Network Analysis / Visualization

Maneesh Agrawala
Jessica Hullman

CS 294-10: Visualization
Fall 2014

Announcements

Final project

Design new visualization method

- Pose problem, Implement creative solution

Deliverables

- Implementation of solution
- 8-12 page paper in format of conference paper submission
- 1 or 2 design discussion presentations

Schedule

- Project proposal: 10/27
- Project presentation: 11/10, 11/12
- Final paper and presentation: TBD, likely 12/1-12/5

Grading

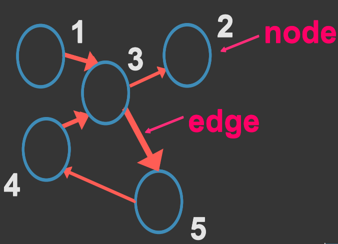
- Groups of up to 3 people, graded individually
- Clearly report responsibilities of each member

Network Analysis & Visualization

*Many slides adapted from E. Adar's / L. Adamic's Network Theory and Applications course slides.

Basic definition

Networks (graphs) are collections of points joined by lines
Edges can be directed / undirected
Edges and nodes can have attributes associated with them

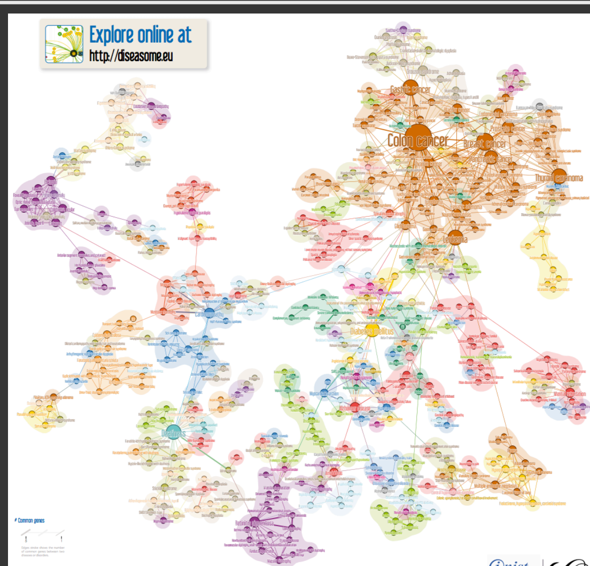


points	lines
vertices	edges, arcs
nodes	links
actors	ties, relations

Adjacency matrix

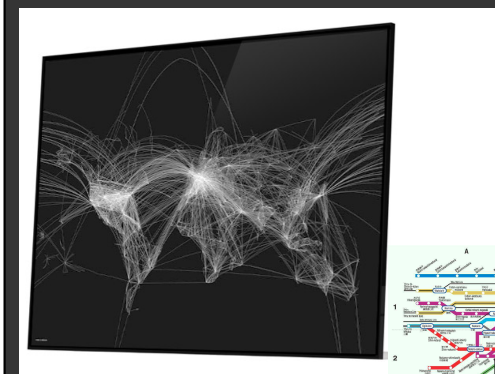
	1	2	3	4	5
1	0	0	1	0	0
2	0	0	1	0	0
3	1	1	0	1	1
4	0	0	1	0	1
5	0	0	1	1	0

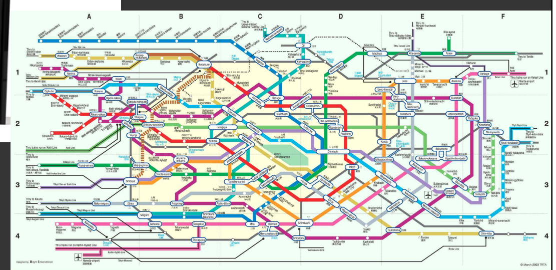
Diseases



http://diseasome.eu/

Transportation

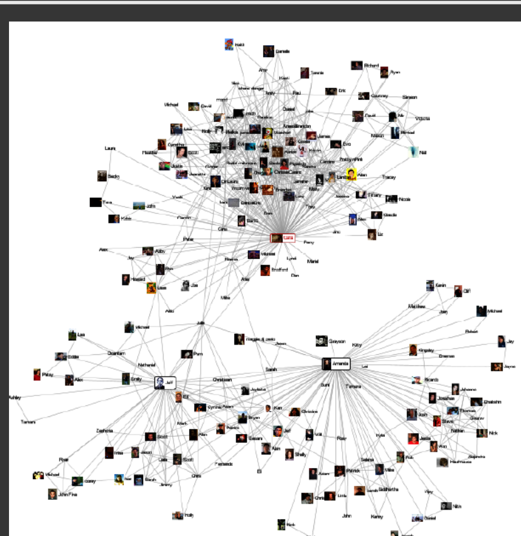




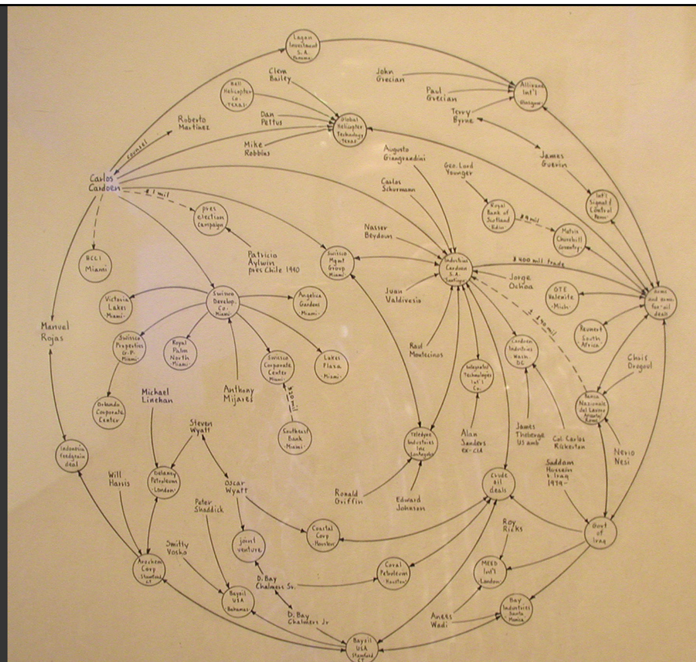
http://www.lx97.com/maps/

Online social networks

Friendster

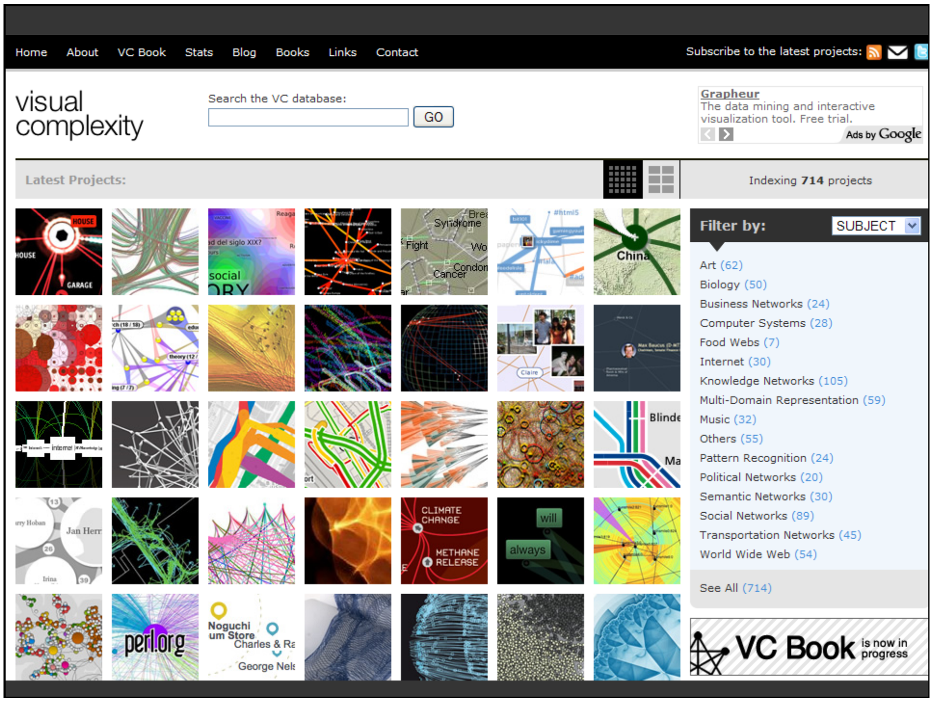


Heer, J., and boyd, d. Vizster: Visualizing Online Social Networks. InfoVis 2005.



Lombardi, M. 'George W. Bush, Harken Energy and Jackson Stephens, ca 1979–90'

Actors and movies (bipartite)

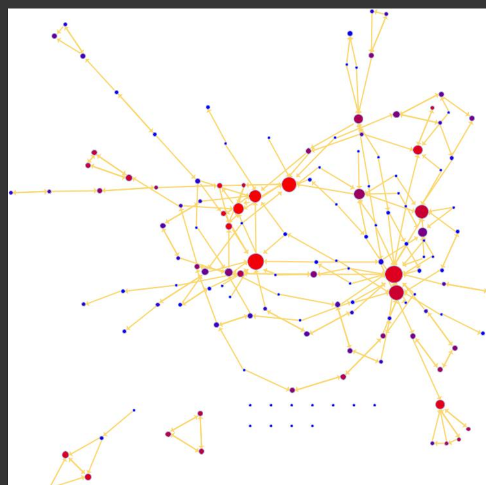


Applications

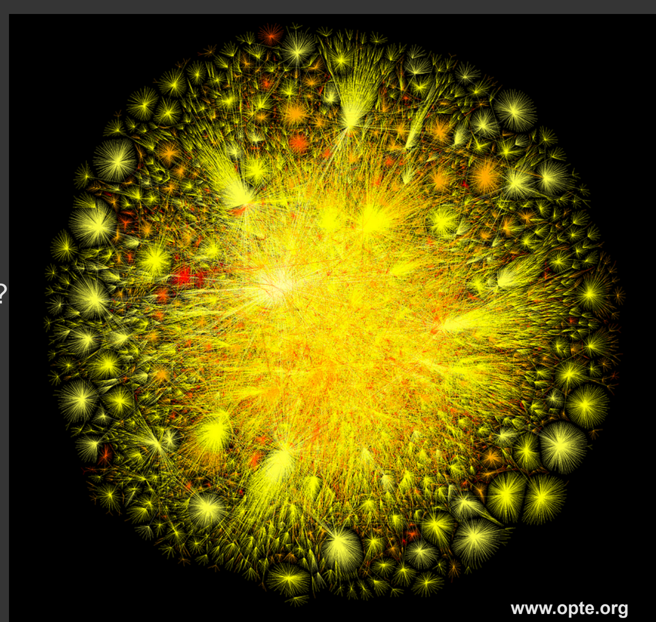
Tournaments
Organization Charts
Genealogy
Text
Biological Interactions (Genes, Proteins)
Computer Networks
Social Networks
Simulation and Modeling
Integrated Circuit Design

Characterizing networks

What does it look like?



Size?
Density?
Centralization?
Clustering?
Components?
Cliques?
Motifs?
Avg. path length?
…



Topics

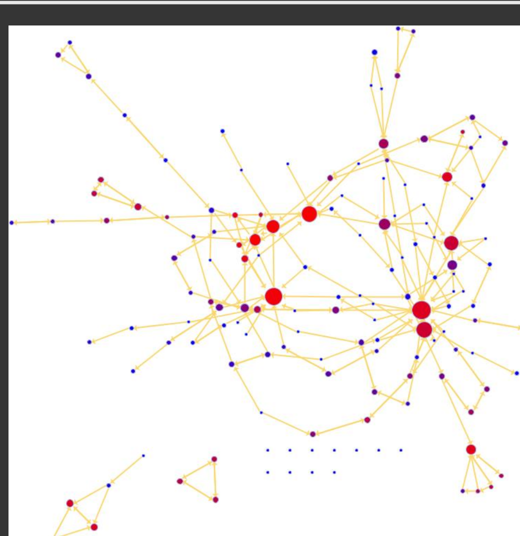
Network Analysis

- Centrality / centralization
- Community structure
- Pattern identification
- Models

Tools for Network EDA

Centrality

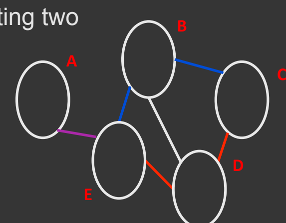
How far apart are things?



Distance: shortest paths

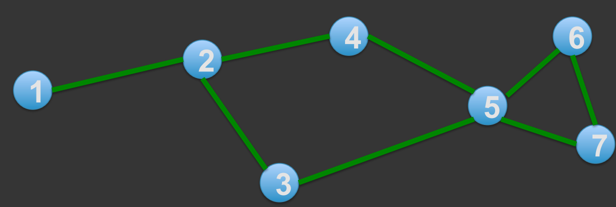
Shortest path (geodesic path)

- The shortest sequence of links connecting two nodes
- Not always unique
- A and C are connected by 2 shortest paths
  - A – E – B - C
  - A – E – D - C



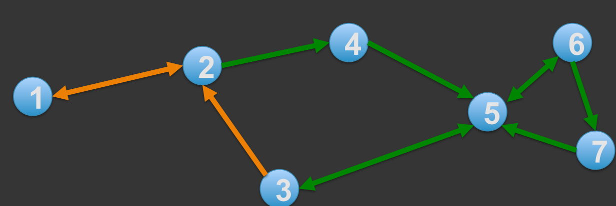
Distance: shortest paths

Shortest path from 2 to 3: 1

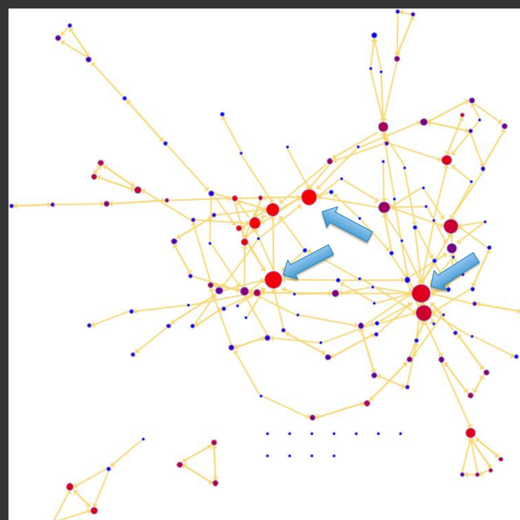


Distance: shortest paths

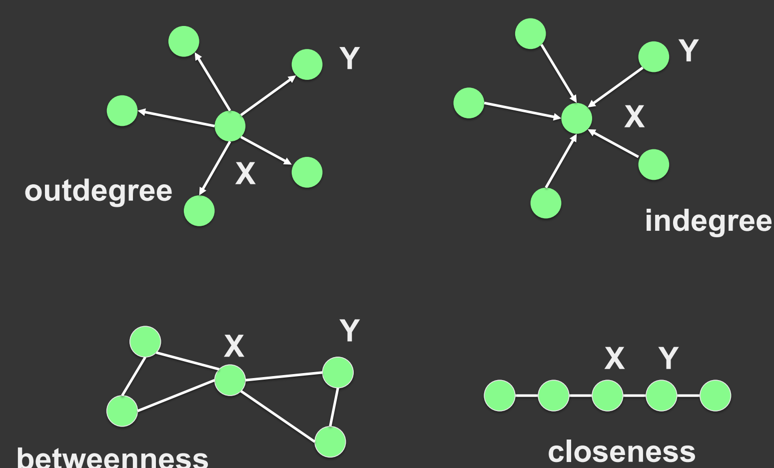
Shortest path from 2 to 3?



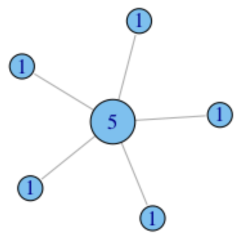
Most important node?



Centrality

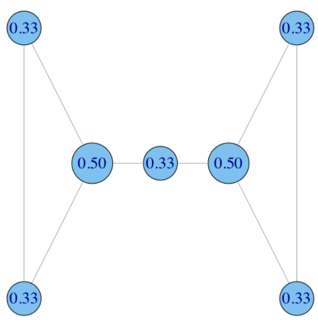


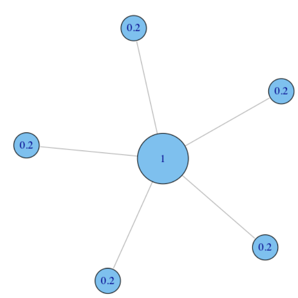
Degree centrality (undirected)



$$C_D = d(n_i) = A_{i+} = \sum_j A_{ij}$$

Normalized degree centrality





$$C_D(i) = \frac{d(i)}{N-1}$$

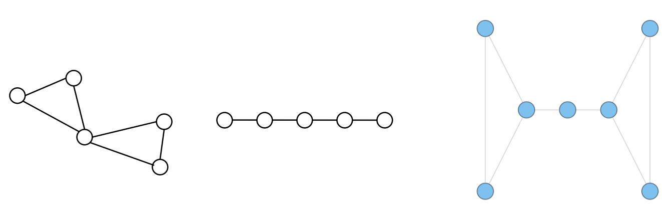
When is degree not sufficient?

ability to broker between groups

likelihood that information originating anywhere in the network reaches you

Betweenness

Assuming nodes communicate using the most direct route, how many pairs of nodes have to pass information through target node?



Betweenness: definition

$$C_B(i) = \sum_{j,k \neq i, j<k} g_{jk}(i) / g_{jk}$$

g_{jk} = the number of geodesics connecting *jk*
g_{jk}(i) = the number that actor *i* is on.

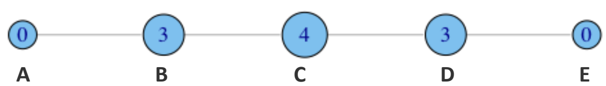
Normalization:

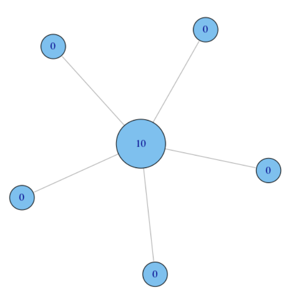
$$C'_B(i) = C_B(i)/[(n-1)(n-2)/2]$$

number of pairs of vertices excluding the vertex itself

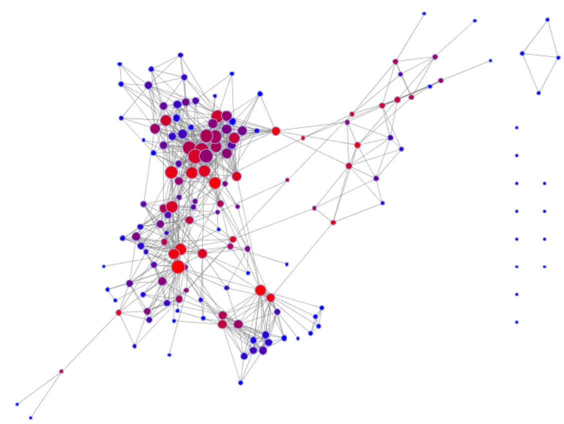
Betweenness - examples

non-normalized:





Lada's facebook network: nodes are sized by degree, and colored by betweenness (blue->red → low->high).



When are C_d, C_b not sufficient?

likelihood that information originating anywhere in the network reaches you

Closeness: definition

Being close to the center of the graph

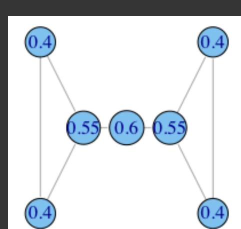
Closeness Centrality:

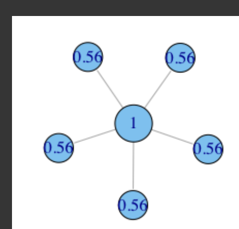
$$C_c(i) = \left[\sum_{j=1, j \neq i}^{N} d(i,j) \right]^{-1}$$

Normalized Closeness Centrality

$$C_C^{'}(i) = (C_C(i)) / (N-1) = \frac{N-1}{\sum_{j=1, j \neq i}^{N} d(i,j)}$$

Examples - closeness





Centrality in directed networks

Prestige

degree centrality ~ indegree centrality

closeness ~ consider nodes from which target node can be reached

influence range ~ nodes reachable from target node

betweenness ~ consider directed shortest paths

Straight-forward modifications to equations for non-directed graphs

Characterizing nodes

	Low Degree	Low Closeness	Low Betweenness
High Degree		Node embedded in cluster that is far from the rest of the network	Node's connections are redundant - communication bypasses him/her
High Closeness	Node links to a small number of important/active other nodes.		Many paths likely to be in network; node is near many people, but so are many others
High Betweenness	Node's few ties are crucial for network flow	Rare. Node monopolizes the ties from a small number of people to many others.	

Centralization – how equal

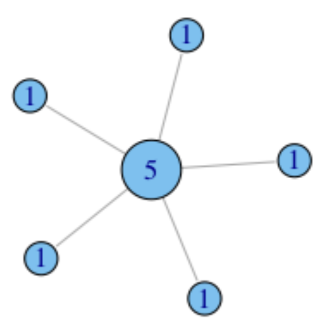
Variation in the centrality scores among the nodes

Freeman's general formula for centralization:

maximum value in the network

$$C_D = \frac{\sum_{i=1}^{g}\left[C_D(n^*) - C_D(i)\right]}{[(N-1)(N-2)]}$$

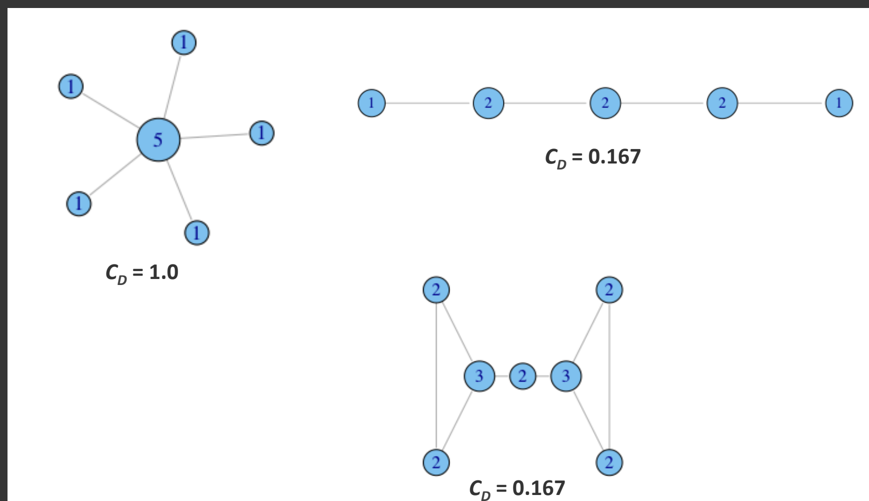
Examples



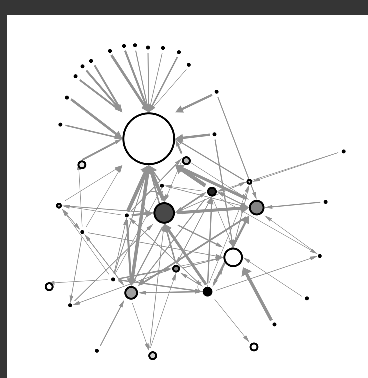
$$C_D = \frac{\sum_{i=1}^{g}\left[C_D(n^*) - C_D(n_i)\right]}{[(N-1)(N-2)]}$$

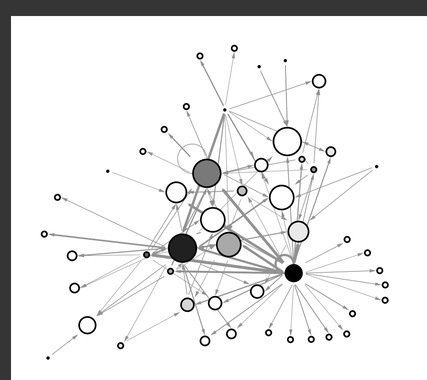
$$C_D = \frac{(5-5)+(5-1)\times 5}{(6-1)(6-2)} = 1$$

Examples



Financial networks

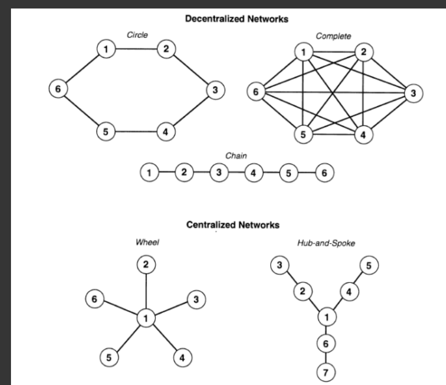




Baker and Faulkner

Price fixing scams in switchgear, transformer, and turbine industries

What network relations facilitate illegal behavior/ conspiracy?



The Social Organization of Conspiracy: Illegal Networks in the Heavy Electrical Equipment Industry, Wayne E. Baker, Robert R. Faulkner. American Sociological Review, Vol. 58, No. 6 (Dec., 1993), pp. 837-860.

Theoretical predictions

Organization Objective	Information-Processing Requirement	
	High	Low
Coordination	Decentralized networks	Centralized networks

Figure 1. Concealment Versus Coordination: Theoretical Expectations

Results

Low information-processing conspiracies are decentralized, high information processing load leads to centralization.

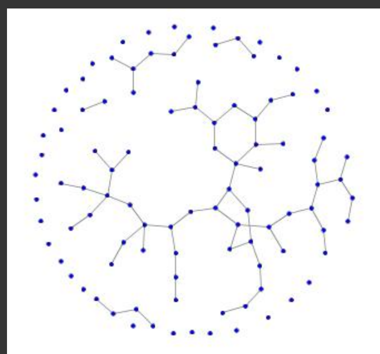
At the individual level, degree centrality predicts verdict.

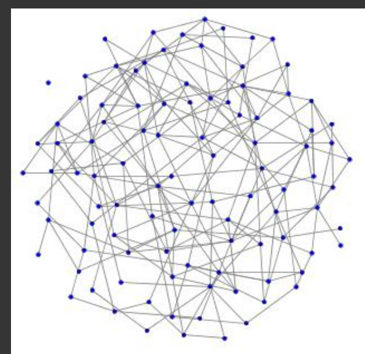
Table 1. Network Characteristics and Outcomes for Three Price-Fixing Conspiracies

Network Characteristic and Outcome	Conspiracy: Switchgear	Transformers	Turbines
Network Characteristic			
Size (number of participants)	33	21	24
Density	23.3	32.4	35.5
Nieminen graph centralization (degree)	41.7	36.1	51.4
Freeman graph centralization (betweenness)	21.3	17.6	24.2
Sabidussi graph centralization (farness)	39.0	37.4	60.8

Community Structure

How dense is it?



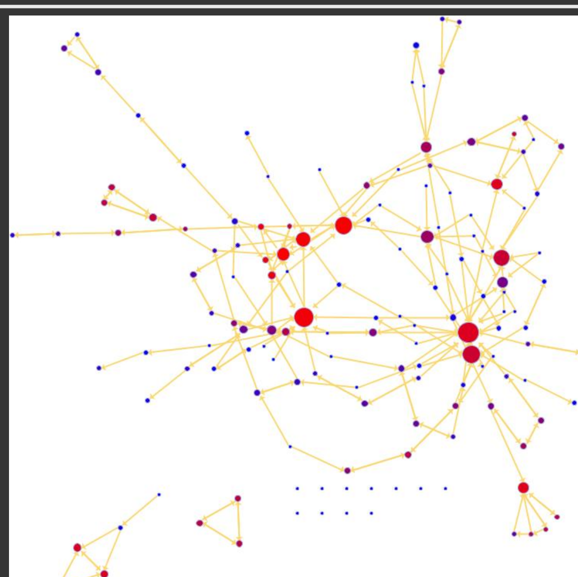


density = e/ e_{max}

Max. possible edges:

- Directed: $e_{max} = n*(n-1)$
- Undirected: $e_{max} = n*(n-1)/2$

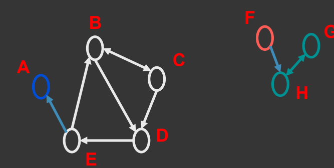
Is everything connected?



Connected components - directed

Strongly connected components

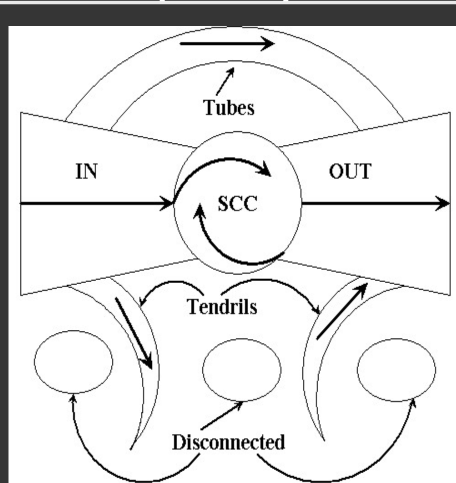
- Each node in component can be reached from every other node in component by following directed links
  - B C D E
  - A
  - G H
  - F



Weakly connected components

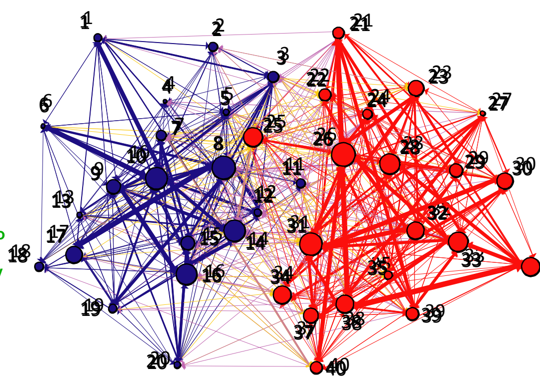
- Each node can be reached from every other node by following links in either direction
  - A B C D E
  - G H F

Broder et al. (1999)



Finding connected components

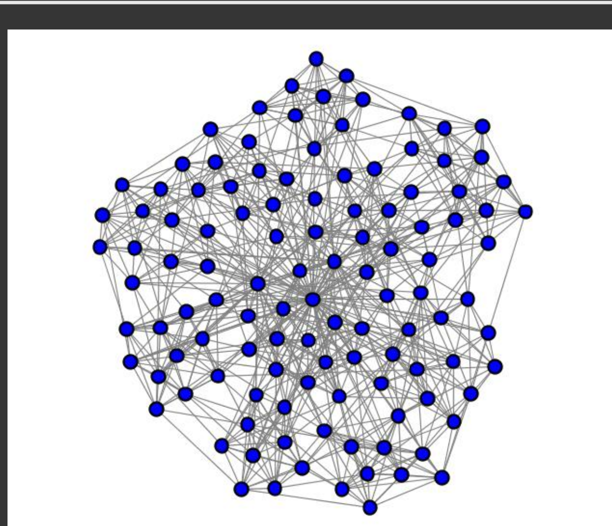
1 Digby's Blog
2 James Walcott
3 Pandagon
4 blog.johnkerry.com
5 Oliver Willis
6 America Blog
7 Crooked Timber
8 Daily Kos
9 American Prospect
10 Eschaton
11 Wonkette
12 Talk Left
13 Political Wire
14 Talking Points Memo
15 Matthew Yglesias
16 Washington Monthly
17 MyDD
18 Juan Cole
19 Left Coaster
20 Bradford DeLong



21 JawaReport
22 Vodka Pundit
23 Roger L Simon
24 Tim Blair
25 Andrew Sullivan
26 Instapundit
27 Blogs for Bush
28 LittleGreenFootballs
29 Belmont Club
30 Captain's Quarters
31 Powerline
32 Hugh Hewitt
33 INDC journal
34 Real Clear Politics
35 Winds of Change
36 Allahpundit
37 Michelle Malkin
38 Wizbang
39 Dean's World
40 Volokh

Adamic, L., and Glance, N. The political blogosphere and the 2004 US election: Divided they blog. Proceedings of the 3rd international workshop on Link discovery, p.36-43, (2005)

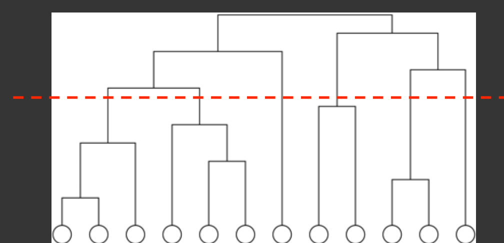
Community finding



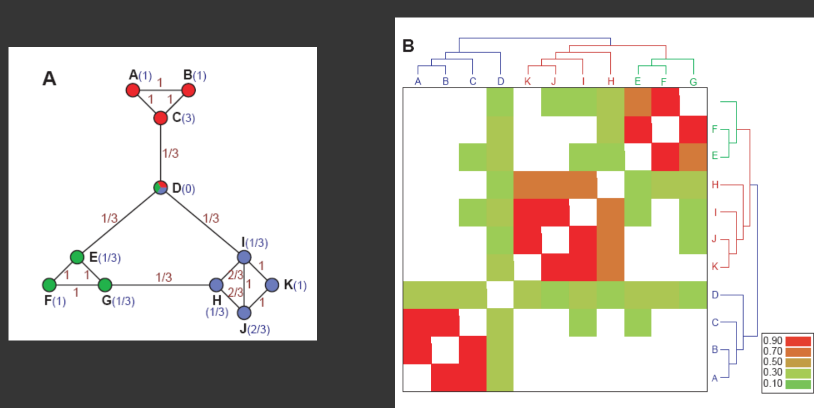
Hierarchical clustering

Process:

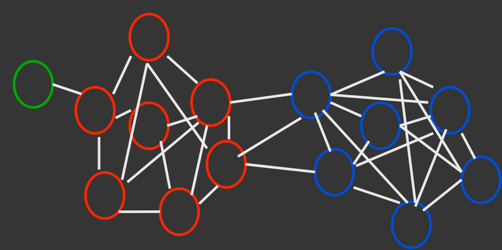
- Calculate weights *W* for all pairs of vertices
- Start: *N* disconnected vertices
- Adding edges (one by one) between pairs in order of decreasing weight
- Result: nested components



Hierarchical clustering



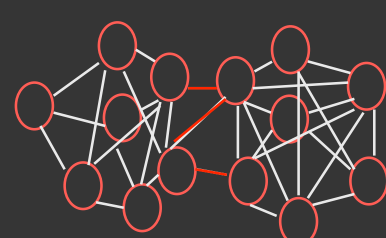
Hierarchical clustering



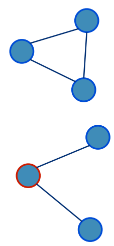
Betweenness clustering

Girvan and Newman 2002 iterative algorithm:

- Compute C_b of all *edges*
- Remove edge *i* where $C_b(i) == max(C_b)$
- Recalculate betweenness

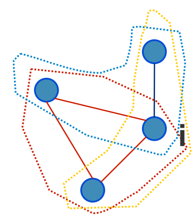


Clustering coefficient



Local clustering coefficient:

$$C_i = \frac{\text{number of closed triplets centered on i}}{\text{number of connected triplets centered on i}}$$



Global clustering coefficient:

$$C_G = \frac{\text{3* number of closed triplets}}{\text{number of connected triplets}}$$

$C_i = 1/3 = 0.33$

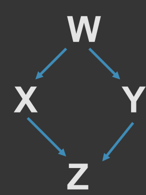
$C_G = 3*1/5 = 0.6$

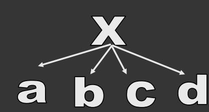
Comparing to random graph

- clustering coefficient
  - compare to a randomized version (conserving node degree)
- degree distribution
- assortativity
  - do high degree nodes connect to other high degree nodes?
- average shortest path
- motif profile

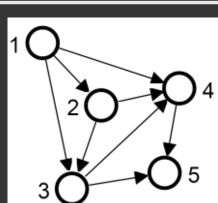
Pattern finding - motifs

Define / search for a particular structure, e.g. complete triads

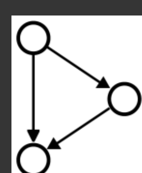




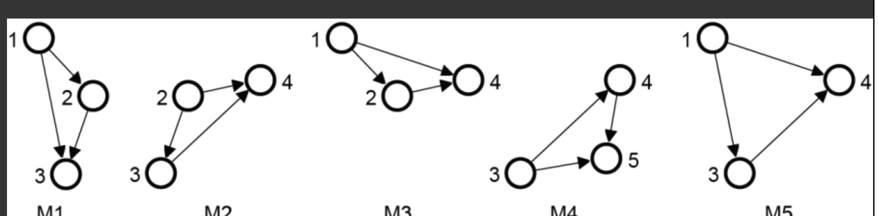
Motifs can overlap in the network



graph



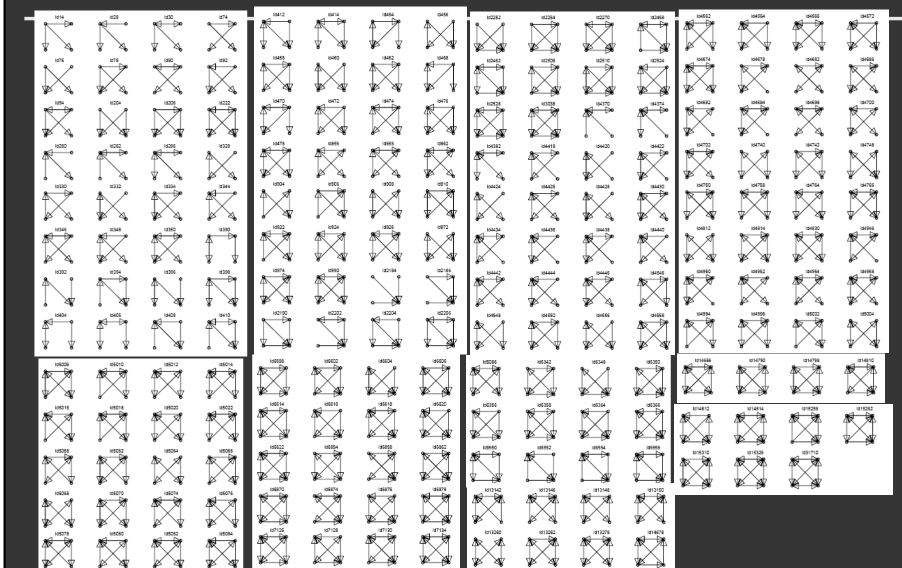
motif to be found



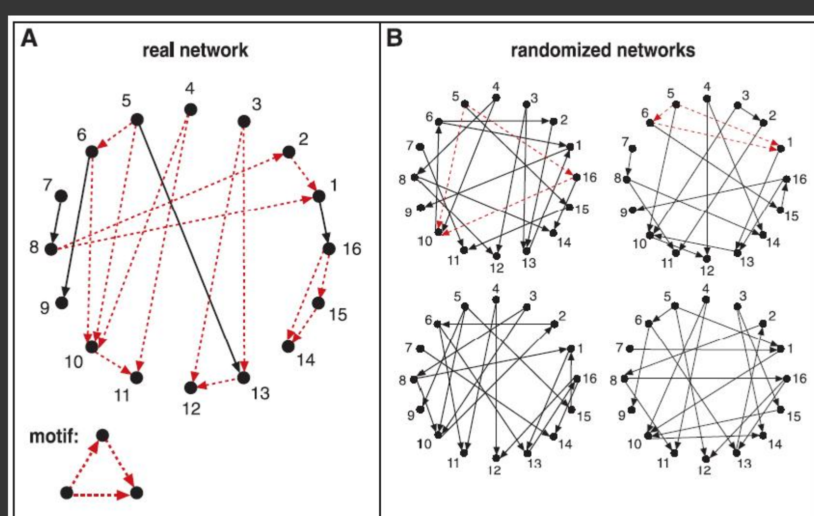
motif matches

http://mavisto.ipk-gatersleben.de/frequency_concepts.html

4 node subgraphs



Motif detection



Tools

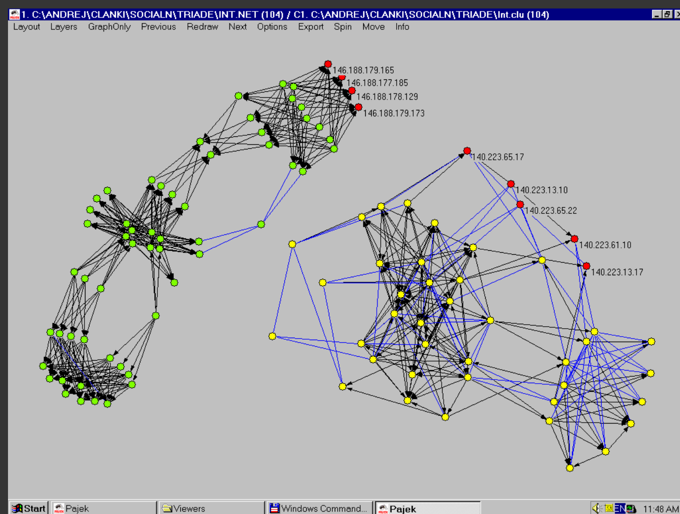
Network EDA

Structure
- Centralization
- Density
- Clustering, components
- Motifs
- Comparison to models

Attributes
- Nodes / links / communities

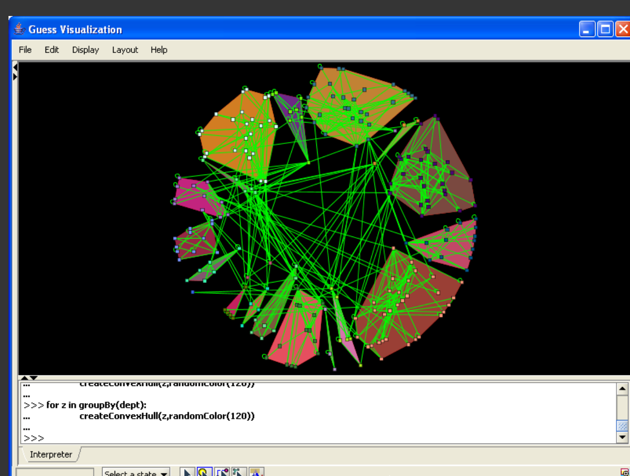
Useful features:
- Associate attributes
- Node/graph level centrality
- Filter on statistics
- Examine distributions
- Identify components, clusters
- Define and search for patterns
- Create random graphs, calculate statistics
- Map statistics to visual features (color, size, weight)
- Track nodes and groups of interest
- Zoom and pan in large graphs

Pajek



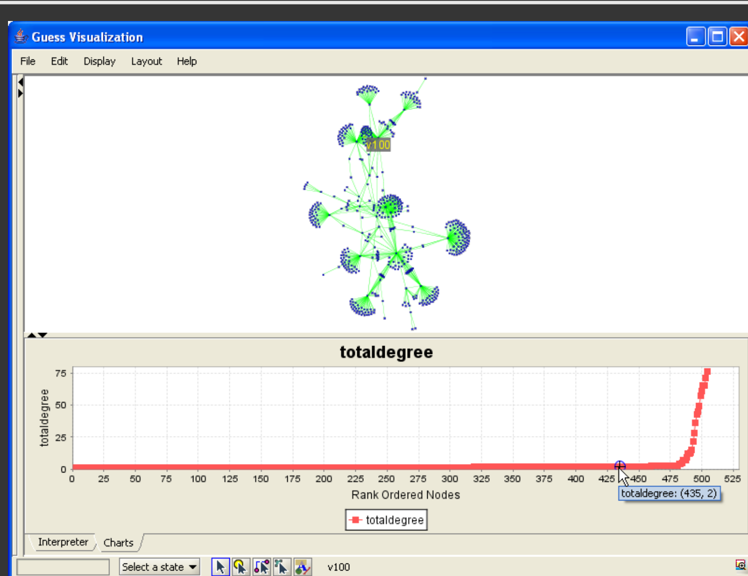
Mrvar, A. Pajek - Program for Large Network Analysis. Connections 21(1998)2, 47-57

GUESS

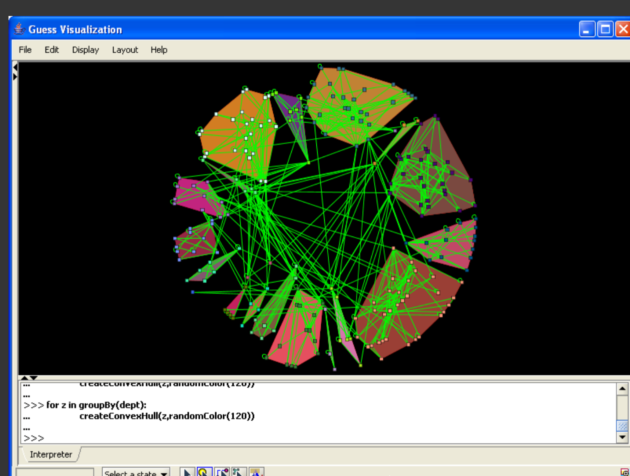


Adar, E. GUESS: The Graph Exploration System. ACM CHI 2006.

GUESS: plotting statistics



GUESS – convex hulls



SocialAction

Challenge:

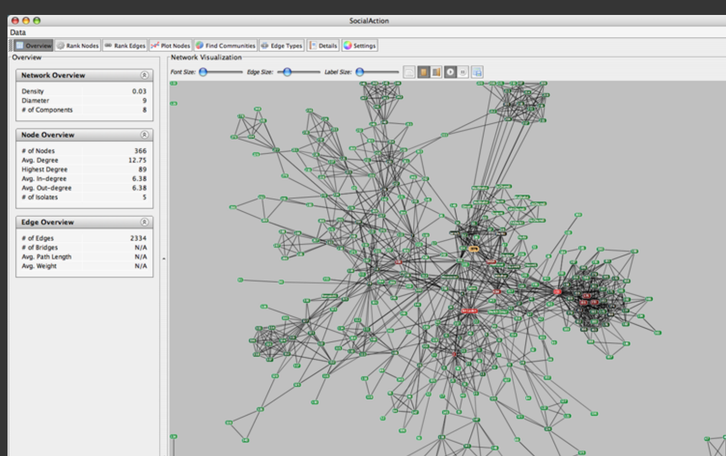
User directedness + number of statistical features leads to opportunistic analysis in most tools

Solution:

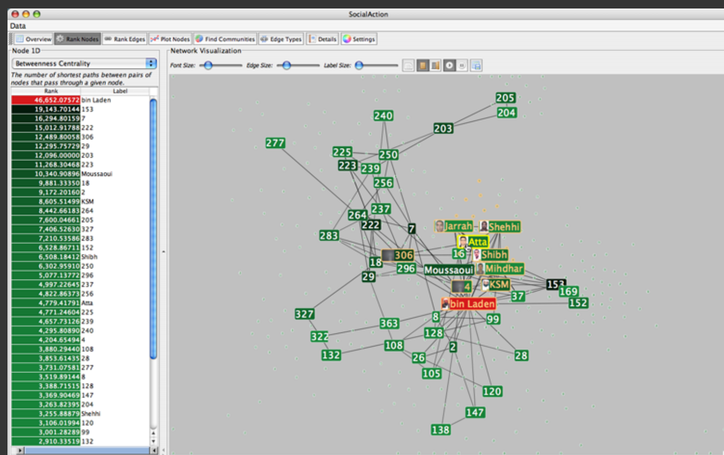
- Provide overview
- Use attribute ranking and coordinated views
- Aggregate networks, identify communities
- View bi-, tripartite (etc.) networks separately
- Access to matrix overview
- Keep nodes in place

Perer, A. and Shneiderman, B. Balancing systematic and flexible exploration of social networks. InfoVis 2006.

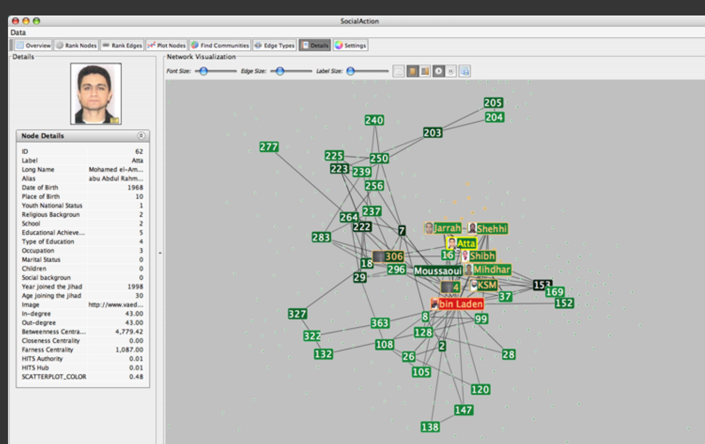
SocialAction



SocialAction



SocialAction



Other tools

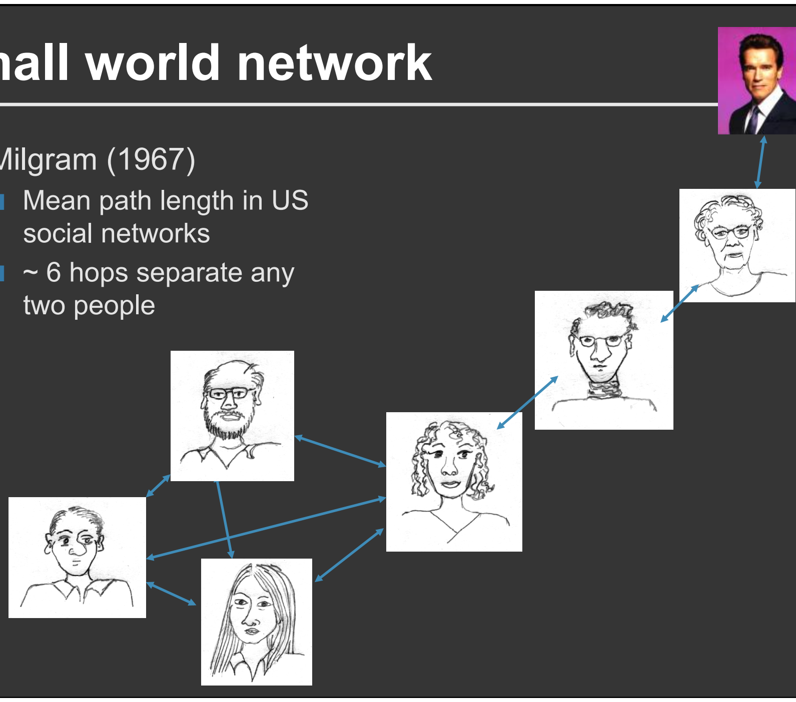
Gephi
Prefuse
TouchGraph
GraphViz
NodeXL
NetLogo

Simulating network models

Small world network

Milgram (1967)

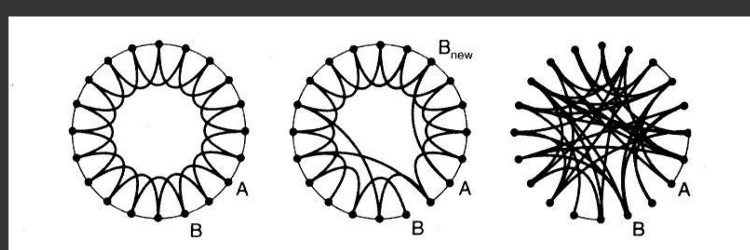
- Mean path length in US social networks
- ~ 6 hops separate any two people



Small world networks

Watts and Strogatz 1998

- a few random links in an otherwise structured graph make the network a small world



regular lattice: my friend's friend is always my friend

small world: mostly structured with a few random connections

random graph: all connections random

Defining small world phenomenon

Pattern:

- high clustering
- low mean shortest path

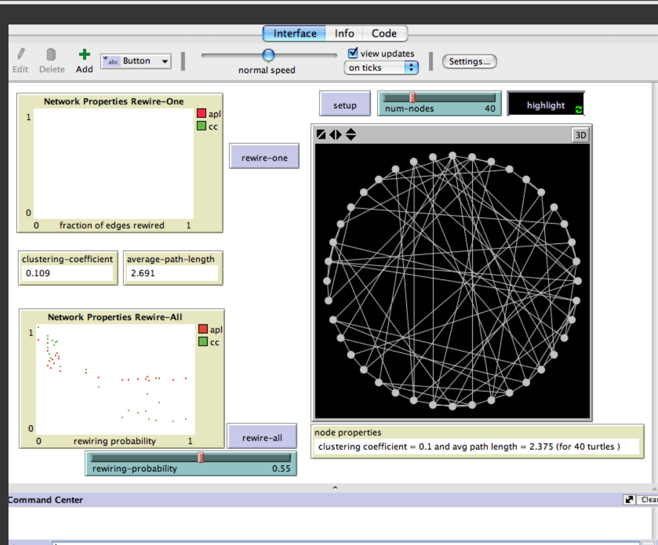
$$C_{\text{network}} >> C_{\text{random graph}}$$

$$l_{\text{network}} \approx \ln(N)$$

Examples

- neural network of C. elegans,
- semantic networks of languages,
- actor collaboration graph
- food webs

NetLogo – small world



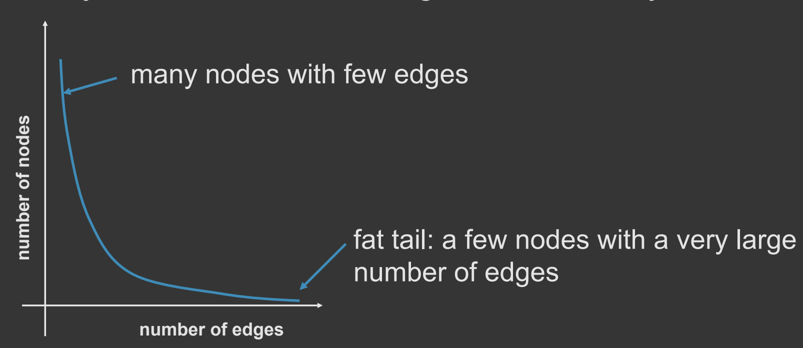
Comparison with "random graph" used to determine whether real-world network is "small world"

Network	size	av. shortest path	Shortest path in fitted random graph	Clustering (averaged over vertices)	Clustering in random graph
Film actors	225,226	3.65	2.99	0.79	0.00027
MEDLINE co-authorship	1,520,251	4.6	4.91	0.56	1.8×10^{-4}
E.Coli substrate graph	282	2.9	3.04	0.32	0.026
C.Elegans	282	2.65	2.25	0.28	0.05

Power law networks

Many real world networks contain hubs: highly connected nodes

Usually the distribution of edges is extremely skewed



Implications

Robustness
Search
Spread of disease
Opinion formation
Spread of computer viruses
Gossip

Summary

Structural analysis

- Centrality
- Community structure
- Pattern finding

→ Widely applicable across domains

Tools for network EDA

- Calculate, filter on statistics
- View graph plus matrix, histograms, etc.
- Overview plus details on demand
- Highlight user-defined nodes of interest, consistent positions