Corso di laurea in
Ingegneria Informatica

Social Networks Analysis
A.A. 2022/23

Graphs and Networks

Graph Theory

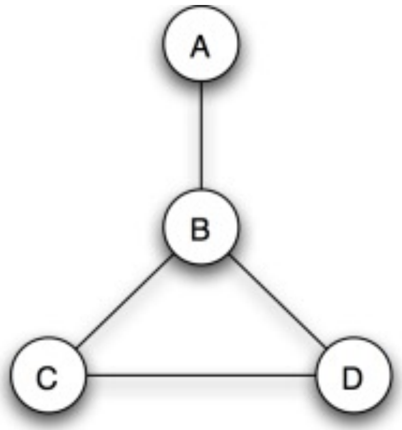
- Graph Theory is a consolidated branch of Mathematics that allows to describe sets of elements together with binary relations between these elements
  - Started in the XVIII century by Euler
    - Solution to the Konigsberg's bridges problem

- Graph theory provides a unifying language to describe the structure of all kinds of networks

- Nowadays, the possibility of gather data on large scale and work with massive data sets pave the way to new approaches and new problems on graphs
  - Recognize statistical properties characterizing the structure of a network and provide methods to measure them
  - Create network models and describe network formation process
  - Predict the behaviour of a network based on their models and structural properties

Graph's Definition

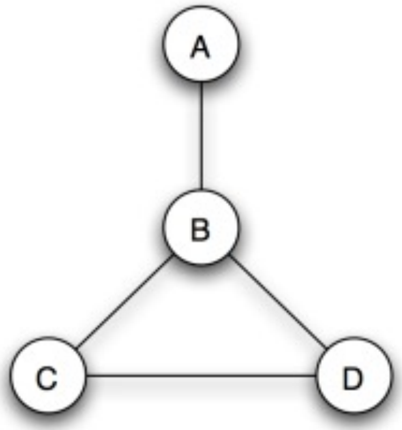
- A *graph* consists of
  - *A set of nodes (vertices)*
  - *A set of edges (links)*
    - *An edge connects two nodes*



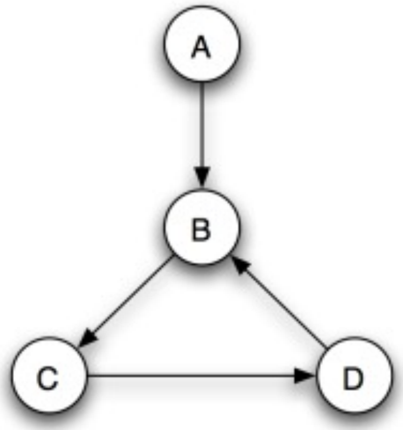
- Two nodes are adjacent (neighbors) if they are connected by an edge
  - C and D are adjacent through the edge (C, D)
  - B is a neighbor of both C and D

DIRECTED AND UNDIRECTED GRAPHS

- Edges can be oriented
  - Directed edge: the relation holds only between the head and the tail of the edge
  - Undirected edge: the relation holds in both the directions



Undirected graphs



Directed graphs

- Directed and undirected edges differ substantially
  - Different models of network formation and maintainance
  - Different algorithms

Graphs as Models of Networks

- Directed and undirected graphs describe different kinds of networks

- Directed graphs
  - The relation between the nodes originates from a unilateral decision
    - Eg: link to a web page, followers, citation of an article

- Undirected graphs
  - The relation between the nodes comes from a decision of both the elements
    - Eg: friendship, alliance, acquaintance, connection, ecc.

Weighted or Signed Graphs

- We can put additional informations to edges
  - sign (friends/enemies)
  - weight (strength of the social connection)
  - distance (connection length)
  - delay (transmission time)
  - reliability (transmission error rate)
  - cost (cost of the link usage)

- In a *weighted* graph each edge has a number associated that is its weight
- In a *signed* graph each edge has a positive/negative sign

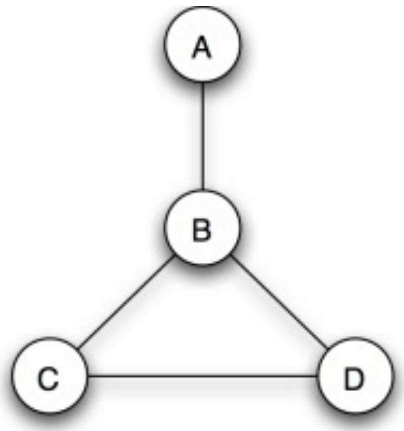
Graph Representation

- Graph Theory is a mathematical theory interesting by itself
  - Studies characteristics and properties of graphs

- A graph is a pair of sets
  - $G = (V, E)$
    - V = vertex set
    - E = edge set

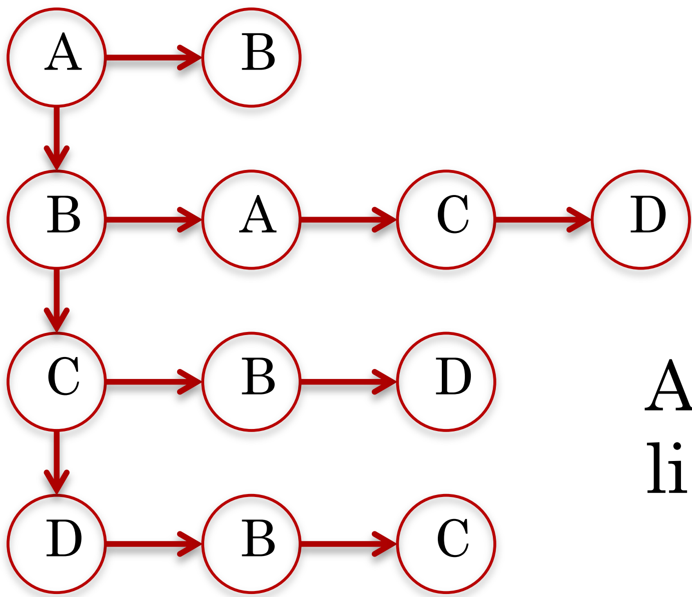
- More used representations
  - Adiacency matrix
    - matrix n x n ($n = |V|$)
    - Element $(i, j) = 1$ if there is an edge between i and j
      - $w_{i,j}$ is the weight of edge (i, j) if the graph is weighted
  - Lists of vertices V and edges E
    - For each vertex v we have the list of vertices adjacent to v
    - For directed graphs we have separate lists of incoming ant outgoing edges
  - Graphic representation

An Example of Graph Representation



$$
\begin{array}{c c}
 & \begin{array}{cccc} A & B & C & D \end{array} \\
\begin{array}{c} A \\ B \\ C \\ D \end{array} & \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}
\end{array}
$$

Adiacency matrix



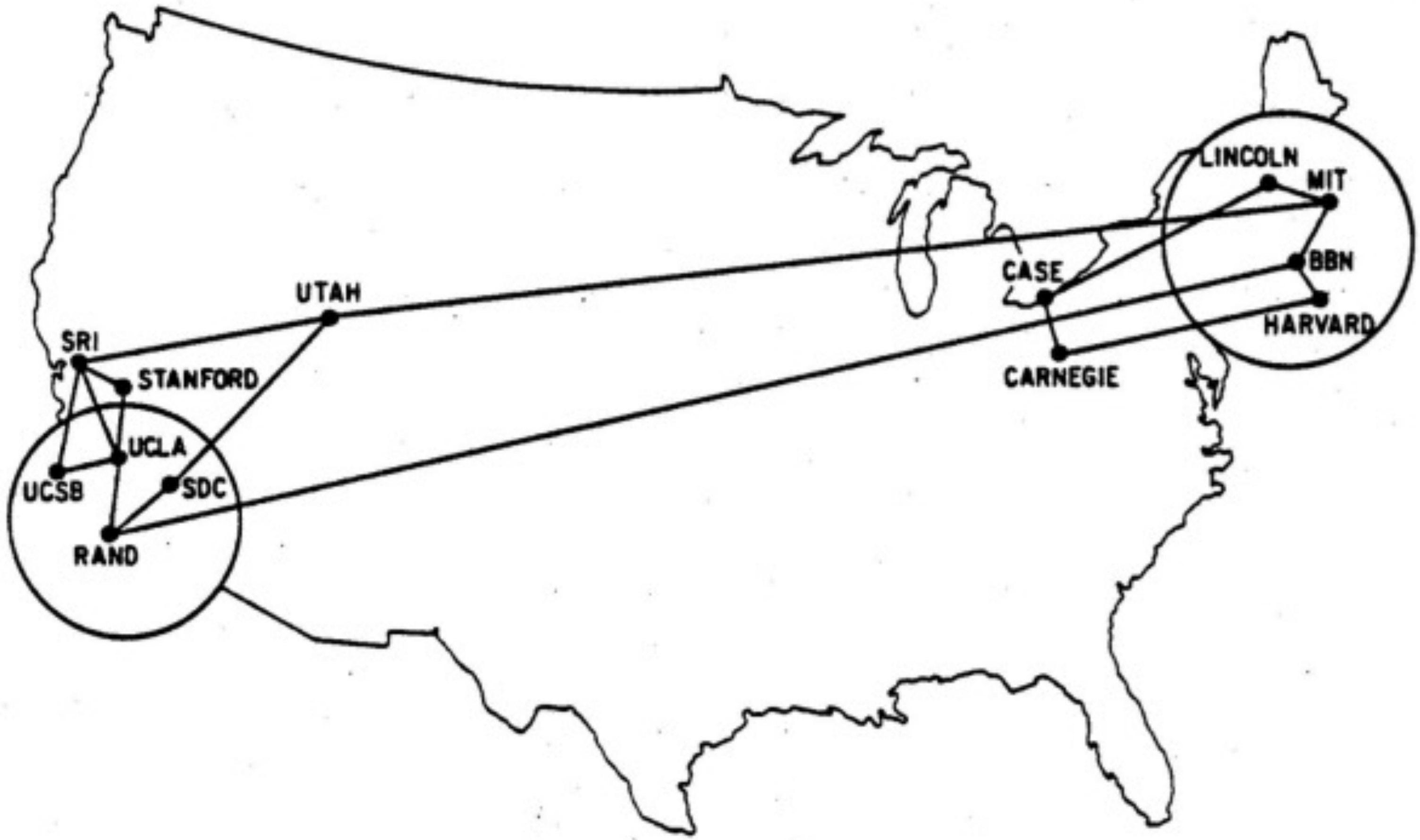
Adiacency lists

Vertices and Edges

- When studying networks vertices and edges can represent entities of the real world
  - Some network abstraction are commonly used

- Some examples
  - Communication networks
    - Network devices, communication links
  - Social networks
    - people, friendships/social connections
    - companies, commercial relations
  - Information networks
    - Web sites, hyperlinks
  - Biological networks
    - Species, predator-prey links
    - molecules, chemical bonds

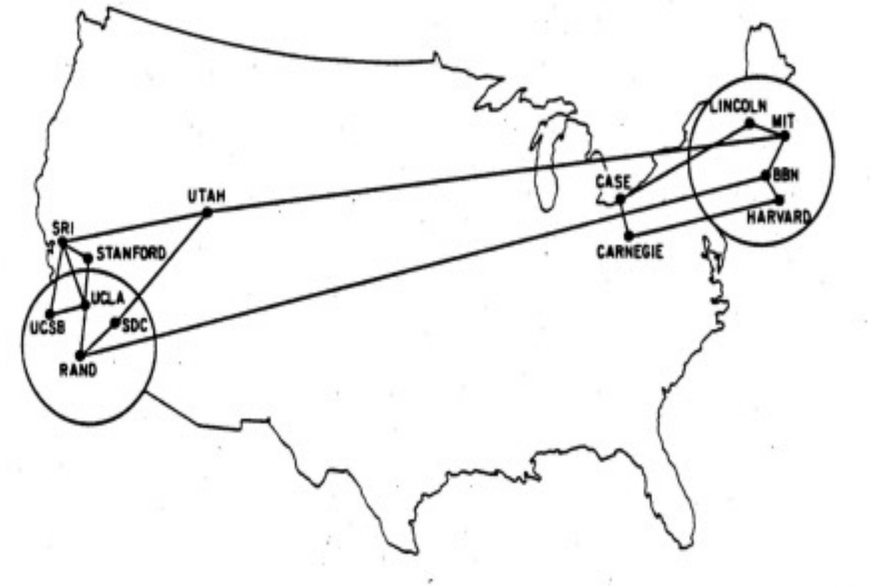
ARPANET: FIRST VERSION OF INTERNET

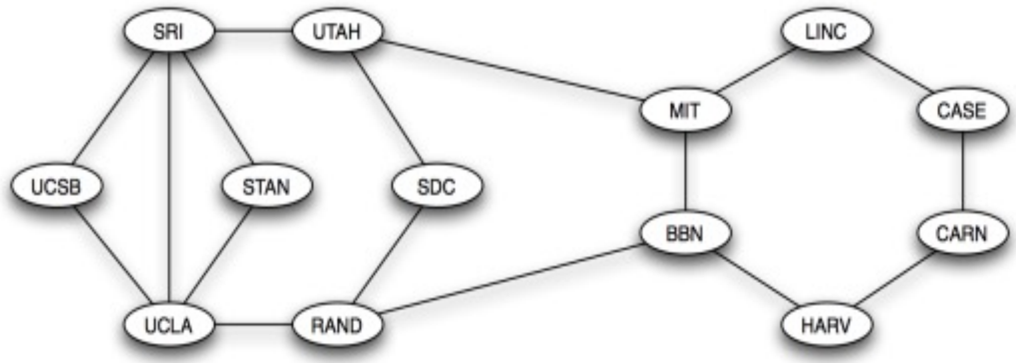


- Created in 1970 with 13 nodes

THE GRAPH OF ARPANET

- We are interestd only in connectivity
  - Distances can be represented as edge weights



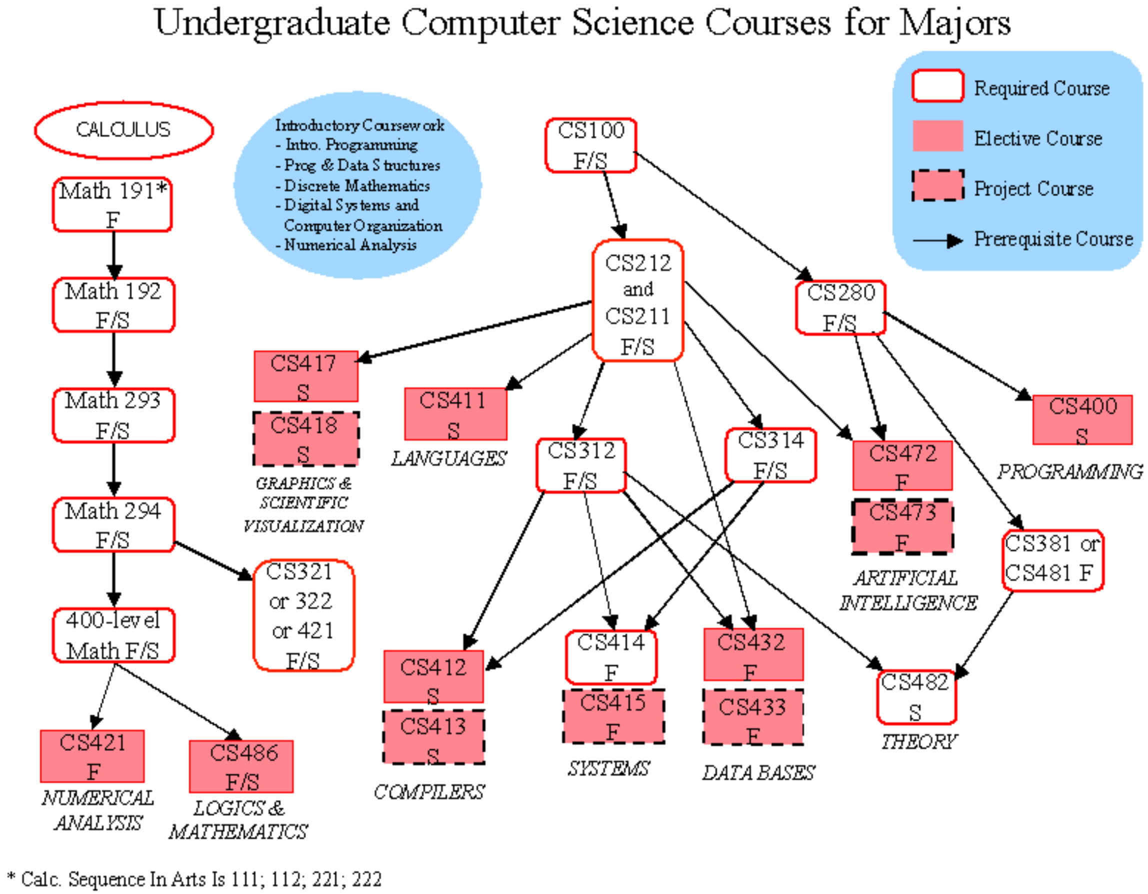


Transportation Networks

- Most of the graph terminology derives form metaphors based on transportation
  - "shortest path", "diameter", "flow"

- Questions we are interested in
  - The network structure can support the required performances? How much robust is it? Is it exposed to cascading failures?

Dependency Graphs



- Nodes are tasks and directed edges are dependencies
- To design complex software systems or industrial processes we need to carefully analize the dependency graph to define a good scheduling policy and avoid deadlocks

Structural Networks

- The internal frameworks of mechanical structures such as buildings, vehicles, or human bodies are based on such networks
- rigidity theory studies the stability of such structures from a graph-based perspective

Most Relevant Concepts on Graphs

- "Graph theory is a terminological jungle in which every newcomer may plant a tree"
  (Social scientist John Barnes)

- We will introduce only concepts that are relevant for the scope of this course
  - paths
  - cicles
  - connectivity
  - components (giant component)
  - distance

Paths

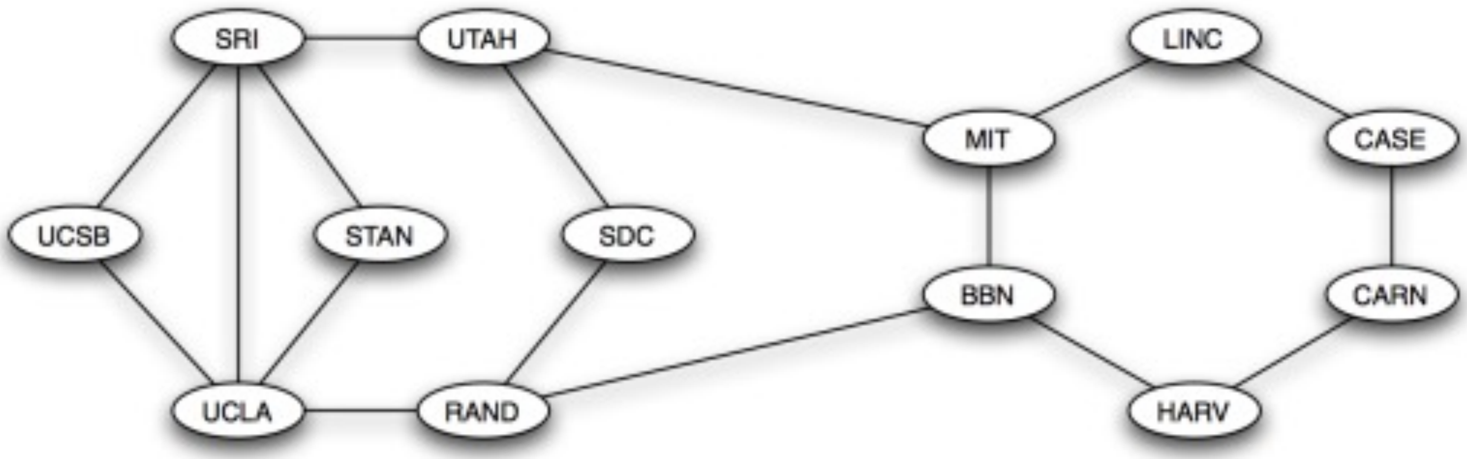
- A characteristic of the social networks is that nodes can influence each other indirectly
  - Influence can travel along the connections of the network

- Several things can travel along edges of a graph
  - Vehicles
  - information
  - influence or popularity
  - diseases

- A *path* is a sequence of nodes such that every two consecutive nodes in the path are connected by an edge
  - You can also see the path as a sequence of edges, where two consecutive edges share an endpoint
  - If there exists a path between u and v then they are in an undirected relation

- Multiplying the adiacency matrix for itself k times we can find how many paths of length k there are between each pair of nodes
  - $M^k(u, v)$ = # paths of length k between u and v

PATHS



- MIT – BBN – RAND – UCLA is a path
- UCSB – UCLA – RAND – MIT is not a path
- A path can go through the same node several times
  - SRI – STAN – UCLA – SRI – UTAH – MIT
- A *simple path* never goes through the same node twice
- A *shortest path* is a path that goes through the minimum number of edges

Cycles

- A *cycle* is a simple path that starts and ends in the same node
  - LINC – CASE – CARN – HARV – BBN – MIT – LINC is a cycle
  - A cycle has at least three edges

- In communication networks and transportation networks each node lies on one or many cycles
  - Redundancy introduced to increase the robustness of the network
  - The network is guaranteed to work even in presence of a limited number of faults

- In a social network cycles are very common but accidental and we don't care of them
  - Es: the bestfriend of the cousin of my wife is the sister of my officemate

Directed Paths and Cycles

- Directed paths and cycles can be defined similarly to the undirected case
  - We have only to take care of the direction of the edges in the path
- Sometimes we consider undirected cycles even in directed graphs
  - We simply ignore the edges' directions
  - Useful if we are interested in the existence of a relation, independently from who activated it

Connectivity

- Two nodes u e v are connected if there exists a path from u to v

- A graph is *connected* if there exists a path between each pair of nodes in the graph
  - The graph of ARPANET is connected
  - Communication networks are usually connected

- In most of the cases graphs are disconnected
  - Social networks
  - Collaboration networks
  - *etc.*

- A directed graph is *strongly connected* if there exists a directed path between each pair of nodes in the graph

Components

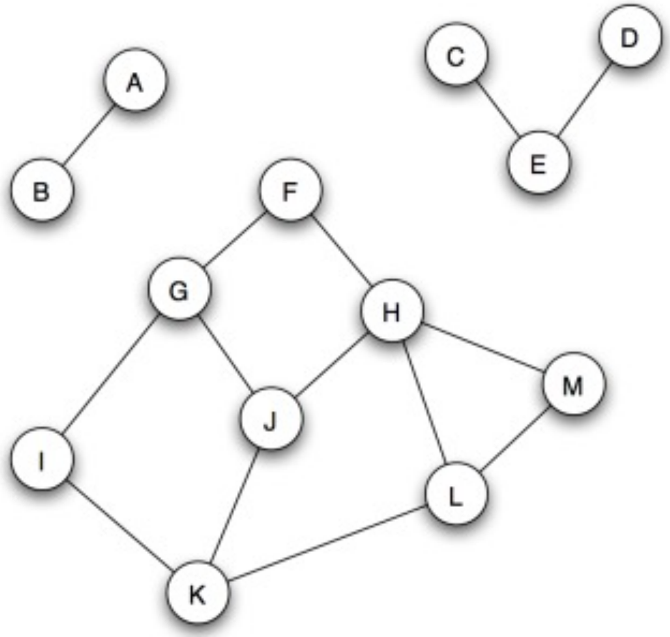
- If a graph is not connected it can be partitioned in subgraphs that are connected

- A *connected component C* in a undirected graph is a subset of nodes such that
  - There exists a path between each pair of nodes in C
  - For each node u not in C, there exists at least one vertex v in C such that there exists no path between u and v

- A *strongly connected component S* in a directed graph is a subset of nodes such that
  - There exists a directed path between each pair of nodes in S
  - For each node u not in S, there exists at least one vertex v in S such that there exists no directed path between u and v

- An edge belongs to a component if both its endpoints belong to the component
- The edge is a bridge if its removal makes the component disconnected

COMPONENTS



- This graph has three connected components
  - {A,B}, {C,D,E}, {F,G,…,M}
- {H, L, M} is not a component
- (D, E) is a bridge

Component Analysis

- Analyzing the components of a graph we can gain useful informations on the global structure of the network
  - Which edges tie different components?
  - How information spreads in the network?
  - Which role has each node?

Graph of the scientific collaborations in a research center

Component Analysis

- In many cases it's crucial to recognize components in a network
- A component is a region of the network densely connected
  - We would like to recognize the most densely connected components and their borders
- Anothe important type of analysis
  - We consider only edges with weigth greater than a given threshold
  - Gradually increasing the threshold the graph will break into several components
  - The remaining components include nodes that are strongly tied

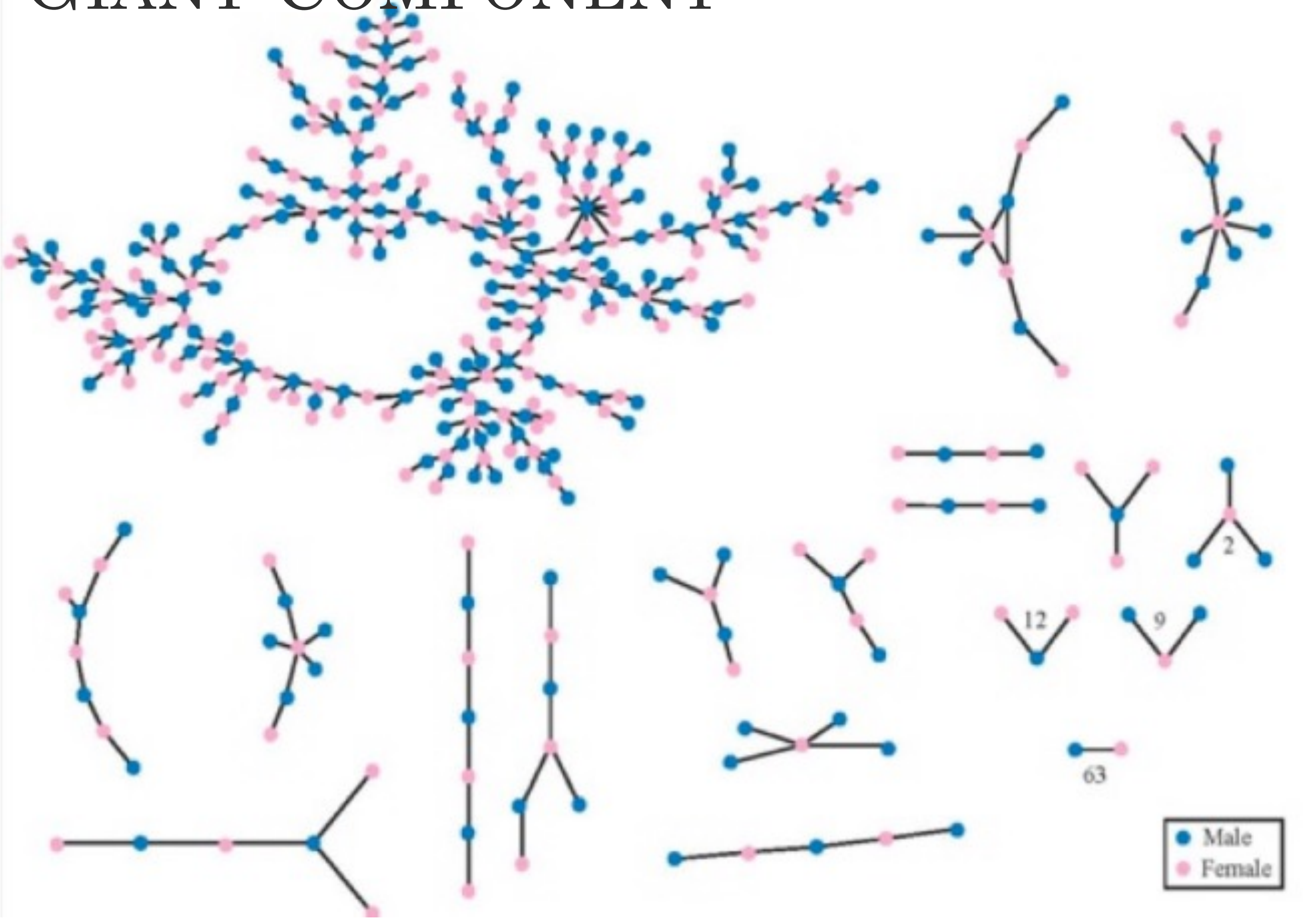
The Giant Component

- Several graphs are not connected byt they include a very large component
  - E.g. The graph or the love relations in a high school, the graph of the Web

- A *giant component* is a connected component containing a large fraction of the nodes
  - Usually the giant component is unique

- When two giant components merge it can give raise to dramatic events
  - The graph of the human relations between populations before the America discover had two giant components
  - Their fusion brought to the extermination of one of the two components

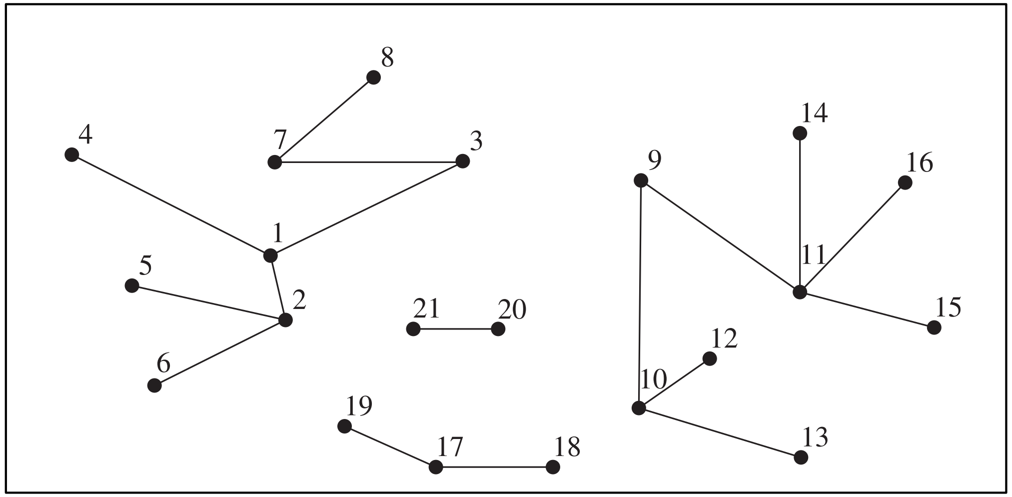
The Giant Component



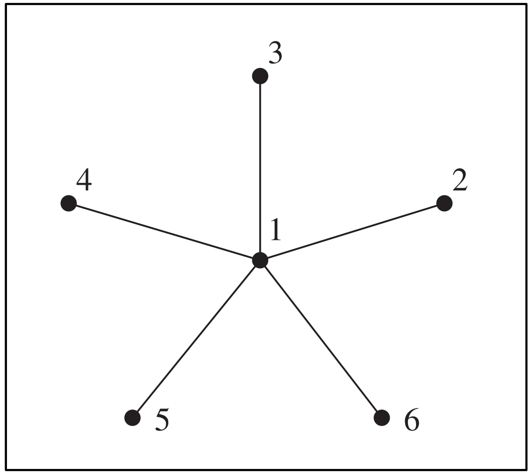
- The existence of a giant component in this network implies a higher risk of diffusion of sexual diseases

Particular Classes of Graphs

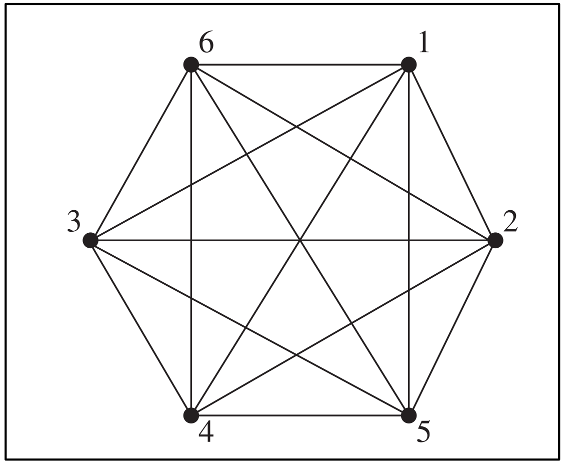
- There are specific topologies that occur very often and there were very well studied



Trees and Forests



Star



Clique

NEIGHBORHOODS

- The neighborhood of a node is the set of nodes that are adjacent to it

- The neighborhood of a set of nodes S is the set of nodes that do not belong to S but they are adjacent to some nodes in S

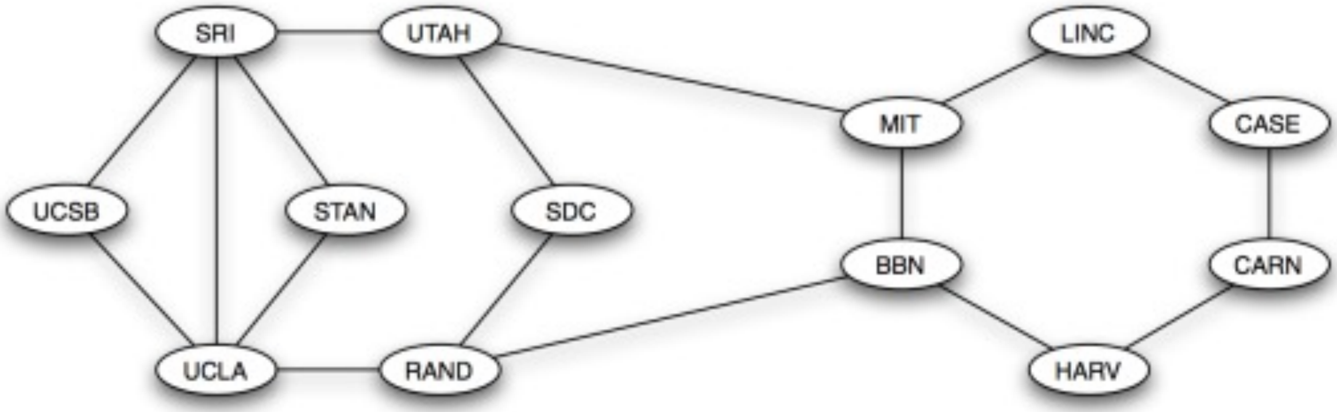
DEGREES

- The *degree* of a node is the number of its neighbors
  - Number of edges adjacent to the node
  - It's equal to the size of the neighborhood
- In a directed graph we distinguish between in-degree and out-degree
  - In-degree: number of incoming edges
  - Out-degree: number of outgoing edges

DISTANCES

- The *distance* between a pair of nodes is the length of the shortest path connecting the nodes
  - We assume each edge has weigth 1

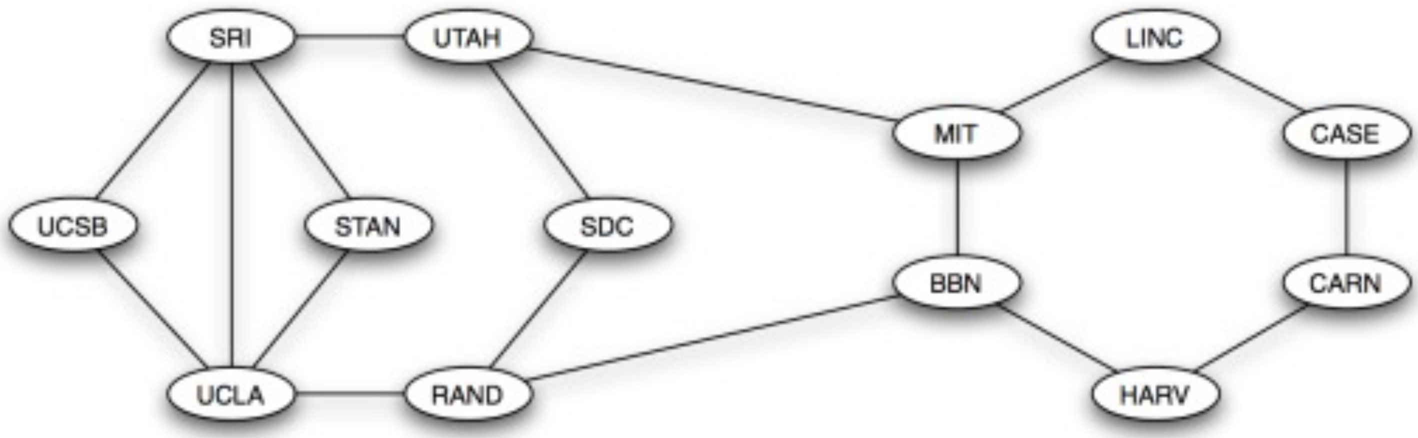
- The *diameter* of a graph is the largest distance between a pair of nodes in the graph
  - Which is the distance between MIT and SDC?
  - Which is the diameter of the network?



How to Compute Minimal Distances

- Given a graph, how can we compute the minimal distances from a node to all the others?
  - We need an efficient algorithm

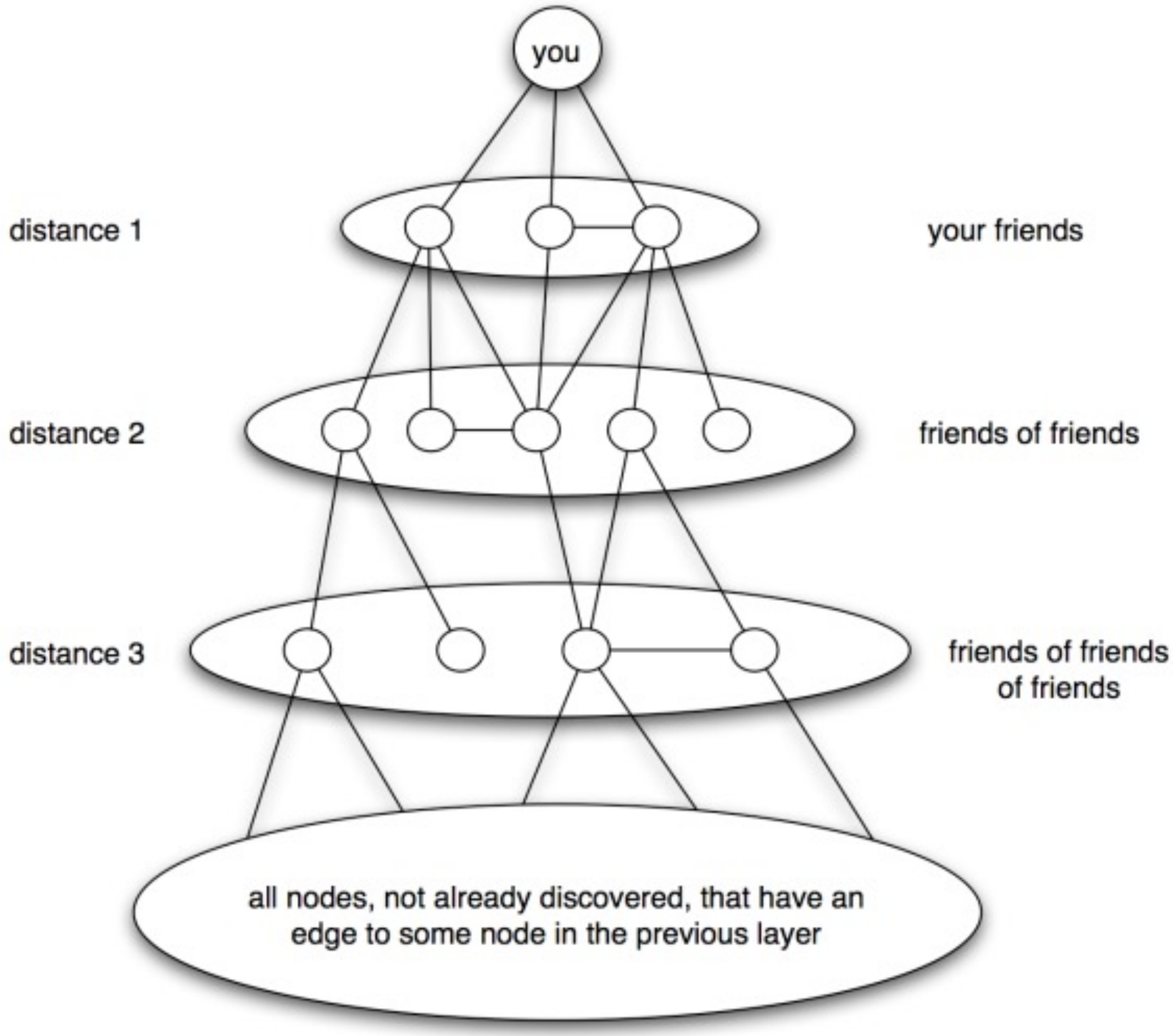
- How can we approach the problem?
  - BFS



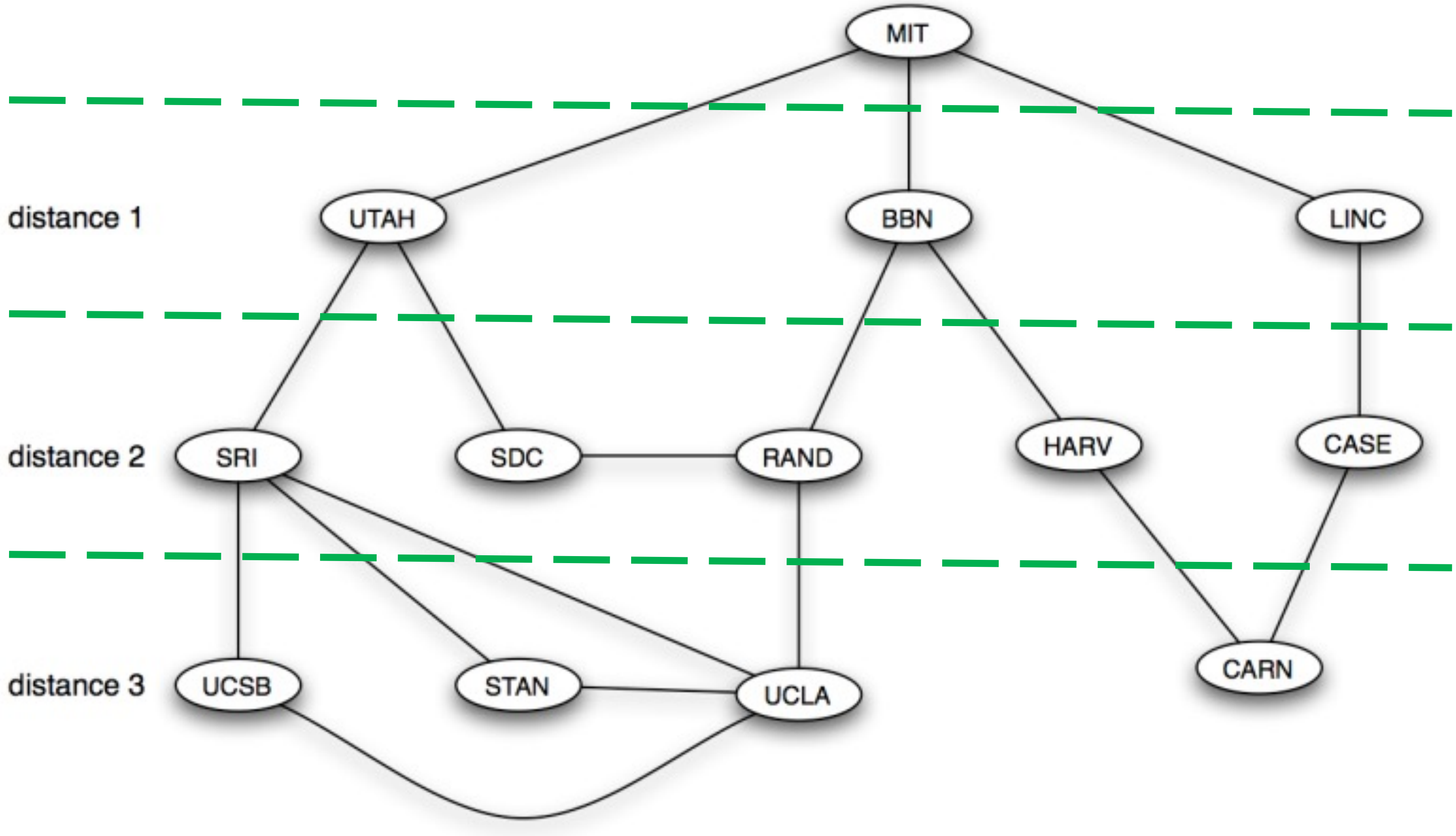
BREADTH-FIRST SEARCH (BFS)

- Starting from the source node (*root*)
  - Find all nodes adjacent to root
    - These nodes are at "distance 1"
  - Find all the nodes that are adjacent to nodes at distance 1 and not yet visited
    - These nodes are at "distance 2"
  - …
  - Find all the nodes that are adjacent to nodes at distance j and not yet visited
    - These nodes are at "distance j+1"
  - Stops when there are no other adjacent vertices not visited

BFS TREE


you
distance 1
your friends
distance 2
friends of friends
distance 3
friends of friends of friends
all nodes, not already discovered, that have an edge to some node in the previous layer

BFS on the Graph of Arpanet



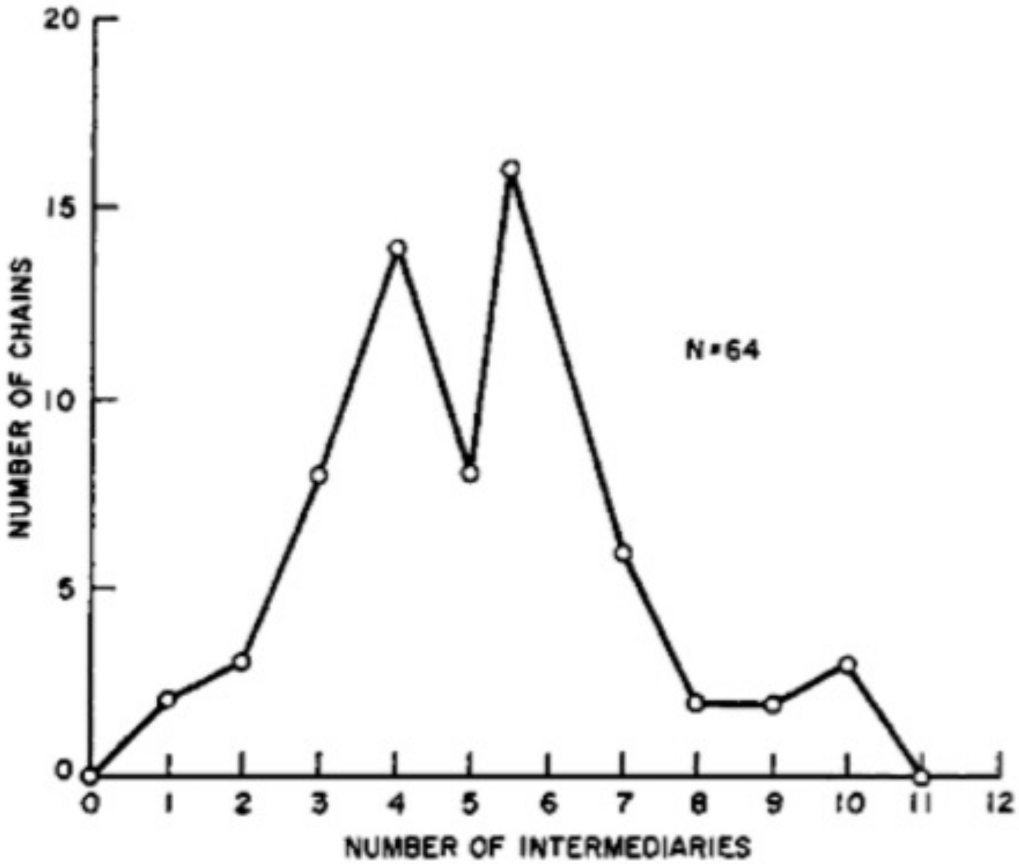
Smallworld phenomenon

- Hypothesis: in large scale networks most of the nodes belong to a giant component and they are connected through very short paths
  - Informations/infections spread very fast

- First experiment realized by Stanley Milgram in 1960s (research budget $680)
  - 296 people, randomly chosen in the USA, was asked to deliver a letter to a given recipient
    - They received a profile of the recipient (address, work, education, place of origin, interests, ecc.)
    - They could send the letter to one of their friends or acquantainces
  - The experiment measured the average number of hops for each letter

Six Degrees of Separation

- In the Milgram's experiment only 64 letters were delivered to the recipient
  - Average number of hops < 6



The experiment has been largely critized in the following years but …

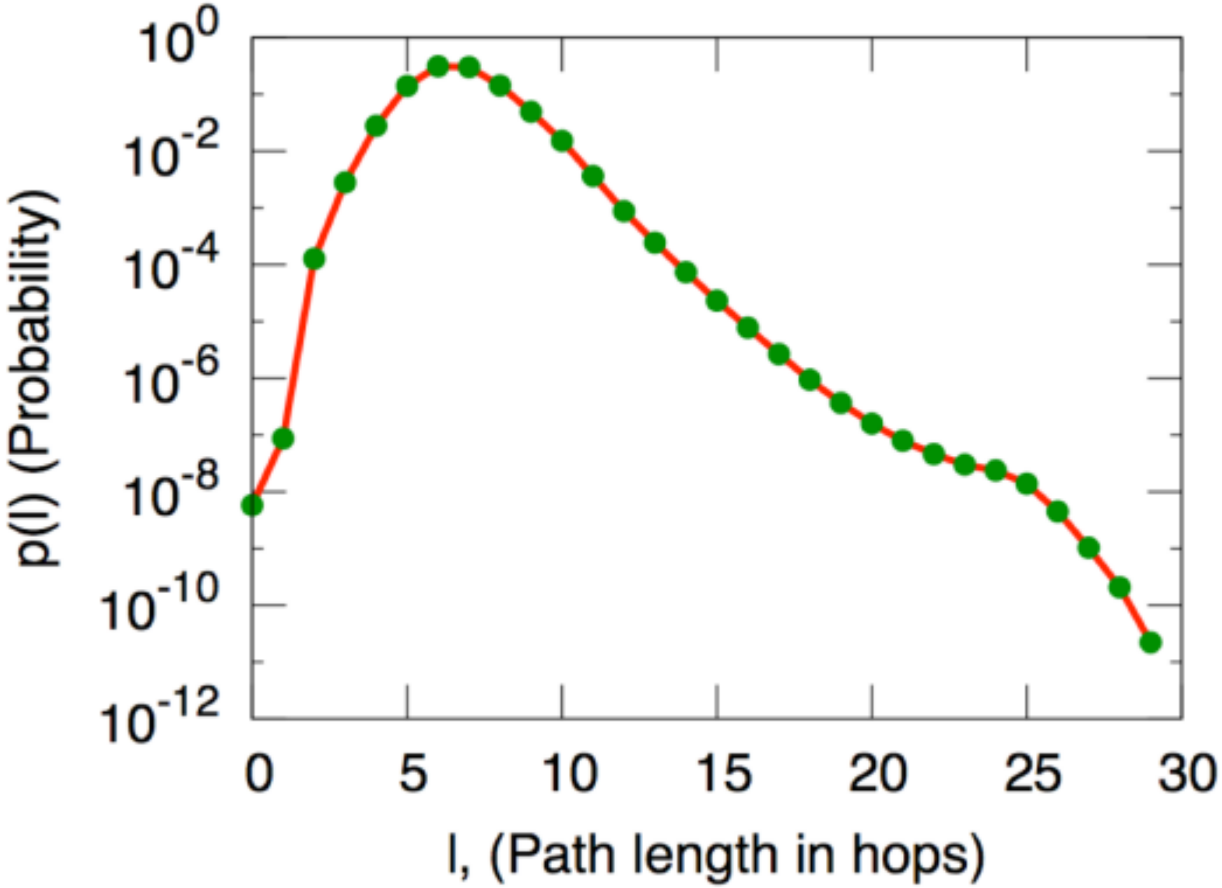
Six Degrees of Separation

- In recent years several experiments confirmed Milgram's results

- In 2008 Leskovec and Horvitz realized a new version of the Milgram's experiment
  - They used data related to Messenger's connections of 240 milion users in a period of 30 days
  - Their graph has a giant component with average distance equal to 6.6

- In each run they selected a random sample of 1000 users and computed minimal distances with BFS

Leskovec and Horvitz Results



expected average distance = 6.6, median = 7

Six degrees of geekiness

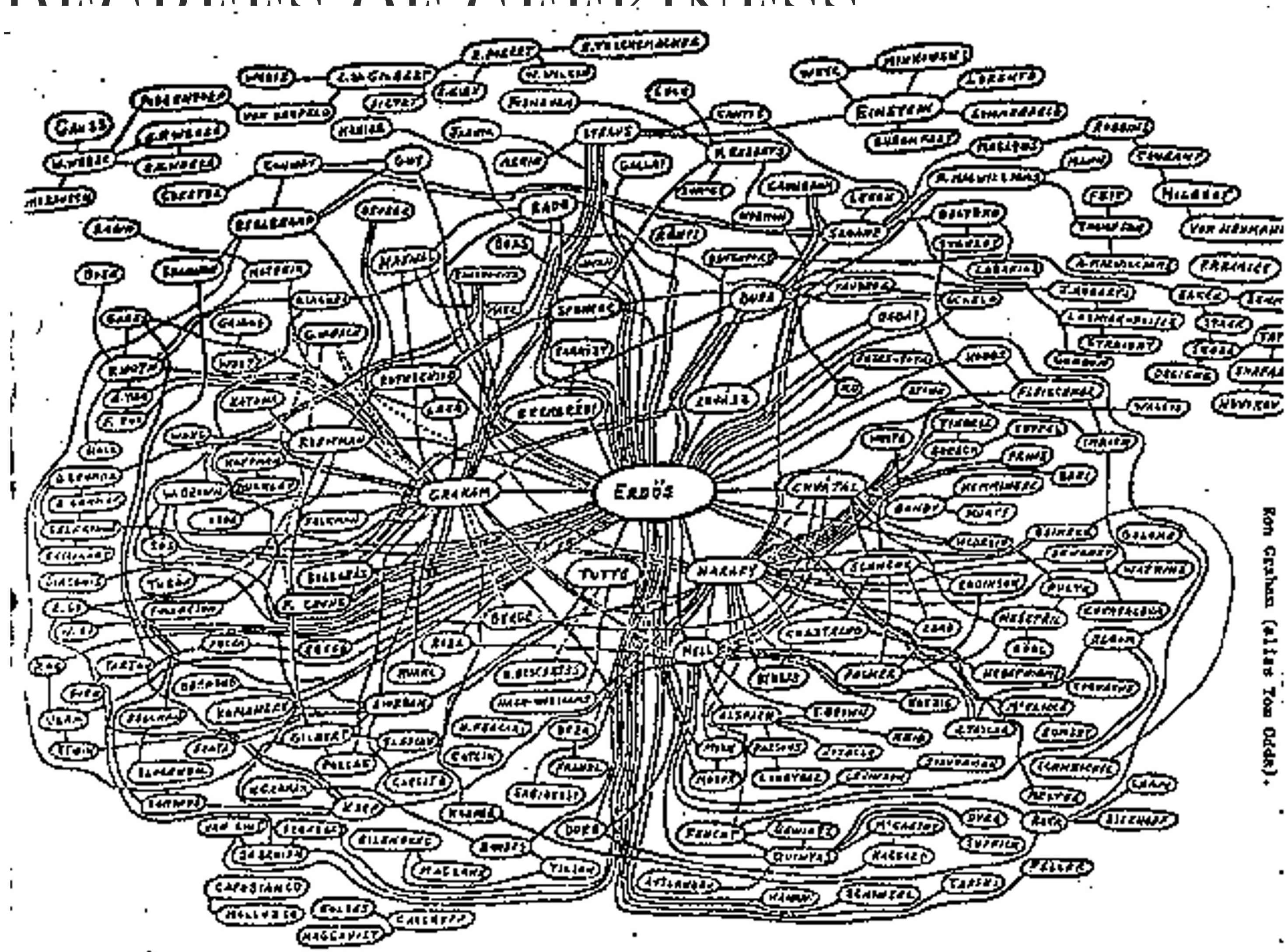


Figure 1
To appear in Topics in Graph Theory (F. Harary, ed.) New York Academy of Sciences (1979).

- Graph of the scientific collaborations rooted in Paul Erdös
  - Most of the mathematicians (and computer scientists) have Erdös numer < 5

How to Describe a Graph Without Representing Explicitly

- A large scale network can have millions of nodes
  - It cannot be represented explicitly
    - We need a set of quantitative parameters that can describe the graph
    - We use these parameters to compare graphs without representing them
- Some of the most used parameters
  - diameter, average path length
  - Clustering indices
  - Centrality measures
  - Node degree distribution

Degree distribution

- The distribution of the node degrees is fundamental characteristic of the network
- Pr(d) = fraction of nodes with degree d
  - Probability that a node randomly chosen has degree d
- k-regular graphs have a degenerate degree distribution
- Random graphs have a Poisson degree distribution
- In several networks the degree distribution is a power law
  - $\Pr(d) = cd^{-\gamma}$

Diameter and Average Path length

- The diameter is an upper bound to the length of the shortest path between each pair of nodes in a connected graph

- The average path length is the average of the shortest path lengths between all pairs of nodes

- Comparing the diameter with the average path length we can obtain useful informations
  - If they are not comparable then there are very few pairs of nodes that are very far apart

CLUSTERING INDICES

- An interesting information about a social network is how much connected and close it is
  - How many of my friends are friends each other?

- These charactesistics can be measured through the clustering indices

- Two alternative definitions
  - Overall clustering: fraction of node pairs that are adjacent and they have a common neighbor
  - Individual clustering (of node u): fraction of pairs of u's neighbors that are adjacent
    - Average Clustering is the average of the individual clusterings of all nodes in the graph

Centrality Measures

- They measure the relevance (centrality) of a node in the network related to a given process
  - We can use them to compare nodes

- There are several centrality measures that can model different processes
  - Degree centrality
  - Closeness centrality
  - Betweenness centrality
  - Katz-prestige centrality
  - Eigenvector centrality

Network Data-Sets

- On the web there are several datasets of large scale netowrks
  - Collaboration grpahs
    - Wikipedia, World of Warcraft, Citation graphs
- Who-talks-to-Whom Graphs
  - Microsoft IM, Cell phone graphs
- Information networks
  - Snapshots of the Web, social netowrks, blogging sites
- Technological networks
  - Power grids, communication links, Internet
- Networks in the Natural world
  - Food webs, neural interconnections, cell metabolism

- SNAP is a general purpose network analys and graph mining library leaded by Jure Leskovec at Stanford Univerisity
  - http://snap.stanford.edu/data
  - There is a repository with lots of data on large scale networks