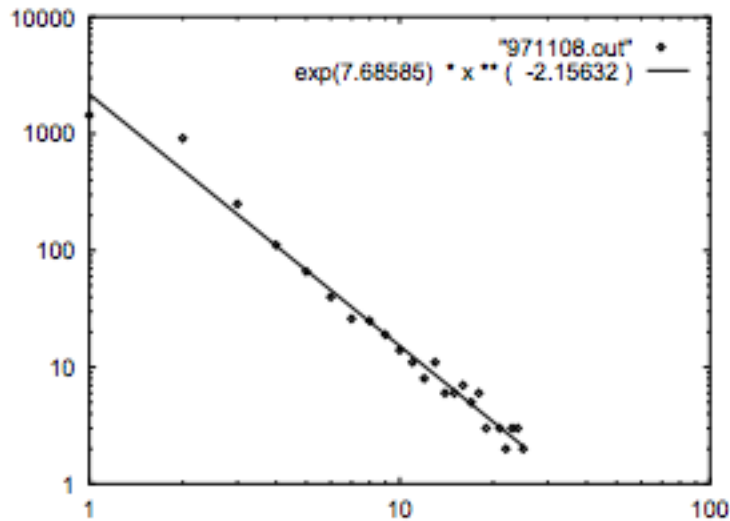
	Int-11-97	Int-04-98	Int-12-98
nodes	3015	3530	4389
edges	5156	6432	8256
avg. outdegree	3.42	3.65	3.76
max. outdegree	590	745	979
diameter	9	11	10
avg. distance	3.76	3.77	3.75

The evolution of the Internet at the inter-domain level.

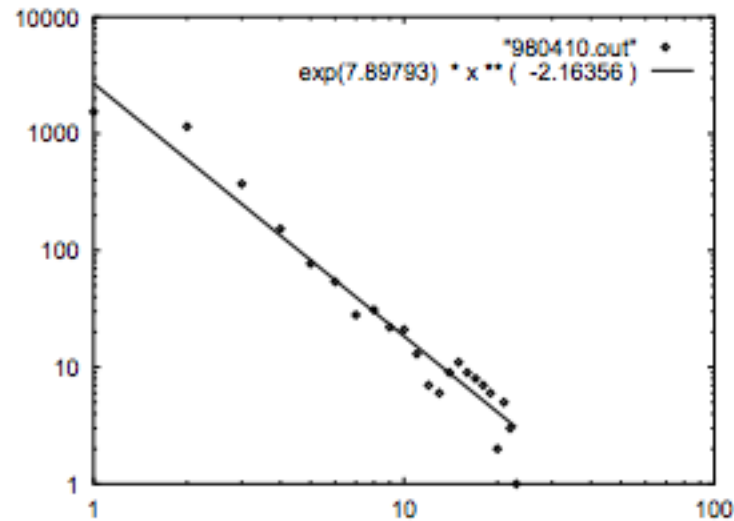
Degree distribution of the inter-domain topology of the Internet



Degree distribution of the inter-domain topology of the Internet



(a) Int-11-97



(b) Int-04-98

The outdegree plots: Log-log plot of frequency f_k versus the outdegree k

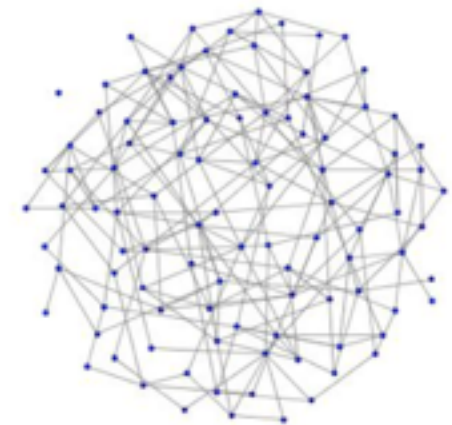
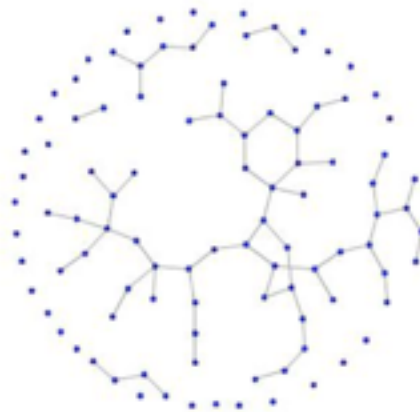
Observation:

- Distribution was linear
- This means that $\log P_k$ can be written as a linear function

Density

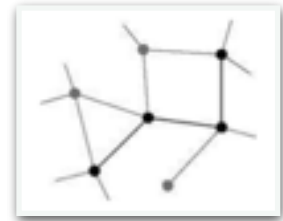
- The density of a network is defined as a ratio of the number of edges E to the number of possible edges.

$$\rho = \frac{2m}{n(n-1)}$$



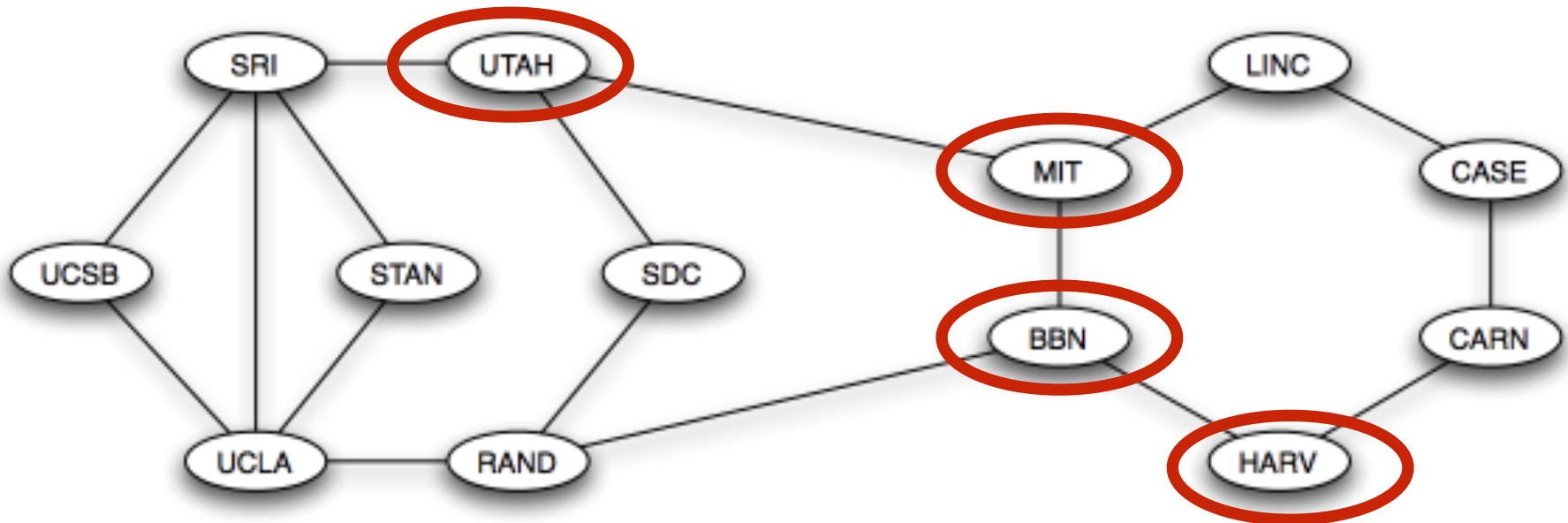
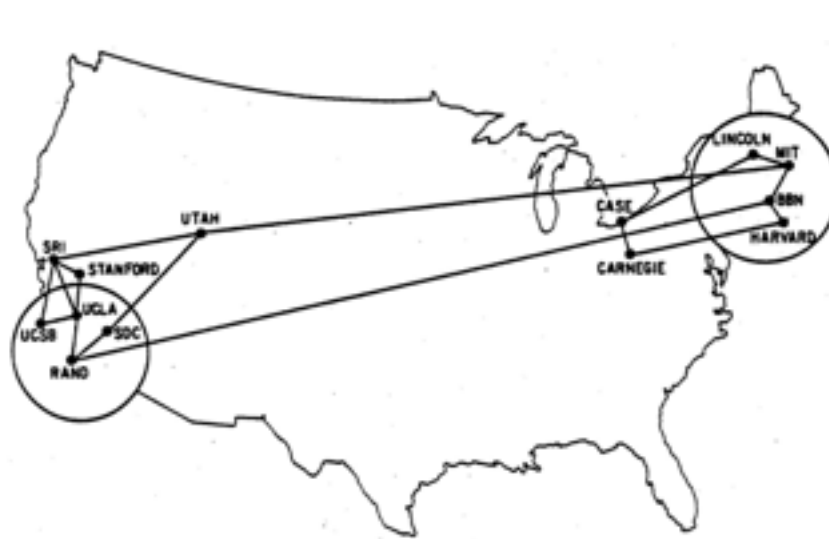
Path

- A **path** in a network is any **sequence of vertices** such that every consecutive pair of vertices in the sequence is connected by an edge in the network
- Paths can be defined for both directed and undirected networks
 - In a directed network, each edge traversed by a path must be traversed in the correct direction for that edge
 - In an undirected network edges can be traversed in either direction
- A path can intersect itself, visiting again a vertex it has visited before, or even running along an edge or set of edges more than once
- Paths that do not intersect themselves are called ***self-avoiding paths***, for example the geodesic path
- The **length of a path** in a network is the **number of edges** traversed along the path (not the number of vertices!)



A path of length three in a network

Internet in 1970



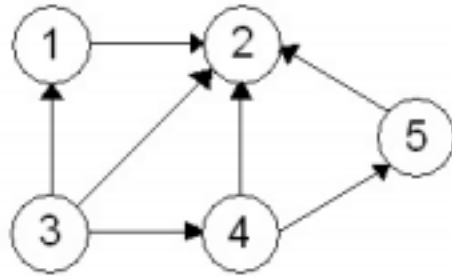
Paths in undirected networks

- The product $A_{ik}A_{kj}$ is 1 iff there is a path between i and j through k (a path of length 2)
- Hence, if we want to find out how many length 2 paths between i and j exist:

$$N_{ij}^{(2)} = \sum_{k=1}^n A_{ik}A_{kj} = [A^2]_{ij}$$

- This can easily be generalized to: $N_{ij}^{(r)} = [A^r]_{ij}$
- $[A^r]_{ii}$ gives the number of length r paths that originate and end at node i (cycles)

Example for determining paths



$$G_1 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

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

$$A^1 = A$$

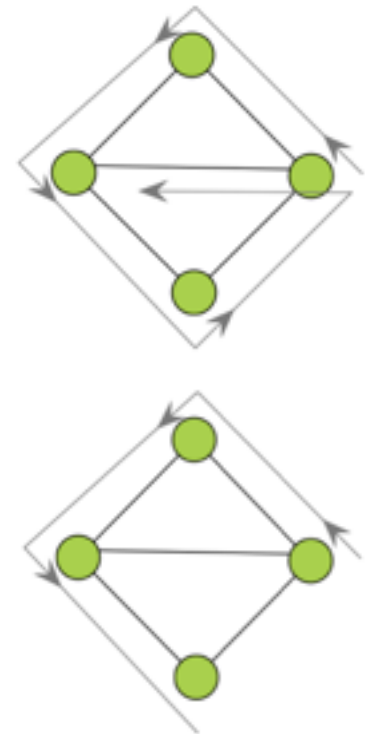
$$A^2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 1 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$A^3 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$A^4 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

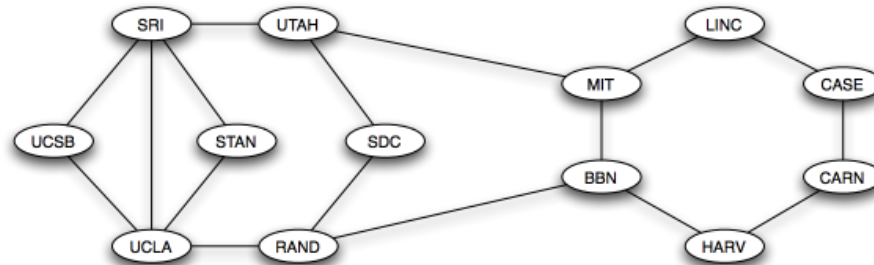
Approaches to find a path

- Eulerian path: traverses each **edge** in a network exactly **once**
- Hamiltonian path: visits each **vertex** in a network exactly **once**
- Possible challenges
 - Two vertices might be not connected by any route in the network because the network is separated in components
 - Geodesic paths are not necessarily unique



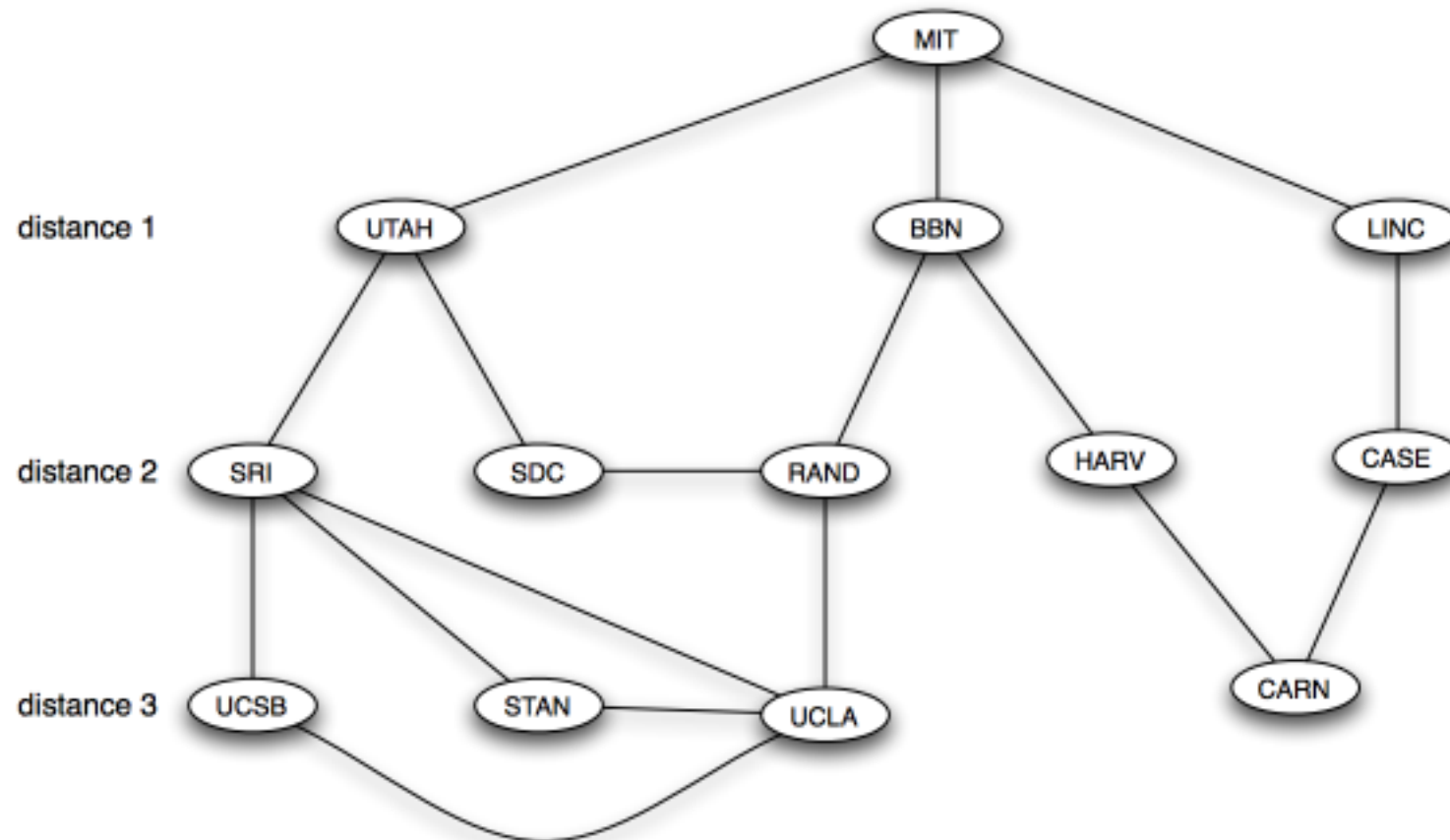
Distance

- The **length of a path** is the number of steps it contains from beginning to end — in other words, the number of edges in the sequence that comprises it.
- Example: MIT, BBN, RAND, UCLA vs. MIT, UTAH



- The **distance** between two nodes in a graph is the length of the shortest path between them a graph
- Geodesic Paths**
 - Length of the geodesic path is often called geodesic distance or shortest distance
 - Minimum is 1, if the graph is complete or maximum $m-1$

Breadth-First Search



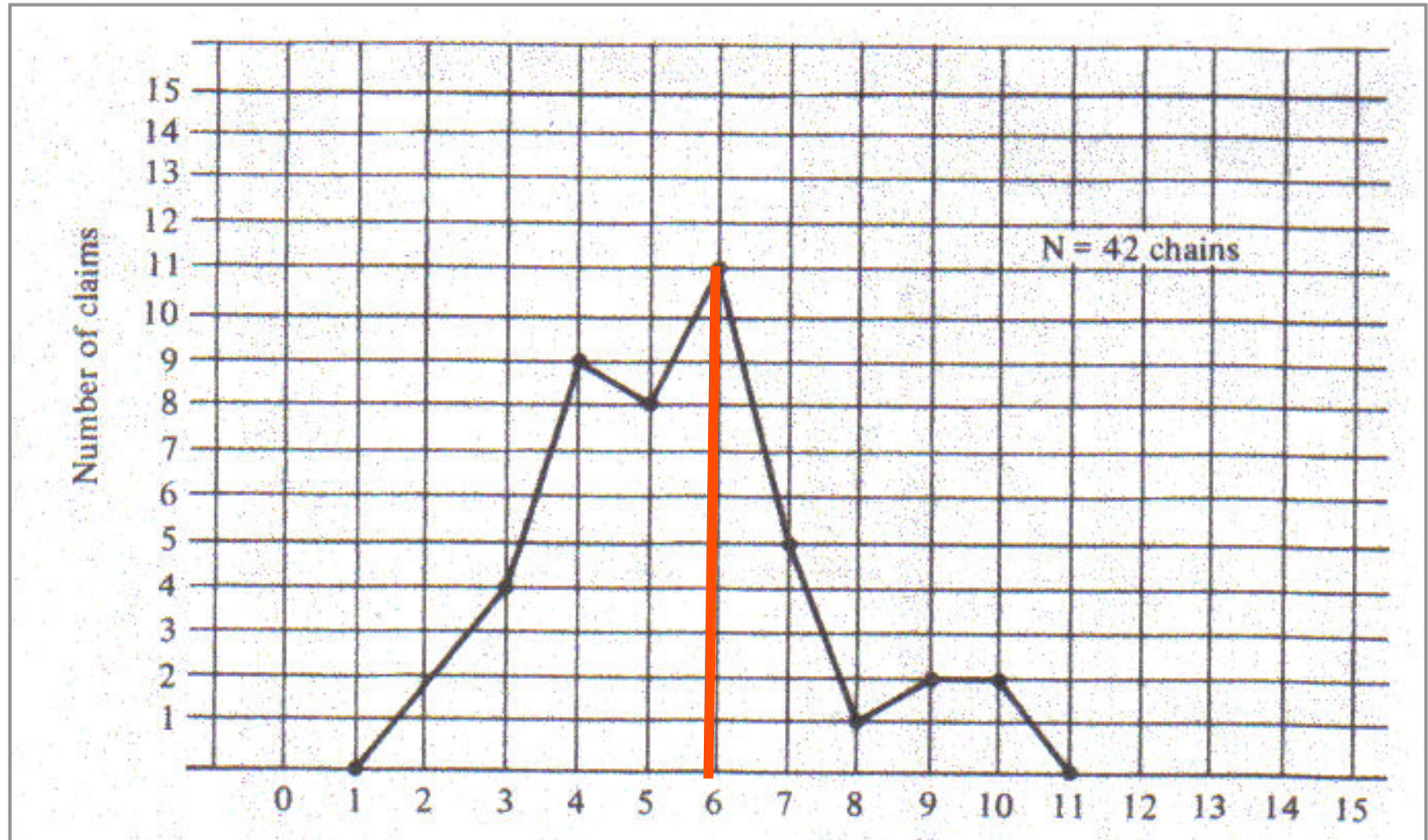
The Small-World Phenomenon

Experiment by Stanley Milgram and his colleagues in the 1960s

- Hypothesis
 - Two random people are connected through only a few intermediate acquaintances
- Set up
 - Choose 300 people at random and ask them to send a letter through friends to a stockbroker near Boston
 - 160 letters: From Wichita (Kansas) and Omaha (Nebraska) to Sharon (Mass)
- Chains consists of approx. six persons („six degrees of separation“), because of so called social hubs, that are persons which have a extremely high number of acquaintances (e.g., chancellor of Germany, Dalai Lama, Boris Becker, ...)

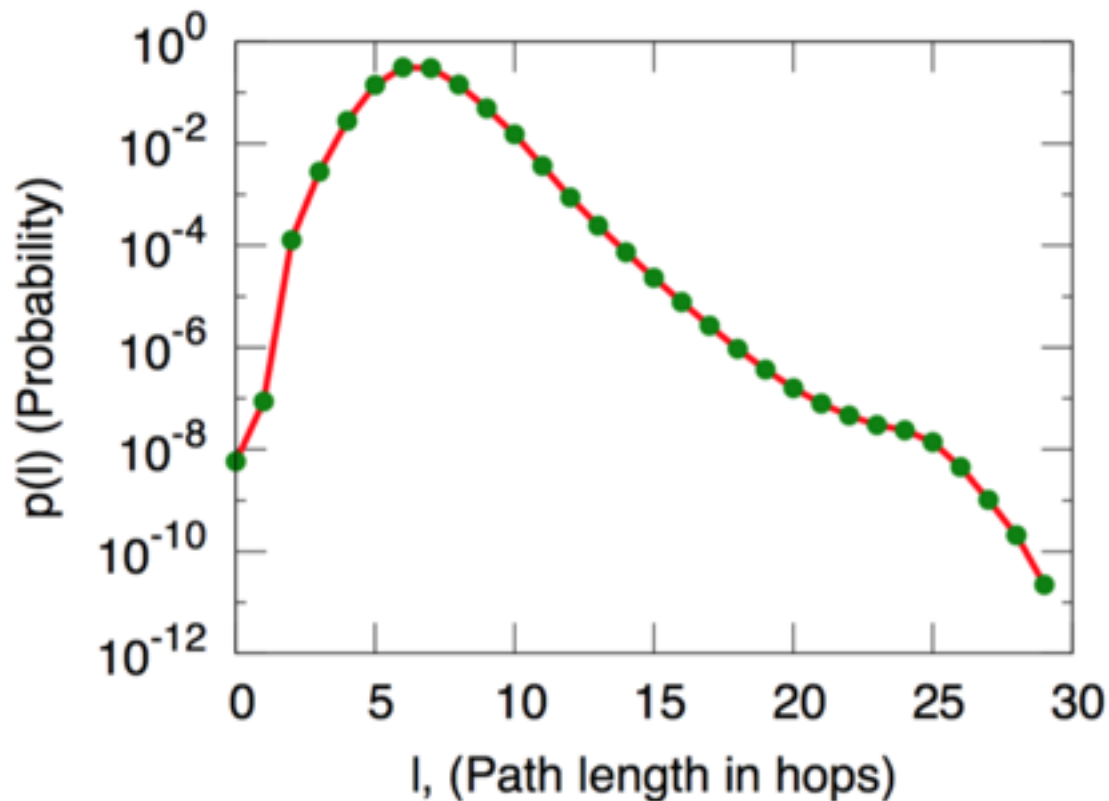


Distribution of path lengths of successful chains



Distribution of path lengths of corporate IM

- Distribution of distances in a graph of all active Microsoft Instant Messenger user accounts, with an edge joining two users if they communicated at least once during a month-long observation period

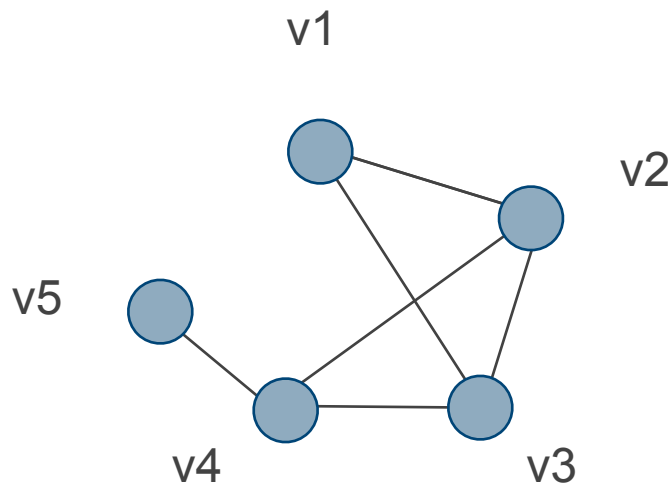


Additional concepts

- The **average path length** is calculated by finding the shortest path between all pairs of nodes, adding them up, and then dividing by the total number of pairs. It shows, on average, the number of steps it takes to get from one member of the network to another.
- The **eccentricity** $\varepsilon(v)$ of a vertex v is the greatest geodesic distance between v and any other vertex.
- The **radius** r of a graph is the minimum eccentricity of any vertex or, in symbols, $r = \min \varepsilon(v)$, where v in V

Diameter of a network

- Diameter of a network is the length of the longest geodesic path between any pair of vertices in the network for which a path actually exists
- The diameter d of a graph is the maximum eccentricity of any vertex in the graph; $d(G) = \max \varepsilon(v)$, where v in V .



$$g(1, 2) = 1$$

$$g(1, 3) = 1$$

$$g(1, 4) = 2$$

$$g(1, 5) = 3$$

$$g(2, 3) = 1$$

$$g(2, 4) = 1$$

$$g(2, 5) = 2$$

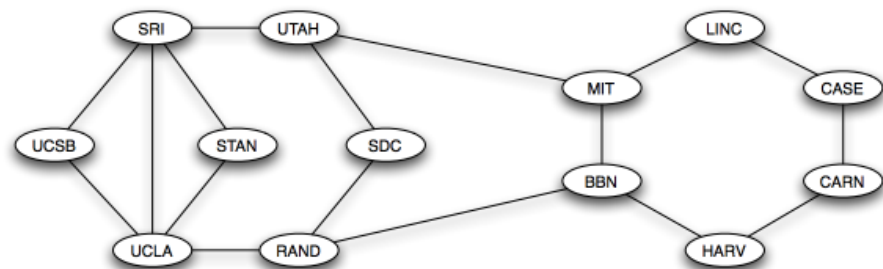
$$g(3, 4) = 1$$

$$g(3, 5) = 2$$

$$g(4, 5) = 1$$

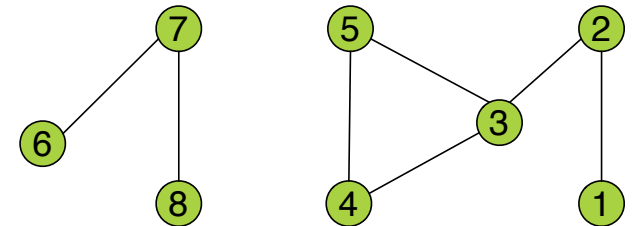
Connectivity

- A graph is connected when there is a path between every pair of vertices. In a connected graph, there are no unreachable vertices. A graph that is not connected is disconnected.
- A graph with just one vertex is connected. An edgeless graph with two or more vertices is disconnected.
- For example the ARPANET is connected but many other networks are not



Components

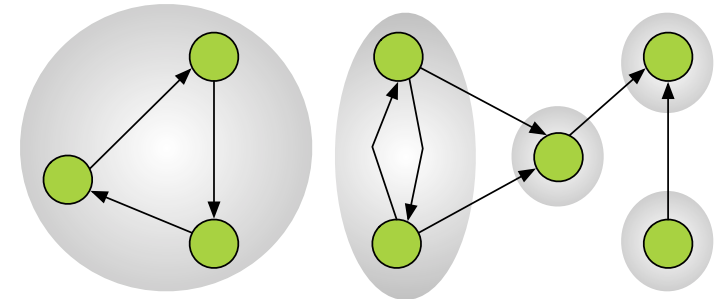
- A network must not be connected, it can be divided in subgroups that are not connected
- Such a network is called “disconnected” and the subgroups are called components
- **Component:** subset of vertices of a network such that there exist at least one path from each member of that subset to each other member
- Adjacency matrix of a network with more than one component can be written in block diagonal form



$$\mathbf{A} = \begin{pmatrix} \boxed{} & 0 & \dots \\ 0 & \boxed{} & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

Components in directed networks

- If we ignore the direction of the edges then we have two components = *weakly connected components*
- Two vertices are in the same weakly connected component if they are connected by one or more paths through the network
- Strongly connected component
 - Two vertices are connected if and only if there exists both a directed path from A to B and a directed path from B to A
 - Every component with more than one vertex must contain at least one cycle
- Other definitions are out-component and in-component



Giant component

- Informal term that describes a connected component that contains a significant fraction of all the nodes
- When a network contains a giant component, it almost always contains only one.



Peter Bearman, James Moody, and Katherine Stovel. Chains of affection: The structure of adolescent romantic and sexual networks. American Journal of Sociology, 110(1):44–99, 2004.

The structure of the Web

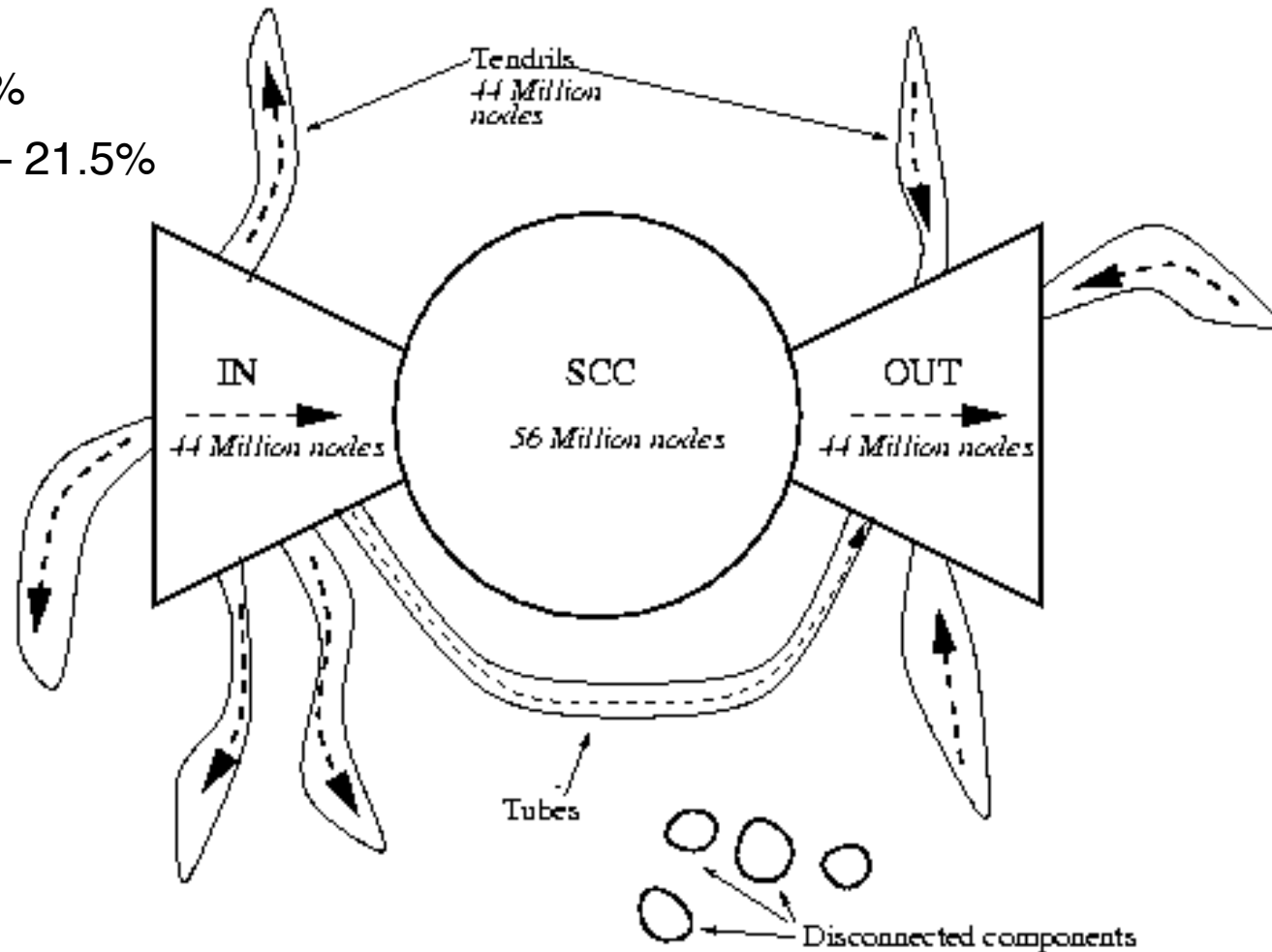
- Data set
 - Web crawl from May 1999 and October 1999
 - Analysis is based on over 200 million pages and 1.5 billion links
 - The Web is a directed graph because webpages link to other webpages
- The connected components tell us what set of pages can be reached from any other just by surfing (no 'jumping' around by typing in a URL or using a search engine)
- If links are treated as undirected edges most (over 90%) of the approximately 203 million nodes form a single connected component
- What about directed links (since hyperlinks are directed)?

The structure of the Web (*cont.*)

- The connected web breaks naturally into four pieces - it forms the *bow-tie-model* of the Web
- This model consists of
 - *SCC (strongly connected component)*: Can reach all nodes from any other by following directed edges
 - *IN*: Can reach SCC from any node in 'IN' component by following directed edges
 - *OUT*: Can reach any node in 'OUT' component from SCC
 - *Tendrils and tubes*: Connect to IN and/or OUT components but not SCC
 - *Disconnected*: Isolated components
- The size of the SCC is relatively small -it comprises about 56 million pages
- Each of the other three sets contain about 44 million pages

The structure of the Web (*cont.*)

- SCC – 27.5%
- IN and OUT – 21.5%
- Tendrils and tubes – 21.5%
- Disconnected – 8%



Broder, A. Z., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., et al. (2000). Graph structure in the Web. *Computer Networks*, 33, 309–320.

Fragen?