GENERAL ASSEMBLY

INTRO TO DATA SCIENCE
LECTURE 16: NETWORK ANALYSIS

Paul Burkard

01/06/2015

LAST TIME:

- ENSEMBLE TECHNIQUES
  - HOW/WHEN MIGHT THEY HELP?
- BAGGING
- BOOSTING

QUESTIONS?

I. NETWORKS
II. NETWORK STATICS
III. NETWORK DYNAMICS
IV. BOOSTING
HANDS-ON: NETWORK ANALYSIS

- What are Networks/Graphs?
  - How might we use them?
- What are some useful quantities in networks?
- What are some common network graph algorithms?

I. NETWORKS

Q: What are **Networks***?*

A: A set of **pairwise relationships** *between* **objects***.*

The ubiquity of social networks gives rise to many interesting data-oriented questions that can be answered with analytical techniques.

Given a large set of social network data, what types of questions do you think would be interesting to ask?

Some natural questions arise when considering social network data, in particular:

- *What is the mathematical language for considering network problems?*
- *What kinds of data structures are well-suited to network analysis?*
- *What does the network look like?*

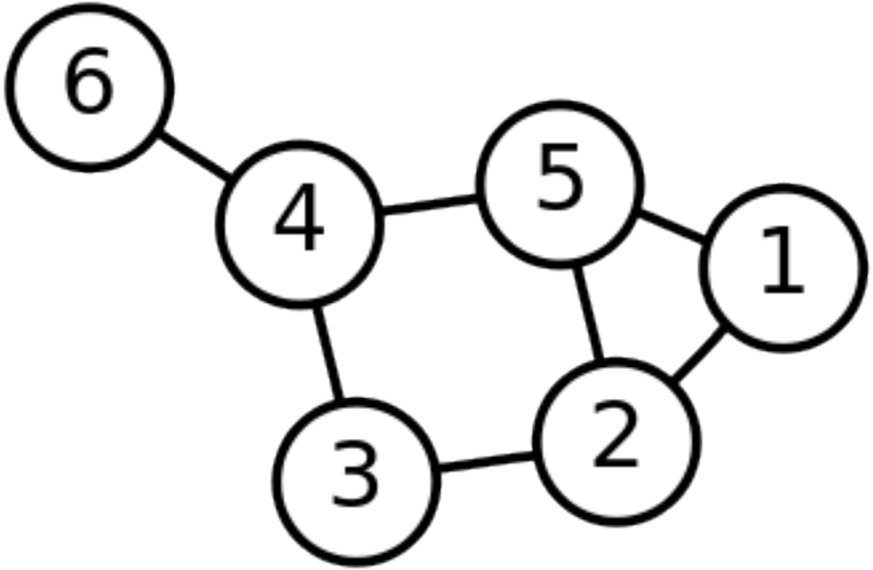
These are questions of **network representation**

Some natural questions arise when considering social network data, in particular:

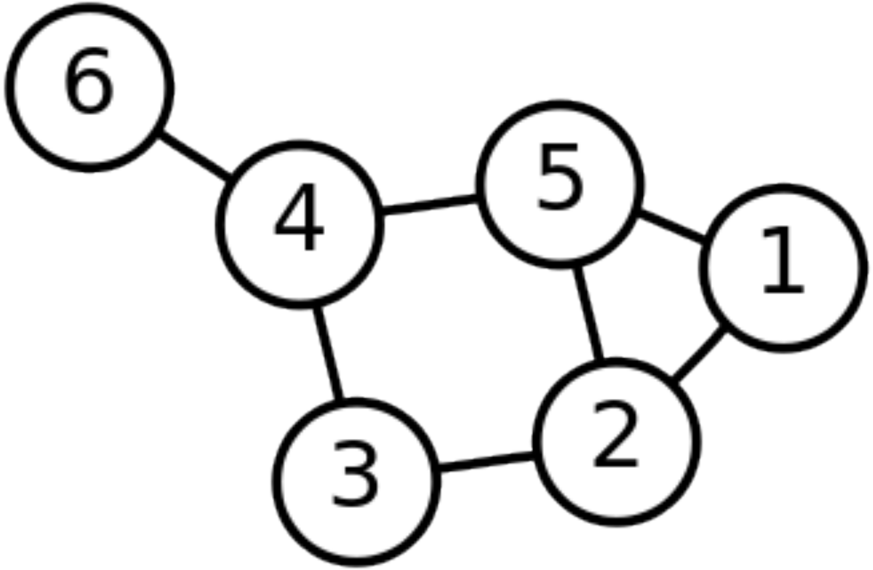
- *How is information propagated through a network?*
- *How does a network acquire or lose members?*
- *How does the structure of the network evolve through time?*
- *How do external events affect the network?*

These are questions about **network behavior**.

The mathematical representation of a network is an object called a **graph**, *which is a configuration of* **nodes** *connected by* **edges**.

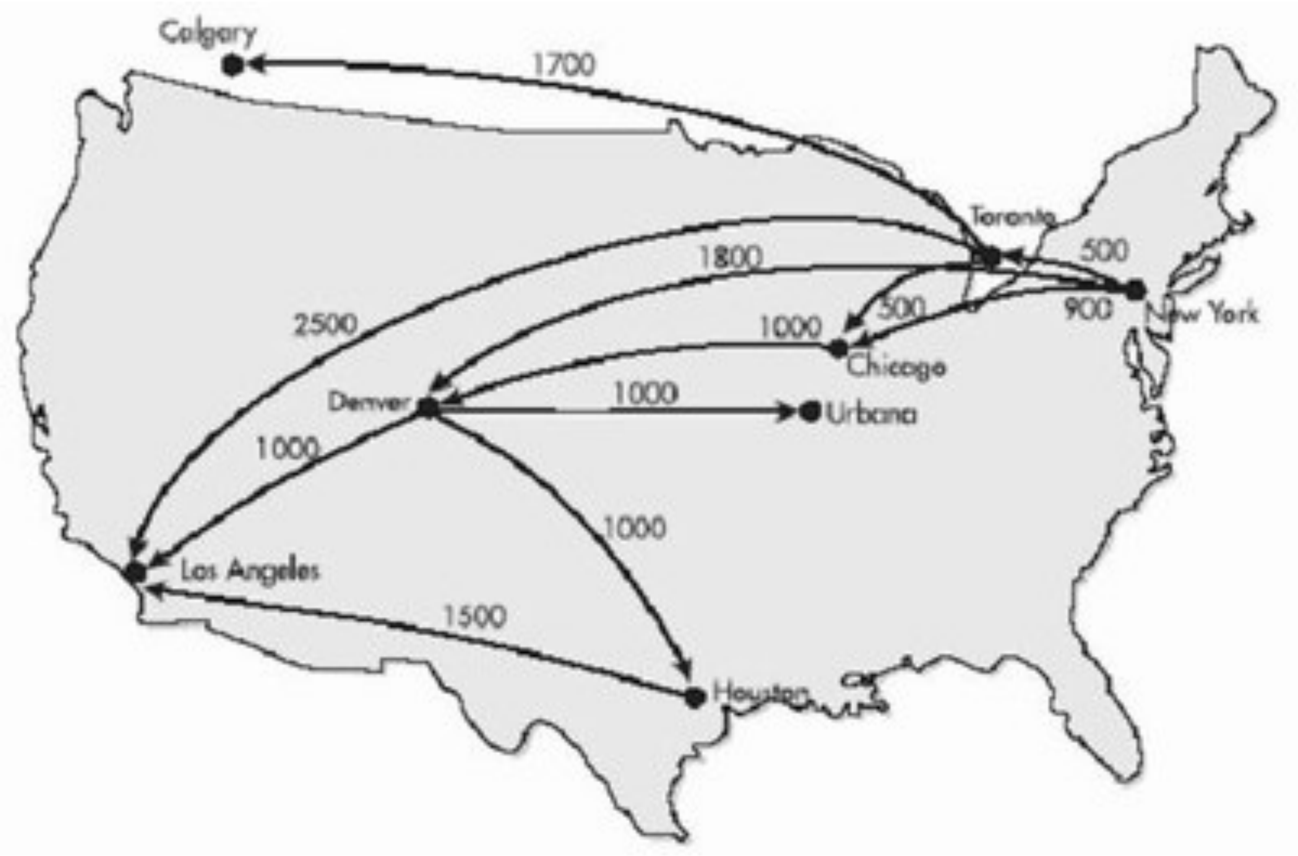


Nodes *represent* **actors** *in the graph, and* **edges** *represent the* **relationships** *between actors.*



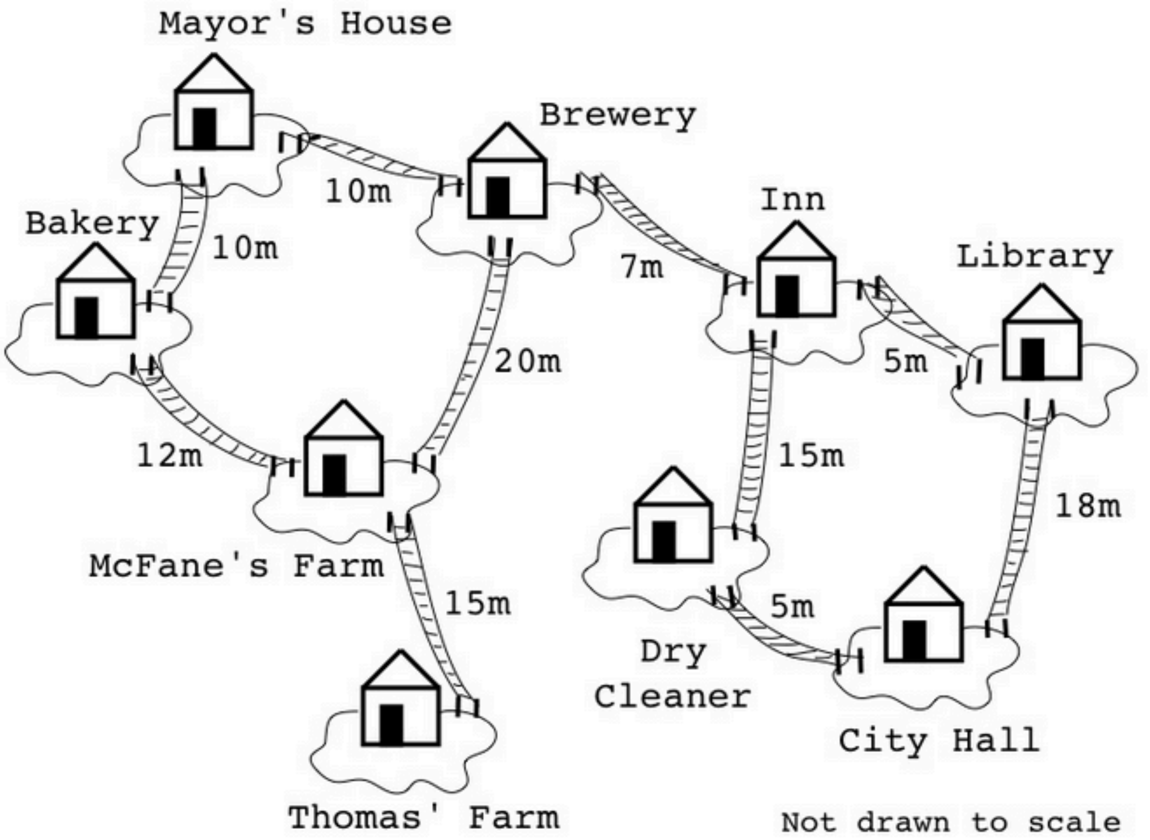
NOTE

An *undirected graph* has no directionality in its edges (bidirectional).



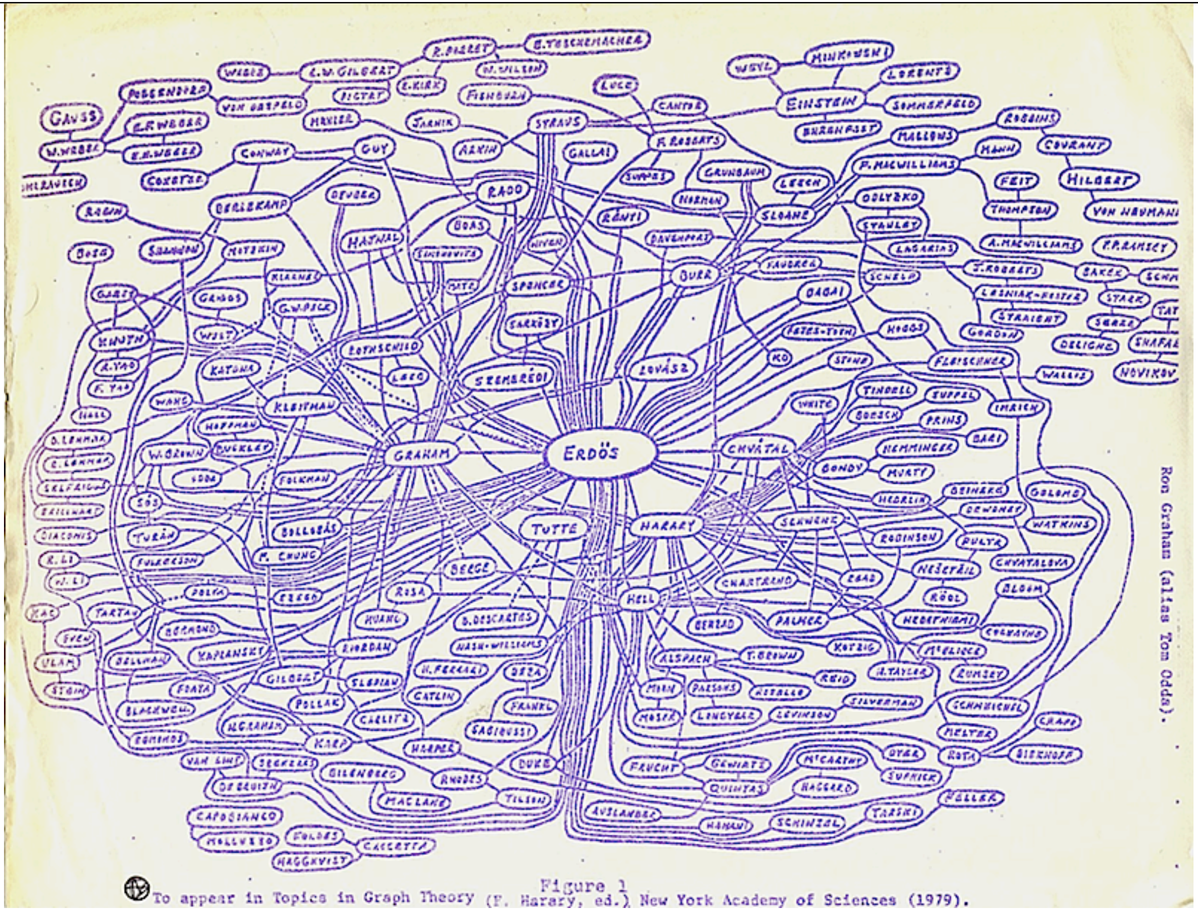
NOTE

A *directed graph* has edges that point from one node to another.



NOTE

A *weighted graph* contains edges associated with real-valued numbers, eg to measure distance or importance.



Erdos Number:

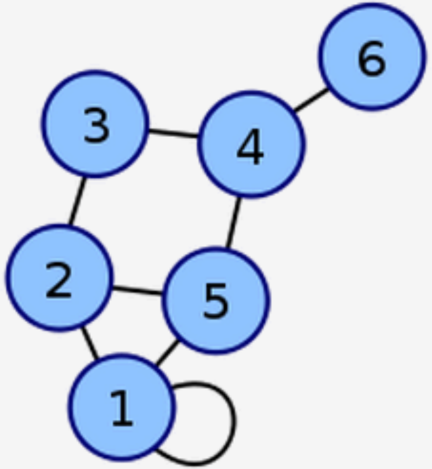
People who have collaborated with famous mathematician Paul Erdos.

. . . The Alternate Route.
OAK GROVE
Malden
Wellington
Sullivan Square
Community College
WONDERLAND
Revere Beach
Beachmont
Suffolk Downs
Orient Heights
Wood Island
Airport
Maverick
AIRPORT TERMINALS
Logan International Airport
ALEWIFE
Davis
Porter
Harvard
Central
Kendall/MIT
Charles/MGH
LECHMERE
Science Park
North Station
BOWDOIN
Haymarket
Gov't Center
Aquarium
State*
Park St
Downtown Crossing
South Station
Courthouse
World Trade Ctr
BOSTON MARINE INDUSTRIAL PARK
CITY POINT
SL1
SL2
SL3
BOSTON COLLEGE
CLEVELAND CIRCLE
RIVERSIDE
Woodland
Waban
Eliot
Newton Highlands
Newton Centre
Chestnut Hill
Reservoir
Beaconsfield
Brookline Hills
Brookline Village
Longwood
Fenway
Kenmore
Hynes/ICA
Copley
Arlington
*Boylston
Prudential
Symphony
Chinatown
NE Medical Center
Back Bay
Mass Ave
Ruggles
Roxbury Crossing
Jackson Sq
Stony Brook
Green St
FOREST HILLS
HEATH
DUDLEY SQ
Broadway
Andrew
JFK/UMass
Savin Hill
Fields Corner
Shawmut
ASHMONT
Cedar Grove
Butler
Milton
Central Ave
Valley Rd
Capen St
MATTAPAN
North Quincy
Wollaston
Quincy Center
Quincy Adams
BRAINTREE
LEGEND
Terminal Station
Transit Station
Wheelchair Accessible
Transfer Station
Parking
Commuter Rail Connection
Commuter Rail Service
Water Transportation Services

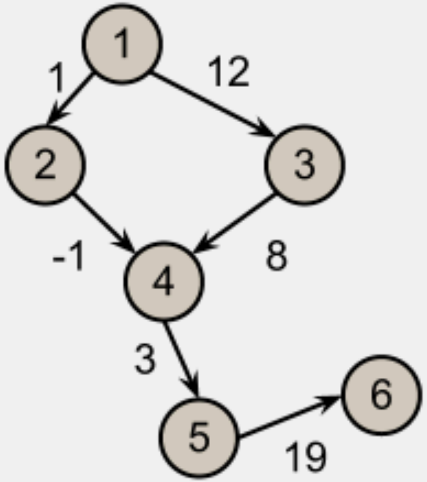
NETWORK REPRESENTATION

In practical terms, we need some data structures to represent and manipulate our network data.

One common graph representation is the **adjacency matrix**. *An n-node undirected graph can be represented by a symmetric n x n adjacency matrix A whose nonzero off-diagonal entries A_{ij} represent an edge between nodes i and j.*

Labeled graph	Adjacency matrix
	$\begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$ Coordinates are 1-6.

Weighted Directed Graph & Adjacency Matrix



Weighted Directed Graph

	1	2	3	4	5	6
1	0	1	12	0	0	0
2	-1	0	0	-1	0	0
3	-12	0	0	8	0	0
4	0	1	-8	0	3	0
5	0	0	0	-3	0	19
6	0	0	0	0	-19	0

Adjacency Matrix

NOTE

A *directed graph* has an asymmetric adjacency matrix. Can you see why?

Another common representation is the **adjacency list**.

In Python, it's just a dict!

```
graph = {'A': ['B', 'C'],
         'B': ['C', 'D'],
         'C': ['D'],
         'D': ['C'],
         'E': ['F'],
         'F': ['C']}
```

Finally there's also the **edge list***.*

It's exactly what it sounds like, we just keep a list of **(out_edge, in_edge, weight)** *tuples, 1 for each edge in the network.*

II. NETWORK STATICS

Q: What do we mean by **Network Statics***?*

A: Any potentially useful quantity that can be calculated from a static (not changing/evolving, aka a snapshot) network.

Can you think of any such questions that might be interesting?

One key concept in the study of network structure is **centrality**. *The* **centrality** *of a node is a measure of its* **importance** *in the network.*

The simplest centrality measure is the **degree** *of a node, which is simply the number of edges connected to it. Using the adjacency matrix notation for an undirected graph, we can express the degree* k_i *of node* i *as:*

$$k_i = \sum_{j=1}^{n} A_{ij}.$$

A more sophisticated measure called **eigenvector centrality** *allows important edges to give larger contributions to centrality:*

$$x_i = \frac{1}{\lambda} \sum_{j=1}^{n} A_{ij} x_j,$$

Here the eigenvector centrality x_i *of node* i *is proportional to the average centrality of* i*'s network neighbors.*

Another useful centrality measure is based on the idea of shortest-distance (or geodesic) paths through the graph.

If σ_{st} is the number of geodesic paths from node s to node t, and $\sigma_{st}(v)$ is the number of these paths that cross node v, then the **betweenness centrality** *of node v is given by:*

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

NOTE

Betweenness centrality measures the proportion of geodesic paths passing through a node.

This gives an idea of the node's *influence* in the network.

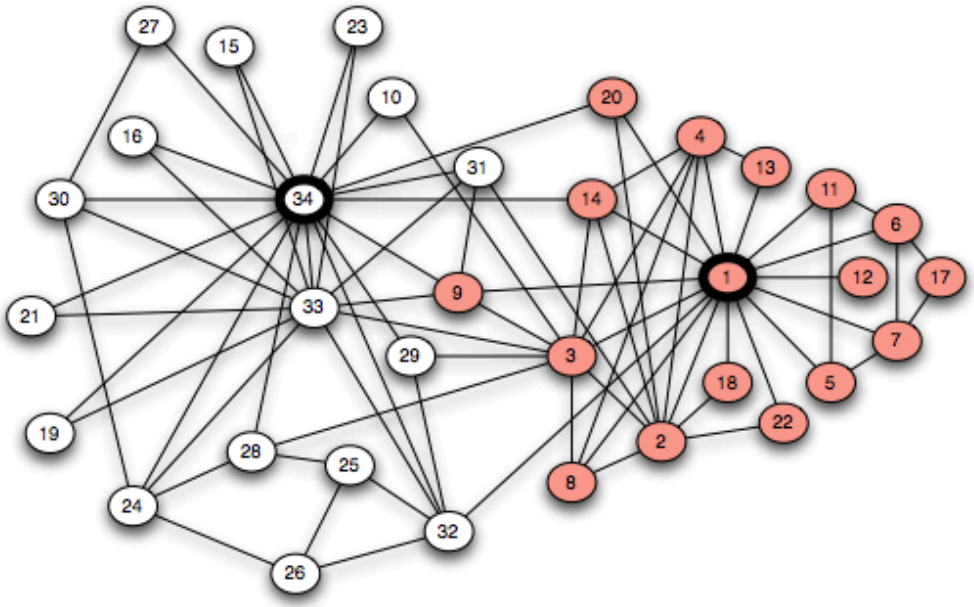


Figure 3.13: A karate club studied by Wayne Zachary [421] — a dispute during the course of the study caused it to split into two clubs. Could the boundaries of the two clubs be predicted from the network structure?

Geodesic paths form the basis of another well-known property of networks called the **small-world effect**.

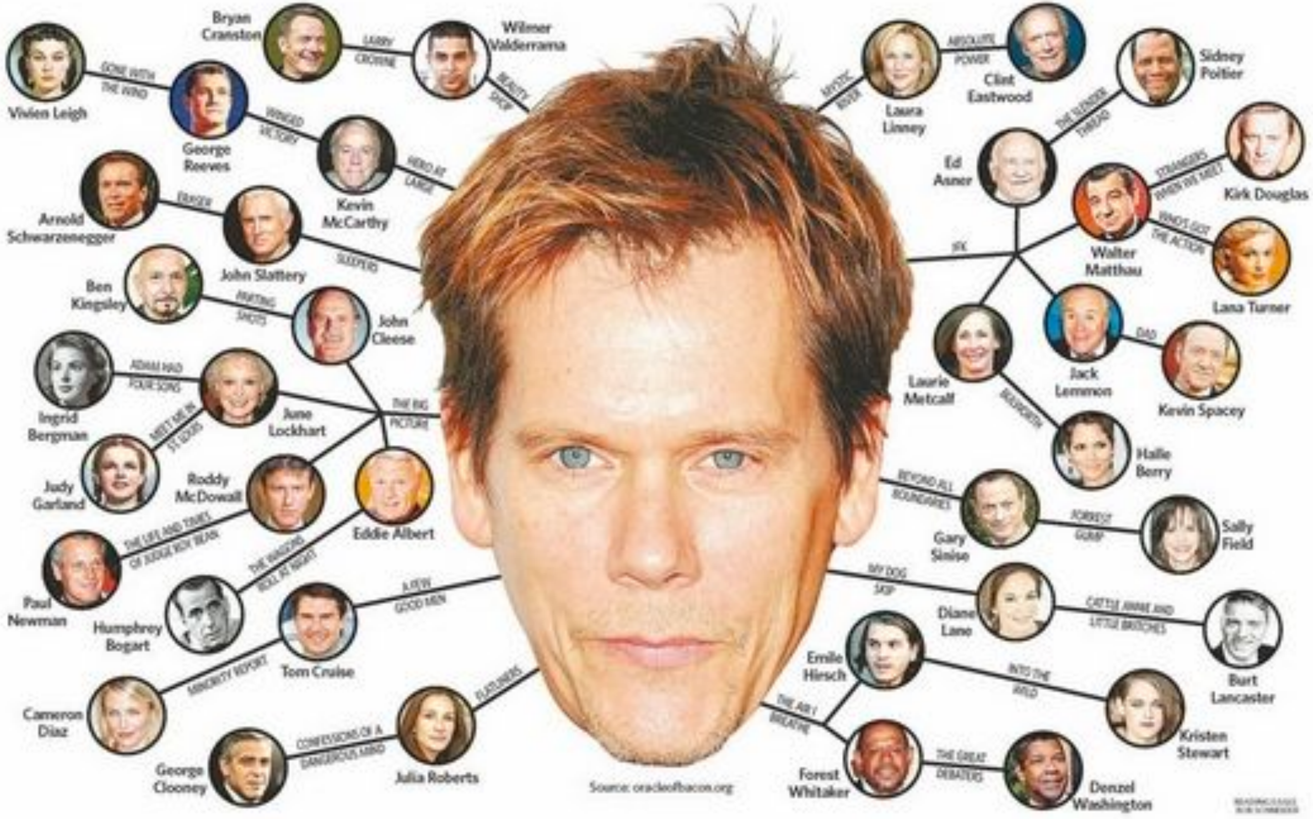
Specifically, most networks have a mean geodesic distance between nodes that is small compared to the network size as a whole.

NOTE

This is the origin of "6 degrees of separation".

A famous study in the 1960s asked participants to try to get a letter to a particular individual by passing it from one acquaintance to another, and found that the mean geodesic distance in this case was about 6.

SMALL-WORLD EFFECT



III. NETWORK DYNAMICS

Q: What do we mean by **Network Dynamics***?*
A: When we want to analyze **how a network evolves** *in time.*

This can be in the context of actors and relationships changing, or…

An even more frequent topic of examination is how **information** *might* **propagate** *through a network.*

Can you think of some use cases for the latter?

Suppose we're interested in the idea of how information (or behavior) spreads through a network, how can this help?:

- *How do members of a social network influence each other to adopt a new technology/product/behavior?*
- *How did information about the Bin Laden raid spread over Twitter?*
- *What's the best way to use a social network to market your product?*

There are two primary methods of influence in social networks:

informational effects – *people observe the decisions of their network neighbors & gain indirect information that lead them to try the innovation themselves*

direct benefit effects – *people may have incentives to use the same products/technology/etc as their network neighbors*

Studies of informational effects have shown that while initial lack of information makes innovations risky to adopt, adopters ultimately benefit.

Furthermore, early adopters share certain common traits (eg higher socio-economic status, wider travel experience), and they influence their neighbors by providing indirect information about the innovation.

Studies of informational effects have shown that while initial lack of information makes innovations risky to adopt, adopters ultimately benefit.

Furthermore, early adopters share certain common traits (eg higher socio-economic status, wider travel experience), and they influence their neighbors by providing indirect information about the innovation.

Consider the following model of information **diffusion** *for two nodes* v*, w and two behaviors* A*,* B *(with payoffs* a*,* b*):*

		w: A	w: B
v	A	a, a	$0, 0$
v	B	$0, 0$	b, b

Figure 19.1: A-B Coordination Game

The question we'd like to answer is, how can v *maximize its payoff given that some of its neighbors adopt* A *& some adopt* B*?*

To start modeling this problem, suppose first that the proportion of v's neighbors selecting A is p, and the proportion selecting B is $(1\text{-}p)$.

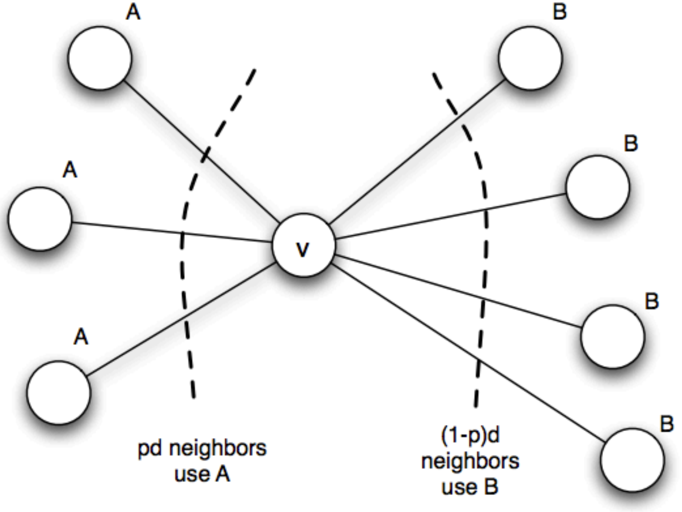


Figure 19.2: v must choose between behavior A and behavior B, based on what its neighbors are doing.

To start modeling this problem, suppose first that the proportion of v's neighbors selecting A is p, and the proportion selecting B is $(1\text{-}p)$.

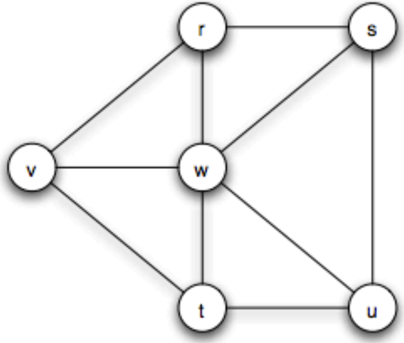
Therefore the payoff to v for choosing A is pda, and the payoff for choosing B is $(1\text{-}p)db$.

And thus v will adopt A if p (meets or) exceeds a threshold q:

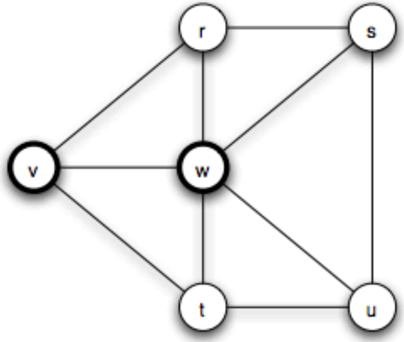
$$pda \geq (1\text{-}p)db \qquad p \geq b/(a+b) = q$$

Adoption depends not only on the relative payoffs, but also on the structure of the network (eg, how many neighbors v has, and which particular nodes these neighbors are).

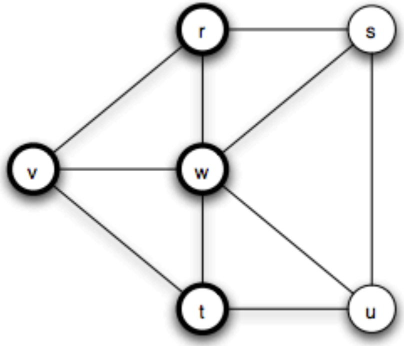
One can imagine a forecaster running a simulation of innovation adoption for the varying options where as adoption (information) flows through the network each actor eventually is faced with the a/b criteria decision just described.



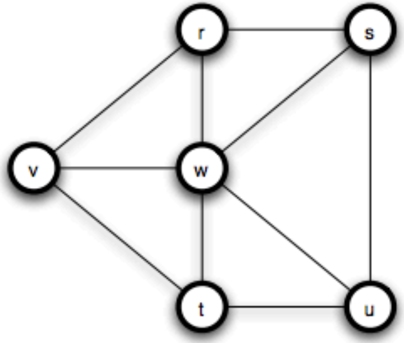
(a) *The underlying network*



(b) *Two nodes are the initial adopters*



(c) *After one step, two more nodes have adopted*

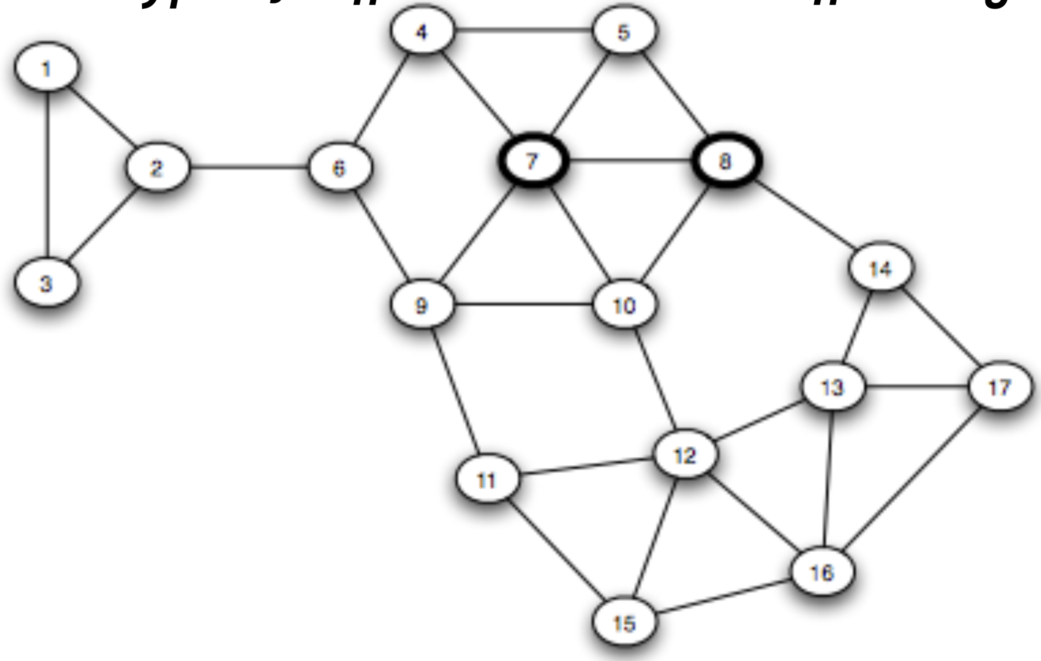


(d) *After a second step, everyone has adopted*

NOTE

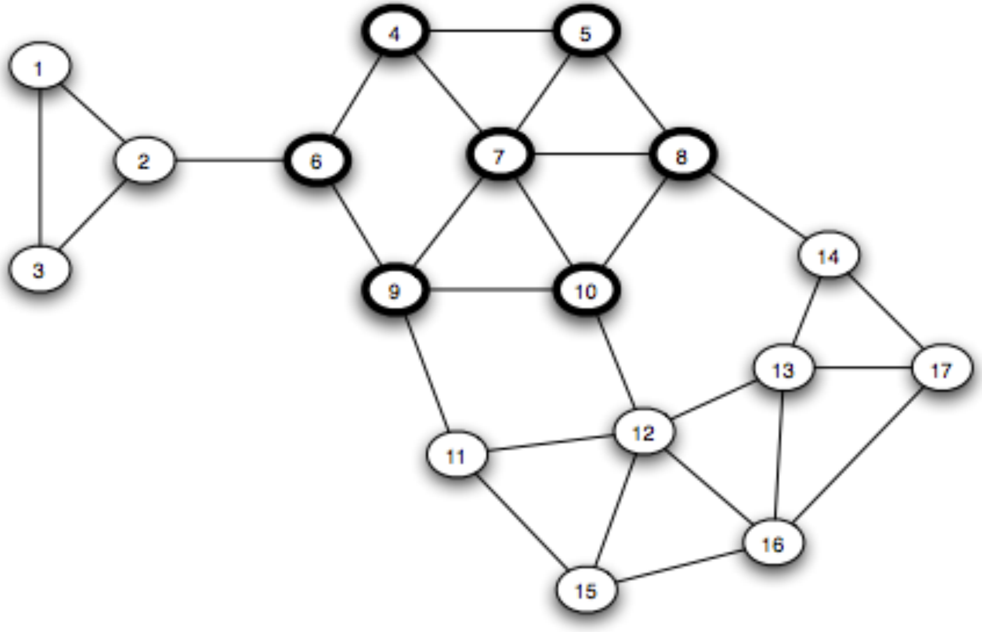
Since all nodes have adopted, this is called a *complete cascade* (at threshold q).

Consider the same type of diffusion now on a different graph:



(a) *Two nodes are the initial adopters*

Since not all nodes adopt, this is called a **partial cascade***.*



(b) *The process ends after three steps*

Here's an interesting question: how can you identify which (non-adopting) nodes are most important to allowing the cascade to continue?

Answering this question effectively is the idea behind **viral marketing**.

IV. MODERN GRAPH PROBLEMS

- *Finding* **shortest paths**
  - *Routing internet traffic or UPS trucks*
- *Finding* **minimum spanning trees**
  - *Telco laying down fiber*
- *Finding* **max flow**
  - *Airline scheduling*
- **Bipartite Matching**
  - *monster.com, match.com*
- *Google* **PageRank!**

HANDS-ON: NETWORK ANALYSIS