Networks

APAM E4990
Modeling Social Data

Jake Hofman

Columbia University

April 5, 2019

History

∼1930s: Relationships as networks

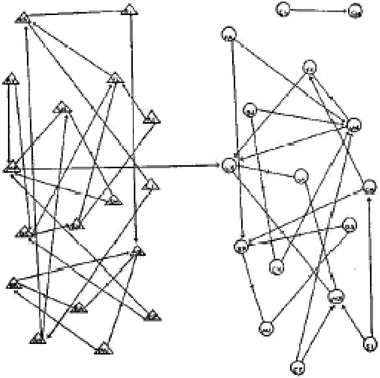

EMOTIONS MAPPED BY NEW GEOGRAPHY

Charts Seek to Portray the Psychological Currents of Human Relationships.

FIRST STUDIES EXHIBITED

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

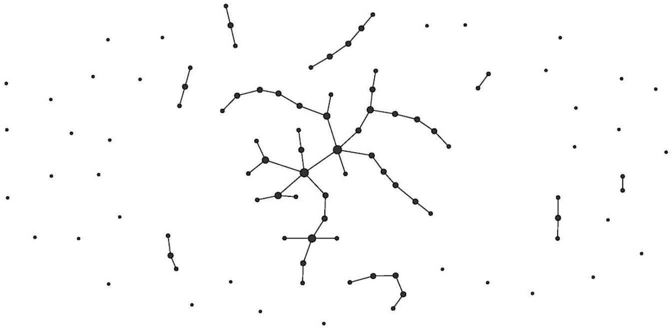
MANY MISFITS REVEALED



Moreno (1933)

http://bit.ly/sociograms

∼1960s: Random graph theory



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

Erdős & Rényi (1959)

∼1970s: Clustering, weak ties

The Strength of Weak Ties[1]

Mark S. Granovetter
Johns Hopkins University

Analysis of social networks is suggested as a tool for linking micro and macro levels of sociological theory. The procedure is illustrated by elaboration of the macro implications of one aspect of small-scale interaction: the strength of dyadic ties. It is argued that the degree of overlap of two individuals' friendship networks varies directly with the strength of their tie to one another. The impact of this principle on diffusion of influence and information, mobility opportunity, and community organization is explored. Stress is laid on the cohesive power of weak ties. Most network models deal, implicitly, with strong ties, thus confining their applicability to small, well-defined groups. Emphasis on weak ties lends itself to discussion of relations *between* groups and to analysis of segments of social structure not easily defined in terms of primary groups.

Granovetter (1973)

∼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,

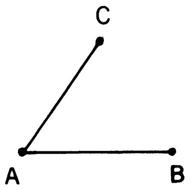


FIG. 1.—Forbidden triad

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

In a random sample of recent professional, technical, and managerial job changers living in a Boston suburb, I asked those who found a new job through contacts how often they *saw* the contact around the time that he passed on job information to them. I will use this as a measure of tie strength.[15] A natural a priori idea is that those with whom one has strong ties are more motivated to help with job information. Opposed to this greater motivation are the structural arguments I have been making: those to whom we are weakly tied are more likely to move in circles different from our own and will thus have access to information different from that which we receive.

I have used the following categories for frequency of contact: often = at least twice a week; occasionally = more than once a year but less than twice a week; rarely = once a year or less. Of those finding a job through contacts, 16.7% reported that they saw their contact often at the time, 55.6% said occasionally, and 27.8% rarely ($N = 54$).[16] The skew is clearly to the weak end of the continuum, suggesting the primacy of structure over motivation.

Granovetter (1973)

∼1970s: Clustering, weak ties

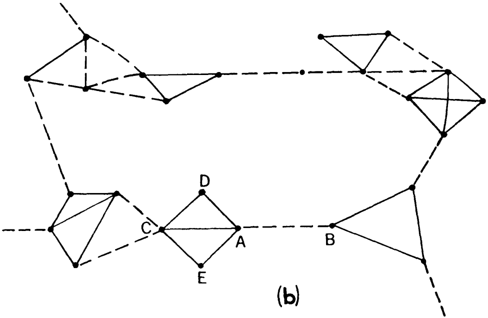
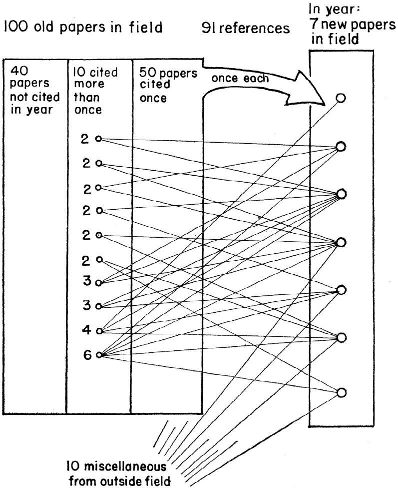


FIG. 2.—Local bridges. *a,* Degree 3; *b,* Degree 13. ——— = strong tie; – – – – = weak tie.

Granovetter (1973)

∼1970s: Cumulative advantage



de Solla Price (1965, 1976)

~1970s: Cumulative advantage

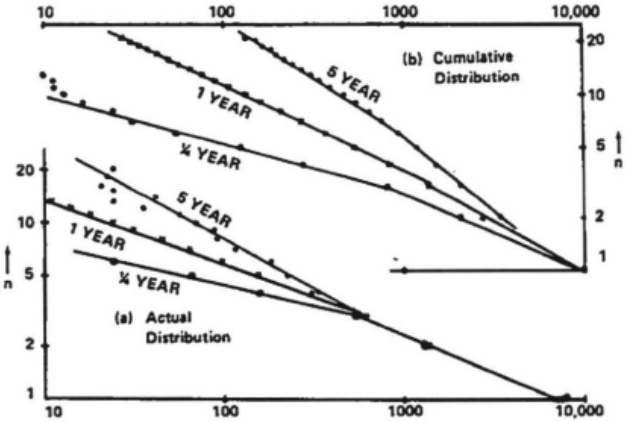
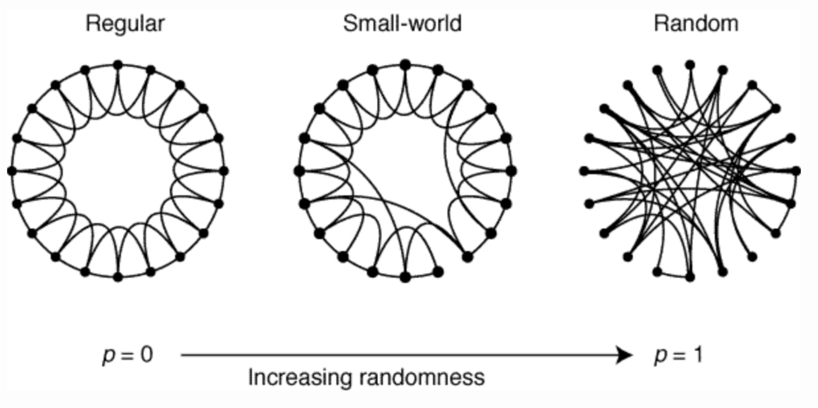


Fig. 1. Number of papers with (a) exactly and (b) at least n citations in ¼, 1, and 5-year indexes.

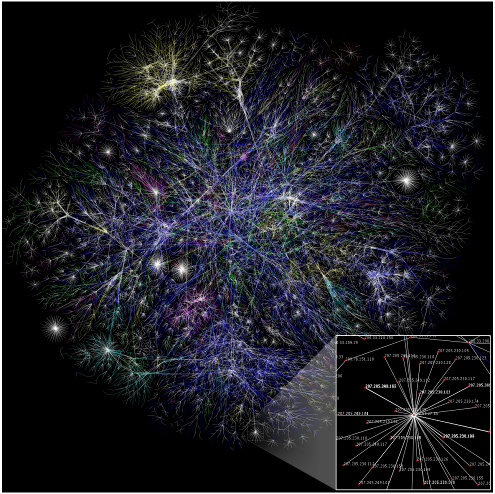
de Solla Price (1965, 1976)

∼1990s: Small-world networks



Watts & Strogatz (1998)

∼1990s: Empirical structure and dynamics of networks



Newman, Barabasi, Watts (2006)

~2000s: Homophily, contagion, and all that

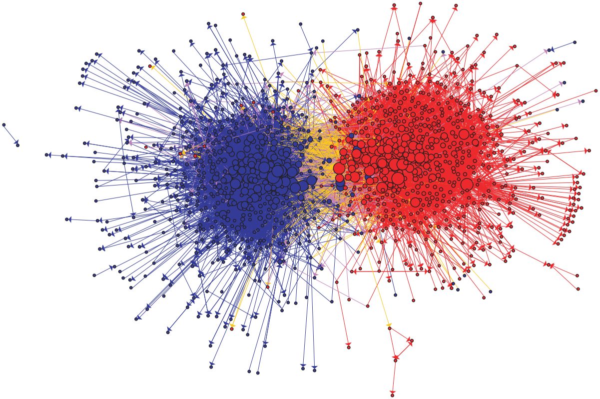


Figure 1: Community structure of political blogs (expanded set), shown using utilizing the GUESS visualization and analysis tool[2]. The colors reflect political orientation, red for conservative, and blue for liberal. Orange links go from liberal to conservative, and purple ones from conservative to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

Adamic & Glance (2005)

Types of networks

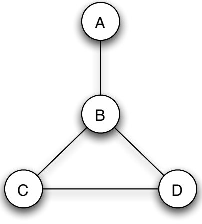
Types of networks

Networks are a useful abstractions for many different types of data

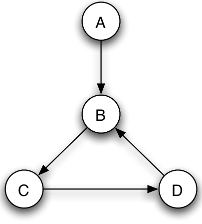
- Social networks (e.g., Facebook)
- Information networks (e.g., the Web)
- Activity networks (e.g., email)
- Biological networks (e.g., protein interactions)
- Geographical networks (e.g., roads)

Representations

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



(a) *A graph on 4 nodes.*

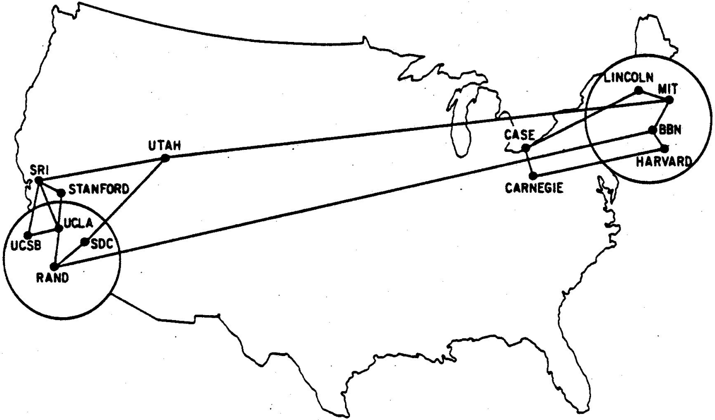


(b) *A directed graph on 4 nodes.*

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.)

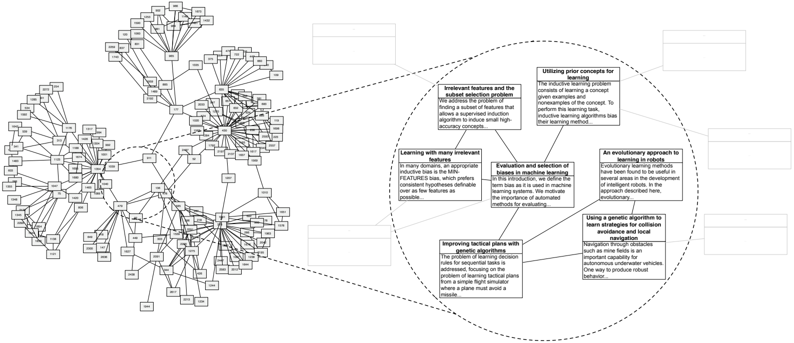


Figure 1: Example data appropriate for the relational topic model. Each document is represented as a bag of words and linked to other documents via citation. The RTM defines a joint distribution over the words in each document and the citation links between them.

Which network?

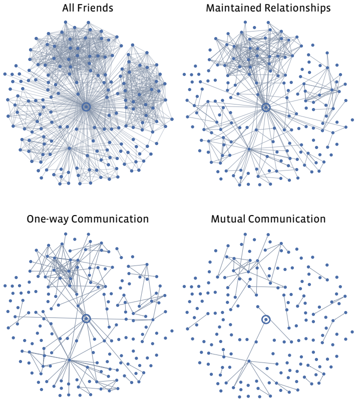


Figure 3.8: Four different views of a Facebook user's network neighborhood, showing the structure of links coresponding respectively to all declared friendships, maintained relationships, one-way communication, and reciprocal (i.e. mutual) communication. (Image from [281].)

Which network?

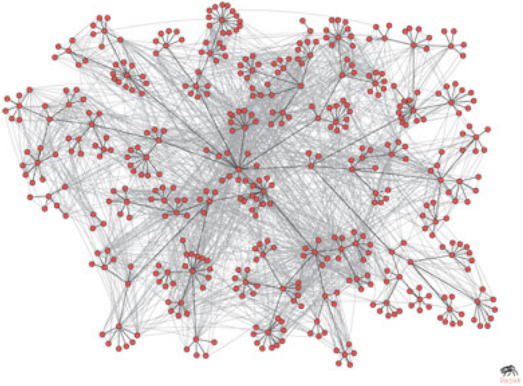


Figure 20.12: The pattern of e-mail communication among 436 employees of Hewlett Packard Research Lab is superimposed on the official organizational hierarchy, showing how network links span different social foci [6]. (Image from http://www-personal.umich.edu/ ladamic/img/hplabsemailhierarchy.jpg)

Which network?

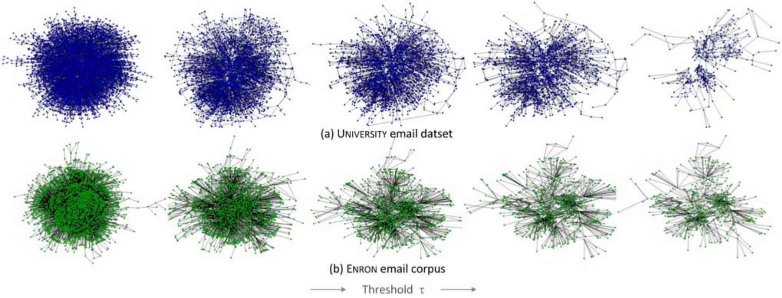
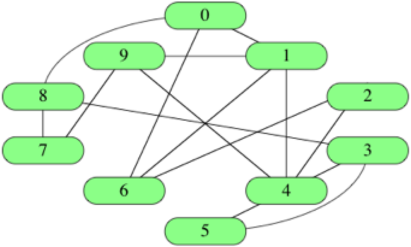


Figure 1: Topology of the largest components over various choices of threshold conditions for (a) a dataset based on email server logs at a US university, and (b) the Enron email corpus. Significant changes in topology are observed as the thresholding condition of the network is varied.

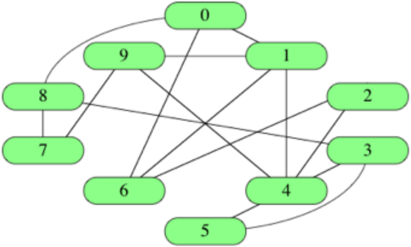
Data structures



[[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

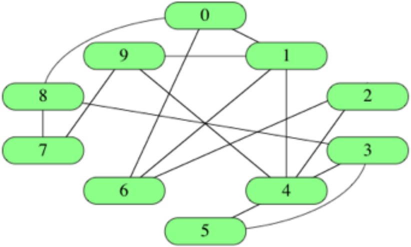


Adjacency matrix

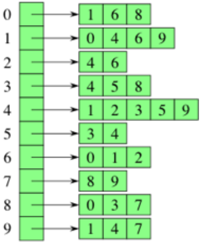
	0	1	2	3	4	5	6	7	8	9
0	0	1	0	0	0	0	1	0	1	0
1	1	0	0	0	1	0	1	0	0	1
2	0	0	0	0	1	0	1	0	0	0
3	0	0	0	0	1	1	0	0	1	0
4	0	1	1	1	0	1	0	0	0	1
5	0	0	0	1	1	0	0	0	0	0
6	1	1	1	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	1	1
8	1	0	0	1	0	0	0	1	0	0
9	0	1	0	0	1	0	0	1	0	0

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