Web Data Mining

Lecture 6: Social Network Analysis

Jaroslav Kuchař & Milan Dojčinovski
jaroslav.kuchar@fit.cvut.cz, milan.dojchinovski@fit.cvut.cz

Czech Technical University in Prague - Faculty of Information Technologies - Software and Web Engineering

Summer semester 2019/2020
Humla v0.3

Overview

- Introduction
- Graph Theory
- The Small-World Phenomenon
- Measures of Centrality
- Strong and Weak Ties
- Network Level Characteristics

Social Network Analysis

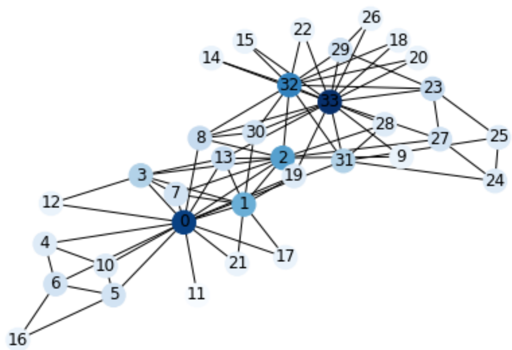
- SNA is not a buzzword attributed to Facebook or Twitter, it is a methodology well known for decades.
  - *A pioneering book: J. Moreno. "Who shall survive?: A new approach to the problem of human interrelations". Nervous and Mental Disease Publishing Co, 1934.*
- Study of human relationships by means of graph theory.
- Analysis of relationships to understand people and groups.
  - *Relationships*
    - → *Binary and Valued Relationships*
      - → *A follows B on Twitter*
      - → *A retweeted 4 tweets from B*
      - → *A talked to B 5 times last week*
    - → *Symmetric and Asymmetric Relationships*
      - → *teacher/student, followers*
      - → *friends, romantic relationships*
    - → *Multimode Relationships*
      - → *student studies at the university*

Example: Karate Club

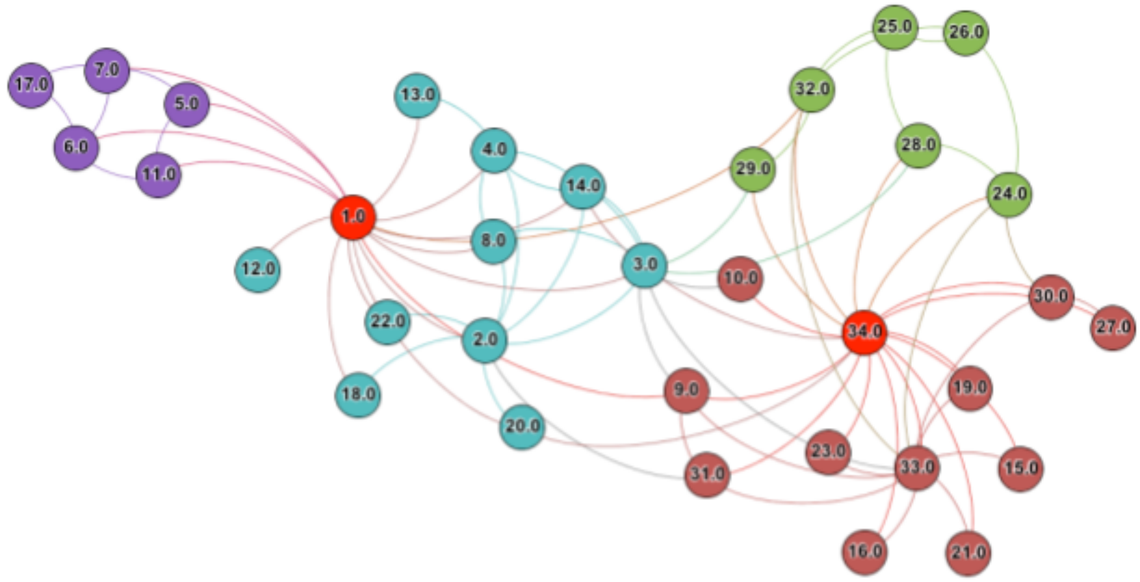
- NetworkX implementation:

```
import networkx as nx
from networkx.drawing.nx_agraph import graphviz_layout
import matplotlib.pyplot as plt
G=nx.karate_club_graph()
pos = graphviz_layout(G, prog="fdp")
nx.draw(G, pos,
        labels={v:str(v) for v in G},
        cmap = plt.get_cmap("Blues"),
        node_color=[G.degree(v) for v in G])
plt.show()
```

- Visualization:



Example: Karate Club (cont.)



- Wayne Zachary. An information flow model for conflict and fission in small groups.Journal of Anthropological Research, 33(4):452–473, 1977.

Applications

- Businesses
  – *analyze and improve communication flow in organization, partners, customers*
- Law enforcement agencies (army)
  – *identify criminal and terrorist networks, key players*
- Web Sites
  – *identify and recommend potential friends*
- Organizations
  – *uncover conflicts of interests (government, lobbies, businesses)*

Network datasets

- Collaboration Graphs
  – *who works with whom*
- Who-talks-to-Whom Graphs
  – *IM, call graphs*
- Information Linkage Graphs
  – *pages and links, citations*
- Technological Networks
  – *computers, power grid*
- Networks in the Natural World
  – *food webs, cascading extinctions, neural connections*

Types of Social Network Analysis

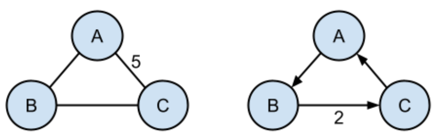
- Sociocentric - whole network analysis
  – *emerged in sociology*
  – *involves quantification of interaction among a socially welldefined group of people*
  – *focus on identifying global structural patterns*
- Egocentric - personal (local) network analysis
  – *emerged in anthropology and psychology*
  – *involves quantification of interactions between an individual (called ego) and all other persons (called alters) related (directly or indirectly) to ego*
- Knowledge Based Network Analysis
  – *emerged in Computer Science*
  – *involves quantification of interaction between individuals, groups and other entities*

Overview

- Introduction
- Graph Theory
- The Small-World Phenomenon
- Measures of Centrality
- Strong and Weak Ties
- Network Level Characteristics

Graphs

- A graph is a way of specifying relationships among a collection of items
- Objects
  - *Nodes - Alice, Bob, ...*
  - *Edges*
    - → *undirected - knows, ...*
    - → *directed - follows, ...*
  - *Values - weights, distances, scores, 0-5 scale, ...*
  - *Attributes - name, time, ...*
- Mathematical models and network structures



Graph Representations

- Adjacency Matrices

/	A	B	C	D	E
A	0	2	0	5	5
B	2	0	0	1	0
C	0	0	0	3	4
D	5	1	3	0	0
E	5	0	4	0	0

- Edge-Lists and Adjacency Lists

from	to	value
A	B	2
A	D	5
…	…	…

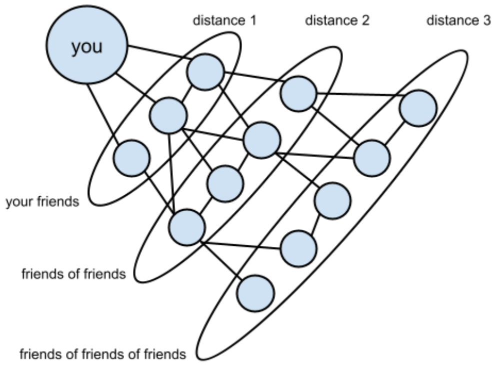
from	edges
A	(B 2), (D 5), (E 5)

Graph Theory

- Graph Theory
  – *"terminological jungle, in which any newcomer may plant a tree"*
    → *John A. Barnes. Social Networks. Number 26 in Modules in Anthropology. Addison Wesley, 1972.*
- Walk
  – *sequence of nodes connected by edges*
  – *open, closed, length*
- Path
  – *open sequence of nodes connected by edges*
- Cycles
  – *path with at least three edges, first and last nodes are the same*
- Connectivity
  – *if exists path between nodes*
- Length
  – *number of edges in the sequence*
- Distance
  – *shortest path between nodes*

Graph Theory

- Shortest path (unweighted graph)
- Cost-based shortest path (weighted graph)
- Depth-first search
- Breadth-first search



Overview

- Introduction
- Graph Theory
- The Small-World Phenomenon
- Measures of Centrality
- Strong and Weak Ties
- Network Level Characteristics

The Small-World Phenomenon

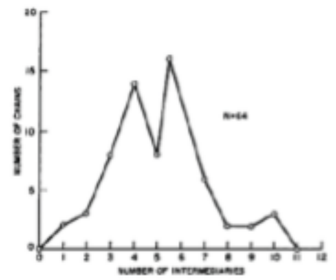
- The idea that the world looks "small"
  - *How short a path of friends it takes to get from you to almost anyone else*
- Six degrees of separation
  - *John Guare. Six Degrees of Separation: A Play. Vintage Books, 1990*
    - → *"I read somewhere that everybody on this planet is separated by only six other people. Six degrees of separation between us and everyone else on this planet."*
  - *There are no more then six connections between any two people on this planet.*

Myth or not?

- Watts and Strogatz
  - *the average path length between two nodes in a random network*
    - → $\frac{ln(N)}{ln(K)}$, *where* N = *total nodes and* K = *acquaintances per node*
    - → $K = 30$, $N = 300{,}000{,}000$ *(90% of the US population)* = 19.5 / 3.4 = 5.7
    - → $K = 30$, $N = 6{,}000{,}000{,}000$ *(90% of the World population)* 22.5 / 3.4 = 6.6
- Dunbar
  - *on average, every person has 150 friends or acquaintances*
  - *within 6 degrees of separation is* 150^6 = *11 trillion*
  - *more then the overall number of people in the world*
    - → *no overlap*
  - *real networks - overlap factor is quite high*
    - → *not enough*
- Generally
  - *many experiments*
  - *exact degree of separation of our society remains unknown*
  - *online networks represent a specific part of our population*
  - *not constant in time*

First experimental study

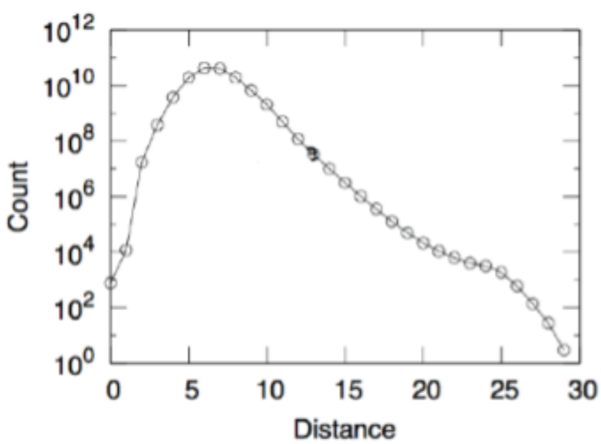
- Stanley Milgram – 1960s
  - *$680 budget*
- Experiment
  - *296 random starters*
    - → *Forward letter to target person*
      - → *Stockbroker who lived in a suburb of Boston*
    - → *Given personal information about target*
    - → *Forward to someone the knew*
      - → *Same instructions*



Milgram experiment (Travers-Milgram 1970)

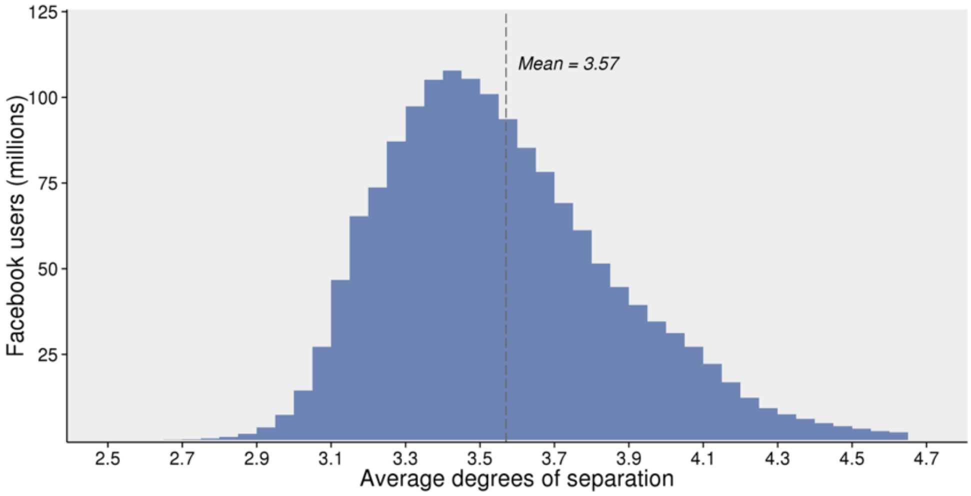
Instant Messaging

- Microsoft Instant Messenger
  - *an edge joining two users if they communicated at least once during a month-long observation period*
  - *240 million active user accounts*
  - *average distance of 6.6*



Facebook

- Facebook
  - *2008: 5.28*
  - *2011: 4.74*
  - *2016: 3.57*



- https://research.fb.com/three-and-a-half-degrees-of-separation/

Paul Erdös

- Itinerant mathematician
- 1500 papers
- Erdös number
  - *the distance from him or her to Erdös – 4 or 5*
  - *Albert Einstein – 2*
  - *Enrico Fermi – 3*
  - *James Watson - 6*

Kevin Bacon

- Movie actors and actress
  - *His or her distance in this graph to Kevin Bacon*
  - *Average = 2.9*
- "I found an incredibly obscure 1928 Soviet pirate film, Plenniki Morya, starring P. Savin with a Bacon number of 7, and whose supporting cast of 8 appeared nowhere else"

Overview

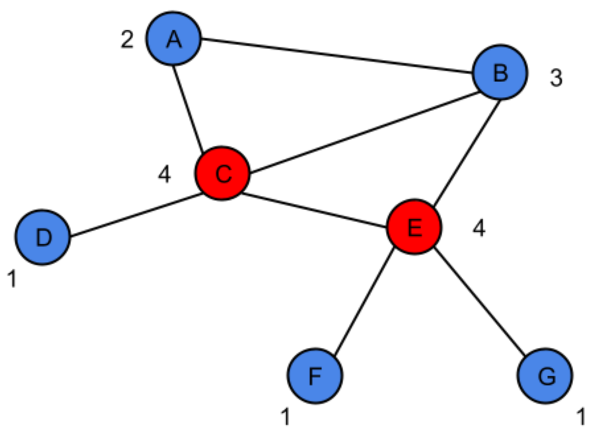
- Introduction
- Graph Theory
- The Small-World Phenomenon
- Measures of Centrality
- Strong and Weak Ties
- Network Level Characteristics

Measures of Centrality

- A variety of different measures exist to measure the importance, popularity, or social capital of a node in a social network.
  – *Measure power, influence, or other individual characteristics of people (based on their connection patterns).*
- Question:
  – *Who is more important in this network?*
  – *Who has the power? ...it depends.*
- Answer:
  – *Degree centrality*
  – *Closeness centrality*
  – *Betweeneness centrality*
  – *Eigenvector centrality*
  – ...

Degree Centrality

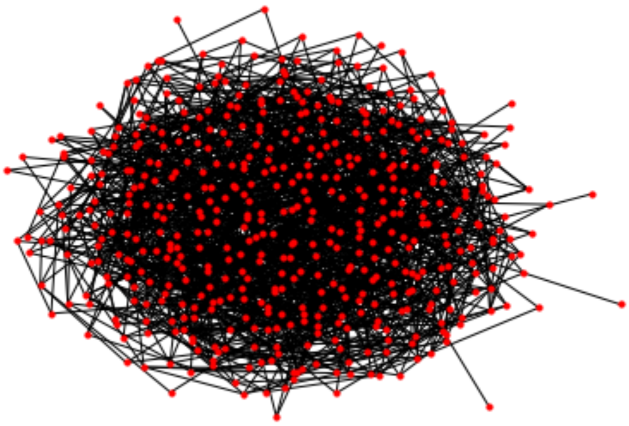
- (in-) or (out-) degree is the number of edges that lead into or out of the node.
  – $Dc_x = deg(x)$
  – *determines the nodes that can quickly spread information to a localized area.*



Degree Centrality - Example

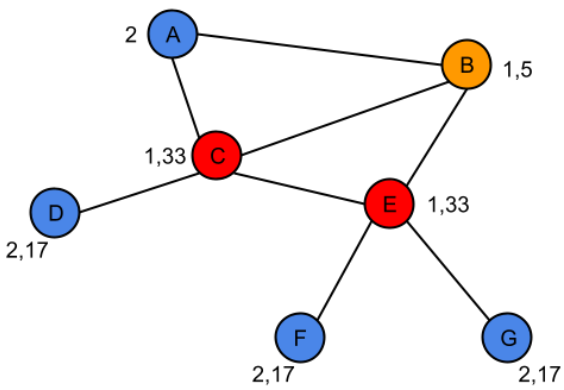
- Dreaded hairball

```
import networkx as nx
from networkx.drawing.nx_agraph import graphviz_layout
import matplotlib.pyplot as plt
G=nx.binomial_graph(500,0.014, seed=5)
pos = graphviz_layout(G)
nx.draw(G, pos, with_labels=False, node_size=10)
plt.show()
```



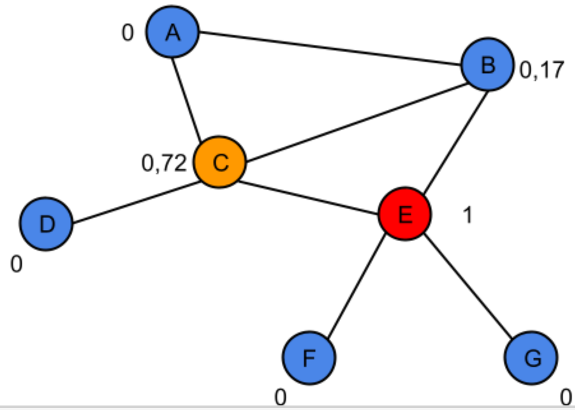
Closeness Centrality

- The mean length of all shortest paths from a node to all other nodes.
  - *Measure of reach, distance to others.*
  - *Horizon of observability (Gossipmongers).*
- Also defined as the inverse of the farness
  - $Cc_x = \frac{1}{\sum_y distance(y,x)}$



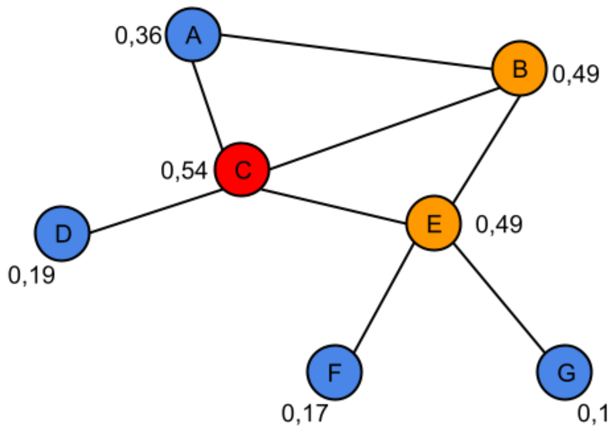
Betweeness Centrality

- The number of shortest paths that pass through a node divided by all shortest paths.
  - *Communication Bottlenecks and/or Community Bridges.*
  - *Boundary spanners - people that act as bridges between two or more communities that otherwise would not be able to communicate to each other.*
- $Bc_x = \sum_{s \neq x \neq t} \frac{\sigma_{sxt}}{\sigma_{st}}$



Eigenvector centrality

- A node with high eigenvector centrality is connected to other nodes with high eigenvector centrality.
  - *Similar to Google PageRank.*
  - *Who is connected to the most connected nodes.*
- Can reveal "Gray Cardinals" - e.g. Don Corleone
  - *"by knowing well-connected people, they can exploit this information and information asymmetry to further their own plans, while staying largely in the shadows"*

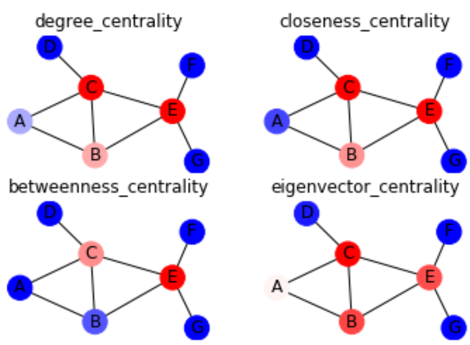


Eigenvector centrality - Algorithm

- Generally
  - *weight each of the links by the degree of the node at the other end of the link*
  - *recursive version of degree centrality*
- Algorithm
  1. *Start by assigning a centrality score of 1 to all Nodes.*
  2. *Recompute the scores of each node as a weighted sum of centralities of all nodes in a node's neighborhood:*
     $v_i = \sum_{j \in N} x_{i,j} * v_j$
  3. *Normalize v by dividing each value by the largest value.*
  4. *Repeat steps 2 and 3 until the values of v stop changing.*

Centralities - Example

```
data= {"B":["A","E"], "C":["A", "B", "D", "E"], "E":["F","G"]}
G=nx.from_dict_of_lists(data)
pos = graphviz_layout(G, prog="fdp")
centralities = [nx.degree_centrality, nx.closeness_centrality,
 nx.betweenness_centrality, nx.eigenvector_centrality]
region=220
for centrality in centralities:
    region+=1
    plt.subplot(region)
    plt.title(centrality.__name__)
    nx.draw(G, pos, labels={v:str(v) for v in G},
      cmap = plt.get_cmap("bwr"), node_color=[centrality(G)[k] for k in centrality(G)])
plt.savefig("centralities.png")
plt.show()
```



Overview

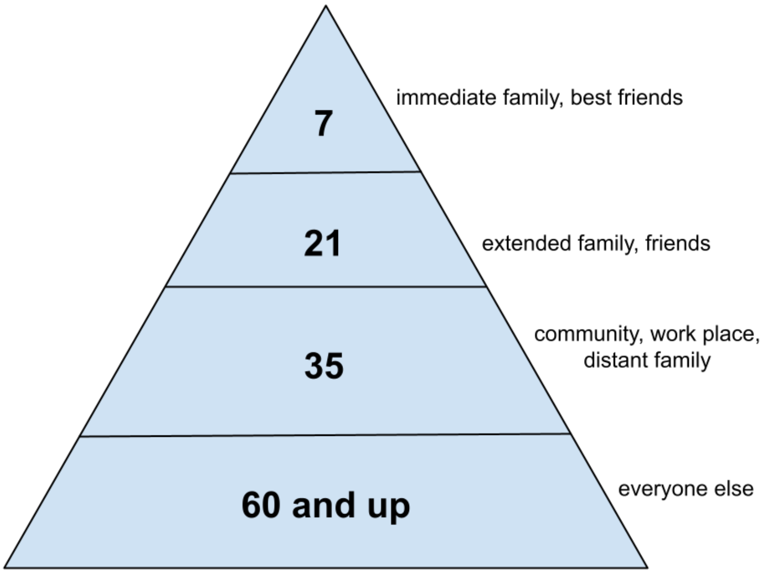
- Introduction
- Graph Theory
- The Small-World Phenomenon
- Measures of Centrality
- Strong and Weak Ties
- Network Level Characteristics

Strong and Weak Ties

- Each edge can influence the network in different way
  - *How information flows through a social network.*
  - *How different nodes can play structurally distinct roles.*
  - *Shape the evolution of the network.*
- Weight of Edge
  - *Frequency of interaction, number of exchanged items, strength of relationship.*

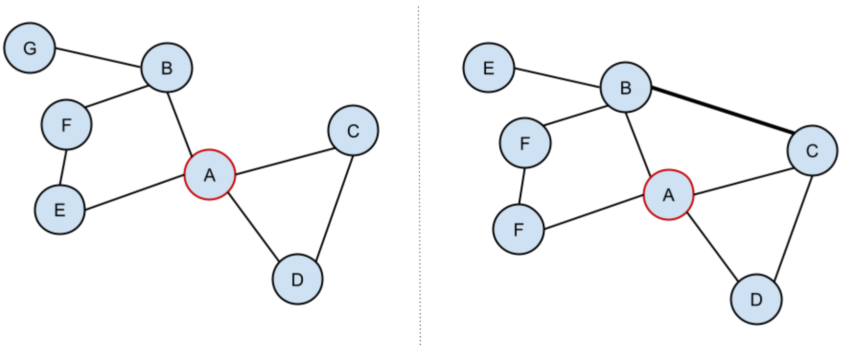
Dunbar Number and Weak Ties

- Average size of a human social network is 150.



Triadic Closure

- Dyad
  – *a pair of actors (connected by a relationship) in the network*
- Triad
  – *a subset of three actors or nodes connected to each other by the social relationship*



- "If two people in a social network have a friend in common, then there is an increased likelihood that they will become friends themselves at some point in the future"

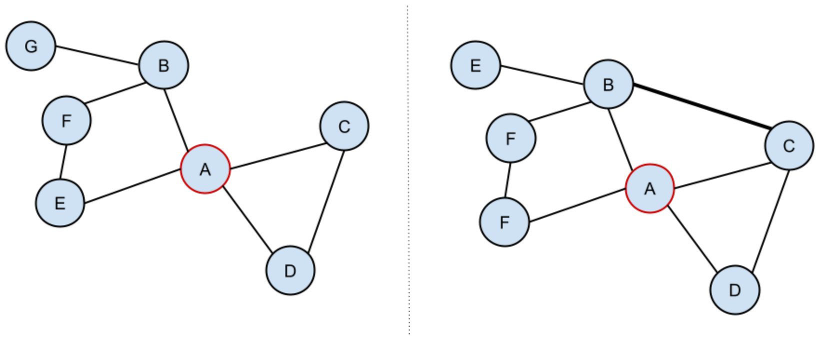
The Clustering Coefficients

- Defined as:
  - *Probability that two randomly selected friends are friends with each other.*
  - *Fraction of pairs of friends that are connected to each other by edges.*
- Vertex v_i has k_i neighbors, $k_i(k_i - 1)/2$ edges can exist.

$$C_i = \frac{2 \| \{e_{jk}\} \|}{k_i(k_i - 1)} : v_j, v_k \in N_i, e_{ij} \in E$$

- Locally indicates how concentrated the neighborhood of a node is, globally indicates level of clustering in a graph.

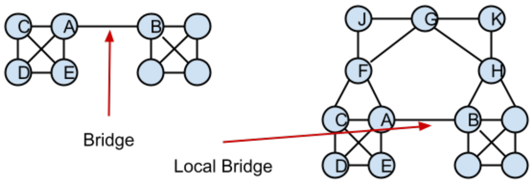
The Clustering Coefficients - Examples



- A - first figure
  - *coefficient = 1/6*
  - *only C-D from all possible (C-D,D-E, B-E, B-C,C-E,D-B)*
- A - second figure
  - *coefficient = 2/6 = 1/3*
  - *C-D and B-C from all possible (C-D,D-E, B-E, B-C,C-E,D-B)*

Bridge and Local Bridges

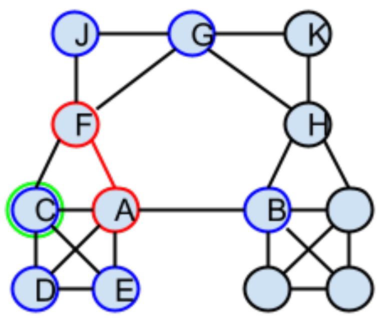
- Bridge
  - *If deleting the edge would cause A and B to lie in two different components. In other words, this edge is literally the only route between its endpoints, the nodes A and B.*
  - *Extremely rare!*
- Local bridge
  - *If its endpoints A and B have no friends in common — in other words, if deleting the edge would increase the distance between A and B to a value strictly more than two.*
  - *When it does not form a side of any triangle in the graph.*



Neighborhood overlap

- Helps identify local bridges
- Local Bridge
  - *NO = 0*

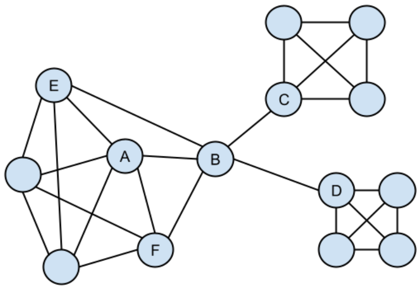
$$NO = \frac{number\ of\ neighbors\ of\ both\ A\ and\ B}{number\ of\ neighbors\ of\ at\ least\ one\ of\ A\ or\ B}$$



- Example:
  - *Edge A-F: 1/6, Edge A-B: 0/8*

Embeddedness

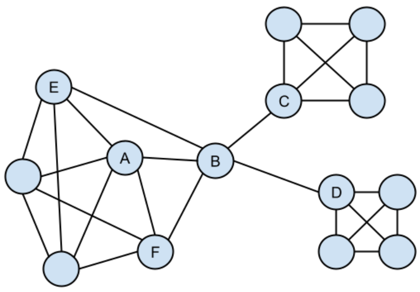
- Embeddedness of an edge
  – *number of common neighbors the two endpoints have*
- Local bridges has embeddedness = 0
- Signifficant embeddedness
  – *easier trust, confidence (social, economic)*



- Example:
  – *A-B = 2*
    → *Common E and F*

Structural hole

- Characteristics
  – *Access information from non-interacting parts*
  – *Interface*
  – *Novel ideas*
  – *Gatekeeping*



- Example
  – *Node B*

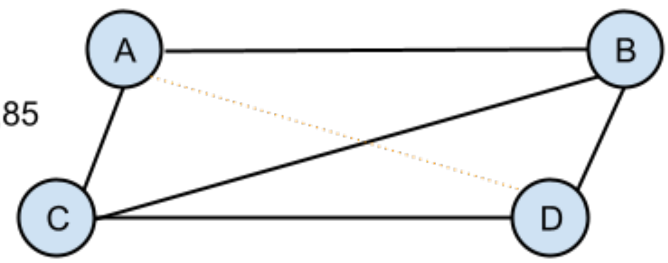
Overview

- Introduction
- Graph Theory
- The Small-World Phenomenon
- Measures of Centrality
- Strong and Weak Ties
- Network Level Characteristics

Density

- The ratio of the number of edges over the total number of possible edges
- Total number
  - *Undirected – n(n-1)/2*
  - *Directed – n(n-1)*

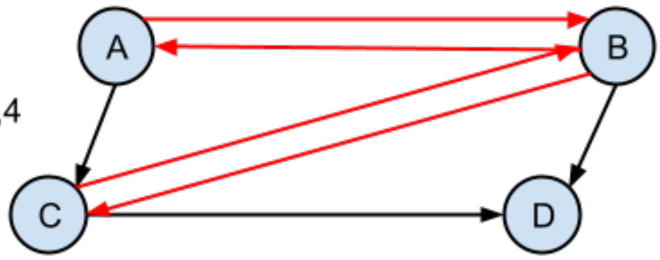
5/6 = 0,85



Reciprocity

- Oriented graphs
- Edge in both directions
- The ration of the number of relations which are reciprocated over the total number of relations

2/5 = 0,4

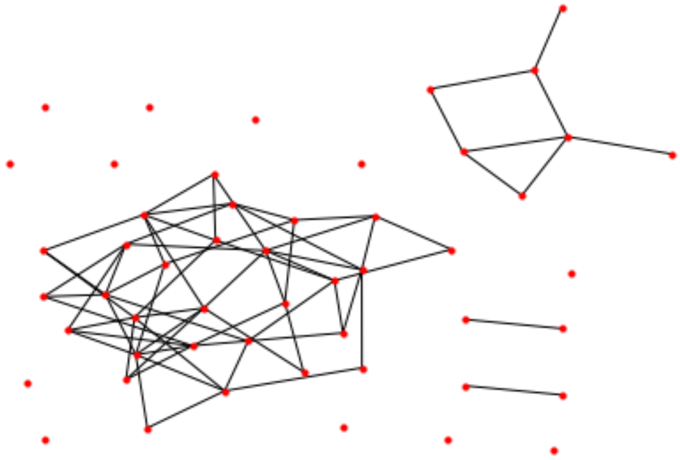


Components

- Components
  - *group of nodes, which are connected*
    - → *every node has a path to every other*
    - → *the group is not part of some larger group*
  - *Giant Components*
    - → *connected component that contains a significant fraction of all the nodes*
  - *Singletons*
    - → *who have no connections and are least central*
  - *Middle region*
    - → *isolated groups which interact amongst themselves, forming isolated stars*

Components - example

```
import networkx as nx
from networkx.drawing.nx_agraph import graphviz_layout
import matplotlib.pyplot as plt
G = nx.random_partition_graph([30,10,5,5,1], 0.15, 0.001, seed=5)
pos = graphviz_layout(G)
nx.draw(G, pos, with_labels=False, node_size=10)
plt.show()
```



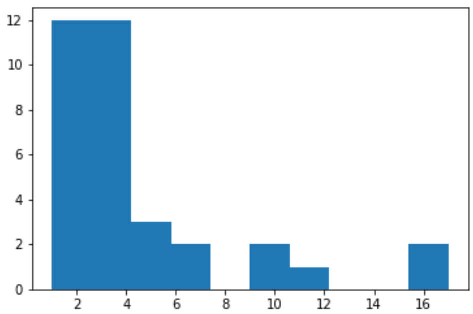
Giant Component Example

- http://www.facebook.com/note.php?note_id=469716398919

Power-Law Distribution

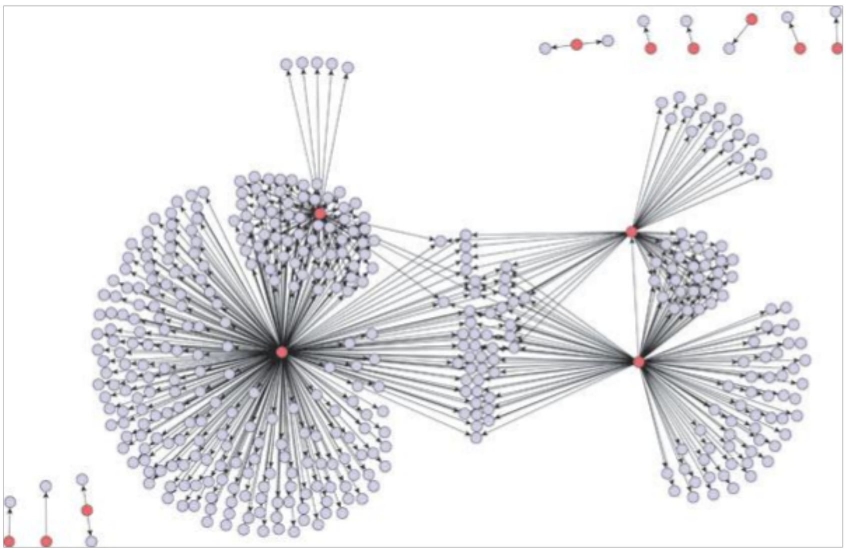
- Degree distribution follows a power law
  - *Scale-free networks*
  - *Some nodes have high number of connections, most nodes have small number of connections*
  - *Characteristics:*
    → *robust against accidental failures*
    → *vulnerable to coordinated attacks*

```
import networkx as nx
import matplotlib.pyplot as plt
G=nx.karate_club_graph()
plt.hist(list(G.degree().values()))
plt.show()
```



Preferential Attachment

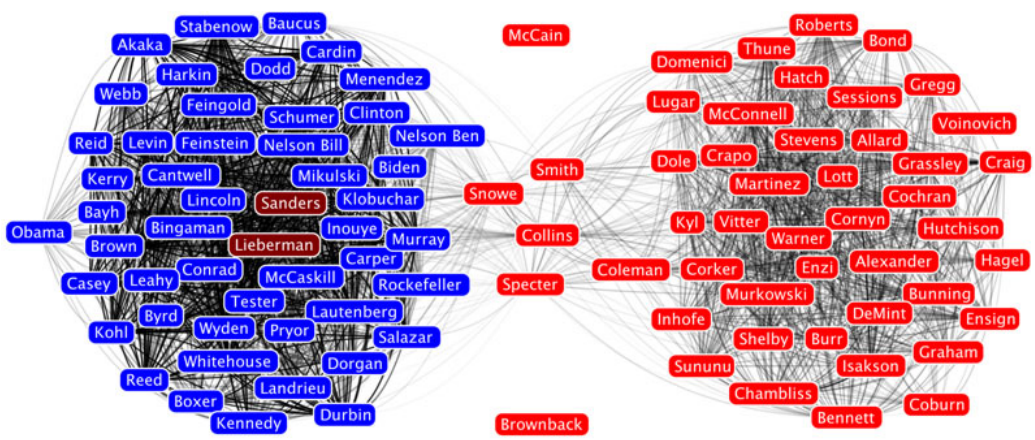
- The great majority of new edges are to nodes with an already high degree.



Homophily

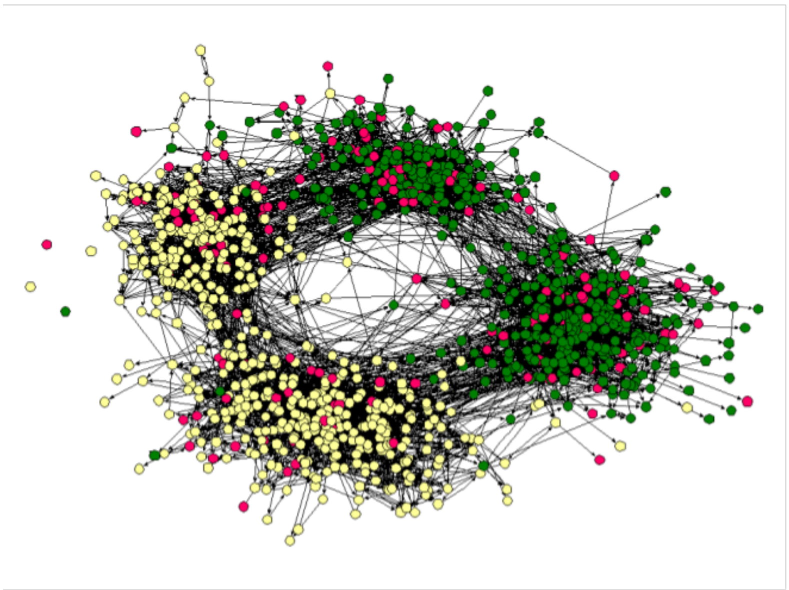
- The principle that we tend to be similar to our friends
  - *Plato*
    → "similarity begets friendship"
  - *Aristotle*
    → people "love those who are like themselves"
  - *Generally*
    → "birds of a feather flock together"
- Similarities
  - *racial, ethnic, age, place, occupation, affluence, interest, opinions, ...*

Homophily - Politics



- Network of senators created by a group in the Human-computer Interaction Lab at the University of Maryland

Homophily - High School

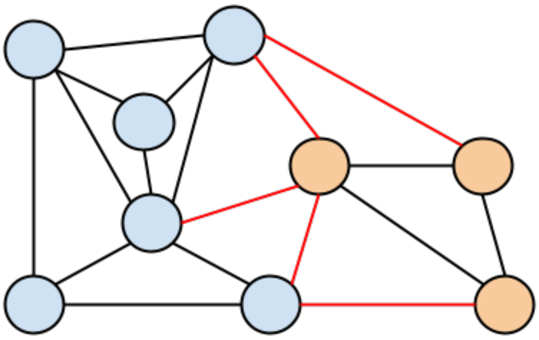


- Homophily: race (different colors of nodes), friendships of students
  – *James Moody: "Race, school integration, and friendship segregation in America," American Journal of Sociology 107, 679-716 (2001)*

Measuring Homophily

- Two classes/similarities
  – *e.g.* male *and* female
- p or q fraction of all that correspond to the first or second class
  – *e.g.* male *or* female
- Probability that for a random edge in a random network both end nodes are of the same class - e.g. same gender
  – $P_s = p^2, q^2$
- Probability of different classes - e.g. cross-gender
  – $P_c = 2 \times p \times q$
- Homophily Test
  – *If the fraction of cross-gender edges is significantly less than* 2pq, *then there is evidence for homophily.*

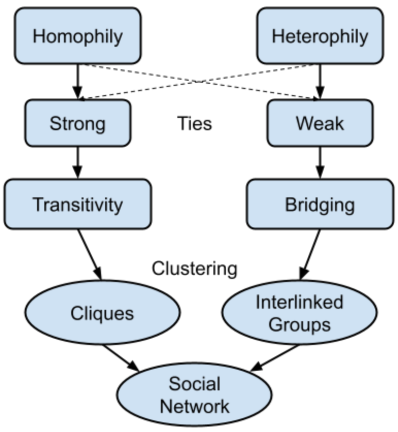
Measuring Homophily - Example



- Example:
  - *5 of 18 cross-gender, p=2/3, q=1/3*
  - *2pq = 4/9 = 8/18*
  - *Test*
    $\rightarrow$ *5/18 < 8/18*

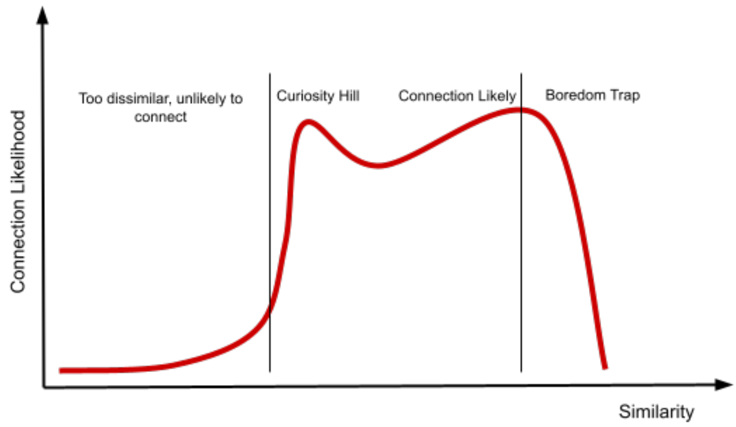
Transitivity

- If there is a tie between A and B and one between B and C, in transitive network A and C will also be connected
- Transitivity and homophily together lead to the formation of cliques (fully connected clusters)



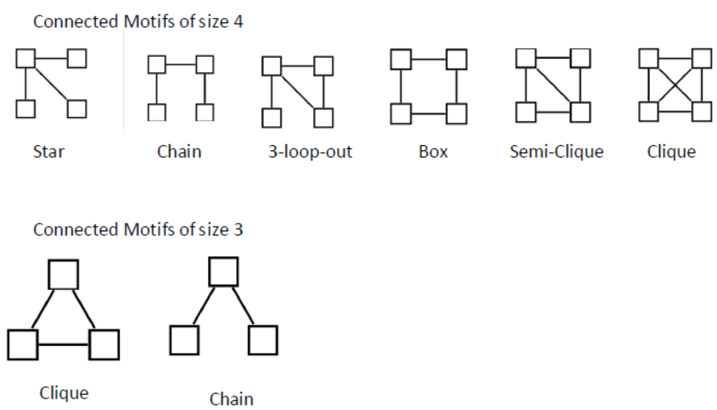
Homophily vs. Curiosity

- Two people are not very similar but not so different as to limit their ability to find topics for conversation
- Boredom trap
  - *Person who is exactly the same as you in every aspect provides no new information or stimulation*

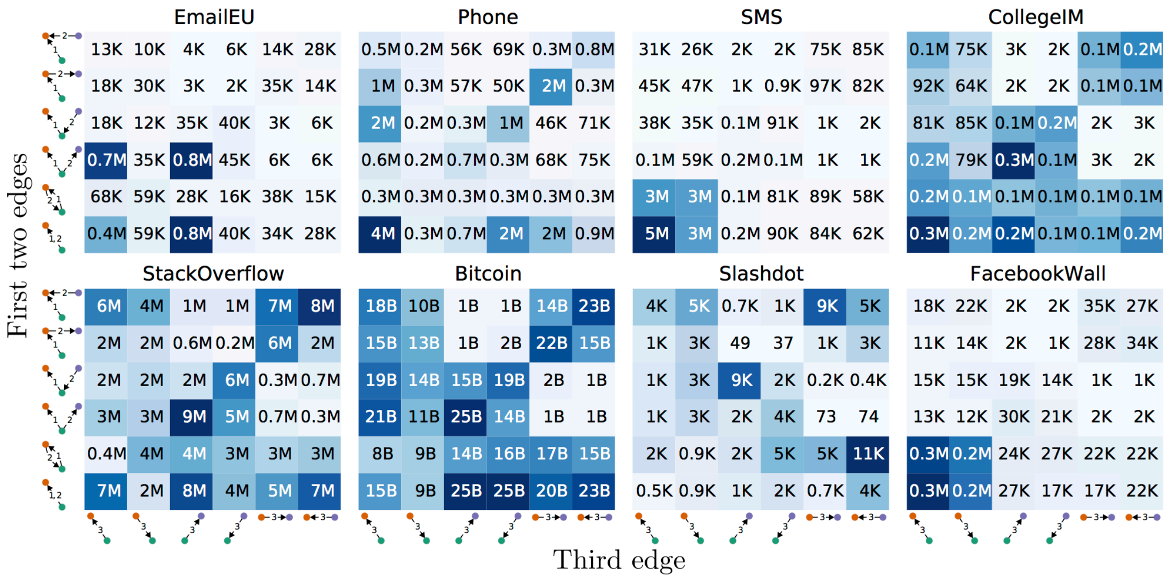


Motifs

- Local property of networks
  - *Defined as recurrent and statistically significant sub-graphs or patterns.*
  - *Networks differ/share specific motifs*
- Useful concept to uncover structural design principles of complex networks.
- Detection is computationally challenging!



Motifs - Example



- A. Paranjape, A. R. Benson, and J. Leskovec: Motifs in Temporal Networks. To appear in WSDM, 2017. http://snap.stanford.edu/temporal-motifs/