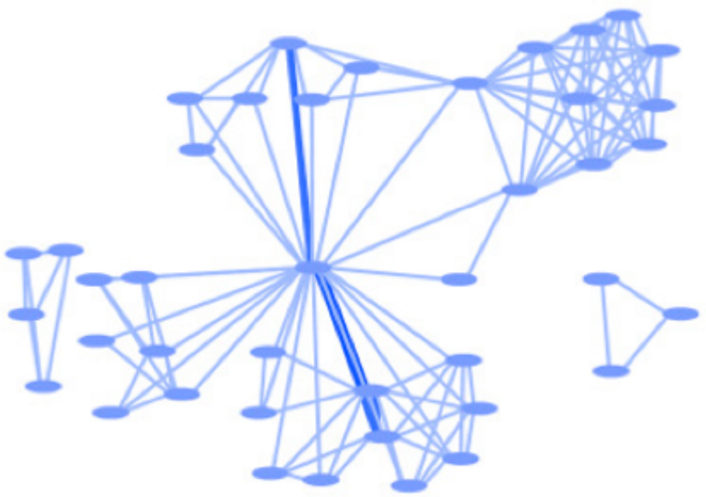
CptS 591: Elements of Network Science

Graph Theory Refresher



Graphs: basic definitions

- Graph theory: the study of network structures
- A graph is a way of specifying binary relationships among collection of items
- It consists of a set of objects called nodes (vertices), with certain pairs of these objects connected by links called edges (links)
- Example:

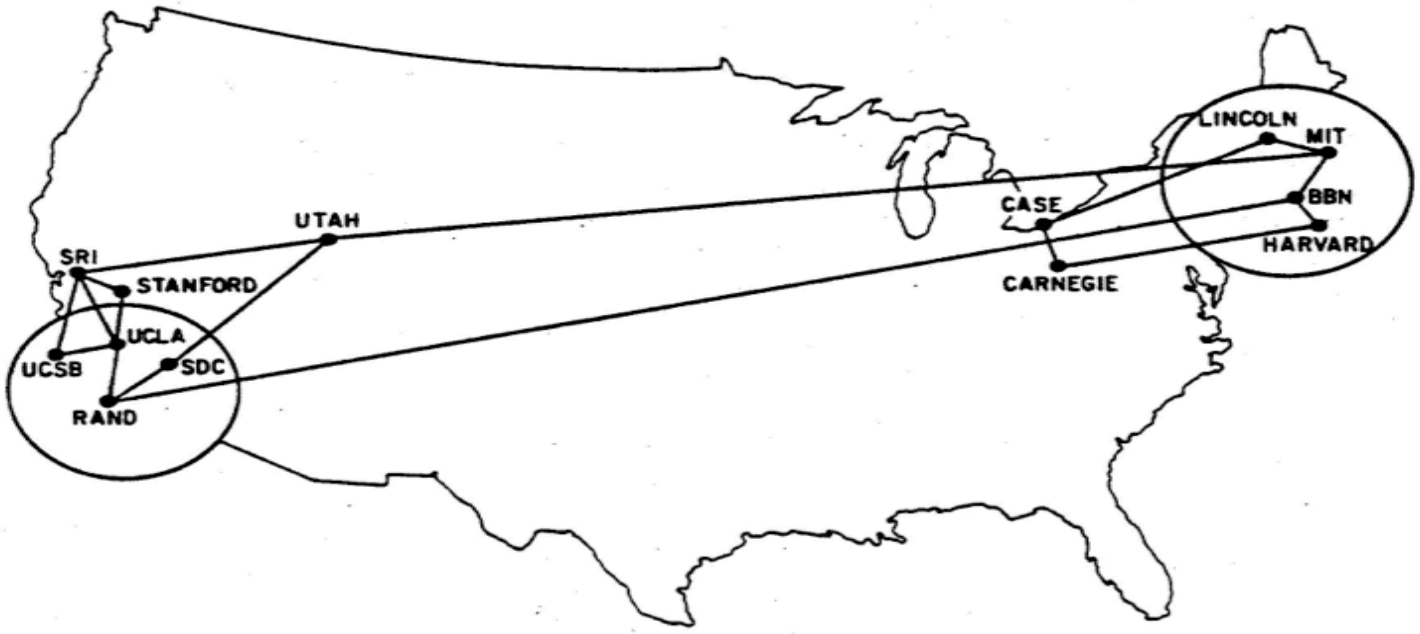
(a) *A graph on 4 nodes.* (b) *A directed graph on 4 nodes.*

Graphs as models of networks

- Graphs are useful because they serve as mathematical models of network structure.
- The ARPANET (Advanced Research Projects Agency Network) example. (1970)

Nodes: computing hosts
Edges: direct communication link

Example of
communication ntk



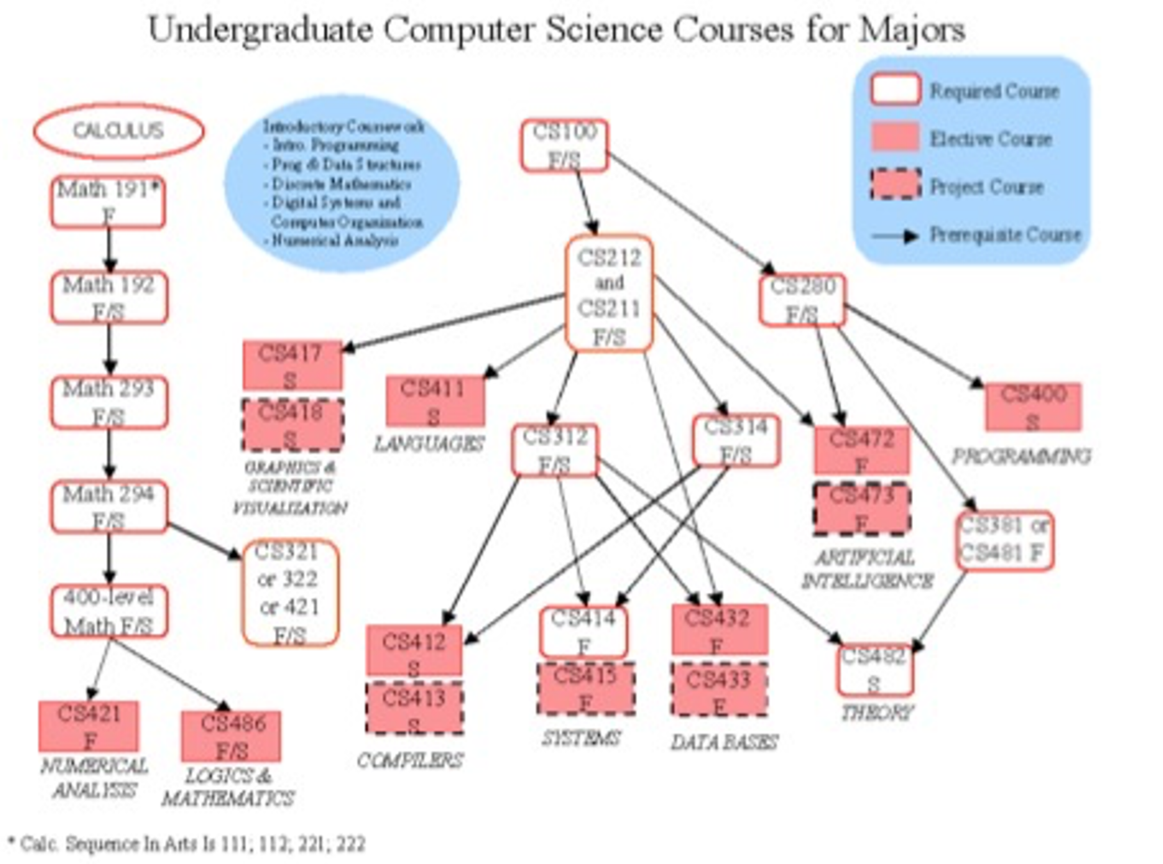
Other examples in which graphs model networks
example 1: transportation network (airlines)

(a) *Airline routes*

Example 2: transportation network (subway)

(b) *Subway map*

Example 3: dependency network



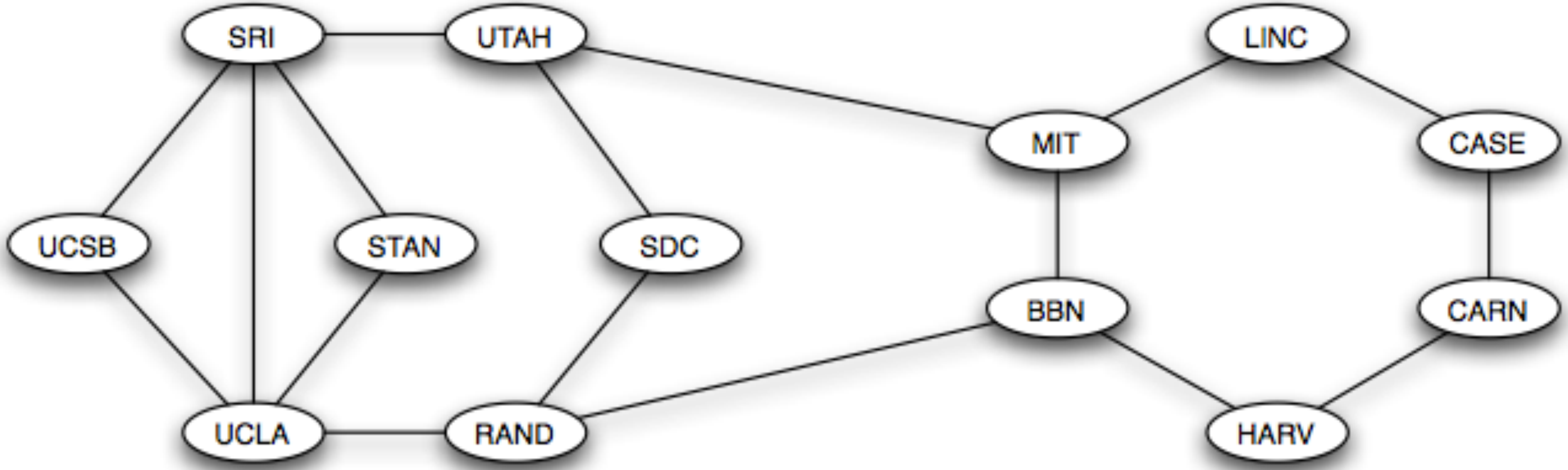
(c) *Flowchart of college courses*

Example 4: structural network

Paths and connectivity

- Graph theoretic notions are enormous, we can go far in network theory with a small set.
- Paths
  - Motivation: things often “travel” across edges.
  - Path: a sequence of nodes with the property that each consecutive pair in the sequence is connected by an edge.
  - Simple path: when no node repeats itself
- Cycles
  - A path with at least 3 edges, in which the first and the last nodes are the same, but otherwise all nodes are distinct
- Example: the ARPANET revisited (next slides)

The ARPANET has many cycles

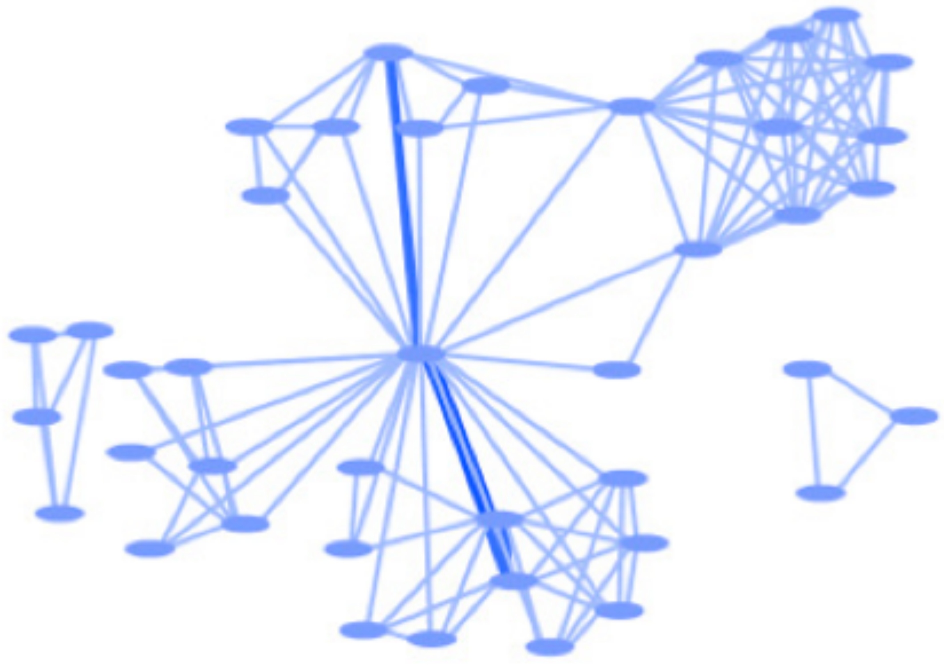


Every edge in the graph belongs to a cycle.

Why do you think it is designed to be so?

Connectivity

- A graph is connected if for every pair of nodes, there is a path between them.
- Not all networks are necessarily connected
  - Motivates definition of components
- A (connected) component of a graph is a subset of nodes such that
  1. Every node in the subset has a path to every other, and
  2. The subset is not part of some larger set with the property that every node can reach every other



Collaboration graph of the biological research center Structural Genomics of Pathogenic Protozoa (SGPP).

Giant components

- Most large, complex networks often have a giant component
  - a connected component that contains a significant fraction of all nodes.
- Moreover, when a network contains a giant component, it almost always contains only one
- To see the latter point, consider this thought experiment:
  - social network of the entire world
- But why only one giant component?

- Jared Diamond's book Guns, Germs, and Steel
  - Network perspective: clash of two giant components

A small scale example of giant component

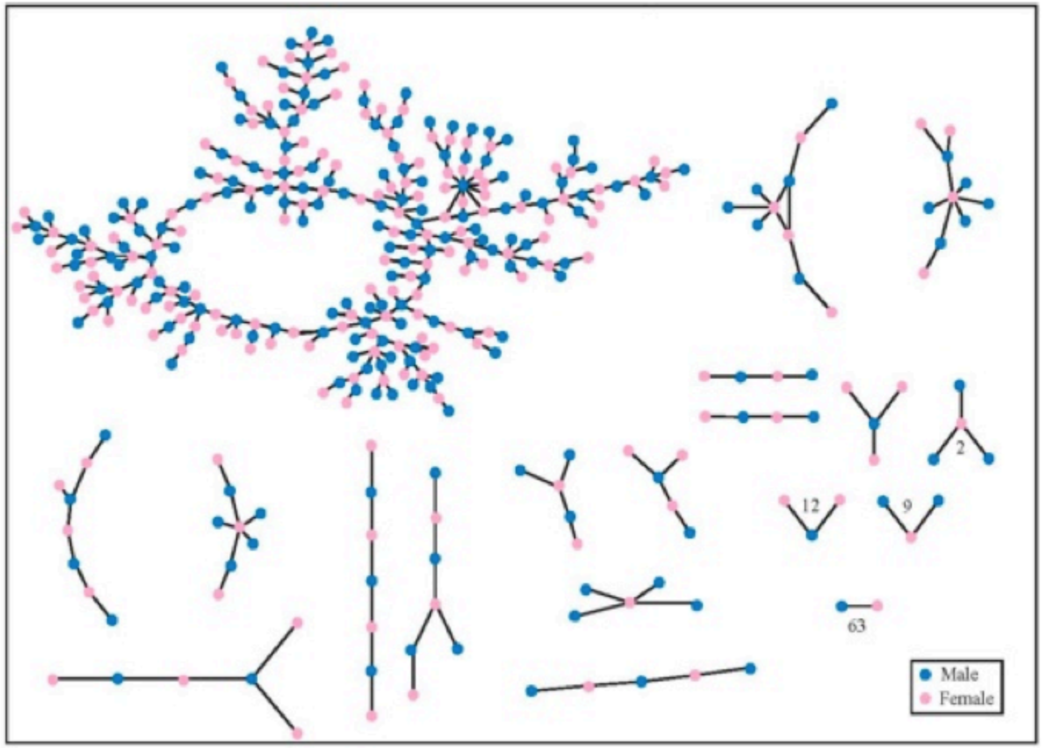
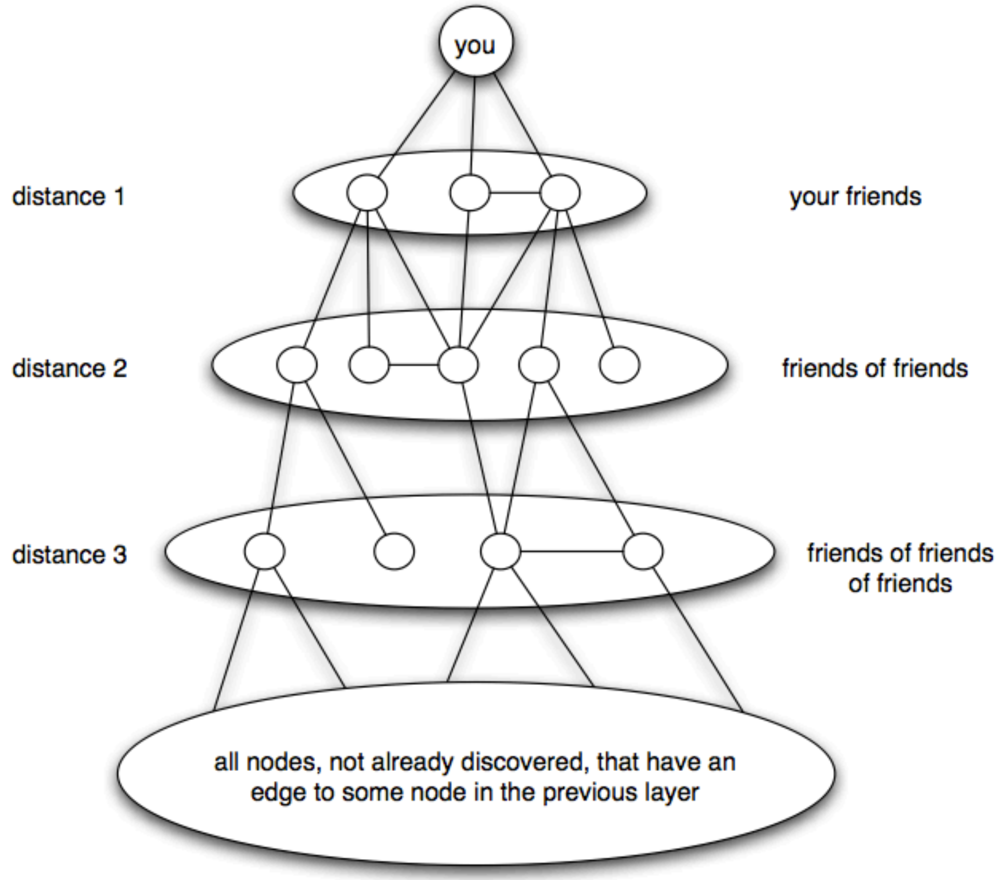


Figure 2.7: A network in which the nodes are students in a large American high school, and an edge joins two who had a romantic relationship at some point during the 18-month period in which the study was conducted [49].

Distance and BFS

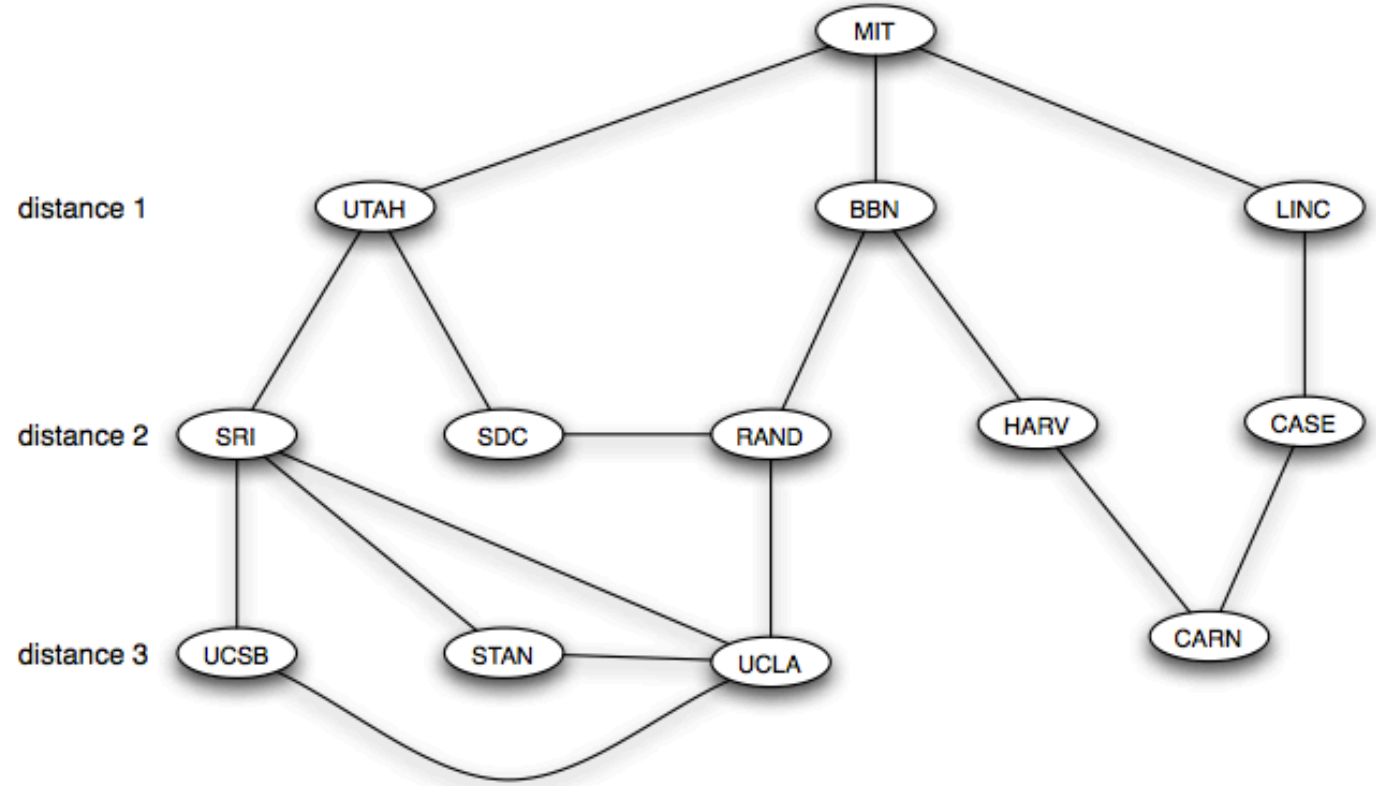
- In addition to simply asking whether two nodes are connected by a path, it is also in most cases interesting to ask how long such a path is.
  - E.g: transportation, Internet communication, spread of news, spread of diseases, etc
- The length of a path is the number of steps it contains from beginning to end (number of edges).
- The distance between two nodes is the length of the shortest path between them.
- Breadth First Search is an efficient way to calculate distances in a graph.

BFS idea illustration
(global friendship network)



BFS discovers distances to nodes one "layer" at a time; each layer is built of nodes that have an edge to at least one node in the previous layer

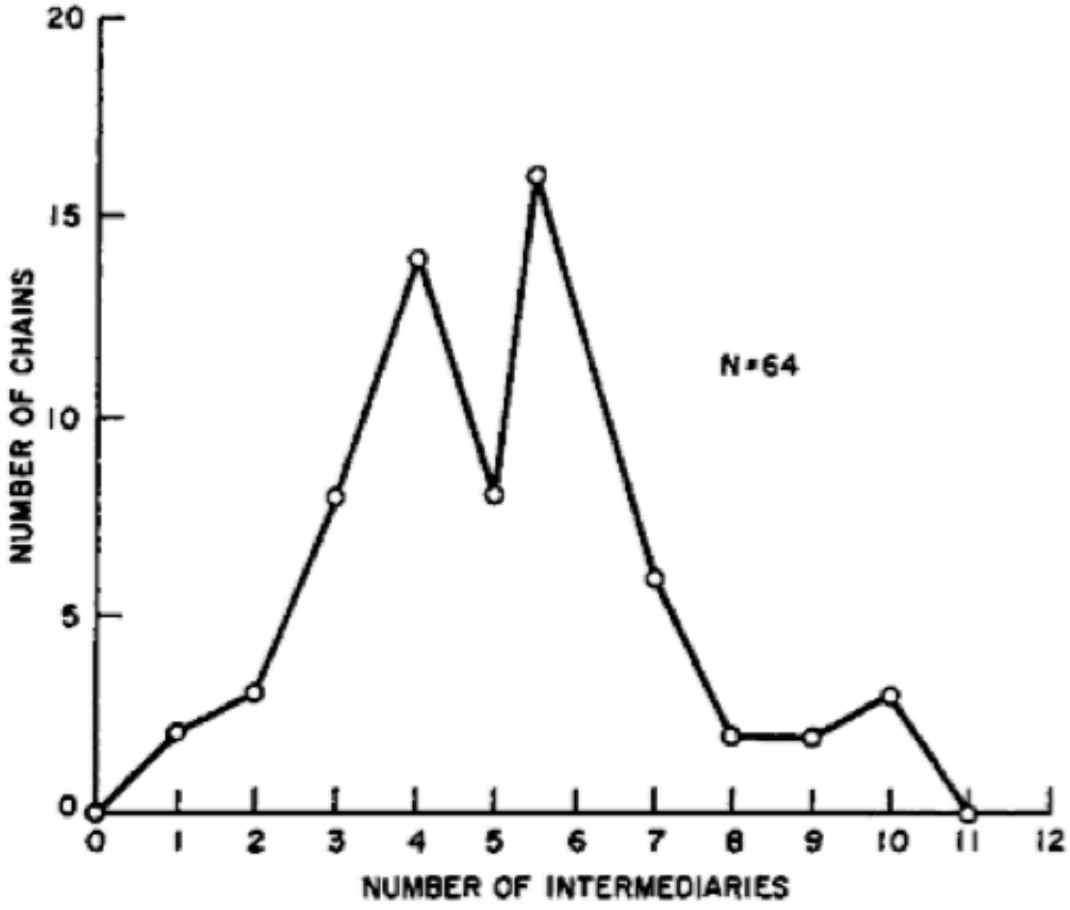
BFS: also useful conceptual framework to organize the structure of a graph (e.g. ARPANET graph)



The small-world phenomenon

- Going back to the thought experiment
  - Not just “giant component” existence, but also “short distance” (stronger)
- Stanley Milgram’s work in the 1960s
  - Budget: $680
  - 296 randomly chosen starters
  - Rule: forward a letter to a “target” person (in Boston). Basic info on target given. Subjects may forward the letter to a person they know on a first-name basis.
  - Goal: get the letter to the target as quickly as possible

Histogram from Travers and Milgram's paper



Caveats:
- Is the experiment sound?
- Does "small distance" imply closeness, socially?

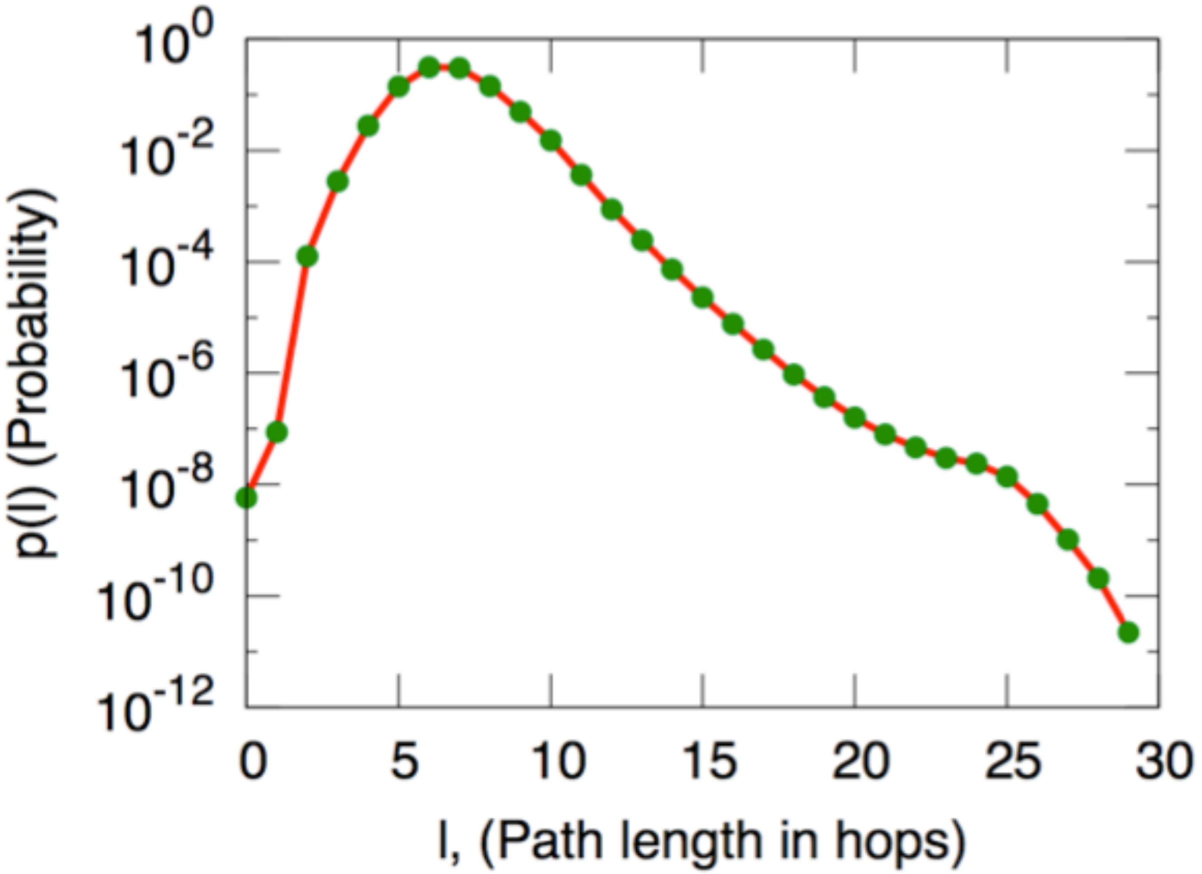
Despite the caveats:

social networks in general tend to have very short paths between essentially arbitrary pairs of people

Example 1 (large-scale)

Distribution of distances in the 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.

Data size: 240 M users (study by Leskovec and Horvitz).



Example 2: small-scale

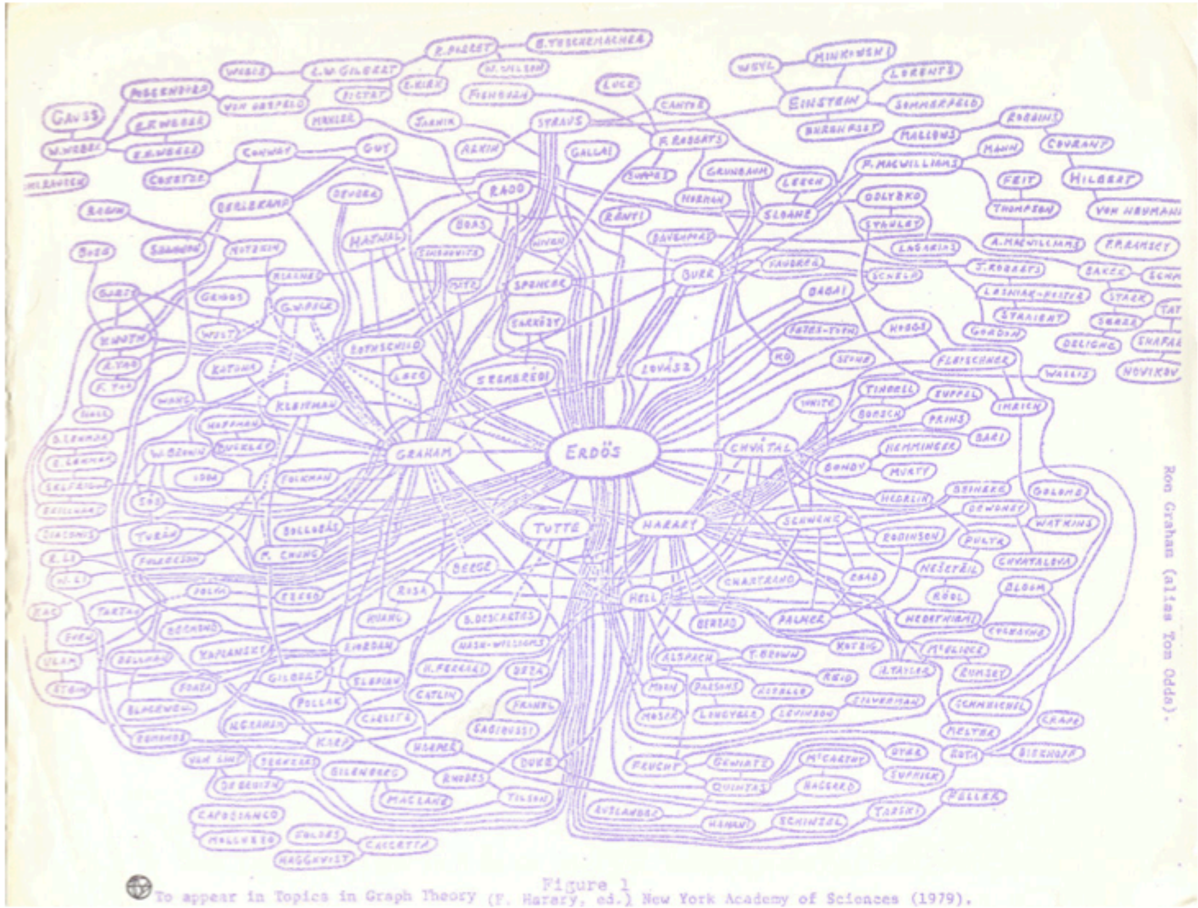


Figure 2.12: Ron Graham's hand-drawn picture of a part of the mathematics collaboration graph, centered on Paul Erdös [189]. (Image from http://www.oakland.edu/enp/cgraph.jpg)

Network datasets

There are several reasons why you might want to study a particular network dataset.

- You may care about the actual domain it comes from
  - Fine-grained details of the data itself are potentially as interesting as the broad picture.
- You are using the dataset as a proxy for a related network that maybe impossible to measure
  - E.g the microsoft IM graph was a proxy for a global friendship (large-scale social) network
- You are trying to look for network properties that appear to be common across many different domains

Some times, all three of these can be at work simultaneously, to varying degrees, in the same piece of research.

Network data sets (large): an overview

- Collaboration graphs
  - Co-authorships among scientists, co-appearance in movies, co-memberships on boards
- Who-talks-to-Whom graphs
  - Email logs, phone calls
- Information linkage graphs
  - Snapshots of the Web graph, linkages among bloggers, linkages among Wikipedia pages, linkages among pages on social-networking sites such as Facebook
- Technological networks
  - Internett, Power grids, Autonomous System (AS) graphs
- Networks in the natural world
  - Food webs, neural networks, metabolic networks