GENERAL ASSEMBLY

INTRO to DATA SCIENCE

Lecture 16: networks & graphs

Last time:

- big data
- map reduce

questions?

I. networks
II. Network statics
III. Network dynamics

I. networks

Q: What is a network?

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

Q: What is a network?
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.

Q: What is a network?
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:

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?

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 about network *representation*.

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

- Who are the most central and/or influential actors in a network?
- Can the network be decomposed into coherent smaller groups?
- Are any of these groups vital to the functioning of the network?
- How does this network compare to other networks?

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

- Who are the most central and/or influential actors in a network?
- Can the network be decomposed into coherent smaller groups?
- Are any of these groups vital to the functioning of the network?
- How does this network compare to other networks?

These are questions about network *structure*.

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

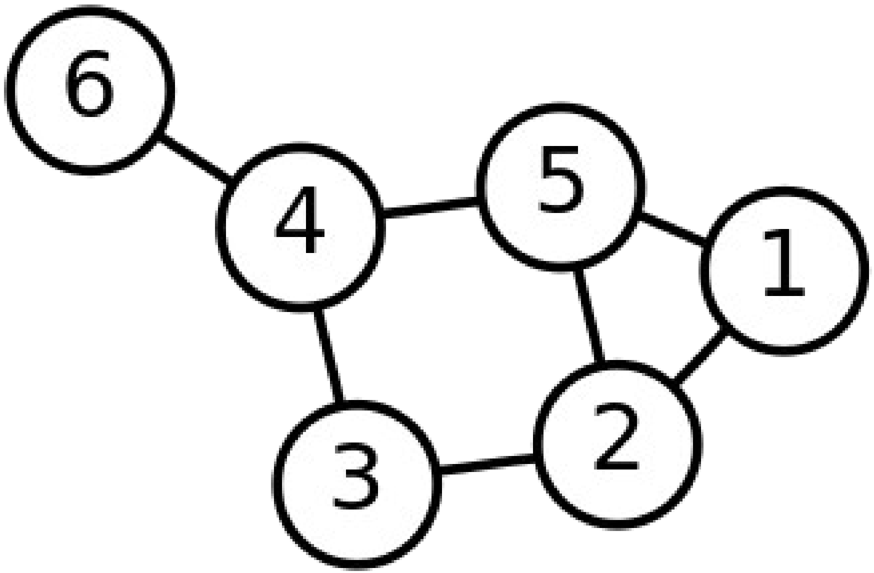
- How is information propagated through a network?
- How does a network acquire or lose members?
- How does the structure of the network evolve through time?
- How do external events affect the network?

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?

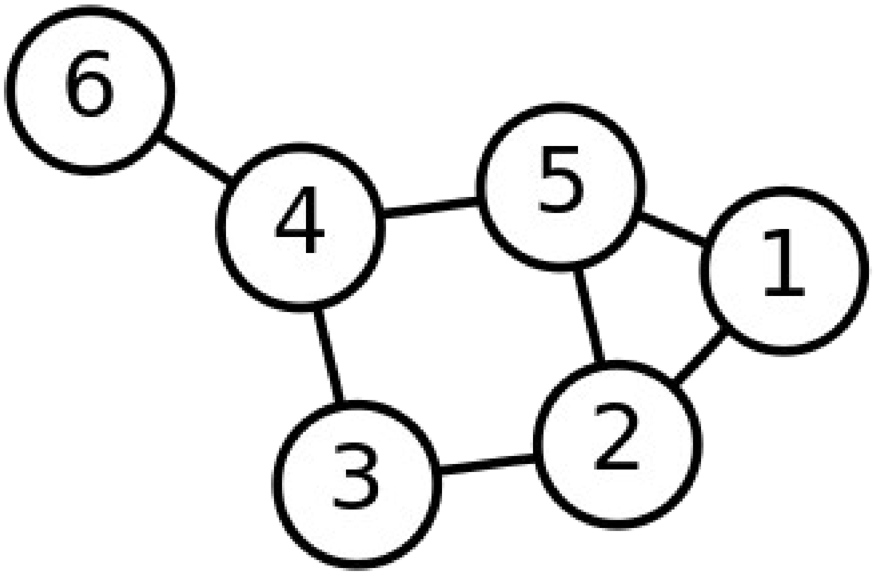
These are questions about network *behavior*.

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

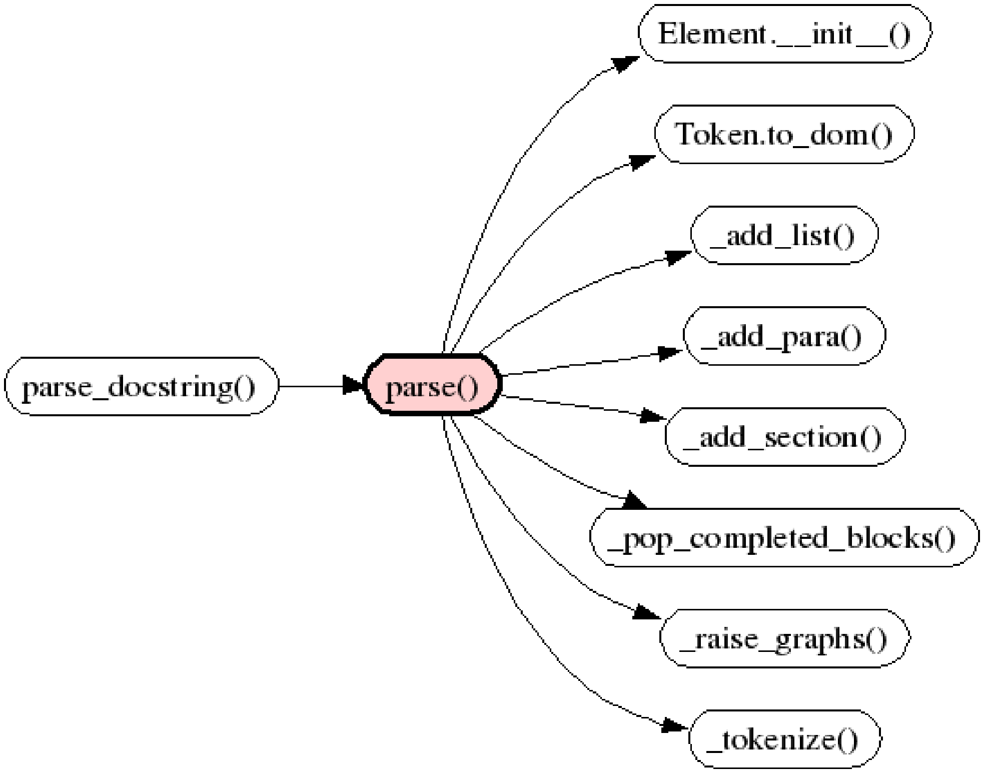


Network representation

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



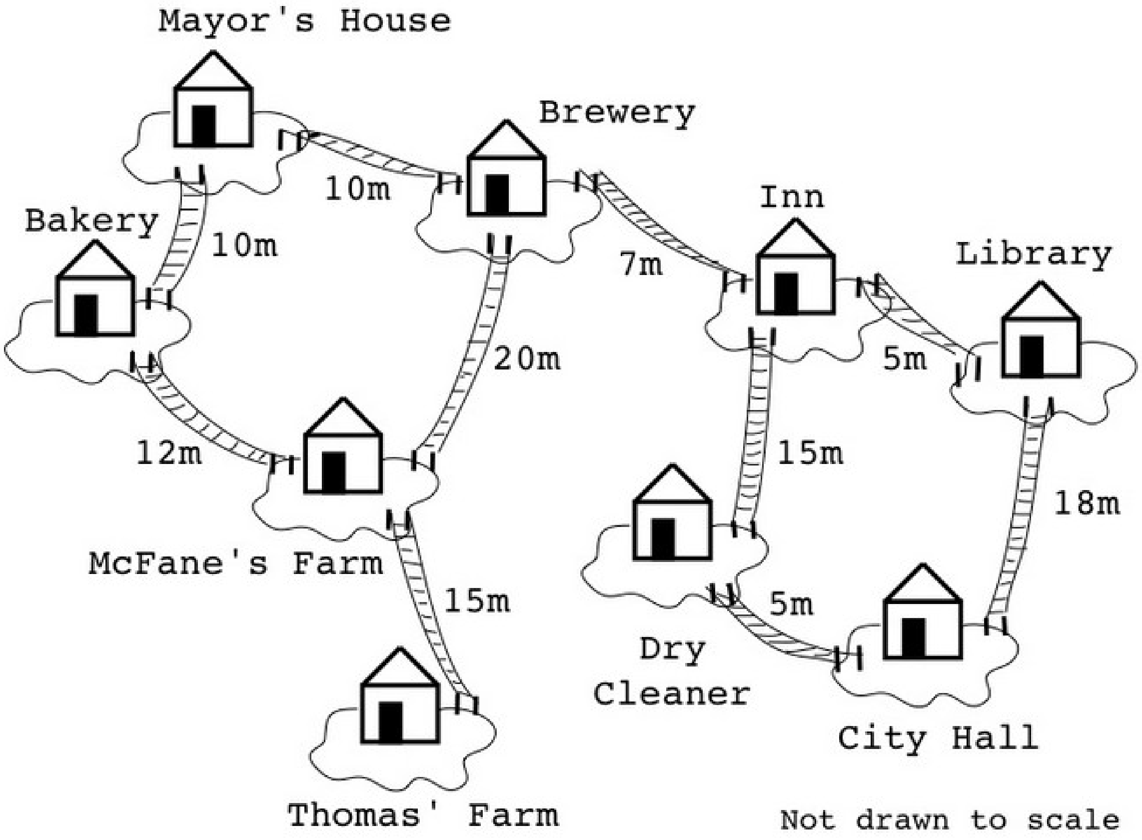
Network representation



NOTE

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

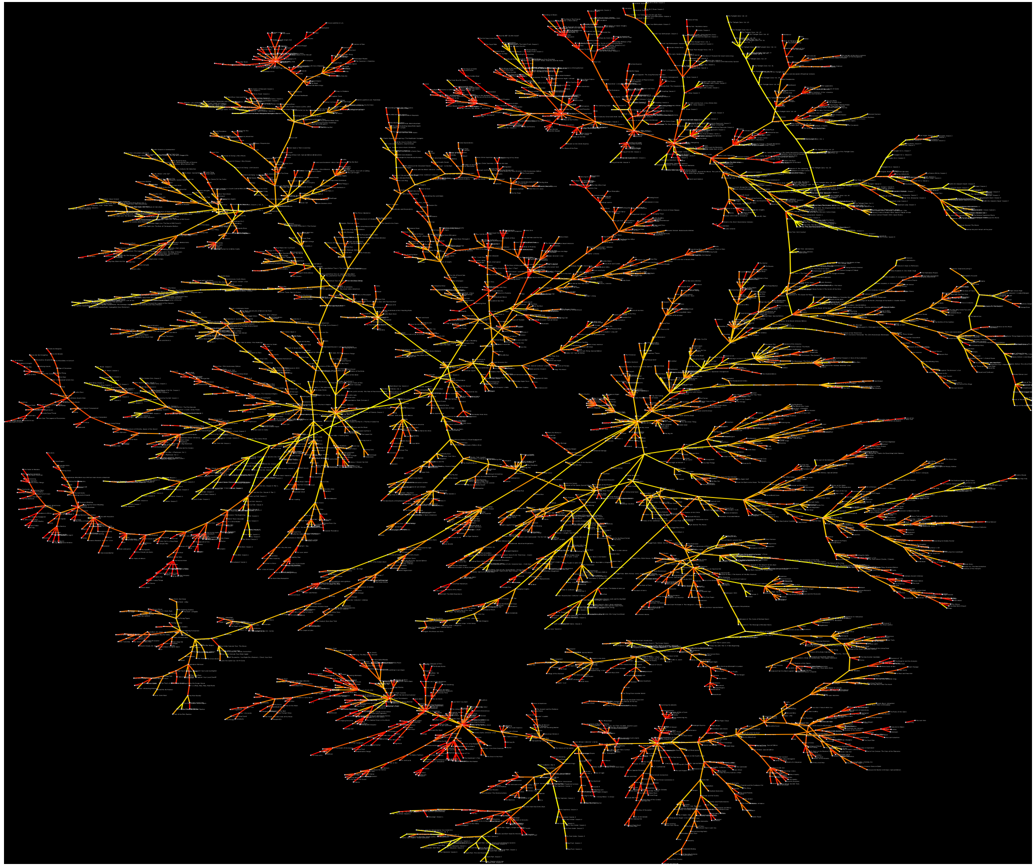
Network representation



NOTE

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

Network representation

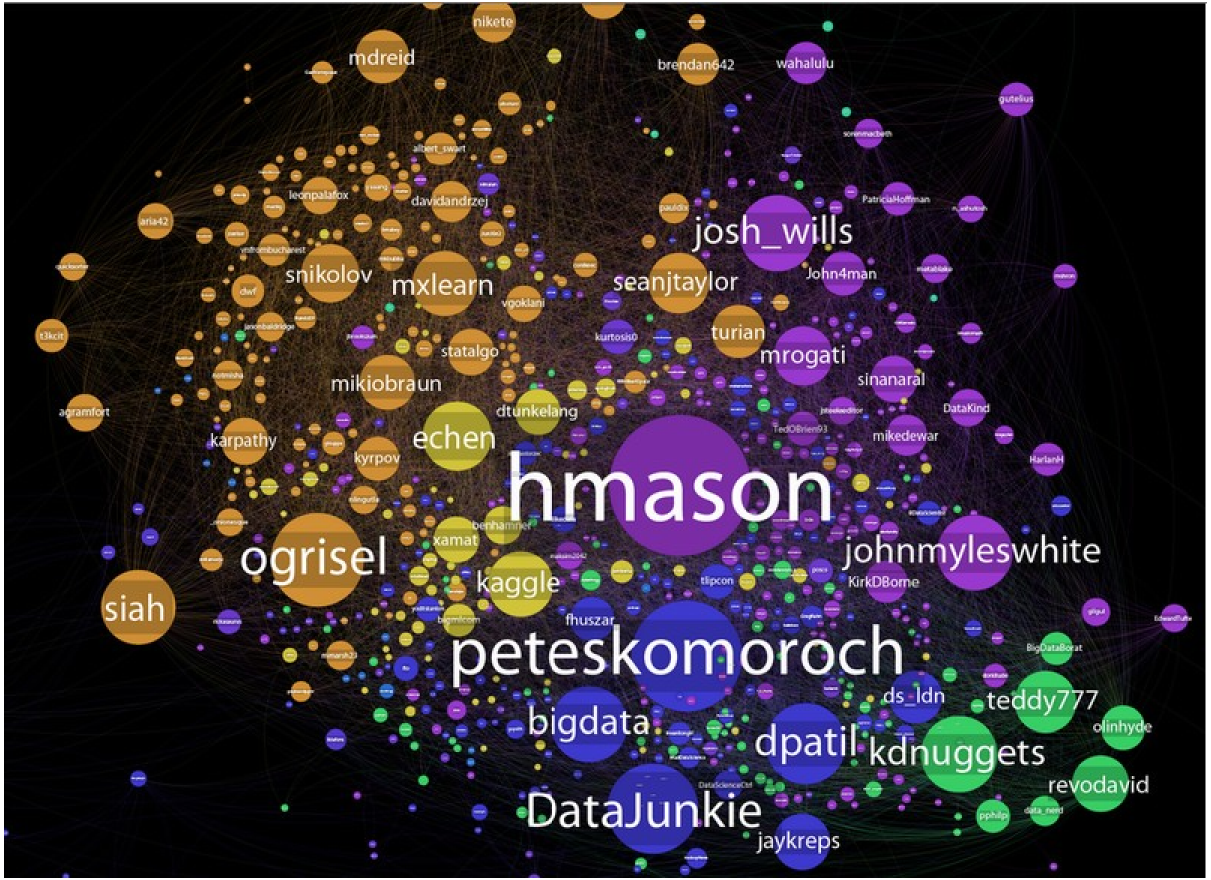


Source: netflix prize team The Ensemble
http://www.flickr.com/photos/chef_ele/3791293142/sizes/o/in/set-72157621825510293/

Network representation

Network representation

Network representation

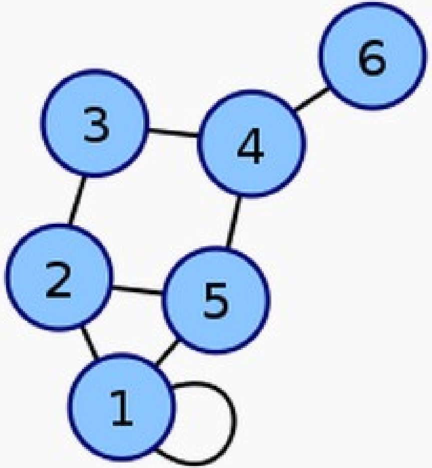


In practical terms, we need some data structures to represent and manipulate our network data.

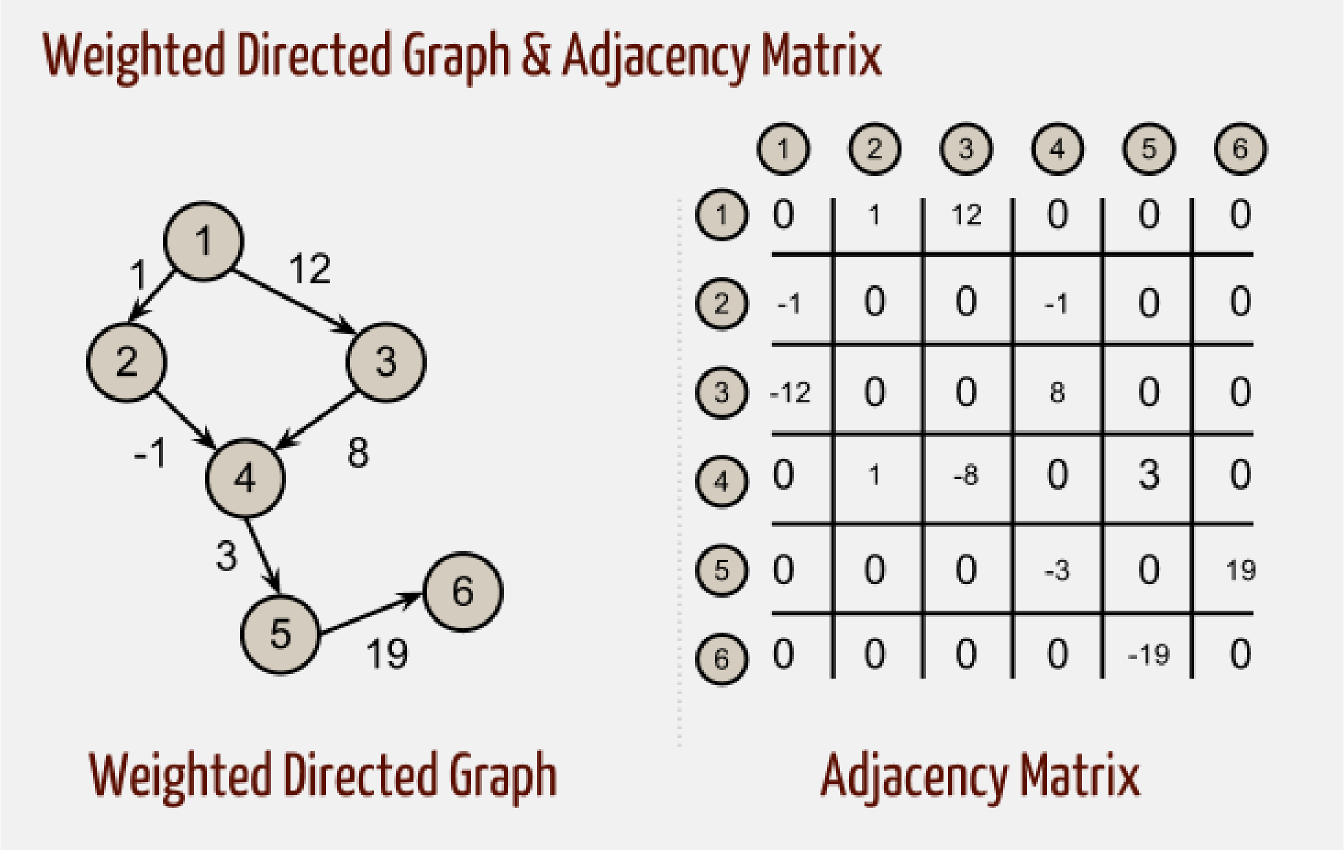
In practical terms, we need some data structures to represent and manipulate our network data.

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

Network representation

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.

Network representation



NOTE

A directed graph has an *asymmetric* adjacency matrix.

Can you see why?

Another useful tool is the adjacency list (actually a dict!):

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

Does this adjacency dict represent a directed or undirected graph? How could you generalize this to represent a weighted graph?

II. network statics

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.

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 *ki* 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:

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

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 neighbours.

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

See http://oracleofbacon.org/movielinks.php

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:

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

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

If σst is the number of geodesic paths from n
node t, and $\sigma st(v)$ is the number of these pat
cross node v, then the betweenness centralit
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.

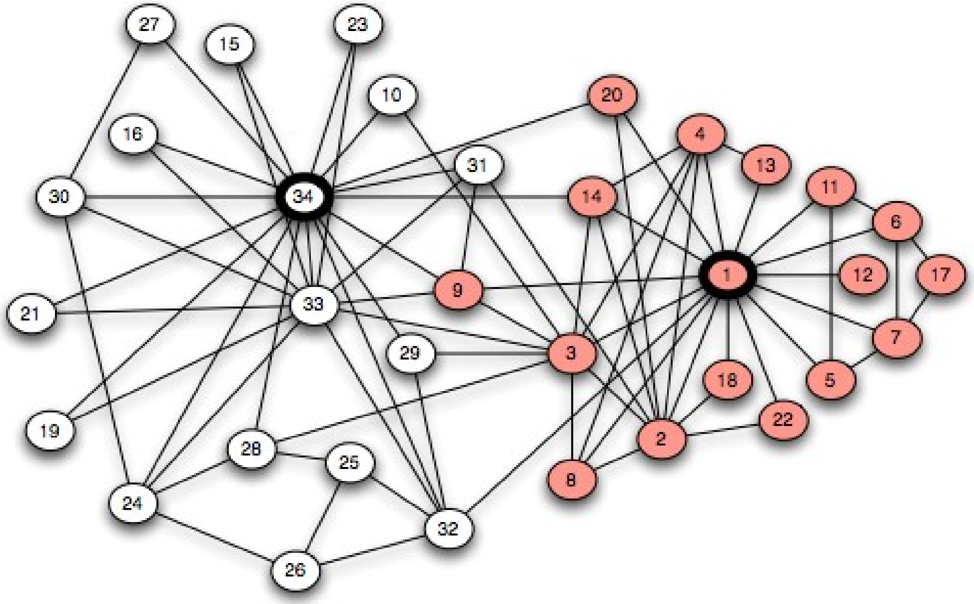


Figure 3.13: A karate club studied by Wayne Zachary [421] — a dispute during the course of the study caused it to split into two clubs. Could the boundaries of the two clubs be predicted from the network structure?

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

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.

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.

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.

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

Specifically, most networks have a mean geo distance between nodes that is small compa the network size as a whole.

NOTE

This is where the phrase “six degrees of separation” comes from.

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.

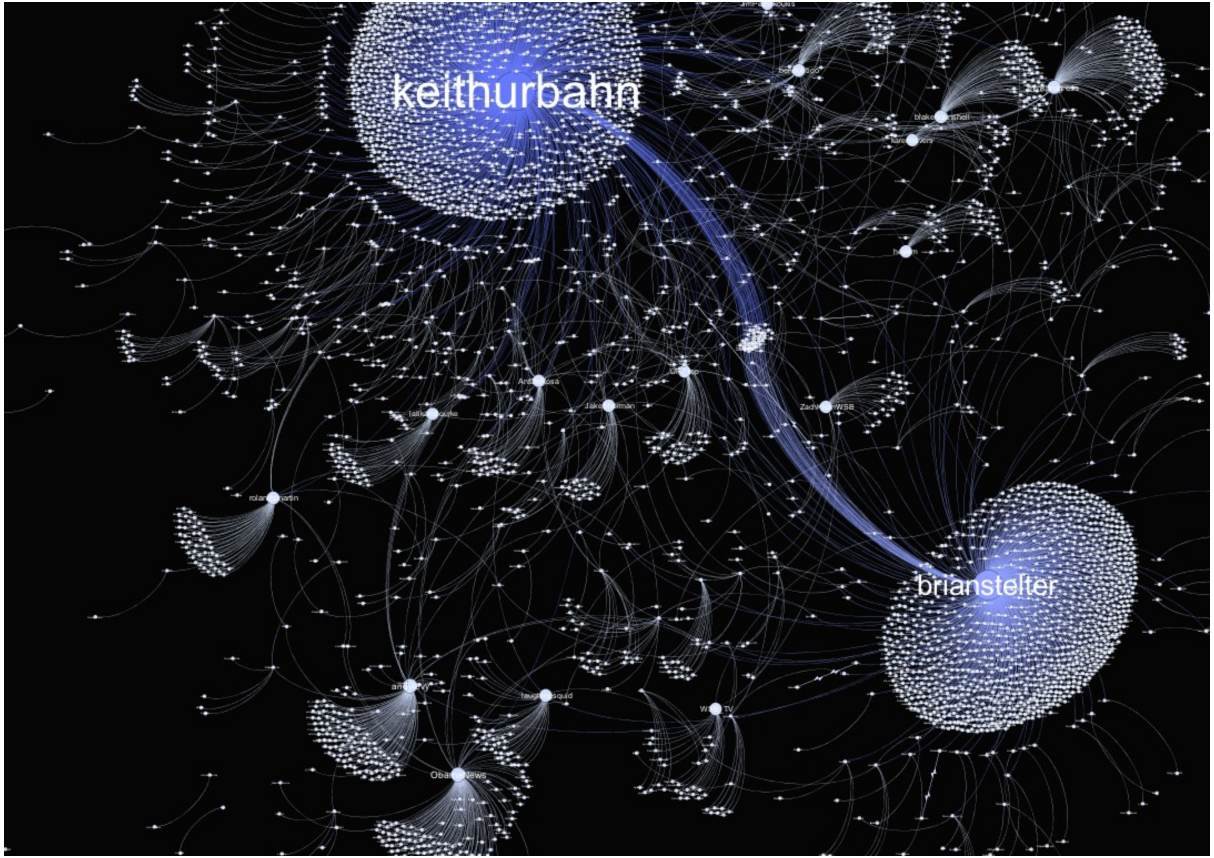
II. network dynamics

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

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

- 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?

Network dynamics



There are two primary methods of influence in social networks:

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

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

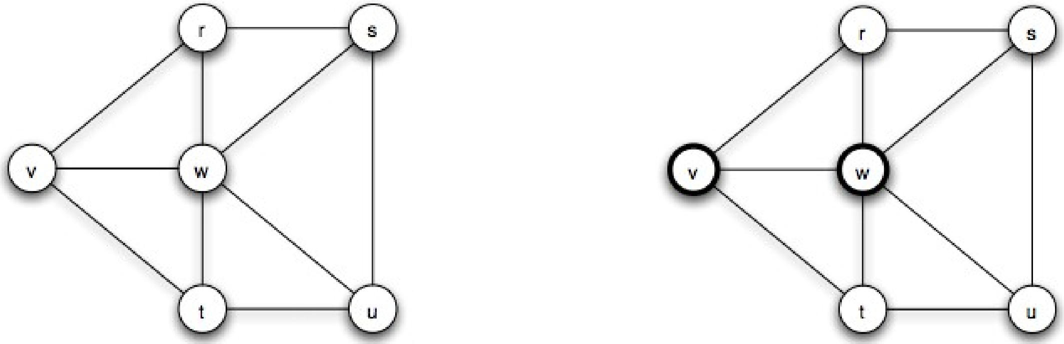
direct benefit effects – people may have incentives to use the same products/technology/etc as their network neighbors

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.

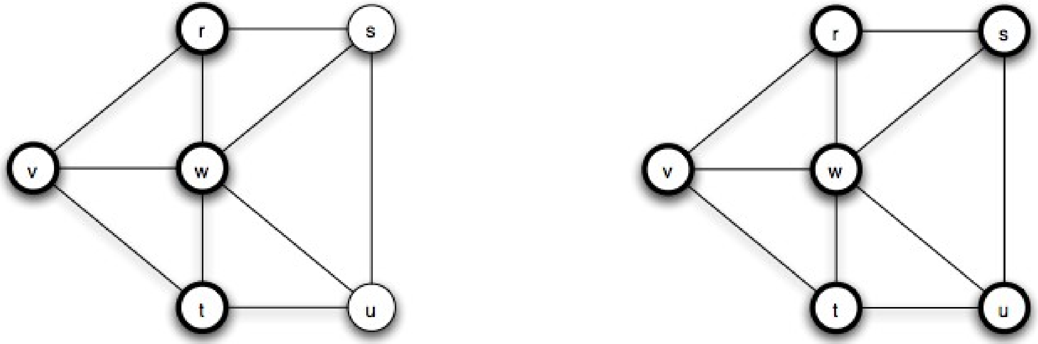
Adoption depends not only on the relative payoffs, but also on the structure of the network (eg, how many neighbors *a node* has, and which particular nodes these neighbors are).

Network dynamics



(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).

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

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*.

Very Good Source on graph Analysis

https://github.com/swinton/Web-Science-Summer-School-2011