Networks

APAM E4990
Modeling Social Data

Jake Hofman

Columbia University

April 7, 2017

History

∼1930s: Relationships as networks

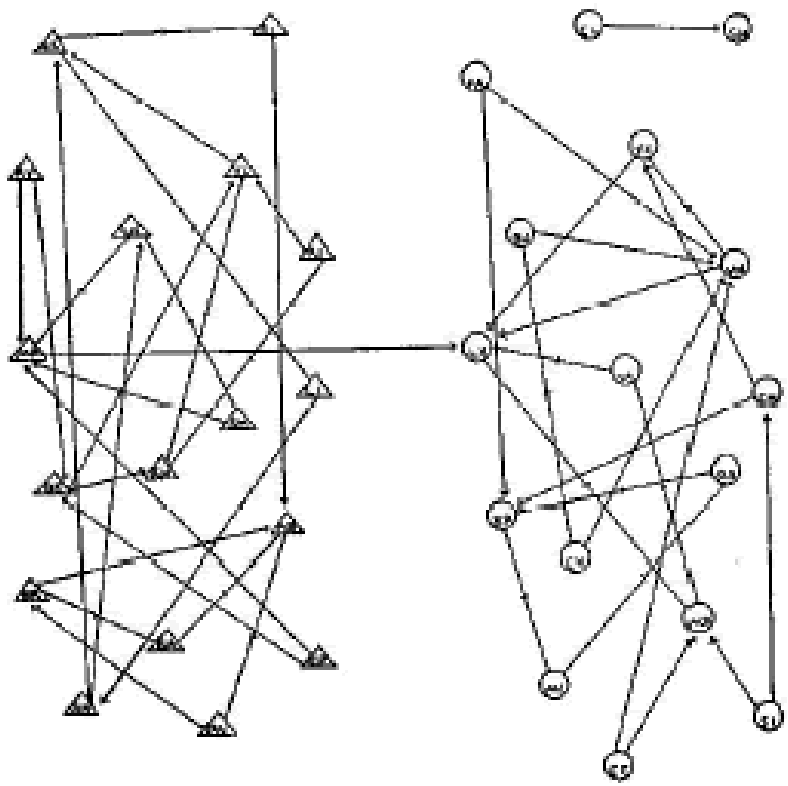
EMOTIONS MAPPED BY NEW GEOGRAPHY

Charts Seek to Portray the Psychological Currents of Human Relationships.

FIRST STUDIES EXHIBITED

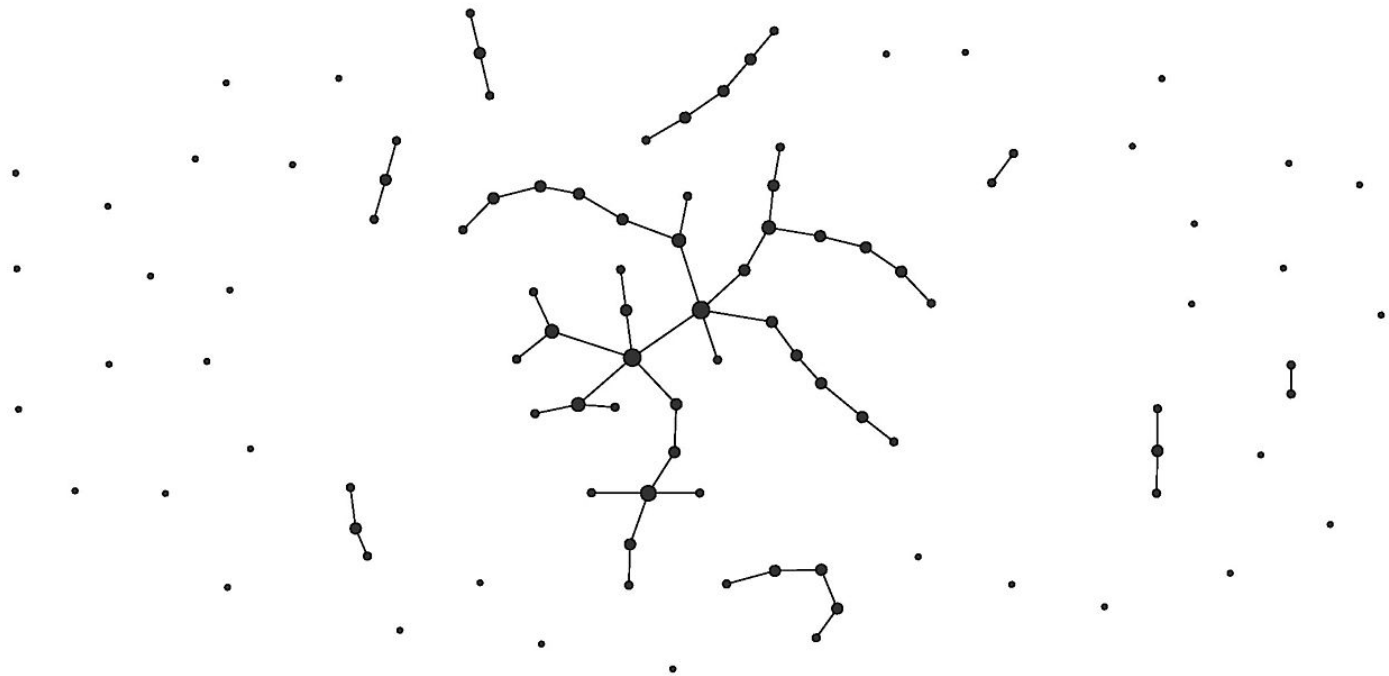
Colored Lines Show Likes and Dislikes of Individuals and of Groups.

MANY MISFITS REVEALED



Moreno (1933)

∼1960s: Random graph theory



$$p > \frac{(1+\epsilon)\ln n}{n}$$

Erdős & Rényi (1959)

∼1970s: Clustering, weak ties

becomes rather involved, however, and it is sufficient for my purpose in this paper to say that the triad which is most *unlikely* to occur, under the hypothesis stated above, is that in which A and B are strongly linked, A has a strong tie to some friend C, but the tie between C and B is absent. This triad is shown in figure 1. To see the consequences of this assertion,

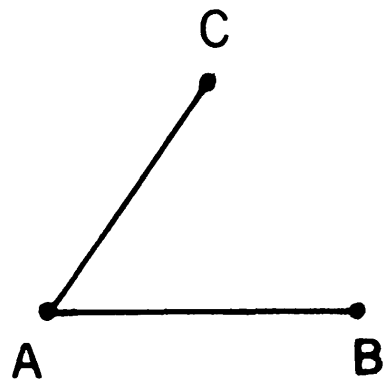


FIG. 1.—Forbidden triad

I will exaggerate it in what follows by supposing that the triad shown *never* occurs—that is, that the B-C tie is always present (whether weak or strong), given the other two strong ties. Whatever results are inferred from this supposition should tend to occur in the degree that the triad in question tends to be absent.

Granovetter (1973)

∼1970s: Clustering, weak ties

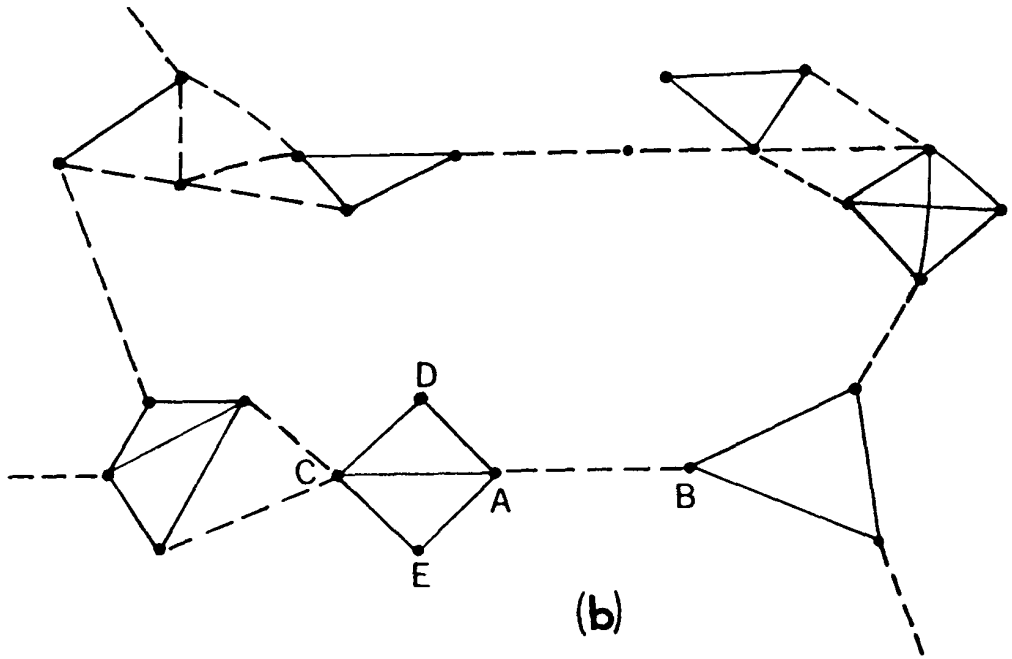
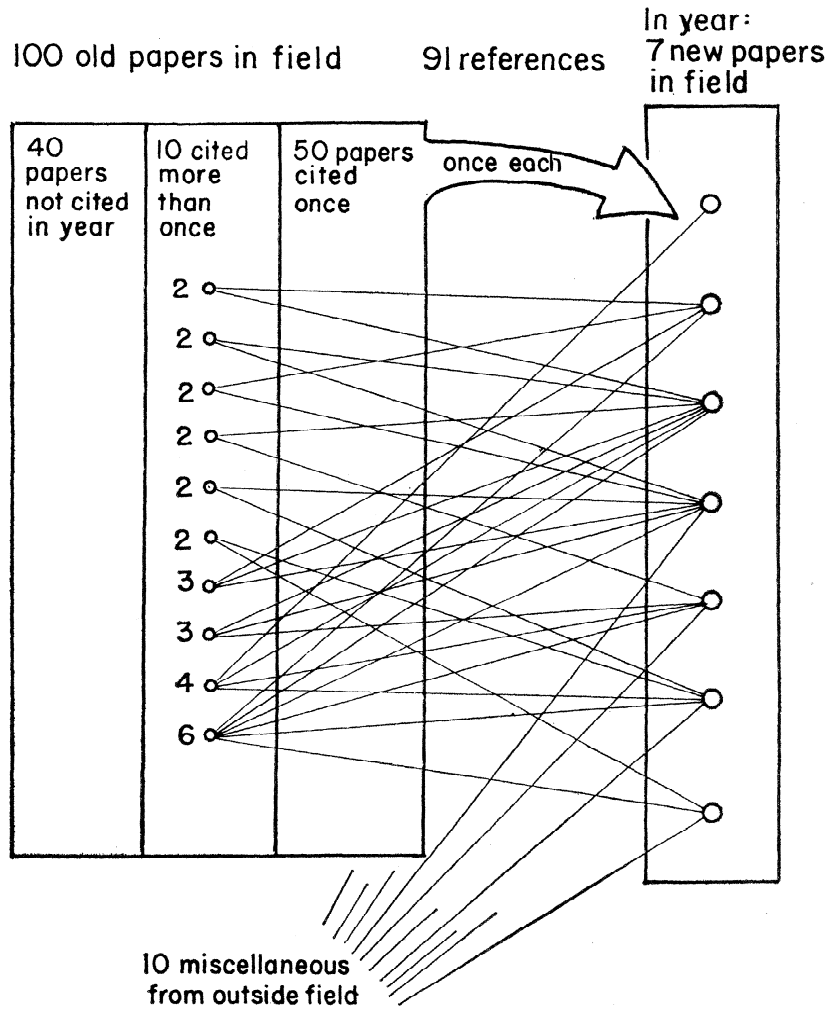


FIG. 2.—Local bridges. *a*, Degree 3; *b*, Degree 13. ——— = strong tie; — — — = weak tie.

Granovetter (1973)

∼1970s: Cumulative advantage



de Solla Price (1965, 1976)

∼1970s: Cumulative advantage

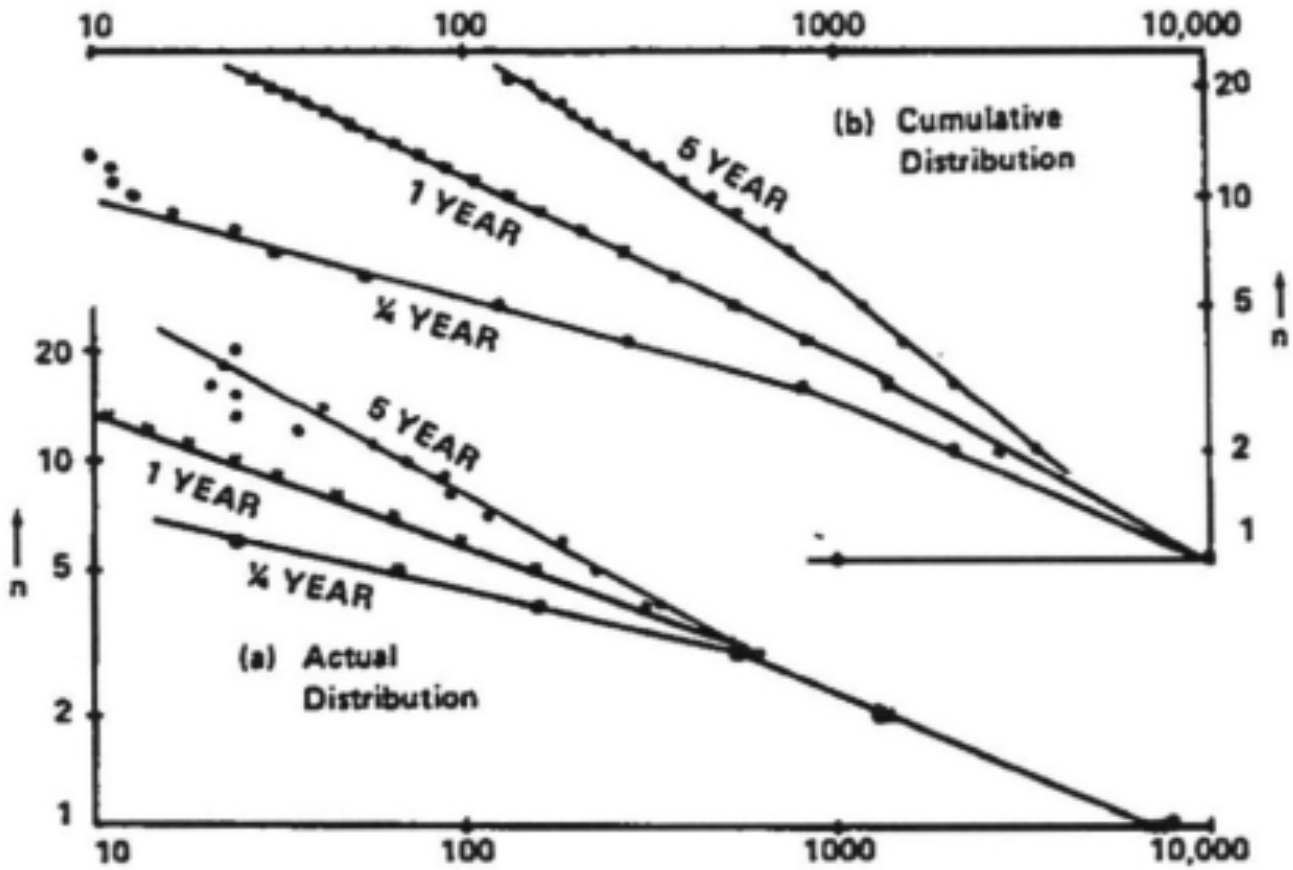
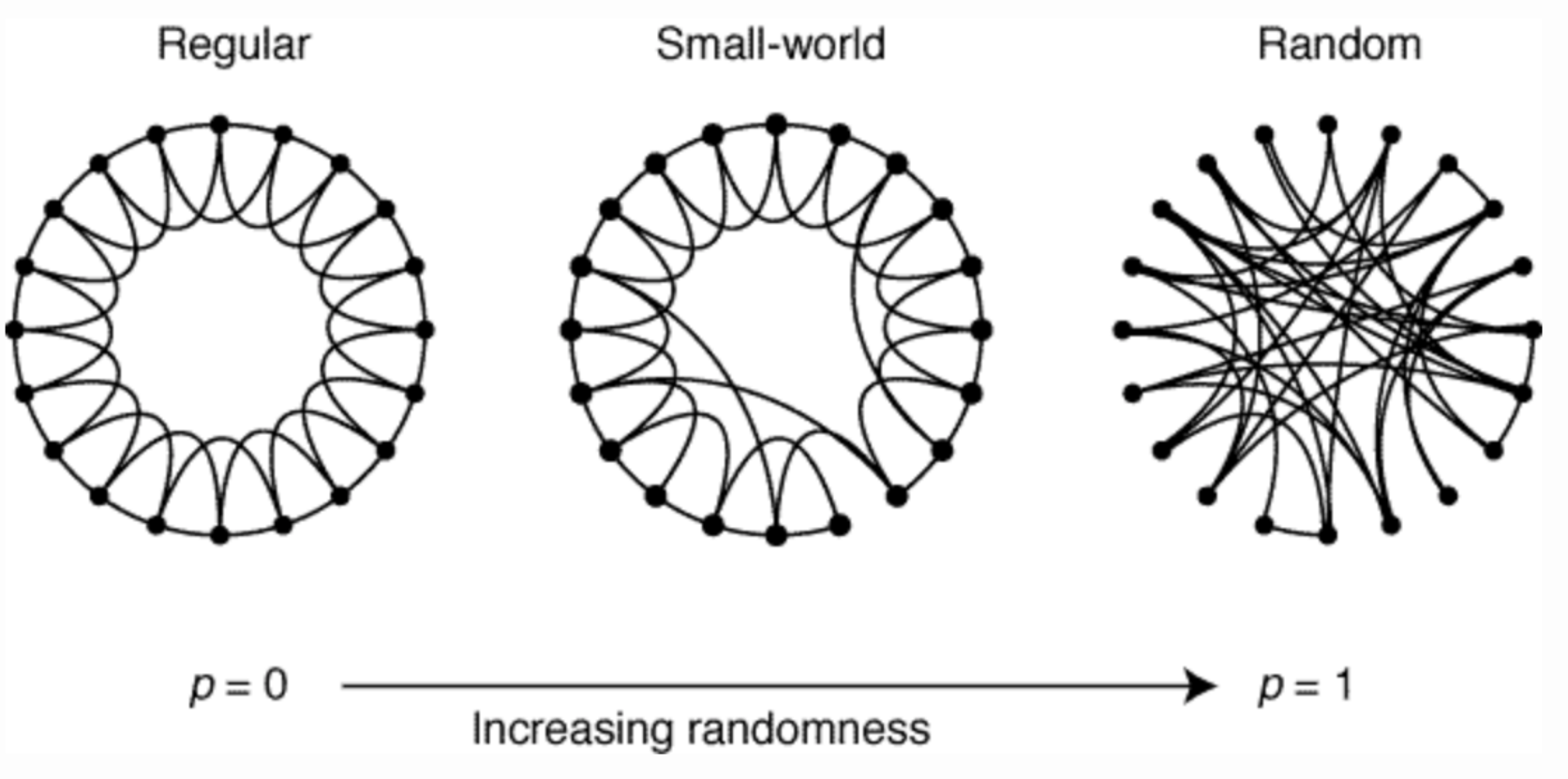


Fig. 1. Number of papers with (a) exactly and (b) at least n citations in ¼, 1, and 5-year indexes.

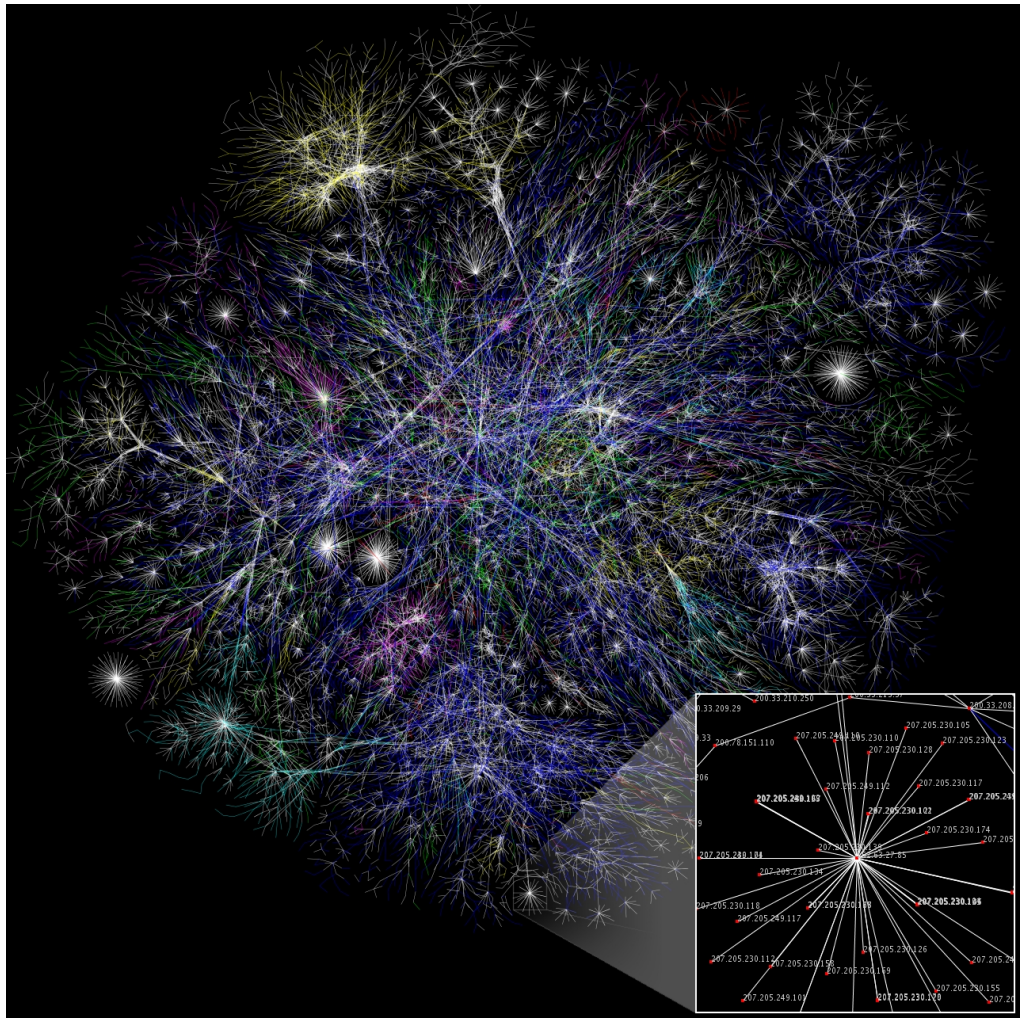
de Solla Price (1965, 1976)

∼1970s: Small-world networks



Watts & Strogatz (1998)

∼1990s: Empirical structure and dynamics of networks



Newman, Barabasi, Watts (2006)

∼2000s: Homophily, contagion, and all that

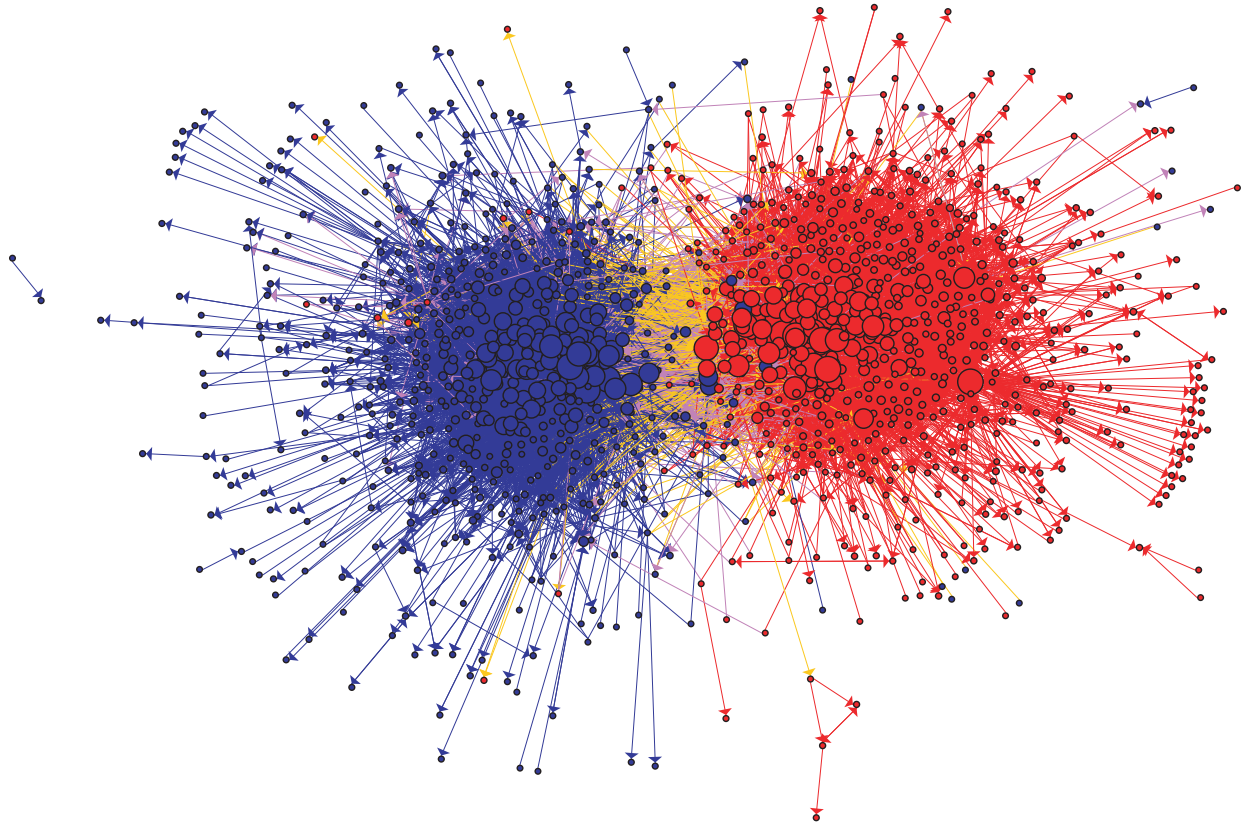


Figure 1: Community structure of political blogs (expanded set), shown using utilizing the GUESS visualization and analysis tool[2]. The colors reflect political orientation, red for conservative, and blue for liberal. Orange links go from liberal to conservative, and purple ones from conservative to liberal. The size of each blog reflects the number of other blogs that link to it.

Adamic & Glance (2005)

Types of networks

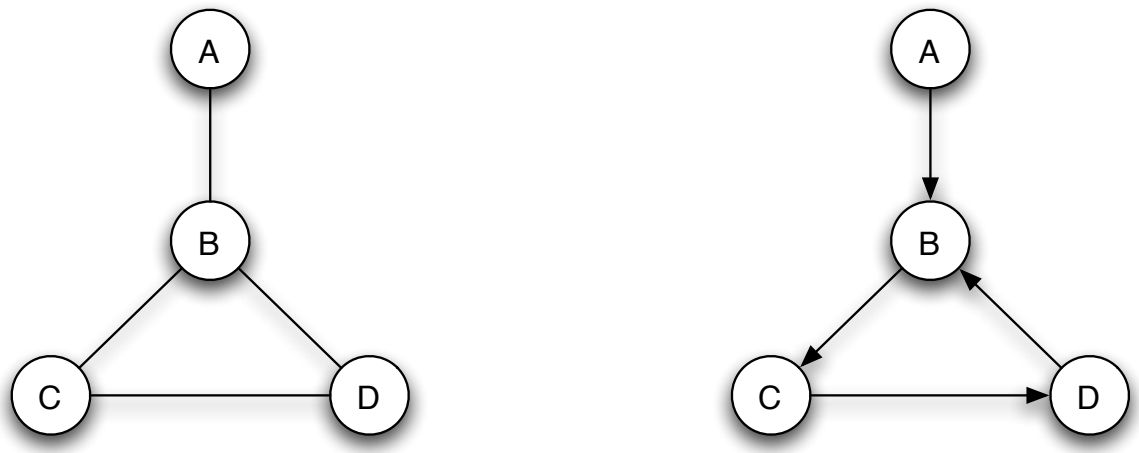
Types of networks

Networks are a useful abstractions for many different types of data

- Social networks (e.g., Facebook)
- Information networks (e.g., the Web)
- Activity networks (e.g., email)
- Biological networks (e.g., protein interactions)
- Geographical networks (e.g., roads)

Representations

There are many different levels of abstraction for representing networks (e.g., directed, weighted, metadata, etc.)

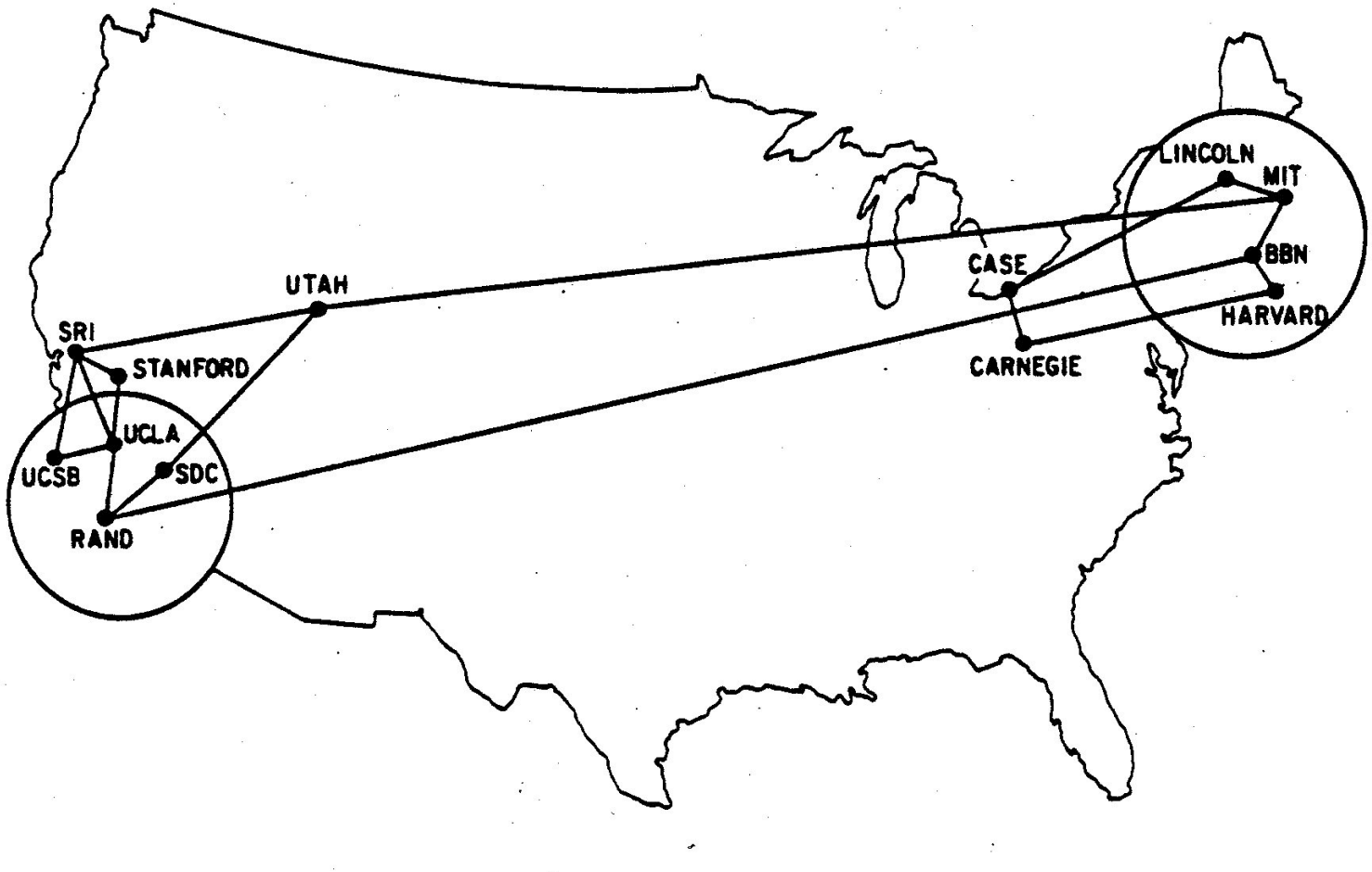


(a) *A graph on 4 nodes.* (b) *A directed graph on 4 nodes.*

Figure 2.1: Two graphs: (a) an undirected graphs, and (b) a directed graph.

Representations

There are many different levels of abstraction for representing networks (e.g., directed, weighted, metadata, etc.)



Representations

There are many different levels of abstraction for representing networks (e.g., directed, weighted, metadata, etc.)

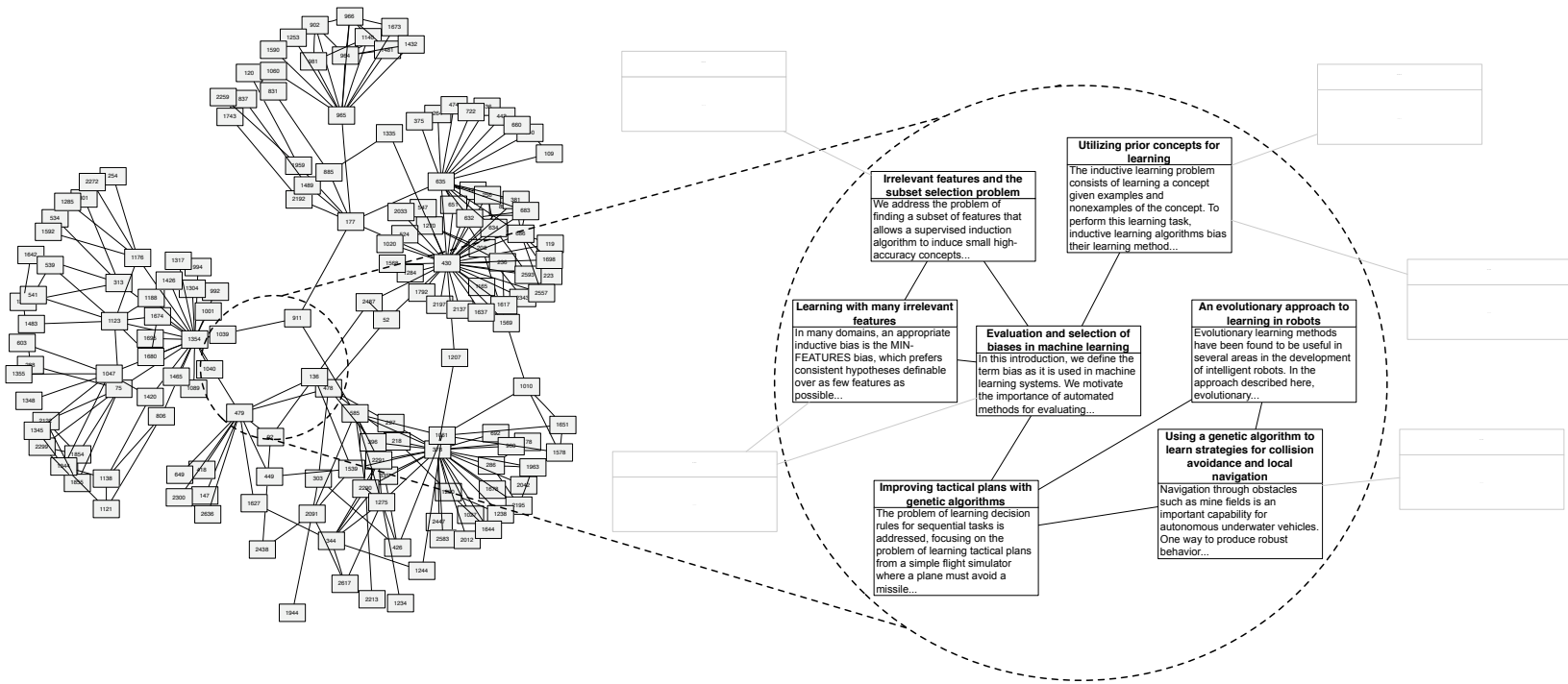


Figure 1: Example data appropriate for the relational topic model. Each document is represented as a bag of words and linked to other documents via citation. The RTM defines a joint distribution over the words in each document and the citation links between them.

Which network?

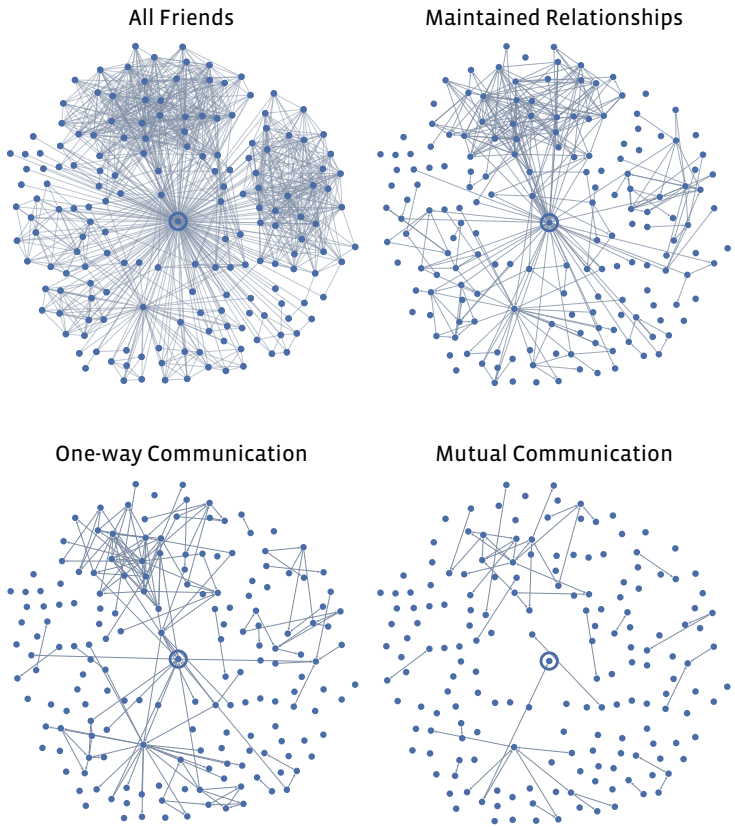


Figure 3.8: Four different views of a Facebook user's network neighborhood, showing the structure of links coresponding respectively to all declared friendships, maintained relationships, one-way communication, and reciprocal (i.e. mutual) communication. (Image from [281].)

Which network?

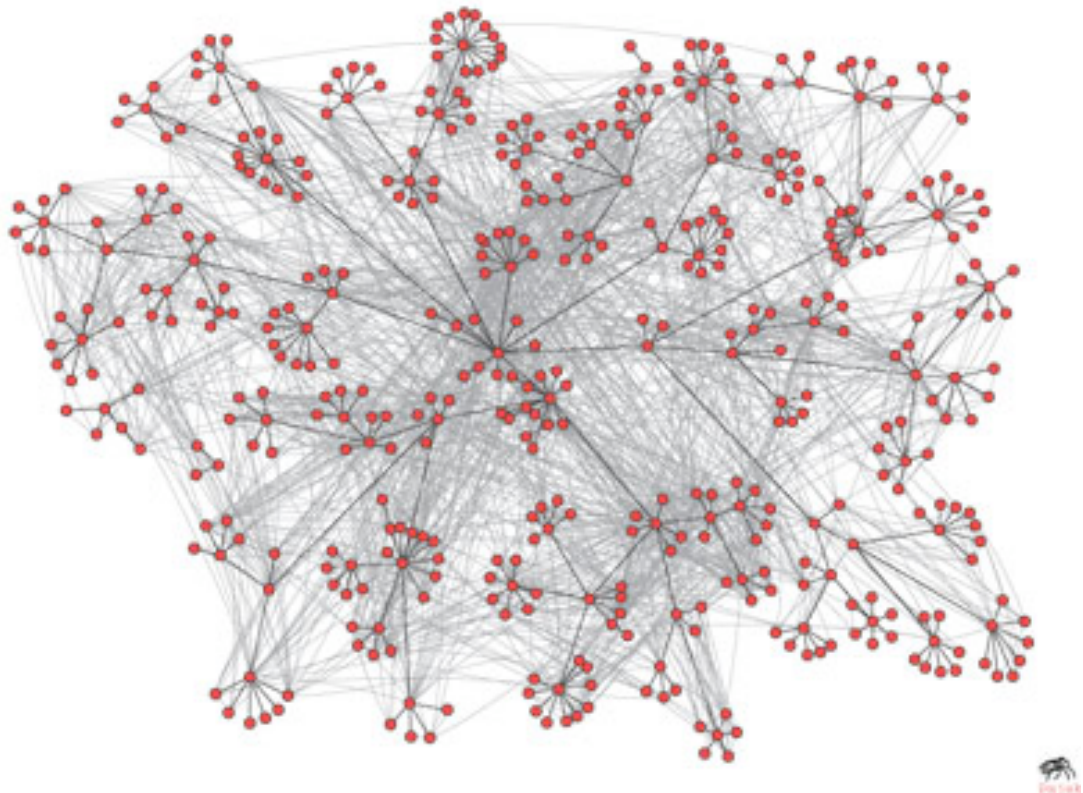


Figure 20.12: The pattern of e-mail communication among 436 employees of Hewlett Packard Research Lab is superimposed on the official organizational hierarchy, showing how network links span different social foci [6]. (Image from http://www-personal.umich.edu/ ladamic/img/hplabsemailhierarchy.jpg)

Which network?

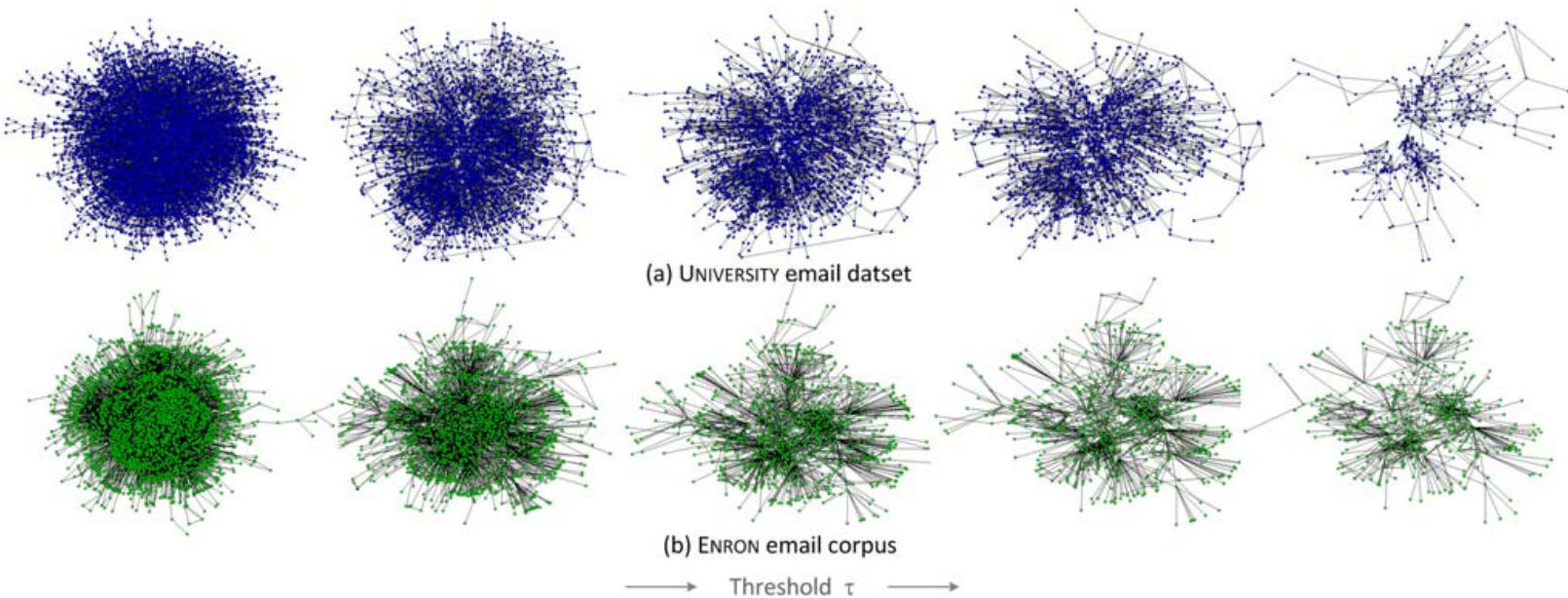
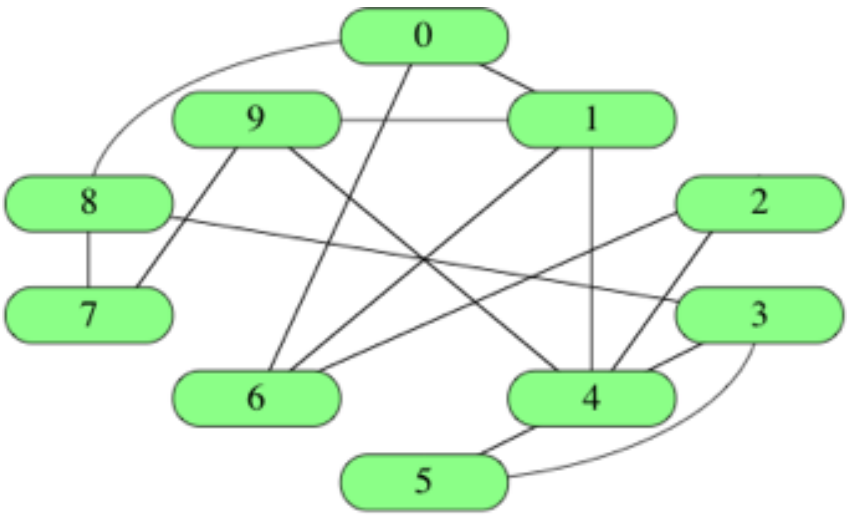


Figure 1: Topology of the largest components over various choices of threshold conditions for (a) a dataset based on email server logs at a US university, and (b) the Enron email corpus. Significant changes in topology are observed as the thresholding condition of the network is varied.

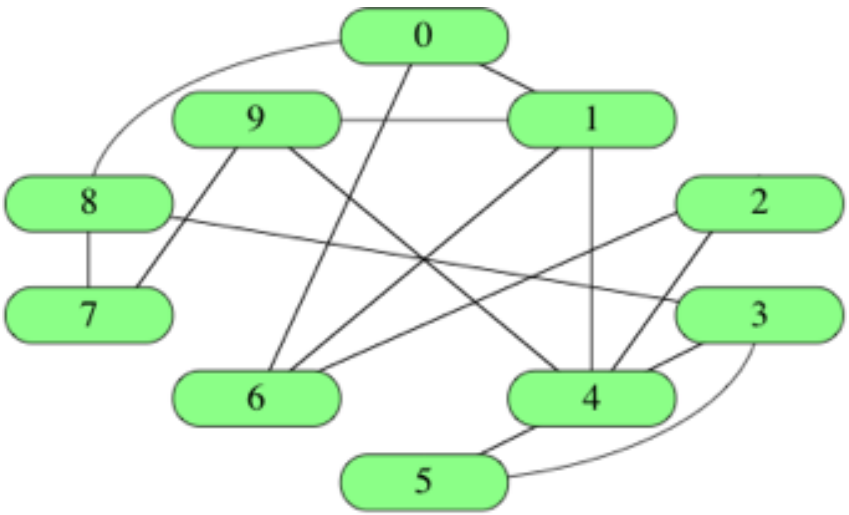
Data structures



[[0,1], [0,6], [0,8], [1,4], [1,6], [1,9], [2,4], [2,6], [3,4], [3,5], [3,8], [4,5], [4,9], [7,8], [7,9]]

Simple for storage, but difficult to compute with

Data structures

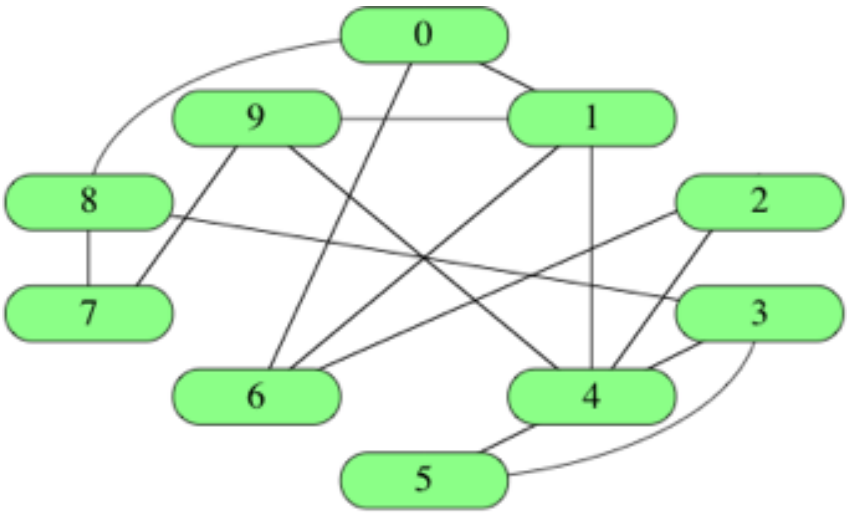


Adjacency matrix

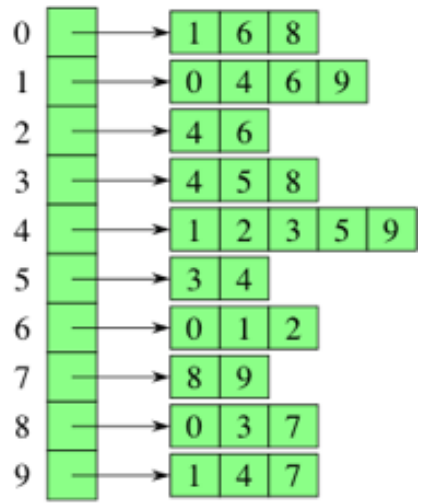
	0	1	2	3	4	5	6	7	8	9
0	0	1	0	0	0	0	1	0	1	0
1	1	0	0	0	1	0	1	0	0	1
2	0	0	0	0	1	0	1	0	0	0
3	0	0	0	0	1	1	0	0	1	0
4	0	1	1	1	0	1	0	0	0	1
5	0	0	0	1	1	0	0	0	0	0
6	1	1	1	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	1	1
8	1	0	0	1	0	0	0	1	0	0
9	0	1	0	0	1	0	0	1	0	0

Quick to check edges, good for linear algebra, often sparse

Data structures



Adjacency list



Good for graph traversal

Describing networks

Descriptive statistics

- Degree: How many connections does a node have?
- Path length: What's the shortest path between two nodes?
- Clustering: How many friends of friends are also friends?
- Components: How many disconnected parts does the network have?

Algorithms for Descriptive statistics

- Degree: How many connections does a node have?
  $\rightarrow$ Degree distributions
- Path length: What's the shortest path between two nodes?
  $\rightarrow$ Breadth first search
- Clustering: How many friends of friends are also friends?
  $\rightarrow$ Triangle counting
- Components: How many disconnected parts does the network have?
  $\rightarrow$ Connected components